RUTCOR
RESEARCH
REPORT

Markov Decision Processes and Stochastic Games with Total Effective Payoff [a]

Endre Boros[b] Khaled Elbassioni[c]
Vladimir Gurvich[b] Kazuhisa Makino[d]

RRR 04-2014, October 2014

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]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).

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

[c]Masdar Institute of Science and Technology, P.O. Box 54224, Abu Dhabi, UAE; (kelbassioni@masdar.ac.ae)

[d]Research Institute for Mathematical Sciences (RIMS) Kyoto University, Kyoto 606-8502, Japan; (makino@kurims.kyoto-u.ac.jp)

Rutcor Research Report
RRR 04-2014, October 2014

Markov Decision Processes and Stochastic Games with Total Effective Payoff [1]

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.

1 Introduction

1.1 Basic concepts

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 $p(u, v) \geq 0$ 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.

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*. We shall denote by $\mathfrak{S}$ the set of all possible strategies and by $\widehat{\mathfrak{S}}$ the set of all pure stationary 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 $t = 0, 1 \ldots$

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 [13, 3]. The second one, called *total payoff* or *total reward*, was introduced by Thuijsman and Vrieze [30], as a "refinement" of the mean payoff. Let us note however that in fact we can show total payoff MDPs to include mean payoff MDPs as a special case (see Lemma 1).

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 [6, 10, 28, 2, 34, 33]; 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.

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

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

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 [15], and it was shown in [5] that it is polynomially equivalent with stochastic games with perfect information [13]. For $\pi = \psi$, this

model is the same as the one introduced in [30] in case of perfect information. The concept was further developed in [9, 31].

1.2 Main results

We first consider total-payoff MDP's and prove the following result.

Theorem 1 *In every MDP with total effective payoff, $\pi = \psi$,* MAX *possesses a uniformly optimal, pure and stationary 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. [18, 3, 6, 26]. 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. [6].

Our approach for proving Theorem 1 is inspired by a result of [31]. 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 1 *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 [20]; we omit the details from this version.*

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

For the mean payoff games with perfect information the above result is well-known [13, 25].

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 pure and stationary strategies. This latter can be shown by using the discounted formulation of the game.

1.3 Applications of the total payoff

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

$$\theta_{\mathfrak{s}}(v_0) \;=\; \liminf_{T\to\infty} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})]. \tag{4}$$

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, 7, 2, 6, 16, 28, 32, 34, 33]. 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.

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 [11]; see a short survey by Israely and Wood [19]. 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. This problem is NP-hard; moreover the inapproximability bound $10\sqrt{5} - 21 \approx 1.36$ was derived in [23] (from the same bound for the Minimum Vertex Cover Problem in graphs obtained by Dinur and Safra [8]).

Unlike the above *total budget* SPIP, the following *vertex-wise budget* SPIP is tractable [24, 23]. In this case, we are given a budget allowing to eliminate (at most) $k(v)$ arcs going

from each state $v \in V$. The problem is reduced to a zero-sum "pseudo-total" payoff game with non-negative local rewards.

The case of non-negative weights (local costs) was considered in [24], 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, we obtain another interesting relation: In this case, the SPIP becomes equivalent [24] with solving the zero-sum mean payoff BW-games. Although the latter problem is known to be in the intersection of NP and co-NP [22, 35], yet, it is not known to be polynomial.

It is also worth noting that BW mean payoff games form a special case of vertex-wise interdiction problems corresponding to $k(v) = 0$ for $v \in V_B$ and $k(v) = outdeg(v) - 1$ for $v \in V_W$. Indeed, White, the maximizer, is allowed to choose any move in a position $v \in V_W$ and cannot restrict the choice of Black in any $v \in V_B$.

Scheduling with and/or precedence constraints [27] 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 games (no random nodes) which have nonnegative local rewards. For this problem [27] provides a polynomial time algorithm.

2 Total rewards generalize mean payoffs

In this section we show that mean payoff games form a special case of total reward games, by a constructive embedding that roughly doubles the size of the underlying graph.

Let us consider a BWR game $\Gamma = (G = (V, E), p, r)$. Let us subdivide every arc $(u, v) \in E$ by a new vertex $w = w_{u,v}$ into two arcs (u, w) and (w, v), denote the obtained digraph $G' = (V', E')$, and define new local rewards by $r'(u, w) = r(u, v)$ and $r'(w, v) = -r(u, v)$. We set $V' = V \cup \{w_{u,v} \mid (u, v) \in E\}$ and $E' = \{(u, w_{u,v}), (w_{u,v}, v) \mid (u, v) \in E\}$. Note that in G' every "new" vertex has only a single outgoing arc, hence it does not matter which player controls these vertices. To make unique definition we define $V'_W = V_W \cup \bigcup_{(u,v)\in E}\{w_{u,v}\}$, $V'_B = V_B$ and $V'_R = V_R$. Finally we define $p'(u, w_{u,v}) = p_{u,v}$ for all $u \in V_R$, $(u, v) \in E$. In this way we obtained a new BWR game $\Gamma' = (G', p', r')$. Since in the "new" nodes MAX

does not have any choice, every strategy of Γ has a unique corresponding strategy in Γ', and conversely, every strategy of Γ' defines a unique strategy in Γ.

Lemma 1 *The equality $\psi_{\Gamma'}(v_0) = \frac{1}{2}\phi_{\Gamma}(v_0)$ holds for all $v_0 \in V$.*

Proof Let us consider an arbitrary pair of strategies $\mathfrak{s}' = (\mathfrak{a}', \mathfrak{b}')$ in Γ', and let $\mathfrak{s} = (\mathfrak{s}_B, \mathfrak{s}_W)$ be the corresponding pair of strategies in Γ. By the above construction we have $\mathbb{E}_{\mathfrak{s}}(r(v_t, v_{t+1})) = \mathbb{E}_{\mathfrak{s}'}(r(v_{2t}, v_{2t+1})) = -\mathbb{E}_{\mathfrak{s}'}(r(v_{2t+1}, v_{2t+2}))$ for all $t = 0, 1 \ldots$. Hence, if in Γ the sequence of the expected rewards is $(\mathbb{E}_{\mathfrak{s}}(r(v_0, v_1), \mathbb{E}_{\mathfrak{s}}(r(v_1, v_2), \ldots) = (a_0, a_1, \ldots)$, then the in Γ' the corresponding sequence of local rewards is $(a_0, -a_0, a_1, -a_1, \ldots)$, and hence, the accumulated rewards are $(a_0, 0, a_1, 0, \ldots)$, which implies the formula $\psi_{\mathfrak{s}'}(v_0) = \frac{1}{2}\phi_{\mathfrak{s}}(v_0)$. Since this holds for all pairs of strategies, the claim follows. □

Remark 2 *In the non-zero sum case Nash equilibria may fail to exist. An example of a two-person non-zero sum BW game with the mean effective payoff and without Nash equilibria in pure stationary strategies was obtained in [14]. The above embedding immediately extends this example to the case of the total effective payoffs. Whether Nash equilibria exist in general (history dependent) pure strategies is an open problem.*

3 Characterization of pure stationary optima in total MDPs

3.1 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:

$$r[x](u, v) = r(u, v) - x(u) + x(v), \quad \text{where } (u, v) \in E. \tag{5}$$

Potential transforms were first introduced in 1958 by Gallai [12], then applied to stochastic games in 1966 by Hoffman and Karp [17] and to B-games in 1978 by Karp [21].

Given a potential transformation x, let us denote by $\phi[x]$ (similarly, $\psi[x]$) the optimal effective payoff vectors in the transformed game $\Gamma[x] = (G, p, r[x])$. Let us further associate to such a potential vector the quantity

$$M(x) \;=\; 2 \max_{v \in V} |x(v)|.$$

Let us also introduce

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

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 $x = 0$ 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., [5]) *There exists a potential y such that, if $v \in V$ is the initial vertex and $t \in \mathbb{Z}_+$, then*

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

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

3.2 Characterization of pure and stationary optima

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

Lemma 2 *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 $\widehat{\psi}_{\mathfrak{s}}(v_0) = -\infty$.*

Proof Assume that $\psi_{\mathfrak{s}}(v_0) \ge \delta > 0$. Then, by the definition of liminf we have a finite threshold T_δ such that

$$\frac{1}{t+1} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})] \;\ge\; \delta \quad \text{for all} \quad t \ge T_\delta.$$

Using $A = \sum_{t=0}^{T_\delta} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})]$ we can write

$$\begin{aligned}
&\frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})] \\
&\qquad = \frac{A}{T+1} + \frac{1}{T+1} \sum_{t=T_\delta+1}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})] \\
&\qquad \ge \frac{A}{T+1} + \frac{1}{T+1} \sum_{t=T_\delta+1}^{T} (t+1)\delta.
\end{aligned}$$

While the nominator of the first term above is bounded, and hence the first term goes to zero as $T \to \infty$, the last term is unbounded, proving our claim.

The second part of the lemma can be shown analogously, interchanging liminf by limsup, reversing the sign of δ and the inequalities above. □

Lemma 3 *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* $\mathfrak{s}$ *and initial vertices* $v_0 \in V$*:*

$$\phi_{\mathfrak{s}}(v_0) = \phi_{\mathfrak{s}}[y](v_0) \ \leq \ 0, \tag{6a}$$

$$\widehat{\phi}_{\mathfrak{s}}(v_0) = \widehat{\phi}_{\mathfrak{s}}[y](v_0) \ \leq \ 0, \tag{6b}$$

and

$$\psi_{\mathfrak{s}}(v_0) \ \leq \ \widehat{\psi}_{\mathfrak{s}}(v_0) \ \leq \ M(y) \ < \ \infty. \tag{6c}$$

Proof By Fact (5) we can write

$$\sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})] \ = \ \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r[y](v_t, v_{t+1})] \ + \ \left(y(v_0) - \mathbb{E}_{\mathfrak{s}}[y(v_{T+1})]\right). \tag{7}$$

Since

$$|y(v_0) - \mathbb{E}_{\mathfrak{s}}[y(v_{T+1})]| \ \leq \ M(y)$$

independently of T and $\mathfrak{s}$, (6a) and (6b) follow, because $\mathbb{E}_{\mathfrak{s}}[r[y](v_t, v_{t+1})] \leq 0$ for all indices t by Fact 1.

Let us write next, using (5) that

$$\frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})] \ = \ \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r[y](v_i, v_{i+1})] + y(v_0) - \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[y(v_{t+1})]. \tag{8}$$

Here we again have

$$\left| y(v_0) - \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[y(v_{t+1})] \right| \ \leq \ M(y) \tag{9}$$

independently of T and $\mathfrak{s}$, thus (6c) follows by Fact 1, that is by the fact that $\mathbb{E}_{\mathfrak{s}}[r[y](v_i, v_{i+1})] \leq 0$ for all indices i. □

Lemma 4 *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* $\mathfrak{s}$*, then* $\widehat{\psi}_{\mathfrak{s}}(v_0) = -\infty$*.*

Proof Let us denote by y a potential as in Fact 1. If $\phi_{\mathfrak{s}}(v_0) \leq -\delta < 0$, then $\phi_{\mathfrak{s}}[y](v_0) \leq -\delta < 0$ follows by (6a), and thus there exists an infinite sequence $T_1 < T_2 < \cdots$ such that

$$\sum_{i=0}^{T_k} \mathbb{E}_{\mathfrak{s}}[r[y](v_i, v_{i+1})] \leq -\delta T_k, \quad \text{for all } k = 1, 2, ...$$

Since $\mathbb{E}_{\mathfrak{s}}[r[y](v_i, v_{i+1})] \leq 0$ holds for all $t = 0, 1, ...$ by Fact 1, it follows that

$$\sum_{i=0}^{T} \mathbb{E}_{\mathfrak{s}}[r[y](v_i, v_{i+1})] \leq -\delta T_k, \quad \text{for all } T \geq T_k \text{ and for all } k = 1, 2, ...$$

from which

$$\widehat{\psi}_{\mathfrak{s}}(v_0) \leq M(y) - \delta T_k$$

follows for all $k = 1, 2, ...$, implying $\psi_{\mathfrak{s}}(v_0) \leq \widehat{\psi}_{\mathfrak{s}}(v_0) = -\infty$. □

For an MDP Γ, payoff function π, and a node u, let us define: $\pi_\Gamma(u) = \sup_{\mathfrak{s} \in \mathfrak{S}} \pi_{\mathfrak{s}}(u)$.

The following corollary of Lemma 2 and Fact 1 states that the total payoff in an MDP is not finite if the mean payoff is not zero.

Corollary 1 *For an MDP and a node u, we have*

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

Proof The first claim follows from Lemma 2, as $\phi_\Gamma(u) > 0$ implies the existence of a strategy $\mathfrak{s} \in \mathfrak{S}$ such that $\phi_{\mathfrak{s}}(u) > 0$ which by Lemma 2 implies that $\psi_{\mathfrak{s}}(u) = +\infty$.

To see the second claim, let y be a potential transformation as in Fact 1. Then (8), (9) and (i) of Fact 1 imply that for all strategies $\mathfrak{s} \in \mathfrak{A}$ and initial position $v_0 = u$, we have

$$\limsup_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})] \leq \limsup_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \phi_\Gamma(u) + M(y) = -\infty,$$

from which $\widehat{\psi}_{\mathfrak{s}}(u) = -\infty$ follows. □

Lemma 5 *Assume that $\sup_{\mathfrak{s}} \phi_{\mathfrak{s}}(v) \leq 0$ for all $v \in V$, and that $\mathfrak{s}$ is a strategy with initial vertex v_0 such that $\psi_{\mathfrak{s}}(v_0)$ is finite. Then we have*

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

Proof Let us denote again by y a corresponding potential, as in Fact 1.

Note next that by Lemma 3, the finiteness of $\psi_{\mathfrak{s}}(v_0)$ also implies the finiteness of $\widehat{\psi}_{\mathfrak{s}}(v_0)$, and that both $\phi_{\mathfrak{s}}(v_0)$ and $\widehat{\phi}_{\mathfrak{s}}(v_0)$ are non-positive.

Now, if $\widehat{\phi}_{\mathfrak{s}}(v_0) < 0$, then $\widehat{\psi}_{\mathfrak{s}}(v_0) = -\infty$ would follow, thus we must have $\widehat{\phi}_{\mathfrak{s}}(v_0) = 0$. Furthermore, since $\psi_{\mathfrak{s}}(v_0)$ is finite, we cannot have $\phi_{\mathfrak{s}}(v_0) < 0$ by Lemma 4, and thus we must have $\phi_{\mathfrak{s}}(v_0) = 0$ too, as claimed. □

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

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

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

Theorem 3 *For a total reward MDP $\Gamma = (G, P, r)$, the following two statements are equivalent:*

(i) the value vector ψ_Γ exists, finite, and MAX possesses a stationary uniformly optimal strategy (optimal among all strategies);

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

$$\mu(u) = M_{E(u)}[r(u,v) + \mu(v)] \qquad \text{for all } u \in V, \tag{10a}$$
$$\mu(u) = M_{E(u)}[\alpha r(u,v) + x(v) - x(u)] \qquad \text{for all } u \in V, \tag{10b}$$
$$\mu(u) = M_{\mathrm{EXT}(u)}[\alpha r(u,v) + x(v) - x(u)] \qquad \text{for all } u \in V_W, \tag{10c}$$

where, for a vertex $u \in V_W$, $\mathrm{EXT}(u)$ denotes the set of arcs in $E(u)$ attaining equality in (10a).

Proof

(ii) $\Rightarrow$ (i): We use a similar argument as in [5, Appendix C]. Let x, μ, α satisfy (10a)-(10c). Let us observe first that (10a)-(10b) imply that for an arbitrary strategy $\mathfrak{s}$ we have

$$\mathbb{E}_\mathfrak{s}[\mu(v_t)] \geq \mathbb{E}_\mathfrak{s}[r(v_t, v_{t+1})] \ + \ \mathbb{E}_\mathfrak{s}[\mu(v_{t+1})] \tag{11a}$$
$$\mathbb{E}_\mathfrak{s}[\mu(v_t)] \geq \mathbb{E}_\mathfrak{s}[\alpha r(v_t, v_{t+1})] \ + \ \mathbb{E}_\mathfrak{s}[x(v_{t+1})] \ - \ \mathbb{E}_\mathfrak{s}[x(v_t)] \tag{11b}$$

for all $t = 0, 1, \ldots$. Furthermore, it follows by (11a) that

$$\phi_\mathfrak{s}(v_0) \ \leq \ \widehat{\phi}_\mathfrak{s}(v_0) \ \leq \ \limsup_{T\to\infty} \frac{1}{T+1} \left(\mu(v_0) - \mathbb{E}_\mathfrak{s}[\mu(v_{T+1})]\right) \ = \ 0. \tag{12}$$

Since the above applies to all strategies, $\sup_\mathfrak{s} \phi_\mathfrak{s}(v_0) \leq 0$ follows for all $v_0 \in V$, and thus Lemmas 3–5 can be applied.

Let us show next that $\mu(v_0) \geq \psi_\mathfrak{s}(v_0)$ holds for an arbitrary strategy $\mathfrak{s}$. Note that by Lemma 3 $\psi_\mathfrak{s}(v_0) \leq \widehat{\psi}_\mathfrak{s}(v_0) < \infty$ follows. If $\psi_\mathfrak{s}(v_0) = -\infty$, then there is nothing to prove. If

$\psi_{\mathfrak{s}}(v_0)$ is finite, then by Lemma 5 we have $\phi_{\mathfrak{s}}(v_0) = \widehat{\phi}_{\mathfrak{s}}(v_0) = 0$. Then, using (11a)-(11b) we can write

$$\begin{aligned}
\psi_{\mathfrak{s}}(v_0) &= \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} \mathbb{E}_{\mathfrak{s}}[r(v_i, v_{i+1})] \\
&\leq \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i=0}^{t} (\mathbb{E}_{\mathfrak{s}}[\mu(v_i)] - \mathbb{E}_{\mathfrak{s}}[\mu(v_{i+1})]) \qquad (13) \\
&= \mu(v_0) + \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} (-\mathbb{E}_{\mathfrak{s}}[\mu(v_{t+1})]) \\
&= \mu(v_0) - \limsup_{T\to\infty} \frac{1}{T+1} \left((\mathbb{E}_{\mathfrak{s}}[\mu(v_{T+1})] - \mu(v_0)) + \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[\mu(v_t)] \right) \\
&= \mu(v_0) - \limsup_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[\mu(v_t)] \\
&\leq \mu(v_0) - \limsup_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} (\alpha \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})] + \mathbb{E}_{\mathfrak{s}}[x(v_{t+1})] - \mathbb{E}_{\mathfrak{s}}[x(v_t)]) \qquad (14) \\
&= \mu(v_0) - \limsup_{T\to\infty} \frac{1}{T+1} \left((\mathbb{E}_{\mathfrak{s}}[x(v_{T+1})] - x(v_0)) + \alpha \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})] \right) \\
&= \mu(v_0) - \alpha \limsup_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})] \\
&= \mu(v_0) - \alpha \widehat{\phi}_{\mathfrak{s}}(v_0) \\
&= \mu(v_0)
\end{aligned}$$

concluding our proof of $\mu(v_0) \geq \psi_{\mathfrak{s}}(v_0)$.

Finally, we show that using (10a)-(10c) we can define a stationary strategy $\widehat{\mathfrak{s}} \in \widehat{\mathfrak{S}}$ such that $\mu(v_0) = \psi_{\widehat{\mathfrak{s}}}(v_0)$ holds for all $v_0 \in V$, implying that $\widehat{\mathfrak{s}}$ is a uniformly optimal stationary strategy: For $u \in V_W$, we set $\widehat{\mathfrak{s}}(u) = (u, v)$ for some arc $(u, v) \in \mathrm{EXT}(u)$ such that $\mu(u) = M_{E(u)}[\alpha r(u, v) + x(v) - x(u)]$ (such an arc exists by (10b) and (10c)). Then, it follows that for $\mathfrak{s} = \widehat{\mathfrak{s}}$ we have equalities in (11a)–(11b), and hence in (13)–(14), and thus the above derivation show that we have $\psi_{\widehat{\mathfrak{s}}}(v_0) = \mu(v_0)$ for all $v_0 \in V$, as claimed.

(i) $\Rightarrow$ (ii): The proof is the same as the one given for Theorem 5.3 in [31].

□

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,

the Theorem 3 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 apply the above theorem with a "maximizer."

4 LP formulation

Our purpose is to show that in a total reward MDP, the optimal solution can always be realized by a stationary 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 [31] let us associate to Γ the following linear programming problem LP(α), where $\alpha \in \mathbb{R}$ is a real parameter. Recall that $N^+(u)$ is the set of out-neighbors of vertex u, and that $E(u) = \{(u,v) \in E \mid v \in N^+(u)\}$.

$$\sum_{u \in V} y(u) \to \min$$

$$y(u) \geq r(u,v) + y(v) \qquad \forall\ u \in V_W,\ (u,v) \in E(u) \tag{15a}$$

$$y(u) \geq \operatorname*{avg}_{v \in N^+(u)} (r(u,v) + y(v)) \qquad \forall\ u \in V_R \tag{15b}$$

$$y(u) \geq \alpha r(u,v) - x(u) + x(v) \qquad \forall\ u \in V_W,\ (u,v) \in E(u) \tag{15c}$$

$$y(u) \geq \operatorname*{avg}_{v \in N^+(u)} (\alpha r(u,v) - x(u) + x(v)) \qquad \forall\ u \in V_R \tag{15d}$$

The main idea is to show that this LP has an optimal solution satisfying conditions (10a)-(10c) of Theorem 3 (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 (15a) and (15c), corresponding to this arc, are tight at (x^*, y^*).

Given a feasible solution (x, y) of LP(α), let us denote by $I^u(y)$ the set of arcs $(u, v) \in E(u)$ for which (15a) holds with equality, and let $J^u_\alpha(x, y)$ denote the set of arcs $(u, v) \in E(u)$ for which (15c) is an equality. Furthermore, let us denote by $I^R(y)$ the set of vertices $u \in V_R$ for which (15b) holds with equality, and let $J^R_\alpha(x, y)$ denote the set of vertices $u \in V_R$ for which (15d) holds with equality.

In view of Theorem 3, it will be enough to show the following:

Theorem 4 *If $\alpha > 0$ is large enough then LP(α) 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 6 *Problem LP(α) is feasible, if $\alpha \geq 0$.*

Proof Let us first consider the Markov Decision Process (MDP) (G, r) on graph G with arc weights r. According to our assumption (A) this MDP has value = 0 for all initial vertices, and hence there exists vertex potentials $y(v)$, $v \in V(G) = V_W$ satisfying inequalities (15a) - (15b) (see, e.g., [5]). Let us note that these inequalities remain valid if we add the same constant to all $y(u)$ potentials. Thus, we can assume, without any loss of generality that

$$y(u) \geq 0 \quad \text{for all} \quad u \in V_W. \tag{16}$$

Let us now fix such a nonegative vertex potential y, define $\widehat{r}(u, v) = \alpha r(u, v) - y(u)$ for all arcs (u, v) of G, and consider the MDP $(G, \widehat{r})$. We claim that this MDP has nonpositive (mean payoff) values form all initial vertices. To see this claim, let us fix an arbitrary strategy $\mathfrak{s}$ of MAX, and consider the play starting from v_0. Then,

$$\begin{aligned}\liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[\widehat{r}(v_t, v_{t+1})] &= \alpha \cdot \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[r(v_t, v_{t+1})] \\ &\quad - \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{\mathfrak{s}}[y(v_t)].\end{aligned}$$

Here the first term is not positive by (A) and by $\alpha \geq 0$, while the second term is not positive since $y(u) \geq 0$ for all vertices. Thus, we can conclude by similar arguments as above that there are potentials $x(u)$, $u \in V_W$ which satisfy the analogues of inequalities (15a) - (15b) with $\widehat{r}$ in place of r, that is inequalities (15c) - (15d). □

Lemma 7 *Problem LP(α) is bounded.*

Proof It is enough to show that the homogenized versions of inequalities (15a) - (15d) (that is when we assume $r(u,v) = 0$ for all arcs in the graph) cannot have a feasible solution with a negative objective function value. To this end, assume that (x, y) is a feasible solution to the homogenized version of inequalities (15a) - (15d), set $\xi = \min_{u \in V_W} y(u)$ and define $L = \{u \in V_W \mid y(u) = \xi\}$. Since $y(v) > \xi$ for all $v \notin L$, inequalities (15a) - (15b) imply that we do not have an arc (u, v) in the graph with $u \in L$ and $v \notin L$ (recall that we assume that all arcs from random vertices have positive probability). For $u \in V_W$ this is implied directly by the homogenized version of (15a). For $u \in V_R$ the homogenized inequalities (15b) mean that the weighted average of the $y(v)$ values for the out-neighbors of u cannot be larger than $y(u) = \xi$. By the minimality of ξ this implies that all out-neighbors of u must also be in L.

Let us now assume that $\xi < 0$, and consider the homogenized version of inequalities (15c) – (15d) for vertices $u \in L$. The left hand side of these inequalities is $\xi < 0$, and thus for every vertex $u \in L$ we have an out-neighbor v of u such that $x(v) - x(u) < 0$. Since no arcs lead to vertices not in L by the above observation, we must have a cycle within L such that for all arcs (u, v) along this cycle we have $x(v) - x(u) < 0$, which is clearly impossible. This contradiction shows that our assumption $\xi < 0$ is incorrect, that is that $y(u) \geq 0$ for all vertices $u \in V_W$. □

Let us denote by $Z(\alpha)$ the optimum value of LP(α).

Corollary 2 *The value $Z(\alpha)$ exists and is finite for all $\alpha \geq 0$.*

Proof Follows directly form Lemmas 6 and 7. □

Let us next observe a simple but very useful property of these linear programs.

Lemma 8 *If (x, y) is a feasible solution of LP(α), then for all $\Delta > 0$, $(x + \Delta y, y)$ is feasible in LP($\alpha + \Delta$), $J^R_{\alpha+\Delta}(x + \Delta y, y) \subseteq I^R(y)$, and $J^u_{\alpha+\Delta}(x + \Delta y, y) \subseteq I^u(y)$ for all $u \in V_W$.*

Proof Let us note that the inequalities in both (15a) and (15c) can be labeled by the arcs of G leaving vertices of V_W. The two inequalities labeled by a particular arc $(u, v) \in E$, $u \in V_W$ we shall call corresponding inequalities. Analogously, in (15b) and (15d) inequalities are labeled by vertices $u \in V_R$, and we shall call the two labeled by the same vertex u as corresponding. Then by adding Δ times the inequalities of (15a) to their corresponding ones in (15c) we get for every arc $(u, v) \in E$, $u \in V_W$ that

$$y(u) \;\geq\; (\alpha + \Delta) r(u,v) - (x(u) + \Delta y(u)) + (x(v) + \Delta y(v)) \,.$$

Analogously, by adding Δ times the inequalities (15b) to their corresponding ones in (15d) we get for all $u \in V_R$ that

$$y(u) \ \geq \ \operatorname*{avg}_{v \in N^+(u)} \left[(\alpha + \Delta) r(u,v) - (x(u) + \Delta y(u)) + (x(v) + \Delta y(v)) \right].$$

The above inequalities together with (15a)-(15b) imply that $x + \Delta y, y)$ is feasible in $LP(\alpha + \Delta)$, as claimed. Let us also note that, since $\Delta > 0$, if any one of a corresponding pair of inequalities is strict for (x, y), then the resulting new inequality is also strict, and hence the relations $J^R_{\alpha+\Delta}(x + \Delta y, y) \subseteq I^R(y)$, and $J^u_{\alpha+\Delta}(x + \Delta y, y) \subseteq I^u(y)$ for $u \in V_W$ follow. □

Corollary 3 *If $\alpha' > \alpha \geq 0$, then $Z(\alpha') \leq Z(\alpha)$.*

Proof Follows directly by Lemma 8. □

Lemma 9 *For any feasible solution (x, y) in $LP(\alpha)$, $\alpha \geq 0$ and for any vertex $u \in V$ and any strategy $\mathfrak{s} \in \mathfrak{A}$ we have*

$$y(u) \ \geq \ \psi_{\mathfrak{s}}(u).$$

Proof As in the proof of Theorem 3. □

Corollary 4 *There exists a real $L \in \mathbb{R}$ such that we have $L \leq Z(\alpha)$ for all $\alpha \geq 0$.*

Proof By Lemma 9 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 $\mathfrak{s} \in \mathfrak{A}$. Let us now fix a uniformly optimal stationary strategy $\mathfrak{s}$ of the mean-payoff MDP (which we know to exist). It was shown in [30] that under the assumption (A) we have $\psi_{\mathfrak{s}}(u)$ finite for all vertices $u \in V$. Consequently, $L = \sum_{u \in V} \psi_{\mathfrak{s}}(u)$ is a finite lower bound (of polynomial length) on the objective function value of any feasible solution of $LP(\alpha)$ for any $\alpha \geq 0$. □

Lemma 10 *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 $\alpha' > \alpha_0$.*

Proof Let us introduce $M = 2(|V_R| + \sum_{u \in v_W} |N^+(u)|)$ nonnegative slack variables $z \in \mathbb{R}^M_+$, rearrange all variables on the left hand side and write the constraints of $LP(\alpha)$ as a system of equations in terms of variables x, y, and z. Let us note that α appears only on the right hand side in some of the inequalities, and all right hand sides are linear functions of α; let us denote it simply by $\alpha b + d$, where $b, d \in \mathbb{R}^M$. Let us also recall from the theory of linear programming that $LP(\alpha)$ has a basic optimal solution. Since we have $N = 2n + M$ variables and M constraints in this system of equations, and since the coefficient matrix is of full row-rank, due to the slack variables, we have at most $\binom{2n+M}{M}$ basic submatrices. For each such submatrix B the nonbasic variables take value zero, while for the basic variables we have $(x, y, z)_B = B^{-1}(\alpha b + d) = \alpha B^{-1} b + B^{-1} d$. Consequently, all variables are linear functions of α and thus we have coefficients f_B and g_B depending only on the input (G, r, p), and independent of α such that

$$\sum_{u \in V} y_B(u) \;=\; \alpha f_B + g_B.$$

For any fixed α value, $Z(\alpha)$ is the minimum of the above linear functions over those bases which are feasible in $LP(\alpha)$, and this minimum will be attained at one of the bases, say at $B(\alpha)$:

$$Z(\alpha) \;=\; \min_{B \text{ is a feasible basis in } LP(\alpha)} \alpha f_B + g_B \;=\; \alpha f_{B(\alpha)} + g_{B(\alpha)}.$$

Let us note next, that if $f_{B(\alpha)} > 0$, then the inequality by Corollary 3

$$Z(0) \;\geq\; Z(\alpha) \;=\; \alpha f_{B(\alpha)} + g_{B(\alpha)}$$

implies that $B(\alpha)$ cannot be the optimal basis for any $\alpha > \left(Z(0) - g_{B(\alpha)}\right) / f_{B(\alpha)}$. Similarly, if $f_{B(\alpha)} < 0$ then the inequality by Corollary 4

$$L \;\leq\; Z(\alpha) \;=\; \alpha f_{B(\alpha)} + g_{B(\alpha)}$$

implies that $B(\alpha)$ cannot be an optimal basis for any $\alpha > \left(L - g_{B(\alpha)}\right) / f_{B(\alpha)}$. The above imply that there exists a finite value $\alpha' \geq 0$ such that

$$\alpha' \;>\; \max \left\{ \frac{Z(0) - g_B}{f_B}, \frac{L - g_B}{f_B} \;\middle|\; \begin{array}{c} B \text{ is a feasible basis in } LP(\alpha) \\ \text{for some } \alpha, \text{ and } f_B \neq 0 \end{array} \right\}$$

and for all $\alpha \geq \alpha'$ we must have $f_{B(\alpha)} = 0$.

Let us then define

$$Z \;=\; \min \left\{ g_B \;\middle|\; \begin{array}{c} B \text{ is a feasible basis in } LP(\alpha) \\ \text{for some } \alpha \geq \alpha', \text{ and } f_B = 0 \end{array} \right\}.$$

Since we have only finitely many different bases, this minimum is well defined. Let us choose now α_0 for which the above minimum is attained. In other words, such that $\alpha_0 \geq \alpha'$,

$f_{B(\alpha_0)} = 0$, $g_{B(\alpha_0)} = Z$ and $B(\alpha_0)$ is a feasible basis in $LP(\alpha_0)$. Then, by the above analysis and by Corollary 3 we have $Z(\alpha) = Z$ for all $\alpha \geq \alpha_0$. It is also clear form the above arguments that α_0 has polynomial bit-length. □

Lemma 11 *Let us consider $\alpha \geq \alpha_0$ and denote by (x^*, y^*) an arbitrary optimal solution of $LP(\alpha)$. Then, we have*

$$I^R(y^*) = V_R \quad \text{and} \quad I^u(y^*) \neq \emptyset \quad \text{for all } u \in V_W.$$

Proof By Lemma 8 the vector (x^{**}, y^*) is feasible in LP$(\alpha + 1)$, where $x^{**} = x^* + y^*$. Furthermore, we have $J^R_{\alpha+1}(x^{**}, y^*) \subseteq I^R(y^*)$, and $J^u_{\alpha+1}(x^{**}, y^*) \subseteq I^u(y^*)$ for all $u \in V_W$. Thus, for a vertex $u \in V_R \setminus I^R(y^*)$ or for $u \in V_W$ with $I^u(y^*) = \emptyset$ we could decrease $y^*(u)$ by some small positive quantity, without violating the feasibility of any of the inequalities in LP$(\alpha + 1)$. This would imply that $Z(\alpha + 1) < Z(\alpha)$, contradicting by Lemma 10 our assumption that $\alpha \geq \alpha_0$. □

To arrive to a proof of Theorem 4, 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, the additional conditions of Theorem 2 calls 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 2.

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:

$$0 \geq \alpha r(u,v) - y^*(u) - x(u) + x(v) \qquad \forall\, u \in V_W,\ (u,v) \in I^u(y^*)$$

$$0 \geq \operatorname*{avg}_{v \in N^+(u)} (\alpha r(u,v) - y^*(u) - x(u) + x(v)) \qquad \forall\, u \in V_R.$$

Note that out of the inequalities of (15c)-(15d) we included only those to which the corresponding inequalities in (15a)-(15b) are tight at y^*. Since $x^* \in X_\alpha(y^*)$, this set is a nonempty, closed convex set.

Lemma 12 *For all $x \in X_\alpha(y^*)$ there exists a finite $\Delta(x)$ such that $(x + \Delta y^*, y^*)$ is feasible in $LP(\alpha + \Delta)$ for all $\Delta \geq \Delta(x)$.*

Proof Indeed, if (x, y^*) is feasible in $LP(\alpha)$, then by Lemma 8 we can choose $\Delta(x) = 0$. Otherwise, (x, y^*) must violate, by the definition of $X_\alpha(y^*)$, some of the inequalities (15c) corresponding to some arcs $(u, v) \notin I^u(y^*)$, $u \in V_W$. For each such arc we have consequently $y^*(u) < \alpha r(u, v) - x(u) + x(v)$ and $y^*(u) > r(u, v) + y^*(v)$, and thus

$$\Delta_{uv} \;=\; \frac{(\alpha r(u, v) - x(u) + x(v)) - y^*(u)}{y^*(u) - (r(u, v) + y^*(v))} \;>\; 0.$$

Let us define $\Delta(x) = \max \Delta_{uv}$, where the maximization runs over all arcs $(u, v) \notin I^u(y^*)$, $u \in V_W$ the corresponding inequality of which in (15c) is violated by (x, y^*). Then, the claim follows along the lines of the proof of Lemma 8. □

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.

Lemma 13 *If $x \in X_\alpha(y^*)$ and $\alpha \geq \alpha_0$, then for all subsets $S \subseteq V \setminus T(x)$ we have either a $u \in S \cap V_R$ such that $N^+(u) \not\subseteq S$, or a $u \in S \cap V_W$ such that for some $(u, v) \in I^u(y^*)$ we have $v \notin S$.*

Proof This claim holds because if $N^+(u) \subseteq S$ for all $u \in S \cap V_R$, and all tight arcs leaving vertices $u \in S \cap V_W$ stay inside S, then defining

$$y(u) = \begin{cases} y^*(u) - \epsilon & u \in S, \\ y^*(u) & \text{otherwise,} \end{cases}$$

for some small $\epsilon > 0$ we would have $(x + \Delta y^*, y)$ as a feasible solution to $\mathrm{LP}(\alpha + \Delta)$ for any $\Delta > \Delta(x)$, by Lemmas 12 and 8 contradicting $Z(\alpha) = Z(\alpha + \Delta)$ for all $\Delta \geq 0$ by Lemma 10. □

Corollary 5 *For all $\alpha \geq \alpha_0$ and $x \in X_\alpha(y^*)$ we have $T(x) \neq \emptyset$.*

Proof Follows immediately from Lemma 13, since no arcs are leaving the set V. □

Lemma 14 *For all $x, x' \in X_\alpha(y^*)$ and $0 < \lambda < 1$ we have $T(\lambda x + (1 - \lambda)x') \subseteq T(x) \cap T(x')$.*

Proof Follows by the definition of $T(x)$, since any inequality of $LP(\alpha)$ which is strict for (x, y^*) or for (x', y^*) will also be strict for $(\lambda x + (1 - \lambda)x', y^*)$. □

Lemma 15 *If $\alpha \geq \alpha_0$, then $U = \bigcap_{x \in X_\alpha(y^*)} T(x) \neq \emptyset$.*

Proof This is because $U = \emptyset$ would imply that for every vertex $v \in V$ there exists a vector $x_v \in X_\alpha(y^*)$ such that $v \notin T(x_v)$. Then, by Lemma 14 we would have $T\left(\frac{1}{|V|}\sum_{v \in V} x_v\right) = \emptyset$, contradicting Corollary 5. □

Lemma 16 *For all vertices $v \in V$ we can test if $v \in U$, and if not we can find $x_v \in X_\alpha(y^*)$ such that $v \notin T(x_v)$ in polynomial time.*

Proof Let us introduce a new real variable $z(u)$ for each vertex $u \in V$, and consider the following linear programming problem:

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

Let us denote by $(\widehat{z}, \widehat{x})$ an optimal solution of this problem. If $\widehat{z}(v) = 0$, then $v \in U$, otherwise for $x_v = \widehat{x}$ we have $v \notin T(x_v)$ by the definitions of $X_\alpha(y^*)$ and the tightness of vertices. □

Corollary 6 *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.*

Proof Let us apply Lemma 16 for each vertex $v \in V$, and set $\overline{x} = \frac{1}{|V \setminus U|}\sum_{v \in V \setminus U} x_v$. The claim follows by Lemmas 14 and 15. □

Lemma 17 *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 $x' \in X_\alpha(y^*)$.

Proof This is because if v is not tight, then variable $x(v)$ appear with a negative coefficient in the system defining $X_\alpha(y^*)$ only in non-tight inequalities. Hence the above change of the feasible x vector will increase the right hand side of those inequalities only when they are not tight, and thus for a small enough positive ϵ they remain valid. □

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 6, and the following linear programming problem:

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

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 only $z(v)$ for $v \in V \setminus U$ are the variables.

Lemma 18 *If $\alpha \geq \alpha_0$ then problem* (LPZ) *has a finite optimum.*

Proof We have to show that (LPZ) is feasible and bounded. Since $\overline{x} \in X_\alpha(y^*)$, $z = 0$ is a feasible solution. To see boundedness, let us consider the homogenized version of (LPZ) (that is in which we set all constants to zero). By the theory of linear programming it is enough to show that this homogenized linear program does not have a solution with a positive objective function value. To see this let us rewrite (for clarity) this homogenized linear program:

$$\begin{array}{lrll}
\max & \displaystyle\sum_{v \in V} z(v) & & \\
s.t. & & & \\
0 & \geq\ z(u) - z(v) & \forall u \in V_W,\ (u,v) \in I^u(y^*), \\
0 & \geq\ z(u) - \displaystyle\sum_{v \in V} p(u,v) z(v) & \forall u \in V_R, \\
0 & =\ z(u) & \forall\ u \in U, \\
0 & \leq\ z(u) & \forall\ u \in V \setminus U.
\end{array}$$

Assume indirectly that the above problem has a feasible solution z with a positive objective function value. Then $\xi = \max_{v \in V} z(v) > 0$, and the set $M = \{v \in V \mid z(v) = \xi\}$ is not empty. By the above inequalities, it follows that for all $u \in M \cap V_R$ we must have $N^+(u) \subseteq M$, and for all $u \in M \cap V_W$ and for all $(u,v) \in I^u(y^*)$ we must have $v \in M$, too. Since M is disjoint from $U = T(\overline{x})$ by its definition, we arrive to a contradiction by Lemma 13. This contradiction thus proves that (LPZ) is bounded. □

Corollary 7 *If $\alpha \geq \alpha_0$, then* (LPZ) *has a finite optimum $\overline{z}$, and we have $T(\overline{x} - \overline{z}) = V$.*

Proof Lemma 18 shows that (LPZ) has a finite optimum, $\overline{z}$, and by the definition of (LPZ) it follows that $\overline{x} - \overline{z} \in X_\alpha(y^*)$. If there were a non tight vertex $v \in V$, then by Lemma 17 we could further decrease the value of $(\overline{x} - \overline{z})(v)$ without leaving the polyhedron $X_\alpha(y^*)$, contradicting the optimality of $\overline{z}$. This contradiction proves that $T(\overline{x} - \overline{z}) = V$. □

Proof of Theorem 4. For an $\alpha' \geq \alpha_0$ value, let y^* be optimal in $LP(\alpha')$, let $\overline{x}$ be as in Corollary 6 and $\overline{z}$ as in Corollary 7, and define

$$x^* = \overline{x} - \overline{z} + \Delta(\overline{x} - \overline{z})y^* \quad \text{and} \quad \alpha = \alpha' + \Delta(\overline{x} - \overline{z}).$$

Then, by Lemmas 11, 12 and Corollary 7 it follows that (x^*, y^*) is an optimal solution in $LP(\alpha)$, satisfying all conditions of the theorem. □

5 General MDP's

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

Lemma 19 *Let u denote a vertex with $\phi_\Gamma(u) = 0$, and let $\mathfrak{s}$ denote a strategy. 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 $\phi_\Gamma(v) > \phi_\Gamma(w)$, then we have $\psi_\mathfrak{s}(v_0) = -\infty$.*

Proof Let y be a potential as in Fact 1. Then by definition, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\sum_{i=0}^{t}\mathbb{E}_\mathfrak{s}[r(v_i, v_{i+1})] = \frac{1}{T+1}\sum_{t=0}^{T}\sum_{i=0}^{t}\mathbb{E}_\mathfrak{s}[r[y](v_i, v_{i+1})] + y(v_0) - \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}_\mathfrak{s}[y(v_{t+1})].$$

Since $\left|y(v_0) - \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}_\mathfrak{s}[y(v_{t+1})]\right|$ is at most $M(y)$ which is independent of T and $\mathfrak{s}$, it is enough to show that $\limsup_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\sum_{i=0}^{t}\mathbb{E}_\mathfrak{s}[r[y](v_i, v_{i+1})] = -\infty$ holds, to complete our proof.

For an arc (v,w), let $\tilde{r}(v,w) = r[y](v,w) - \phi_\Gamma(v)$. Then we have $\tilde{r}(v,w) \leq 0$ for any arc (v,w), and hence

$$\begin{aligned}\frac{1}{T+1}\sum_{i=0}^{T}\sum_{j=0}^{i}\mathbb{E}_{\mathfrak{s}}[r[y](v_j,v_{j+1})] &= \frac{1}{T+1}(\sum_{i=0}^{T}\sum_{j=0}^{i}\mathbb{E}_{\mathfrak{s}}[\tilde{r}(v_j,v_{j+1})] \\ &\quad + \sum_{i=0}^{T}\sum_{j=0}^{i}\mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_j)]) \\ &\leq \frac{1}{T+1}\sum_{i=0}^{T}\sum_{j=0}^{i}\mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_j)]. \qquad (17)\end{aligned}$$

Recall that $\phi_\Gamma(v) = \max_{w\in E(v)}\phi_\Gamma(w)$ if $v \in V_W$, and $\phi_\Gamma(v) = \text{ avg }_{w\in E(v)}\phi_\Gamma(w)$ if $v \in V_R$. Thus, we have

$$0 = \phi_\Gamma(v_0) \geq \mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_j)] \geq \mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_{j+1})] \text{ for } j = 1,2,...$$

Let j^* be the time when $\mathfrak{s}$ uses with positive probability an arc (v,w) such that $v \in V_W$ and $\phi_\Gamma(v) > \phi_\Gamma(w)$. Then $\mathbb{E}_{\mathfrak{s}}[\phi_\Gamma(v_{j^*+1})] = -\epsilon$ holds for some constant $\epsilon > 0$. This implies that (17) is at most $\frac{1}{T+1}\sum_{i=j^*+1}^{T} -\epsilon(i-j^*)$, which goes to $-\infty$ as $T \to +\infty$. This completes the proof. □

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, $u \in V$. 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 holds with respect to the total reward payoff function ψ.

Lemma 20 *Fix an initial vertex $v_0 = u$ such that $\phi_\Gamma(u) = 0$. Then any strategy $\mathfrak{s}$ in Γ' satisfies $\mathbb{E}_{\mathfrak{s}}[r'(v_j,v_{j+1})] = \mathbb{E}_{\mathfrak{s}}[r(v_j,v_{j+1})]$.*

Proof We note that $\phi_\Gamma(v) = \max_{w\in E(v)}\phi_\Gamma(w)$ if $v \in V_W$, and $\phi_\Gamma(v) = \text{ avg }_{w\in E(v)}\phi_\Gamma(w)$ if $v \in V_R$. By construction, $\phi_\Gamma(v) = \phi_\Gamma(w)$ for all $w \in E'(v)$, if $v \in V_W$, and $\phi_\Gamma(v) =$

avg $_{w\in E'(v)}\,\phi_\Gamma(w)$ if $v \in V_R$. Since $\phi_\Gamma(u) = 0$, and hence $\mathbb{E}_\mathfrak{s}[\phi_\Gamma(v_j)] = 0$ for all j, it follows that $\mathbb{E}_\mathfrak{s}[r'(v_j, v_{j+1})] = \mathbb{E}_\mathfrak{s}[r(v_j, v_{j+1})] - \mathbb{E}_\mathfrak{s}[\phi_\Gamma(v_j)] = \mathbb{E}_\mathfrak{s}[r(v_j, v_{j+1})]$. □

Lemma 21 *Any optimal strategy in Γ' with respect to the total payoff ψ is also optimal in Γ.*

Proof Let u be an initial vertex, and let $\mathfrak{s}$ denote an optimal strategy in Γ' with respect to the total payoff ψ. By Lemmas 19 and 20, $\mathfrak{s}$ 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 $w \in E'(v)$, and $\phi_\Gamma(v) =$ avg $_{w\in E'(v)}\,\phi_\Gamma(w)$ if $v \in V_R$. By Theorems 3 and 4 we know that $\mathfrak{s}$ is pure and stationary. By our construction we known that $\phi_{\Gamma'}(v) = 0$ for all $v \in V$. Thus, by [30] it follows that $\psi_\mathfrak{s}(v)$ is finite for all $v \in V$. Since by our construction we have

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

the equality $\psi_\Gamma = +\infty$ (resp., $-\infty$) follows. □

6 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 reward payoff.

6.1 Discounted BWR-games

Let β be a number in $\in (0, 1]$ called *discount factor*. *Discounted mean payoff* stochastic games were introduced by Shapley [29] and have payoff function:

$$\phi^\beta_\mathfrak{s}(v_0) \;=\; (1-\beta)\sum_{j=0}^{\infty}\beta^j\mathbb{E}_\mathfrak{s}[r_\mathfrak{s}(v_t, v_{t+1})], \tag{18}$$

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

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 [35]

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 [29].

Fact 2 ([29]) *A BWR-game with the discounted mean payoff function ϕ^β has a saddle point in uniformly optimal, strategies pure and stationary, for all $0 < \beta < 1$.*

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

6.2 Existence of a Saddle Point in Stationary 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 stationary 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

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

Finally, we will denote by $Q_{\mathfrak{s}}$ the (unique) limiting average probability matrix satisfying $Q_{\mathfrak{s}} P_{\mathfrak{s}} = P_{\mathfrak{s}} Q_{\mathfrak{s}} = Q_{\mathfrak{s}}$.

Proposition 1 ([31]) *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 5 *Consider an undiscounted BWR-game Γ such that $\phi_\Gamma = \mathbf{0}$. Then there is a uniformly optimal pair of pure and stationary strategies $(\mathfrak{s}_B, \mathfrak{s}_W)$ satisfying:*

$$\pi_{\mathfrak{s}^*_B, \mathfrak{s}_W}(v_0) \leq \pi_{\mathfrak{s}^*_B, \mathfrak{s}^*_W}(v_0) \leq \pi_{\mathfrak{s}_B, \mathfrak{s}^*_W}(v_0) \quad \text{for all } \mathfrak{s}_B \in \widehat{\mathfrak{S}}_B \text{ and } \mathfrak{s}_W \in \widehat{\mathfrak{S}}_W. \tag{19}$$

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 $\beta = 1 - \frac{1}{(nD)^{O(n)}R}$.

Proof We start with the following claim.

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

$$\lim_{\beta \to 1^-} \frac{\psi_{\mathfrak{s}} - (I - \beta P_{\mathfrak{s}})^{-1} r_{\mathfrak{s}}}{1 - \beta} = P_{\mathfrak{s}} (I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-2} r_{\mathfrak{s}}. \tag{20}$$

Proof We note that the numerator in the left-hand side of (20) vanishes as $\beta \to 1^-$, since

$$\begin{aligned}
\lim_{\beta \to 1^-} (I - \beta P_{\mathfrak{s}})^{-1} r_{\mathfrak{s}} &= \lim_{\beta \to 1^-} \sum_{i=0}^{\infty} \beta^i P_{\mathfrak{s}}^i r_{\mathfrak{s}} \\
&= \lim_{\beta \to 1^-} \sum_{i=0}^{\infty} \beta^i (P_{\mathfrak{s}}^i - Q_{\mathfrak{s}}) r_{\mathfrak{s}} \\
&= \left((I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1} - Q_{\mathfrak{s}} \right) r_{\mathfrak{s}} \\
&= (I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1} r_{\mathfrak{s}},
\end{aligned}$$

where the third equality follows by the following result of Blackwell [3]:

$$\lim_{\beta \to 1^-} \sum_{i=0}^{\infty} \beta^i (P_{\mathfrak{s}}^i - Q_{\mathfrak{s}}) = (I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1} - Q_{\mathfrak{s}}. \tag{21}$$

Thus, we may apply L'Hôpital's rule:

$$\begin{aligned}
\lim_{\beta\to 1^-} \frac{\psi_{\mathfrak{s}} - (I-\beta P_{\mathfrak{s}})^{-1} r_{\mathfrak{s}}}{1-\beta} &= \lim_{\beta\to 1^-} \sum_{i=0}^{\infty} i\beta^{i-1} P_{\mathfrak{s}}^i r_{\mathfrak{s}} = \lim_{\beta\to 1^-} \sum_{i=0}^{\infty}\sum_{j=0}^{i-1} \beta^{i-1} P_{\mathfrak{s}}^i r_{\mathfrak{s}} \\
&= \lim_{\beta\to 1^-} \sum_{j=0}^{\infty}\sum_{i=j+1}^{\infty} \beta^{i-1} P_{\mathfrak{s}}^i r_{\mathfrak{s}} \\
&= \lim_{\beta\to 1^-} \sum_{j=0}^{\infty} \beta^j P_{\mathfrak{s}}^{j+1} \sum_{i=0}^{\infty} \beta^i P_{\mathfrak{s}}^i r_{\mathfrak{s}} \\
&= \lim_{\beta\to 1^-} \sum_{j=0}^{\infty} \beta^j (P_{\mathfrak{s}}^j - Q_{\mathfrak{s}}) P_s \sum_{i=0}^{\infty} \beta^i (P_{\mathfrak{s}}^i - Q_{\mathfrak{s}}) r_{\mathfrak{s}} + \sum_{j=0}^{\infty} \beta^j Q_{\mathfrak{s}} P_{\mathfrak{s}} \sum_{i=0}^{\infty} \beta^i P_{\mathfrak{s}}^i r_{\mathfrak{s}} \\
&= \lim_{\beta\to 1^-} \sum_{j=0}^{\infty} \beta^j (P_{\mathfrak{s}}^j - Q_{\mathfrak{s}}) P_s \sum_{i=0}^{\infty} \beta^i (P_{\mathfrak{s}}^i - Q_{\mathfrak{s}}) r_{\mathfrak{s}} + \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} \beta^{i+j} Q_{\mathfrak{s}} r_{\mathfrak{s}} \\
&= \lim_{\beta\to 1^-} \sum_{j=0}^{\infty} \beta^j (P_{\mathfrak{s}}^j - Q_{\mathfrak{s}}) P_s \sum_{i=0}^{\infty} \beta^i (P_{\mathfrak{s}}^i - Q_{\mathfrak{s}}) r_{\mathfrak{s}} \\
&= \lim_{\beta\to 1^-} \sum_{j=0}^{\infty} \beta^j (P_{\mathfrak{s}}^j - Q_{\mathfrak{s}}) P_s \cdot \lim_{\beta\to 1^-} \sum_{i=0}^{\infty} \beta^i (P_{\mathfrak{s}}^i - Q_{\mathfrak{s}}) r_{\mathfrak{s}} \\
&= \left((I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1} - Q_{\mathfrak{s}}\right) P_{\mathfrak{s}} \left((I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1} - Q_{\mathfrak{s}}\right) r_{\mathfrak{s}} \\
&= P_{\mathfrak{s}} (I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-2} r_{\mathfrak{s}},
\end{aligned}$$

due to the commutation of $(I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-1}$ and $P_{\mathfrak{s}}$. □

Let

$$\begin{aligned}
\gamma &= \min_{u\in V} \min_{\mathfrak{s},\mathfrak{s}'\in\widehat{\mathfrak{S}}:\ \psi_{\mathfrak{s}}(u)\neq\psi_{\mathfrak{s}'}(u)} |\psi_{\mathfrak{s}}(u) - \psi_{\mathfrak{s}'}(u)| \\
\kappa &= \max_{u\in V} \max_{\mathfrak{s}\in\widehat{\mathfrak{S}}} |P_{\mathfrak{s}}(I - P_{\mathfrak{s}} + Q_{\mathfrak{s}})^{-2} r_{\mathfrak{s}}|.
\end{aligned}$$

Standard estimation arguments (see, e.g., [4]) give $\gamma \geq \frac{1}{D^{O(n)}}$ and $\kappa \leq (nD)^{O(n)} R$.

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

$$\|(1-\beta(\epsilon))\psi_{\mathfrak{s}} - \phi_{\mathfrak{s}}^{\beta(\epsilon)}\|_\infty < (1-\beta(\epsilon))^2(\epsilon+\kappa) \leq 2(1-\beta(\epsilon))^2\kappa. \tag{22}$$

Let us choose ϵ such that $\beta(\epsilon) > 1 - \frac{\gamma}{4\kappa}$. Then for any two pairs of stationary 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 (22), 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 5 implies the existence of saddle point $\mathfrak{s}^* = (\mathfrak{s}_B^*, \mathfrak{s}_W^*)$, among uniformly optimal stationary and pure 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 stationary, 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 5.

6.3 A sufficient and necessary condition for BWR-games

Thuijsman and Vrieze (Theorem 5.3 in [31]) gave 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*. We strengthen this result for BWR games as follows. First, we show that, as is the case for mean payoff BWR-games, only one vector of potentials $x \in \mathbb{R}^V$ is enough in the system of equations in part (ii) of the theorem below (in contrast, as in Theorem 5.3 in [31], two vectors maybe necessary in the case of general stochastic games). Second, we show that the existence of a solution for the system in (ii) implies the existence of an optimal solution in stationary strategies, even if each player is allowed to choose from the space of all, possibly history-dependent, strategies.

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

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

Theorem 6 *For a total reward BWR-game $\Gamma = (G, P, r)$, the following two statements are equivalent:*

(i) the value vector ψ_Γ exists and is finite, and each player possesses a uniformly optimal, pure and stationary strategy (optimal among all strategies);

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

$$\mu(u) = M_{E(u)}[r(u,v) + \mu(v)] \qquad \text{for all } u \in V, \tag{23a}$$
$$\mu(u) = M_{E(u)}[\alpha r(u,v) + x(v) - x(u)] \qquad \text{for all } u \in V, \tag{23b}$$
$$\mu(u) = M_{\mathrm{EXT}(u)}[\alpha r(u,v) + x(v) - x(u)] \qquad \text{for all } u \in V_W \cup V_B, \tag{23c}$$

where, for a vertex $u \in V_W \cup V_B$, $\mathrm{EXT}(u)$ denotes the set of arcs in $E(u)$ attaining equality in (23a).

Proof (ii) $\Rightarrow$ (i): This can be shown using essentially the same argument as in Theorem 3, but applied twice, once for each player.

(i) $\Rightarrow$ (ii): The proof is more or less the same as the one given for Theorem 5.3 in [31]; for completeness, we reproduce it here and point out the required changes.

Let $s^* = (\mathfrak{s}_B^*, \mathfrak{s}_W^*)$ be a uniformly optimal pair of pure and stationary strategies in Γ; thus, $\psi_{\mathfrak{s}_B^*,\mathfrak{s}_W} \leq \psi_{\mathfrak{s}_B^*,\mathfrak{s}_W^*} = \psi_\Gamma \leq \psi_{\mathfrak{s}_W^*,\mathfrak{s}_B}$ for all stationary strategies $\mathfrak{s}_B \in \widehat{\mathfrak{S}}_B$ and $\mathfrak{s}_W \in \widehat{\mathfrak{S}}_W$.

By assumption, $\psi_{\mathfrak{s}^*}$ is finite, and hence, it satisfies necessarily the (also called Shapley) equations (23a). For $u \in V_W \cup V_B$, let $\mathrm{EXT}(u) \subseteq E(u)$ be the set of arcs for which equality is achieved in these equations. Until almost the end of this proof, we delete all the arcs not in $\mathrm{EXT}(u)$, and assume for the moment therefore that $\mathrm{EXT}(u) = E(u)$ for all $u \in V_W \cup V_B$.

Since $\psi_{\mathfrak{s}^*}$ is finite, Proposition 1 implies that $\phi_{\mathfrak{s}^*} = \mathbf{0}$. As in the the proof of Proposition 1, we can show that $Q_{\mathfrak{s}^*}\psi_{\mathfrak{s}^*} = Q_{\mathfrak{s}^*}r_{\mathfrak{s}^*} = 0$. The same holds for any pair of stationary strategies $\mathfrak{s}$ such that $\psi_{\mathfrak{s}}$ is finite. On the other hand, for a pair of stationary strategies $\mathfrak{s}$, if $\psi_{\mathfrak{s}} = -\infty$ (resp., $\psi_{\mathfrak{s}} = \infty$), then again by Proposition 1 we have $\phi_{\mathfrak{s}} = Q_{\mathfrak{s}}r_{\mathfrak{s}} < 0$ (resp., $\phi_{\mathfrak{s}} = Q_{\mathfrak{s}}r_{\mathfrak{s}} > 0$). Since the number of stationary strategies is finite, we conclude from $\psi_{\mathfrak{s}} \leq \psi_{\mathfrak{s}^*} = \psi_\Gamma \leq \psi_{\mathfrak{s}'}$ that, for some $\alpha' \geq 0$,

$$Q_{\mathfrak{s}}(\alpha' r_{\mathfrak{s}} - \psi_\Gamma) \leq 0 = Q_{\mathfrak{s}^*}(\alpha' r_{\mathfrak{s}^*} - \psi_\Gamma) \leq Q_{\mathfrak{s}'}(\alpha' r_{\mathfrak{s}'} - \psi_\Gamma), \tag{24}$$

for all pairs of stationary strategies $\mathfrak{s} = (\mathfrak{s}_B^*, \mathfrak{s}_W)$ and $\mathfrak{s}' = (\mathfrak{s}_B, \mathfrak{s}_W^*)$. It follows that the mean payoff BWR-game Γ' with local rewards $r'(u,v) = \alpha' r(u,v) - \psi_\Gamma(u)$ for $(u,v) \in E$ has value vector $\phi_{\Gamma'} = \mathbf{0}$. By Theorem 4 in [5], there exists a potential vector x' such that for all $u \in V_R$, $\mathbf{0} = M_{E(u)}[r'[x'](u,v)]$, and for all $u \in V_W \cup V_B$, $\mathbf{0} = M_{\mathrm{EXT}(u)}[r'[x'](u,v)]$. In other words:

$$\begin{aligned} \psi_\Gamma &= M_{E(u)}[\alpha' r(u,v) + x'(v) - x'(u)] \text{ for all } u \in V_R, \\ \psi_\Gamma &= M_{\mathrm{EXT}(u)}[\alpha' r(u,v) + x'(v) - x'(u)] \text{ for all } u \in V_W \cup V_B. \end{aligned}$$

Finally, let us consider the (deleted) arcs in $(u,v) \in E(u) \setminus \mathrm{EXT}(u)$ for $u \in V_W \cup V_B$. Suppose that, for some $u \in V_W$ and $(u,v) \in E(u)$, we have

$$\psi_\Gamma(u) < \alpha' r(u,v) + x(v) - x(u). \tag{25}$$

Let $\alpha'' \geq 0$ and $\alpha = \alpha' + \alpha''$. For $u \in V$, define $x(u) = x'(u) + \alpha''\psi_\Gamma(u)$. Then

$$\alpha r(u,v) + x(v) - x(u) = \alpha' r(u,v) + x'(v) - x'(u) + \alpha''(r(u,v) + \psi_\Gamma(v) - \psi_\Gamma(u)).$$

It follows that, if we choose α'' large enough, we will have

$$\alpha r(u,v) + x(v) - x(u) \begin{cases} = \alpha' r(u,v) + x(v) - x(u) & \text{if } (u,v) \in EXT(u), \\ \leq \alpha' r(u,v) + x(v) - x(u)) & \text{if } (u,v) \notin EXT(u). \end{cases}$$

A similar adjustment can be done for $u \in V_B$. Thus the equations in (ii) can be satisfied with $\mu = \psi_\Gamma$, x, and α. □

7 Open Problems

It is a challenge to find other effective payoff functions (in addition to the mean and total ones) for which Theorems 1 and 2 hold.

Assigning to an arbitrary real sequence its lim sup and lim inf, we define the lim sup and lim inf mean and total effective payoffs. Theorems 1 and 2 can be extended to the lim sup and lim inf mean payoffs, but if these theorems are extendable to the lim sup and lim inf total payoffs is an open problem.

In Appendix II, we give an example that supports the last conjecture for the BW games.

Another open question is Nash-solvability of stochastic (non-zero-sum) games with perfect information in general (history dependent) pure strategies for mean or total effective payoff. In both cases, the answer is negative already for the BW games if we restrict the players to their pure and *stationary* strategies; see Remark 2. However, if we allow history dependent pure strategies then numerous open questions arise: whether Nash equilibria exist for BW, BWR, or n-person games with or without random positions; for mean or total effective payoff.

The following subcase of the mean payoff case is answered in the positive. Given a (non-zero sum) mean payoff BWR game, in which both players Black and White are maximizers, let us consider the corresponding two zero-sum games Γ_B and Γ_W. In each of them the players have uniformly optimal pure stationary strategies (that form a saddle point for every given initial position). The optimal strategies of Black in Γ_B and White in Γ_W will be called *cautious*, while the optimal strategies of Black in Γ_W and White in Γ_B will be called *punishing*. Let us assume additionally that both games are *ergodic*, that is, not only the optimal strategies, but also the values do not depend on an initial position. Then, it is not difficult to see that the following pair of combined strategies form a Nash equilibrium: Each player applies the cautious strategy until the opponent is doing the same and switches to the punishing strategy as soon as the opponent deviates. Let us note that all assumptions (both games are ergodic and the effective payoff is mean) are essential.

Let us finally notice that the above construction can be naturally extended to the n-person games with random positions. Indeed, due to perfect information, the player that has to be punished (if any) is unambiguously defined.

References

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

[2] Dimitri P. Bertsekasy and Huizhen Yuz. Stochastic shortest path problems, under weak conditions, lids report 2909. Technical report, MIT, 2013.

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

[4] 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.

[5] E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. On canonical forms for zero-sum stochastic mean payoff games. *Dynamic Games and Applications*, 2013.

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

[7] John N. Tsitsiklis Dimitri P. Bertsekas. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

[8] I. Dinur and S. Safra. On the hardness of approximating minimum vertex cover. *Annals of Mathematics*, 162:439–485, 2005.

[9] Jerzy Filar and Koos Vrieze. *Competitive Markov Decision Processes*. Springer, Berlin, 1996.

[10] 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.

[11] D.R. Fulkerson and G.C. Harding. Maximizing the minimum source-sink path subject to a budget constraint. *Mathematical Programming*, 13:116–118, 1977.

[12] T. Gallai. Maximum-minimum Sätze über Graphen. *Acta Mathematica Academiae Scientiarum Hungaricae*, 9:395–434, 1958.

[13] 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.

[14] V. Gurvich. A stochastic game with perfect information and without nash equilibria in pure stationary strategies. *Uspehi Mat. Nauk (in Russian)*, 260:135–136, 1988, English transl. in Russian Mathematical Surveys 43(2) (1988).

[15] V. Gurvich, A. Karzanov, and L. Khachiyan. Cyclic games and an algorithm to find minimax cycle means in directed graphs. *USSR Computational Mathematics and Mathematical Physics*, 28:85–91, 1988.

[16] O. Onésimo Hernández-Lerma and Jean-Bernard Lasserre. *Further topics on discrete-time Markov control processes*. Applications of mathematics. Springer, New York, 1999.

[17] A. J. Hoffman and R. M. Karp. On nonterminating stochastic games. *Management Science, Series A*, 12(5):359–370, 1966.

[18] R. A. Howard. *Dynamic programming and Markov processes*. Technology press and Willey, New York, 1960.

[19] E. Israeli and R. K. Wood. Shortest-path network interdiction. *Networks*, 40(2):97–111, 2002.

[20] R. M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Math.*, 23:309–311, 1978.

[21] R. M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Math.*, 23:309–311, 1978.

[22] A. V. Karzanov and V. N. Lebedev. Cyclical games with prohibition. *Mathematical Programming*, 60:277–293, 1993.

[23] 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.

[24] 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.

[25] T. M. Liggett and S. A. Lippman. Stochastic games with perfect information and time-average payoff. *SIAM Review*, 4:604–607, 1969.

[26] H. Mine and S. Osaki. *Markovian decision process*. American Elsevier Publishing Co., New York, 1970.

[27] Rolf H. Möhring, Martin Skutella, and Frederik Stork. Scheduling with and/or precedence constraints. *SIAM J. Comput.*, 33(2):393–415, 2004.

[28] S. D. Patek and D. P. Bertsekas. Stochastic shortest path games. *SIAM Journal on Control and Optimization*, 37:804–824, 1997.

[29] L. Shapley. Stochastic games. *Proc. Nat. Acad. Sci. USA*, 39:1095–1100, 1953.

[30] F. Thuijsman and O. J. Vrieze. The bad match, a total reward stochastic game. *Operations Research Spektrum*, 9:93–99, 1987.

[31] 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.

[32] Peter Whittle. *Optimization over Time*. John Wiley & Sons, Inc., New York, NY, USA, 1982.

[33] Huizhen Yu and Dimitri P. Bertsekas. Q-learning and policy iteration algorithms for stochastic shortest path problems. *Annals OR*, 208(1):95–132, 2013.

[34] Huizhen Yuz. Stochastic shortest path games and q-learning, lids report 2875. Technical report, MIT, 2011.

[35] U. Zwick and M. Paterson. The complexity of mean payoff games on graphs. *Theoret. Comput. Sci.*, 158(1-2):343–359, 1996.

Appendix I: An Example with $\phi_{\mathfrak{s}}(u) < 0$ and $\psi_{\mathfrak{s}}(u) = +\infty$

We demonstrate by an example that for a reward sequence $\mathbf{r}$ corresponding to a general history dependent strategy, the inequality $\phi(\mathbf{r}) < 0$ may not imply that $\psi(\mathbf{r}) = -\infty$. We construct below such a reward sequence where in fact we have $\phi(\mathbf{r}) < 0$ and $\psi(\mathbf{r}) = +\infty$. This reward sequence corresponds to a (history dependent) strategy in the two node MDP shown in Figure 1.

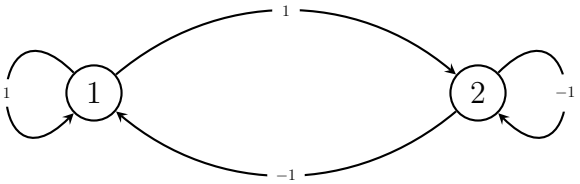

Figure 1: A simple MDP with two nodes, both controlled by MAX.

To describe the construction, we need to introduce some notations. Given integers m and x, let $(m(x))$ denote the sequence of length m in which each element is x, that is

$(m(x)) = x\mathbf{e}_m$, where $\mathbf{e}_m \in \mathbb{R}^m$ is the vector of all 1s of dimension m. Given two finite sequences $\mathbf{a}$ and $\mathbf{b}$, we denote by $(\mathbf{a}, \mathbf{b})$ their concatenation.

Let $p_i = 11 \cdot 4^i$ and $q_i = 14 \cdot 4^i$ for $i = 0, 1, ...$, define $\mathbf{a}_i = (p_i(+1))$ and $\mathbf{b}_i = (q_i(-1))$ for $i = 0, 1, ...$, and set $\mathbf{c}_i = (\mathbf{a}_i, \mathbf{b}_i)$. For instance, in $\mathbf{c}_0$ we have 11 ones followed by 14 negative ones. In general, $\mathbf{c}_i$ is of length $25 \cdot 4^i$, contains $11 \cdot 4^i$ consecutive $+1$s followed by $14 \cdot 4^i$ consecutive -1s.

Let us define finally an infinite sequence by

$$\mathbf{r} = (+1, +1, -1, -1, -1, \mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, ...).$$

This sequence consists of subsequences of consecutive ones and negative ones. The consecutive negative ones terminate in positions $n_i = 5 \cdot 4^i$, where we have

$$\sum_{j=1}^{n_i} \mathbf{r}_j \;=\; -4^i$$

for all $i = 0, 1, ...$ Consequently, we have

Fact 3

$$\phi(\mathbf{r}) \;=\; \liminf_{t\to\infty} \frac{1}{t} \sum_{j=1}^{t} \mathbf{r}_j \;=\; \lim_{i\to\infty} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{r}_j \;=\; -\frac{1}{5}.$$

Furthermore, the double sums $\sum_{t=1}^{T} \sum_{j=1}^{t} \mathbf{r}_j$ terminate their decreasing subsequences at $T = m_i = 6 \cdot 4^i$ for $i = 0, 1, ...$, and we have

$$\sum_{t=m_{i-1}+1}^{m_i} \sum_{j=1}^{t} \mathbf{r}_j \;=\; 9 \cdot 4^{2(i-1)} \quad \text{for all } i \geq 1.$$

Consequently, we have

Fact 4

$$\psi(\mathbf{r}) \;=\; \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{t} \mathbf{r}_j \;=\; \lim_{i\to\infty} \frac{9 \cdot 4^{2(i-1)}}{6 \cdot 4^i} \;=\; +\infty.$$

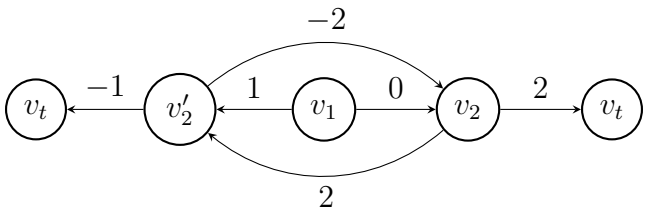

Figure 2: A saddle point free BW game with the "pseudo-total" effective payoff. Player MAX controls v_1, player MIN controls v_2 and v_2', that is, $V_W = \{v_1\}, V_B = \{v_2, v_2'\}, V_R = \emptyset$, while v_t is a terminal position; a loop with the local reward 0 is assumed in v_t. In addition, the graph contains only one dicycle. It consists of two arcs with the local rewards 2 and -2.

Appendix II: Solvability fails for the "pseudo-total" effective payoff but remains open for the $\liminf$ and $\limsup$ total effective payoffs

For simplicity let us consider a BW game ($V_R = \emptyset$) and restrict both players to their pure stationary strategies. Then, obviously, any play is a "lasso" L that consists of an initial path P and a cycle C repeated infinitely. If this play is finite (that is, C is a loop with the local reward zero) then the negated effective payoff $-\psi(L)$ can be naturally interpreted as the cost or the length of the path P. Two players MAX and MIN want to minimize and, respectively, to maximize this cost.

Yet, we have to define the effective payoffs for the other directed cycles too. If $\psi(C) = \sum_{e \in C} r(e)$ is strictly positive or strictly negative then the effective cost $\Psi(L)\psi(L)$ is defined as $-\infty$ and $+\infty$), respectively.

Still, a problem remains when $\psi(C) = 0$. It seems "natural" to define the effective payoff of such a play as $\Psi(L) = \Psi(P) = \sum_{e \in P} r(e)$. Let us call such effective payoff *pseudo-total*. Yet, in this case solvability in the pure stationary strategies fails. The example is given in Figure 2 and Table 1.

In contrast, any BWR game with the *total effective payoff* has a saddle point, by Theorem 2.

Let us also notice that Nash-solvability may hold (this is an open question) after the following "minor" modifications of the above pseudo-total effective payoff. Given a lasso L that consists of P and C, let us define $\Psi(L)$ as (a) the maximum or (b) the minimum of the sums of local rewards over all paths P' that consist of P and in addition may include also next several successive arcs along C.

For the above example in case (a), the effective payoff for the lasso $\langle v_1, v_2, v_2', v_2 \rangle$ becomes $0 + 2 = 2$ rather than 0, while for the other lasso $\langle v_1, v_2', v_2, v_2' \rangle$ the effective payoff remains 1.

	$v_2\, v_2'$ $v_2'\, v_2$	$v_2\, v_2'$ $v_2'\, v_t$	$v_2\, v_t$ $v_2'\, v_2$	$v_2\, v_t$ $v_2'\, v_t$
$v_1\, v_2$	$v_1v_2v_2'v_2$ 0	$v_1v_2v_2'v_t$ 1	$v_1v_2v_t$ 2	$v_1v_2v_t$ 2
$v_1\, v_2'$	$v_1v_2'v_2v_2'$ 1	$v_1v_2'v_t$ 0	$v_1v_2'v_2v_t$ 1	$v_1v_2'v_t$ 0

Table 1: The corresponding normal form: the rows and columns are associated with the pure stationary strategies of the players MAX and MIN, respectively. The corresponding "lasso" plays with their pseudo-total effective payoff values are given. The game has no saddle point: $0 = \max_{col} \min_{row} < \min_{row} \max_{col} = 1$.

It is easy to verify that if we replace the 0 entry $(1,1)$ in the upper left corner of the matrix by 2 than $(1,2)$ becomes a saddle point with the value 1.

In case (b), the effective payoff for the lasso $\langle v_1, v_2', v_2, v_2' \rangle$ the effective payoff becomes $1-2=-1$ rather than 1, while for the other lasso $\langle v_1, v_2, v_2', v_2 \rangle$ the effective payoff remains 0. It is easy to verify that if we replace the 1 entry $(2,1)$ by -1 than the upper left corner $(1,1)$ becomes a saddle point with the value 0.