Markov Decision Processes and Stochastic Games with Total Effective Payoff *

Endre Boros[1], Khaled Elbassioni[2], Vladimir Gurvich[1], and Kazuhisa Makino[4]

1 **MSIS Dep. of RBS and RUTCOR, Rutgers University; 100 Rockafeller Road, Piscataway, NJ 08854-8054, USA**
{endre.boros,vladimir.gurvich}@rutgers.edu

2 **Masdar Institute of Science and Technology, P.O. Box 54224, Abu Dhabi, UAE**
kelbassioni@masdar.ac.ae

3 **Research Institute for Mathematical Sciences (RIMS) Kyoto University, Kyoto 606-8502, Japan**
makino@kurims.kyoto-u.ac.jp

Abstract

We consider finite Markov decision processes (MDPs) with undiscounted *total* effective payoff. We show that there exist uniformly optimal pure stationary strategies that can be computed by solving a polynomial number of linear programs. We apply this result to two-player zero-sum stochastic games with perfect information and undiscounted total effective payoff, and derive the existence of a saddle point in uniformly optimal pure stationary strategies.

1998 ACM Subject Classification G.3 Probability and Statistics, G.1.6 Optimization

Keywords and phrases Markov decision processes, undiscounted stochastic games, linear programming, mean payoff, total payoff

Digital Object Identifier 10.4230/LIPIcs.STACS.2015.103

1 Introduction

1.1 Basic concepts

1.1.1 Markov decision proccesses

We will consider Markov decision processes (MDPs) with *total effective payoff*. Let $G = (V, E)$ be a *finite* directed graph (digraph) in which loops and multiple arcs are allowed. The vertices $v \in V$ are called positions (or states) and the arcs $e \in E$ are called *moves* (or transitions). The vertex-set V is partitioned into two subsets $V = V_W \cup V_R$ that correspond to white and random positions, controlled respectively, by a player (decision maker), who will be called Max, and by nature. Let us denote by $E(u)$ the set of arcs leaving u and assume that $E(u) \neq \emptyset$ in every position $u \in V$.

For all random positions $u \in V_R$ we are given probabilities $p(u, v) > 0$ for all random moves $(u, v) \in E(u)$ such that $\sum_{(u,v)\in E(u)} p(u, v) = 1$. There is also a *local reward* function $r : E \to \mathbb{Z}$ given. The the triplet $\Gamma = (G, p, r)$ will be called an MDP.

* This research was partially supported by the Scientific Grant-in-Aid from Ministry of Education, Science, Sports and Culture of Japan. The first author also thanks for partial support the National Science Foundation (Grant and IIS-1161476).

32nd Symposium on Theoretical Aspects of Computer Science (STACS 2015).
Editors: Ernst W. Mayr and Nicolas Ollinger; pp. 103–115

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1.1.2 Strategies

The vertices represent the states of a finite state dynamical system. If at time t the system is in state $v_t = u \in V_W$ then the controller (MAX) chooses (as an action) one of the outgoing arcs $(u, v) \in E(u)$ with some probability and the system moves with this probability to $v_{t+1} = v$. If $v_t = u \in V_R$, then the system moves to $v_{t+1} = v$ with probability $p(u, v)$ (MAX has no influence over this move.)

A strategy (policy) of MAX is a mapping $\mathfrak{s}$ that for every possible $v_t = u \in V_W$ provides a probability distribution over $E(u)$. These probabilities may depend, in general, not only on u and t but also on the entire history of the system up to time t. If these probabilities take only values 0 and 1, then the strategy $\mathfrak{s}$ is called *pure*; if these probabilities depend only on the current state u, then $\mathfrak{s}$ is called *stationary*. A pure stationary strategy is also called *positional*. We shall denote by $\mathfrak{S}$ the set of all possible strategies and by $\widehat{\mathfrak{S}}$ the set of all positional strategies.

Once MAX chooses a strategy $\mathfrak{s} \in \mathfrak{S}$, and we fix an initial state v_0, the above process produces a series of states $v_t(\mathfrak{s}) \in V$, $t = 0, 1, \ldots$, which generally are random variables for $t > 0$. We associate to such a process the sequence of expected local rewards

$$a_t(\mathfrak{s}) \;=\; \mathbb{E}_{\mathfrak{s}}[r(v_t(\mathfrak{s}), v_{t+1}(\mathfrak{s})] \quad \text{for } t = 0, 1, \ldots,$$

and set $a(\mathfrak{s}) = \langle a_0(\mathfrak{s}), a_1(\mathfrak{s}), \ldots\rangle$. For simplicity we will omit in the sequel the argument $\mathfrak{s}$ and write v_t and $\mathbb{E}_{\mathfrak{s}}(r(v_t, v_{t+1})$ rather than $v_t(\mathfrak{s})$ and $\mathbb{E}_{\mathfrak{s}}[r(v_t(\mathfrak{s}), v_{t+1}(\mathfrak{s}))]$ for $t = 0, 1 \ldots$

1.1.3 Effective payoffs

We consider an effective payoff function $\pi : \mathbb{R}^* \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and $\mathbb{R}^*$ standardly denotes the set of all real sequences. The objective of MAX is to find a strategy $\mathfrak{s} \in \mathfrak{S}$ such that $\pi(a(\mathfrak{s})) = \pi_{\mathfrak{s}}(v_0)$ is as large as possible. A strategy $\mathfrak{s}$ is called *uniformly optimal* if $\pi_{\mathfrak{s}}(v_0) \geq \pi_{\mathfrak{s}'}(v_0)$ for any strategy $\mathfrak{s}' \in \mathfrak{S}$ and any initial position $v_0 \in V$.

In this paper we consider the following two effective payoff functions:

$$\phi_{\mathfrak{s}}(v_0) \;=\; \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})], \tag{1}$$

$$\psi_{\mathfrak{s}}(v_0) \;=\; \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{j=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_j, v_{j+1})]. \tag{2}$$

The first one, called *mean payoff*, is classic [12, 4]. The second one, called *total payoff* or *total reward*, was introduced by Thuijsman and Vrieze [27], as a "refinement" of the mean payoff. Let us note however that in fact total payoff MDPs can be shown to include mean payoff MDPs as a special case.

We note that in many earlier works the effective payoff of a play was defined as the *sum* of all local rewards assigned to the moves of this play. Yet, evaluation of the infinite plays may constitute a problem. For that reason, in most of the papers an assumption has to be made such as termination with probability one [7, 9, 25, 3, 31, 30]; in fact definition (2) is a generalization of the sum of local rewards, taking properly into account how to handle cycling in an infinite (non-terminating) play; see Section 1.3.

For an MDP Γ, payoff function π, and a node u, let us define

$$\pi_\Gamma(u) = \sup_{\mathfrak{s}\in\mathfrak{S}} \pi_{\mathfrak{s}}(u),$$

as the *value* of the MDP at node u.

1.1.4 Stochastic games with perfect information: The BWR model

We also consider the following natural and standard generalization. Assume that the finite vertex set V of a given finite directed graph $G = (V, E)$ is partitioned into three (rather than two) subsets $V = V_B \cup V_W \cup V_R$ that correspond to *black*, *white*, and *random* positions, controlled respectively, by two players, MIN and MAX, and nature.

Analogously to MDPs, we can define strategies for the players, and denote by $\mathfrak{S}_B$ and $\mathfrak{S}_W$ the sets of strategies of MIN and MAX, respectively. Given a pair of strategies $\mathfrak{s} = (\mathfrak{s}_B, \mathfrak{s}_W)$ of the players and an initial vertex $v_0 \in V$, we can associate a sequence of expected rewards $\mathbb{E}_{\mathfrak{s}}[r(v_t(\mathfrak{s}), v_{t+1}(\mathfrak{s}))]$ to these, just like we did for MDPs. The objectives of MIN and MAX are to minimize and respectively maximize the expected effective payoff $\pi_{\mathfrak{s}_B,\mathfrak{s}_W}(v_0) = \pi_{\mathfrak{s}}(v_0)$.

Given a stochastic game with a fixed initial position v_0, a *saddle point* is defined as a pair of strategies $\mathfrak{s}^*_B \in \mathfrak{S}_B$ and $\mathfrak{s}^*_W \in \mathfrak{S}_W$ such that

$$\pi_{\mathfrak{s}^*_B,\mathfrak{s}_W}(v_0) \le \pi_{\mathfrak{s}^*_B,\mathfrak{s}^*_W}(v_0) \le \pi_{\mathfrak{s}_B,\mathfrak{s}^*_W}(v_0) \quad \text{for all} \quad \mathfrak{s}_B \in \mathfrak{S}_B \text{ and } \mathfrak{s}_W \in \mathfrak{S}_W. \tag{3}$$

If such a pair exists, the quantity $\pi_{\mathfrak{s}^*_B,\mathfrak{s}^*_W}(v_0)$ is called the value of the game at node v_0. The saddle point $(\mathfrak{s}^*_B, \mathfrak{s}^*_W)$ is called *uniform* (*subgame perfect*) if the above inequalities hold for all initial positions $v_0 \in V$.

For $\pi = \phi$, such a model was first mentioned in [13], and it was shown in [6] that it is polynomially equivalent with stochastic games with perfect information [12]. For $\pi = \psi$, this model is the same as the one introduced in [27] in case of perfect information. The concept was further developed in [8, 28].

1.2 Main results

We first consider total-payoff MDPs and prove the following result.

▶ **Theorem 1.** *In every MDP with total effective payoff, $\pi = \psi$,* MAX *possesses a uniformly optimal positional strategy. Moreover, such a strategy, together with the optimal value can be found in polynomial time.*

For mean payoff MDPs, the analogous result is well-known, see, e.g. [16, 4, 7, 23]. In fact there are several known approaches to construct the optimal stationary strategies. For instance, a polynomial-time algorithm to solve mean payoff MDPs is based on solving two associated linear programs, see, e.g., [7].

Our approach for proving Theorem 1 is inspired by a result of [28]. We extend this result to characterize the existence of pure and stationary optima within *all* possible strategies by the feasibility of an associated linear system. Next, we show that this system is always feasible and a solution can be obtained by solving a polynomial number of linear programming problems.

▶ Remark. If there are no random nodes in the MDP, then a uniformly optimal stationary strategy can be found by a combinatorial algorithm that solves a polynomial number of minimum mean-cycle problems [18]; we omit the details from this version.

▶ **Theorem 2.** *Every BWR-game with total effective payoff, $\pi = \psi$, has a saddle point in uniformly optimal positional strategies.*

For the mean payoff games with perfect information the above result is well-known [12, 22].

Let us note that there may be no stationary best response against a non-stationary strategy of the opponent. However, for the case of total payoff BWR-games, Theorem 1

implies that for any stationary strategy of a player there is a pure stationary best response (among all strategies) of the opponent. This fact implies that it is enough to construct a saddle point within the family of positional strategies. This latter can be shown by using the discounted formulation of the game.

1.3 Applications of the total payoff

1.3.1 Total payoff MDPs/games with a terminating condition

This is the special case of MDPs/ stochastic games with one special *terminal* state, which is absorbing (that is, $p(t,t) = 1$) and cost free (that is, $r(t,t) = 0$). The payoff function, which is also sometimes called "Total Payoff' is defined as the sum

$$\theta_{\mathfrak{s}}(v_0) \;=\; \liminf_{T\to\infty} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})].$$

This type of MDPs/games have been considered under different names, such as *stochastic shortest path problems/games*, *first passage problems* and *transient programming problems* [1, 2, 3, 7, 14, 25, 29, 31, 30]. This can be thought of as a generalization of the classical (deterministic) shortest path problem on graphs, with the difference that, at each node, one should select a probability distribution over successor nodes, out of a given set of probability distributions. The objective is that the chosen *random* path leads to the terminal node with probability one and with the smallest expected cost. In order to establish the existence of optimal stationary strategies that can derived by the solutions of Bellman-type equations, several assumptions have been made in earlier works, most notably, the existence of a *proper* stationary strategy, i.e., one that guarantees termination from every state with probability 1. Note that for such a proper strategy $\mathfrak{s}$, the resulting Markov chain contains exactly one absorbing class, namely the terminal node, and this case, it is not hard to see that the values obtained from sum payoff $\theta_{\mathfrak{s}}$ and the total payoff $\psi_{\mathfrak{s}}$ are the same. Thus total payoff MDPs/games considered in this paper can be thought of as a generalization of shortest path problems/games, when we do not assume that there is a single terminal.

1.3.2 The shortest path interdiction problem (SPIP)

This is the special case of shortest path games when there are no random nodes. More precisely, in this problem, edges have positive lengths and there is a dedicated terminal vertex to which the minimizer tries to find a short path, while the opponent tries to block such paths. It is easy to see that if we add a loop with zero length on the terminal vertex then the total payoff ψ will be exactly the length of the path for every terminating path, and will be $+\infty$ otherwise.

The problem was introduced by Fulkerson and Harding [10]; see a short survey by Israely and Wood [17]. The simplest version is as follows: Given a digraph $G = (V, E)$, with weighted arcs $r : E \to \mathbb{Z}_+$, and two vertices $s, t \in V$, eliminate (at most) k arcs of E to maximize the length of a shortest (s,t)-path. While this problem is APX-hard [20], the following *vertex-wise budget* SPIP is tractable [21, 20]: we are given a budget allowing to eliminate (at most) $k(v)$ arcs going from each state $v \in V$. This version was considered in [21], where an efficient interdiction algorithm was obtained. Given a digraph $G = (V, E)$, an integer-valued local cost function $r : E \to \mathbb{Z}_+$, a constraint $k(v)$ in every vertex $v \in V$, and an initial vertex s, this algorithm finds in quadratic time an interdiction that maximizes simultaneously the

lengths of all shortest paths from s to each vertex $v \in V$. The execution time is just slightly larger than for the classic Dijkstra shortest path algorithm.

Waving the non-negativity condition from the latter version, we obtain another interesting relation: In this case, the SPIP becomes equivalent [21] with solving mean payoff BW-games (no random nodes). Although the latter problem is known to be in the intersection of NP and co-NP [19, 32], yet, it is not known to be polynomial.

1.3.3 Scheduling with and/or precedence constraints

[24] is another application of the total payoff with $r \geq 0$. Given a digraph $G = (V, E)$, whose states are interpreted as jobs, the and/or precedence constraints require that some jobs $u \in V$ cannot be started before *all* immediate predecessors (v such that $(v, u) \in E$) are completed, while some other jobs $w \in V$ cannot be started before *at least one* immediate predecessor is complete. It is easy to see that this model is equivalent with a total reward BW-game which has nonnegative local rewards. For this problem [24] provides a polynomial time algorithm.

2 Characterization of pure stationary optima in total MDPs

Our proof of Theorem 1 is based on strengthening a result of Thuijsman and Vrieze (Theorem 5.3 in [28]) which gives a sufficient and necessary condition for a general total reward stochastic game to have a saddle point when both players are *restricted to stationary strategies*. In case of MDPs, this amounts to the feasibility of a linear program of the form that will be described in Section 3. In this section, we show that the existence of a solution for this LP implies in fact the existence of an optimal solution in positional strategies, even if each player is allowed to choose from the space of all, possibly history-dependent, strategies. Our proof relies heavily on the concept of a *potential transformation* and relating the total and mean effective payoffs of a transformed game to those in the original game.

2.1 Potential transformation

Let us consider a mapping $x : V \to \mathbb{R}$, whose values $x(v)$ will be called *potentials*, and define the transformed reward local function $r[x] : E \to \mathbb{R}$ as:

$$r[x](u, v) = r(u, v) - x(u) + x(v), \quad \text{where} \quad (u, v) \in E. \tag{4}$$

Potential transforms were first introduced in 1958 by Gallai [11], then applied to stochastic games in 1966 by Hoffman and Karp [15] and to B-games (that is, min mean-cycles) in 1978 by Karp [18].

Given a potential transformation x, and an MDP $\Gamma = (G, p, r)$, let us denote by $\phi[x]$ (similarly, $\psi[x]$) the optimal effective payoff vectors in the transformed MDP $\Gamma[x] = (G, p, r[x])$. Let us further associate to such a potential vector the quantity $\mathcal{M}(x) = 2 \max_{v \in V} |x(v)|$.

Let us also introduce

$$\widehat{\phi}_{\mathfrak{s}}[x](v_0) = \limsup_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r[x](v_t, v_{t+1})], \quad \text{and}$$

$$\widehat{\psi}_{\mathfrak{s}}[x](v_0) = \limsup_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r[x](v_i, v_{i+1})],$$

and for $x = \mathbf{0}$ write $\widehat{\phi}_{\mathfrak{s}}[0](v_0) = \widehat{\phi}_{\mathfrak{s}}(v_0)$, and analogously $\widehat{\psi}_{\mathfrak{s}}[0](v_0) = \widehat{\psi}_{\mathfrak{s}}(v_0)$.

▶ Fact 1 (see, e.g., [6]). For any MDP Γ, there exists a potential y such that, if $v_0 = v \in V$ is the initial vertex and $t \in \mathbb{Z}_+$, then

(i) $\mathbb{E}_{\mathfrak{s}}[r[y](v_t, v_{t+1})] \leq \phi_\Gamma(v)$ for any arbitrary strategy $\mathfrak{s}$;

(ii) $\mathbb{E}_{\mathfrak{s}^*}[r[y](v_t, v_{t+1})] = \phi_\Gamma(v)$ for some positional strategy $\mathfrak{s}^*$.

2.2 Characterization of pure and stationary optima

Let us start with a few useful properties connecting mean payoff and total payoff values.

▶ **Lemma 3.** *If for a strategy $\mathfrak{s} \in \mathfrak{S}$ and initial vertex $v_0 \in V$ we have $\phi_{\mathfrak{s}}(v_0) > 0$, then $\psi_{\mathfrak{s}}(v_0) = +\infty$. Analogously, if we have $\widehat{\phi}_{\mathfrak{s}}(v_0) < 0$, then $\widehat{\psi}_{\mathfrak{s}}(v_0) = -\infty$.*

▶ **Lemma 4.** *Assume that $\sup_{\mathfrak{s}} \phi_{\mathfrak{s}}(v) \leq 0$ for all $v \in V$, and denote by y a corresponding potential transformation as in Fact 1. Then we have the following relations hold for all strategies $\mathfrak{s} \in \mathfrak{S}$ and initial vertices $v_0 \in V$:*

$$\phi_{\mathfrak{s}}(v_0) = \phi_{\mathfrak{s}}[y](v_0) \leq 0, \tag{5a}$$

$$\widehat{\phi}_{\mathfrak{s}}(v_0) = \widehat{\phi}_{\mathfrak{s}}[y](v_0) \leq 0, \tag{5b}$$

and

$$\psi_{\mathfrak{s}}(v_0) \leq \widehat{\psi}_{\mathfrak{s}}(v_0) \leq \mathcal{M}(y) < \infty. \tag{5c}$$

▶ **Lemma 5.** *Assume that $\sup_{\mathfrak{s}} \phi_{\mathfrak{s}}(v) \leq 0$ for all $v_0 \in V$. Then if $\phi_{\mathfrak{s}}(v_0) < 0$ for a strategy $\mathfrak{s} \in \mathfrak{S}$, then $\widehat{\psi}_{\mathfrak{s}}(v_0) = -\infty$.*

The following corollary of Lemma 3 and Fact 1 states that the total payoff in an MDP is not finite if the mean payoff is not zero.

▶ **Corollary 6.** *For an MDP and a node u, we have*

$$\begin{aligned} \phi_\Gamma(u) > 0 &\implies \psi_\Gamma(u) = \widehat{\psi}_\Gamma(u) = +\infty, \\ \phi_\Gamma(u) < 0 &\implies \psi_\Gamma(u) = \widehat{\psi}_\Gamma(u) = -\infty. \end{aligned}$$

▶ **Lemma 7.** *Assume that $\sup_{\mathfrak{s}} \phi_{\mathfrak{s}}(v) \leq 0$ for all $v \in V$, and that $\mathfrak{s} \in \mathfrak{S}$ is a strategy with initial vertex v_0 such that $\psi_{\mathfrak{s}}(v_0)$ is finite. Then we have*

$$\phi_{\mathfrak{s}}(v_0) = \widehat{\phi}_{\mathfrak{s}}(v_0) = 0.$$

For brevity, we will use the following notation throughout the rest of this section: Given a mapping $f : E(u) \to \mathbb{R}$ and a subset $F \subseteq E(u)$ we write

$$M_{(u,v)\in F}[f] = \begin{cases} \max_{(u,v)\in F} f(u,v), & \text{for } u \in V_W, \\ \operatorname{avg}_{(u,v)\in F} f(u,v), & \text{for } u \in V_R, \end{cases}$$

where $\operatorname{avg}_{(u,v)\in F}(f(v,u)) := \sum_{(u,v)\in F} p(u,v)\, f(u,v)$.

▶ **Theorem 8.** *For a total reward MDP $\Gamma = (G, P, r)$, the following two statements are equivalent:*

(i) *the value vector ψ_Γ exists, is finite, and* Max *possesses a uniformly optimal positional strategy (optimal among all strategies);*

(ii) *the following set of equations has a (finite) solution for variables* $\mu, x \in \mathbb{R}^V$, $\alpha \in \mathbb{R}_+$*:*

$$\mu(u) = M_{(u,v)\in E(u)}[r(u,v) + \mu(v)] \qquad \text{for all } u \in V, \tag{6a}$$
$$\mu(u) = M_{(u,v)\in E(u)}[\alpha r(u,v) + x(v) - x(u)] \qquad \text{for all } u \in V, \tag{6b}$$
$$\mu(u) = M_{(u,v)\in \mathrm{EXT}(u)}[\alpha r(u,v) + x(v) - x(u)] \qquad \text{for all } u \in V_W, \tag{6c}$$

where, for a vertex $u \in V_W$, $\mathrm{EXT}(u)$ *denotes the set of arcs in* $E(u)$ *attaining equality in (6a).*

Let us remark that the series of Lemmas we used to prove the above theorem remain true if we replace in the definitions of ϕ and ψ the operator lim inf with lim sup. Thus, Theorem 8 also holds with this modified definition, too. Consequently, switching the controller to a "minimizer" an analogous theorem will hold, since we can obtain this situation by changing the sign of all local rewards, switching to lim sup in the definitions of ϕ and ψ, and then applying the above theorem with a "maximizer." This observation will be useful for using the "symmetry" between the players in proving Theorem 2.

3 LP formulation

Our purpose in this section is to show that in a total reward MDP, the optimal solution can always be realized by a positional strategy that can be obtained in polynomial time. One of the main ingredients in this proof is the treatment of the case when

$$\phi_\Gamma(u) = 0 \quad \forall u \in V. \tag{A}$$

In this section we shall assume that the above condition holds, and show that in this case the optimal solution can be obtained via solving a small series of linear programs. To arrive to the proof of this statement, we need a series of technical lemmas.

Based on the idea of [28] let us associate to Γ the following linear programming problem $\mathrm{LP}(\alpha)$, where $\alpha \in \mathbb{R}$ is a real parameter. Recall that $E(u) = \{(u,v) \in E \mid v \in N^+(u)\}$, where $N^+(u)$ is the set of out-neighbors of vertex u.

$$\min \sum_{u\in V} y(u)$$
$$s.t. \tag{7a}$$
$$y(u) \geq r(u,v) + y(v) \qquad \forall\ u \in V_W,\ (u,v) \in E(u) \tag{7b}$$
$$y(u) \geq \operatorname*{avg}_{v\in N^+(u)} (r(u,v) + y(v)) \qquad \forall\ u \in V_R \tag{7c}$$
$$y(u) \geq \alpha r(u,v) - x(u) + x(v) \qquad \forall\ u \in V_W,\ (u,v) \in E(u) \tag{7d}$$
$$y(u) \geq \operatorname*{avg}_{v\in N^+(u)} (\alpha r(u,v) - x(u) + x(v)) \qquad \forall\ u \in V_R. \tag{7e}$$

The main idea is to show that this LP has an optimal solution satisfying conditions (6a)-(6c) of Theorem 8 (with $y(u) = \mu(u)$). For this we need to show that, starting from an arbitrary optimal solution (x,y), we can construct another optimal solution (x^*, y^*) such that for all $u \in V_W$, there is an arc $(u,v) \in E$ such that the inequalities (7b) and (7d), corresponding to this arc, are tight at (x^*, y^*).

Given a feasible solution (x,y) of $\mathrm{LP}(\alpha)$, let us denote by $I^u(y)$ the set of arcs $(u,v) \in E(u)$ for which (7b) holds with equality, and let $J^u_\alpha(x,y)$ denote the set of arcs $(u,v) \in E(u)$ for which (7d) is an equality. Furthermore, let us denote by $I^R(y)$ the set of vertices $u \in V_R$ for

which (7c) holds with equality, and let $J^R_\alpha(x,y)$ denote the set of vertices $u \in V_R$ for which (7e) holds with equality.

In view of Theorem 8, it will be enough to show the following:

▶ **Theorem 9.** *Under Assumption (A), if $\alpha > 0$ is large enough then $LP(\alpha)$ has an optimal solution (x^*, y^*) such that*

$$\emptyset \neq J^u_\alpha(x^*,y^*) \subseteq I^u(y^*) \text{ and } J^R_\alpha(x^*,y^*) = I^R(y^*) = V_R.$$

To arrive to the proof of this claim, we need several technical lemmas. Let us first show that this linear program has a finite optimum whenever α is nonnegative. We break this claim into two lemmas:

▶ **Lemma 10.** *Problem $LP(\alpha)$ is feasible, if $\alpha \geq 0$.*

▶ **Lemma 11.** *Problem $LP(\alpha)$ is bounded.*

Let us then denote by $Z(\alpha)$ the optimum value of $\mathrm{LP}(\alpha)$.

▶ **Corollary 12.** *The value $Z(\alpha)$ exists and is finite for all $\alpha \geq 0$.*

Strengthening Lemma 11, we can get an explicit lower bound on $Z(\alpha)$, of *polynomial bit-length* in terms of the input size (assuming rational input), as follows.

▶ **Lemma 13.** *For any feasible solution (x,y) in $LP(\alpha)$, $\alpha \geq 0$, and for any vertex $u \in V$ and any strategy $\mathfrak{s} \in \mathfrak{S}$, we have $y(u) \geq \psi_{\mathfrak{s}}(u)$.*

▶ **Corollary 14.** *There exists a real $L \in \mathbb{R}$ such that we have $L \leq Z(\alpha)$ for all $\alpha \geq 0$.*

Proof. By Lemma 13 we have $\sum_{u\in V} y(u) \geq \sum_{u\in V} \psi_{\mathfrak{s}}(u)$, for all feasible solutions (x,y) of $LP(\alpha)$, $\alpha \geq 0$ and for any strategy $\mathfrak{s} \in \mathfrak{S}$. Let us now fix a uniformly optimal stationary strategy $\mathfrak{s}$ of the mean-payoff MDP (which we know to exist, see, e.g., [23]). It was shown in [28] that under the assumption (A) we have $\psi_{\mathfrak{s}}(u)$ finite for all vertices $u \in V$ (see Proposition 1 in Section 5.2 for an explicit formula). Consequently, $L = \sum_{u\in V} \psi_{\mathfrak{s}}(u)$ is a finite lower bound (of polynomial bit-length) on the objective function value of any feasible solution of $LP(\alpha)$ for any $\alpha \geq 0$. ◀

▶ **Lemma 15.** *There exists a finite real α_0 (of polynomial bit-length in terms of the input size), such that $Z(\alpha) = Z(\alpha_0)$ for all $\alpha > \alpha_0$.*

▶ **Lemma 16.** *Let us consider $\alpha \geq \alpha_0$ and denote by (x^*, y^*) an arbitrary optimal solution of $LP(\alpha)$. Then, we have $I^R(y^*) = V_R$ and $I^u(y^*) \neq \emptyset$ for all $u \in V_W$.*

To arrive to a proof of Theorem 9, which is the main aim of this section, it will not be enough simply to take an optimal solution of $LP(\alpha)$ for a large enough value of α, e.g., for $\alpha \geq \alpha_0$. While the optimal values in y^* will be indeed optimal in the MDP, the additional conditions of Theorem 9 call for a careful selection of an optimal x^*. In fact $LP(\alpha)$ typically has many optimal solutions, even if we fix the values in y^*, and the rest of the proof will focus on showing how can we find efficiently an appropriate x^* satisfying all conditions of Theorem 9.

To this end let us fix an optimal solution (x^*, y^*) of $LP(\alpha)$ for some $\alpha \geq \alpha_0$, and consider the polyhedron $X_\alpha(y^*)$ defined as the set of feasible $x \in \mathbb{R}^V$ vectors in the following system of inequalities:

$$\begin{aligned} 0 &\geq \alpha r(u,v) - y^*(u) - x(u) + x(v) && \forall\, u \in V_W,\ (u,v) \in I^u(y^*) \\ 0 &\geq \operatorname*{avg}_{v\in N^+(u)} (\alpha r(u,v) - y^*(u) - x(u) + x(v)) && \forall\, u \in V_R. \end{aligned}$$

Note that out of the inequalities of (7d) we have included only those to which the corresponding inequalities in (7b)-(7c) are tight at y^*. Since $x^* \in X_\alpha(y^*)$, this set is a nonempty, closed convex set.

▶ **Lemma 17.** *For all $x \in X_\alpha(y^*)$ there exists a finite $\Delta(x) \geq 0$ such that $(x + \Delta y^*, y^*)$ is feasible in $LP(\alpha + \Delta)$ for all $\Delta \geq \Delta(x)$.*

Given a vector $x \in X_\alpha(y^*)$ let us call a vertex $u \in V_R$ *tight* if $u \in J^R_\alpha(x, y^*)$. Analogously, we call a vertex $u \in V_W$ tight if $0 = \alpha r(u,v) - y^*(u) - x(u) + x(v)$ for some arc $(u,v) \in I^u(y^*)$. Let us finally denote by $T(x)$ the set of tight vertices. We will be done if we show that there is a potential vector $x \in X_\alpha(y^*)$ such that $T(x) = V$, and which can be found by linear programming.

Let us define the set of vertices which belong to all tight sets:

$$U \;=\; \bigcap_{x \in X_\alpha(y^*)} T(x).$$

▶ **Lemma 18.** *If $\alpha \geq \alpha_0$, then $U \neq \emptyset$.*

▶ **Lemma 19.** *For all vertices $w \in V$ we can test if $w \in U$, and if not, find $x_w \in X_\alpha(y^*)$ such that $w \notin T(x_w)$ in polynomial time.*

▶ **Corollary 20.** *For each $\alpha \geq 0$ we can find the set $U \subseteq V$, and a vector $\overline{x} \in X_\alpha(y^*)$ such that $U = T(\overline{x})$ in polynomial time.*

▶ **Lemma 21.** *For all $x \in X_\alpha(y^*)$ and for all $v \notin T(x)$ there exists a small $\epsilon > 0$ such that for the vector*

$$x'(u) = \begin{cases} x(u) & \text{if } u \neq v, \\ x(u) - \epsilon & \text{if } u = v \end{cases}$$

we have $x' \in X_\alpha(y^*)$.

We shall prove next, with the above lemma in mind, that there exists a vector in $X_\alpha(y^*)$, if $\alpha \geq \alpha_0$, at which all vertices are tight. To this end let us consider the set U and the vector $\overline{x}$, as in Corollary 20, and the following linear programming problem:

$$\max \sum_{u \in V} z(u) \;\; s.t. \;\; (\overline{x} - z) \in X_\alpha(y^*), \; z \geq 0, \; z(u) = 0 \quad \forall\, u \in U. \tag{LPZ}$$

Let us note that in this linear program α, r, y^*, and $\overline{x}$ are all constants, just like the $z(u) = 0$ values for $u \in U$, and hence $z(v)$ for $v \in V \setminus U$ are the only variables.

▶ **Lemma 22.** *If $\alpha \geq \alpha_0$ then problem* (LPZ) *has a finite optimum.*

▶ **Corollary 23.** *If $\alpha \geq \alpha_0$, and $\overline{z}$ is an optimum solution of* (LPZ)*, then $T(\overline{x} - \overline{z}) = V$.*

Proof of Theorem 9. For an $\alpha' \geq \alpha_0$, let y^* be optimal in $LP(\alpha')$, let $\overline{x}$ be as in Corollary 20 and $\overline{z}$ as in Corollary 23, and define $x^* = \overline{x} - \overline{z} + \Delta(\overline{x} - \overline{z})y^*$ and $\alpha = \alpha' + \Delta(\overline{x} - \overline{z})$. Then, by Lemma 17 and Corollary 23 it follows that (x^*, y^*) is an optimal solution in $LP(\alpha)$, satisfying all conditions of the theorem. ◀

4 General MDPs

In this section we extend the result of the previous section to the more general case when $\phi_\Gamma(u) \neq 0$ for some $u \in V$.

▶ **Lemma 24.** *Let u denote a vertex with $\phi_\Gamma(u) \leq 0$, and let $\mathfrak{s}$ denote a strategy in $\mathfrak{S}$. If, starting from initial vertex $v_0 = u$ strategy $\mathfrak{s}$ uses with positive probability an arc (v, w) such that $v \in V_W$ and $0 \geq \phi_\Gamma(v) > \phi_\Gamma(w)$, then we have $\psi_{\mathfrak{s}}(v_0) = -\infty$.*

Let us introduce a new MDP $\Gamma' = (G' = (V, E'), p, r')$ obtained from $\Gamma = (G, p, r)$ as follows:

1. Delete all the arcs (u, v) from G such that $u \in V_W$ and $\phi_\Gamma(u) > \phi_\Gamma(v)$
2. Define $r'(u, v) = r(u, v) - \phi_\Gamma(u)$ for all the remaining arcs (u, v).

Let us denote by E' the set of arcs of G', and by $E'(u)$ the set of arcs in G' leaving vertex $u \in V$. Clearly, $E'(u) = E(u)$ for $u \in V_R$.

Let us note that we have $\phi_\Gamma(u) = \phi_\Gamma(v)$ for all $(u, v) \in E'(u)$, $u \in V_W$, since Max could not have an arc $(u, v) \in E(u)$, $u \in V_W$ such that $\phi_\Gamma(u) < \phi_\Gamma(v)$, and all arcs going down in value are removed in Γ'. Let us also note that $\phi_{\Gamma'}(u) = 0$ for all vertices u.

It is easy to see that an optimal strategy with respect to the mean payoff function ϕ in Γ' is also optimal in Γ. We shall prove below in two lemmas that the same essentially holds in positional strategies with respect to the total payoff function ψ.

▶ **Lemma 25.** *Fix an initial vertex $v_0 = u$ such that $\phi_\Gamma(u) = 0$. Then any strategy $\mathfrak{s}$ in Γ' satisfies $\mathbb{E}_{\mathfrak{s}}[r'(v_j, v_{j+1})] = \mathbb{E}_{\mathfrak{s}}[r(v_j, v_{j+1})]$.*

Since $\phi_{\Gamma'}(u) = 0$ for all $u \in V$, Theorems 8 and 9 imply that Γ' possesses a uniformly optimal positional strategy $\mathfrak{s}^*$ with respect to the total payoff function ψ.

▶ **Lemma 26.** *$\mathfrak{s}^*$ is also optimal in Γ.*

Proof. Let u be an initial vertex. By Lemmas 24 and 25, $\mathfrak{s}^*$ is optimal in Γ, if u satisfies $\phi_\Gamma(u) = 0$. On the other hand, if $\phi_\Gamma(u) > 0$ (resp., < 0), let us note that $\phi_\Gamma(v) = \phi_\Gamma(w)$ if $v \in V_W$ and $(v, w) \in E'(v)$, and $\phi_\Gamma(v) = \operatorname{avg}_{(v,w)\in E'(v)} \phi_\Gamma(w)$ if $v \in V_R$. This implies that $\mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_t)] = \phi_\Gamma(v_0)$ for all t. Since by our construction we have, for any strategy $\mathfrak{s} \in \mathfrak{S}$,

$$\mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})] = \mathbb{E}_{\mathfrak{s}}[r'(v_t, v_{t+1})] + \mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_t)] = \mathbb{E}_{\mathfrak{s}}[r'(v_t, v_{t+1})] + \phi_\Gamma(v_0) \tag{8}$$

and $\psi_{\mathfrak{s}}(v)$ is finite for all $v \in V$, the equality $\psi_\Gamma(v_0) = +\infty$ (resp., $-\infty$) follows. Indeed, if $\phi_{\Gamma'}(v_0) > 0$, then (8) implies for $\mathfrak{s} = \mathfrak{s}^*$ that $\psi_{\mathfrak{s}^*}(v_0) = +\infty$; on the other hand, if $\phi_{\Gamma'}(v_0) < 0$, then (8) together with Lemma 24 imply for any $\mathfrak{s} \in \mathfrak{S}$ that $\psi_{\mathfrak{s}}(v_0) = -\infty$. ◀

5 Two-player zero-sum games with perfect information (BWR-games)

We now turn our attention to two-person zero-sum stochastic games with perfect information and total effective payoff.

5.1 Discounted BWR-games

Let β be a number in $\in (0,1]$ called the *discount factor*. *Discounted mean payoff* stochastic games were introduced by Shapley [26] and have payoff function:

$$\phi^{\beta}_{\mathfrak{s}}(v_0) \;=\; (1-\beta)\sum_{j=0}^{\infty}\beta^j\mathbb{E}_{\mathfrak{s}}[r_{\mathfrak{s}}(v_t,v_{t+1})], \tag{9}$$

where $a(\mathfrak{s}) = \langle \mathbb{E}_{\mathfrak{s}}[r_{\mathfrak{s}}(v_0,v_1)], \mathbb{E}_{\mathfrak{s}}[r_{\mathfrak{s}}(v_1,v_2)], \ldots\rangle$ is the sequence of expected rewards incurred at steps $0,1,\ldots$ of the play, according to the pair of strategies $\mathfrak{s} = (\mathfrak{s}_B, \mathfrak{s}_W)$.

Discounted games, in general, are easier to solve, due to the fact that a standard value iteration is in fact a fast converging contraction. Hence, they are widely used in the literature of stochastic games together with the above limit equality. In fact, for mean payoff BW-games with n vertices and *integral* rewards of maximum absolute value R it is known [32] that for two pairs of stationary strategies $\mathfrak{s}, \mathfrak{s}' \in \widehat{\mathfrak{S}}$ we have $\phi^{\beta}_{\mathfrak{s}}(u) < \phi^{\beta}_{\mathfrak{s}'}(u)$ if and only if $\phi_{\mathfrak{s}}(u) < \phi_{\mathfrak{s}'}(u)$ whenever $1-\beta \leq \frac{1}{4n^3R}$.

If the discount factor β is strictly less than 1, we obtain the following result, which follows essentially from [26].

▶ Fact 2 ([26]). A BWR-game with the discounted mean payoff function ϕ^{β} has a saddle point in uniformly optimal positional strategies, for all $0 < \beta < 1$.

We show in the next subsection that the same pair of stationary strategies form a uniform Nash equilibrium with respect to the total payoff ψ, if β is sufficiently close enough to 1.

5.2 Existence of a saddle point in positional strategies

When the mean payoff values are zero, there is an explicit formula for computing the total reward values, corresponding to a stationary strategy, as a function of the limiting probability matrix. To write this formula, we need first to introduce some notation. Given a BWR-game $\Gamma = (G,p,r)$ and a pair of positional strategies $\mathfrak{s} = (\mathfrak{s}_B, \mathfrak{s}_W)$, we obtain a weighted Markov chain $\Gamma_{\mathfrak{s}} = (P_{\mathfrak{s}}, r)$ with transition matrix $P_{\mathfrak{s}}$ in the obvious way:

$$p_{\mathfrak{s}}(u,v) = \begin{cases} 1 & \text{if } u \in V_W \cup V_B \text{ and } (u,v) \text{ is chosen by } \mathfrak{s};\\ 0 & \text{if } u \in V_W \cup V_B \text{ and } (u,v) \text{ is not chosen by } \mathfrak{s};\\ p(v,u) & \text{if } v \in V_R. \end{cases}$$

We define the *expected local reward* $r_{\mathfrak{s}} : V \to \mathbb{R}$, corresponding to the pair $\mathfrak{s}$ as

$$r_{\mathfrak{s}}(u) = \begin{cases} r(u,v) & \text{if } u \in V_W \cup V_B \text{ and } (u,v) \text{ is chosen by } \mathfrak{s};\\ \sum_{(u,v)\in E(u)} p(u,v)r(u,v) & \text{if } v \in V_R. \end{cases}$$

Finally, we will denote by $Q_{\mathfrak{s}}$ the (unique) limiting average probability matrix satisfying $Q_{\mathfrak{s}}P_{\mathfrak{s}} = P_{\mathfrak{s}}Q_{\mathfrak{s}} = Q_{\mathfrak{s}}$. Note that $\phi_{\mathfrak{s}} = Q_{\mathfrak{s}}r_{\mathfrak{s}}$ and $\phi^{\beta}_{\mathfrak{s}} = (1-\beta)(I-\beta P_{\mathfrak{s}})^{-1}r_{\mathfrak{s}}$.

▶ Proposition 1 ([28]). If $\mathfrak{s}$ is stationary strategy such that $\phi_{\mathfrak{s}} = \mathbf{0}$, then $\psi_{\mathfrak{s}} = (I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1}r_{\mathfrak{s}}$, where I is the $|V| \times |V|$ identity matrix.

To prove our main result for BWR-games (Theorem 2), it will be enough to consider games in which $\phi_{\Gamma}(u) = 0$ for all $u \in V$.

▶ **Theorem 27.** *Consider an undiscounted BWR-game Γ such that $\phi_{\Gamma} = \mathbf{0}$. Then there is a uniformly optimal pair of positional strategies $(\mathfrak{s}_B, \mathfrak{s}_W)$ satisfying:*

$$\pi_{\mathfrak{s}^*_B,\mathfrak{s}_W}(v_0) \leq \pi_{\mathfrak{s}^*_B,\mathfrak{s}^*_W}(v_0) \leq \pi_{\mathfrak{s}_B,\mathfrak{s}^*_W}(v_0) \quad \textit{for all } \mathfrak{s}_B \in \widehat{\mathfrak{S}}_B,\ \mathfrak{s}_W \in \widehat{\mathfrak{S}}_W \textit{ and for all } v_0 \in V.$$

If $|V| = n$, all rewards are integral with maximum absolute value R, and all transition probabilities are rational with maximum common denominator $D > 0$, then such a saddle point can be found by solving a discounted game with $\beta = 1 - \frac{1}{(nD)^{O(n^2)}R}$.

Proof. We start with the following claim.

▶ Claim 1. Let $\mathfrak{s} = (\mathfrak{s}_B, \mathfrak{s}_W)$ be a pair of positional strategies such that $\phi_{\mathfrak{s}}(v) = 0$ for all $v \in V$. Then, we have

$$\lim_{\beta \to 1^-} \frac{\psi_{\mathfrak{s}} - (I - \beta P_{\mathfrak{s}})^{-1} r_{\mathfrak{s}}}{1 - \beta} = P_{\mathfrak{s}}(I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-2} r_{\mathfrak{s}}.$$

$$\text{Let} \quad \gamma = \min_{u \in V} \min_{\substack{\mathfrak{s}, \mathfrak{s}' \in \widehat{\mathfrak{S}} \\ \psi_{\mathfrak{s}}(u) \neq \psi_{\mathfrak{s}'}(u)}} |\psi_{\mathfrak{s}}(u) - \psi_{\mathfrak{s}'}(u)| \quad \text{and} \quad \kappa = \max_{u \in V} \max_{\mathfrak{s} \in \widehat{\mathfrak{S}}} |P_{\mathfrak{s}}(I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-2} r_{\mathfrak{s}}|.$$

Standard estimation arguments (see, e.g., [5]) give $\gamma \geq \frac{1}{(nD)^{O(n^2)}}$ and $\kappa \leq (nD)^{O(n^2)}R$.

Claim 1 implies that, for any sufficiently small $\epsilon > 0$, there exists a $\beta(\epsilon) \in (0, 1)$ such that, for all pairs of positional strategies $\mathfrak{s} \in \widehat{\mathfrak{S}}$, we have

$$\|(1 - \beta(\epsilon))\psi_{\mathfrak{s}} - \phi_{\mathfrak{s}}^{\beta(\epsilon)}\|_\infty < (1 - \beta(\epsilon))^2(\epsilon + \kappa) \leq 2(1 - \beta(\epsilon))^2\kappa. \tag{10}$$

Let us choose ϵ such that $\beta(\epsilon) > 1 - \frac{\gamma}{4\kappa}$. Then for any two pairs of positional strategies $\mathfrak{s}, \mathfrak{s}' \in \widehat{\mathfrak{S}}$, such that $\psi_{\mathfrak{s}}(u) > \psi_{\mathfrak{s}'}(u)$, we have $\psi_{\mathfrak{s}}(u) - \psi_{\mathfrak{s}'}(u) \geq \gamma$. On the other hand, by (10), we get

$$\left|(1 - \beta(\epsilon))\psi_{\mathfrak{s}}(u) - \phi_{\mathfrak{s}}^{\beta(\epsilon)}(u)\right| < 2(1-\beta(\epsilon))^2\kappa \text{ and } \left|(1 - \beta(\epsilon))\psi_{\mathfrak{s}'}(u) - \phi_{\mathfrak{s}'}^{\beta(\epsilon)}(u)\right| < 2(1-\beta(\epsilon))^2\kappa.$$

Consequently, by our choice of ϵ, $\phi_{\mathfrak{s}}^{\beta(\epsilon)} > \phi_{\mathfrak{s}'}^{\beta(\epsilon)}$ follows, proving the claim of the theorem. ◀

Proof of Theorem 2. First assume that $\phi_\Gamma(u) = 0$ for all $u \in V$. Then Theorem 27 implies the existence of saddle point $\mathfrak{s}^* = (\mathfrak{s}_B^*, \mathfrak{s}_W^*)$, among uniformly optimal positional strategies. Since, by Theorem 1, the best response in the MDP obtained by fixing Max's strategy to $\mathfrak{s}_W^*$ (resp., Min's strategy to $\mathfrak{s}_B^*$) is positional, it follows that $\mathfrak{s}^*$ is a saddle point among all strategies of the two players. The case when $\phi_\Gamma(u) \neq 0$ for some $u \in V$ is handled using the same approach used in Section 4. ◀

References

1 D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1987.

2 D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

3 D. P. Bertsekas and H. Yuz. Stochastic shortest path problems, under weak conditions, lids report 2909. Technical report, MIT, 2013.

4 D. Blackwell. Discrete dynamic programming. *Ann. Math. Statist.*, 33:719–726, 1962.

5 E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. A pumping algorithm for ergodic stochastic mean payoff games with perfect information. In *Proc. 14th IPCO*, volume 6080 of *LNCS*, pages 341–354. Springer, 2010.

6 E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. On canonical forms for zero-sum stochastic mean payoff games. *Dynamic Games and Applications*, 3(2):128–161, 2013.

7 C. Derman. *Finite State Markov decision processes.* Academic Press, New York and London, 1970.

8 J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer, Berlin, 1996.
9 O. Friedmann, T. D. Hansen, and U. Zwick. Subexponential lower bounds for randomized pivoting rules for the simplex algorithm. In *STOC*, pages 283–292, 2011.
10 D.R. Fulkerson and G.C. Harding. Maximizing the minimum source-sink path subject to a budget constraint. *Mathematical Programming*, 13:116–118, 1977.
11 T. Gallai. Maximum-minimum Sätze über Graphen. *Acta Mathematica Academiae Scientiarum Hungaricae*, 9:395–434, 1958.
12 D. Gillette. Stochastic games with zero stop probabilities. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contribution to the Theory of Games III*, volume 39 of *Annals of Mathematics Studies*, pages 179–187. Princeton University Press, 1957.
13 V.A. Gurvich, A.V. Karzanov, and L.G. Khachiyan. Cyclic games and an algorithm to find minimax cycle means in directed graphs. *USSR Comput. Math. Math. Phys.*, 28:85–91, 1988.
14 O. O. Hernández-Lerma and J.-B. Lasserre. *Further topics on discrete-time Markov control processes*. Applications of mathematics. Springer, New York, 1999.
15 A. J. Hoffman and R. M. Karp. On non-terminating stochastic games. *Management Science*, 12:359–370, 1966.
16 R. A. Howard. *Dynamic programming and Markov processes*. Technology press and Willey, New York, 1960.
17 E. Israeli and R. K. Wood. Shortest-path network interdiction. *Networks*, 40(2):97–111, 2002.
18 R. M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Math.*, 23:309–311, 1978.
19 A. V. Karzanov and V. N. Lebedev. Cyclical games with prohibition. *Mathematical Programming*, 60:277–293, 1993.
20 L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, V. Gurvich, G. Rudolf, and J. Zhao. On short paths interdiction problems: Total and node-wise limited interdiction. *Theory Comput. Syst.*, 43(2):204–233, 2008.
21 L. Khachiyan, V. Gurvich, and J. Zhao. Extending dijkstra's algorithm to maximize the shortest path by node-wise limited arc interdiction. In *CSR*, pages 221–234, 2006.
22 T. M. Liggett and S. A. Lippman. Stochastic games with perfect information and time-average payoff. *SIAM Review*, 4:604–607, 1969.
23 H. Mine and S. Osaki. *Markovian decision process*. American Elsevier Publishing Co., New York, 1970.
24 R. H. Möhring, M. Skutella, and F. Stork. Scheduling with and/or precedence constraints. *SIAM J. Comput.*, 33(2):393–415, 2004.
25 S. D. Patek and D. P. Bertsekas. Stochastic shortest path games. *SIAM Journal on Control and Optimization*, 37:804–824, 1997.
26 L. Shapley. Stochastic games. *Proc. Nat. Acad. Sci. USA*, 39:1095–1100, 1953.
27 F. Thuijsman and O. J. Vrieze. The bad match, a total reward stochastic game. *Operations Research Spektrum*, 9:93–99, 1987.
28 F. Thuijsman and O. J. Vrieze. Total reward stochastic games and sensitive average reward strategies. *Journal of Optimization Theory and Applications*, 98:175–196, 1998.
29 P. Whittle. *Optimization over Time*. John Wiley & Sons, Inc., New York, NY, USA, 1982.
30 H. Yu and D. P. Bertsekas. Q-learning and policy iteration algorithms for stochastic shortest path problems. *Annals OR*, 208(1):95–132, 2013.
31 H. Yuz. Stochastic shortest path games and q-learning, lids report 2875. Technical report, MIT, 2011.
32 U. Zwick and M. Paterson. The complexity of mean payoff games on graphs. *Theoretical Computer Science*, 158(1-2):343 – 359, 1996.