RUTCOR
RESEARCH
REPORT

Discounted approximations of undiscounted stochastic games and Markov decision processes are already poor in the almost deterministic case

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

RRR 24-2012, September 2012

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]RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway NJ 08854-8003;
(boros@rutcor.rutgers.edu)

[b]Max-Planck-Institute for Informatics; Stuhlsatzenhausweg 85, 66123, Saarbruecken, Germany (elbassio@mpi-sb.mpg.de)

[c]RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway NJ 08854-8003;
(gurvich@rutcor.rutgers.edu)

[d]Graduate School of Information Science and Technology, University of Tokyo, Tokyo, 113-8656, Japan; (makino@mist.i.u-tokyo.ac.jp)

Rutcor Research Report
RRR 24-2012, September 2012

Discounted approximations of undiscounted stochastic games and Markov decision processes are already poor in the almost deterministic case

Abstract. It is shown that the discount factor needed to solve an undiscounted mean payoff stochastic game to optimality is exponentially close to 1, even in one-player games with a single random node and polynomially bounded rewards and transition probabilities. On the other hand, for the class of the so-called irreducible games with perfect information and a constant number of random nodes, we obtain a pseudo polynomial algorithm using discounts.

Keywords: discounted and undiscounted stochastic games and Markov decision processes, Black, White, and Random positions, Cesaro and Abel sums

1 Introduction and motivation

We consider two-person zero-sum stochastic games with perfect information and mean payoff: Let $G = (V, E)$ be a digraph whose vertex-set V is partitioned into three subsets $V = V_B \cup V_W \cup V_R$ that correspond to black, white, and random positions, controlled respectively, by two players, BLACK - the *minimizer* and WHITE - the *maximizer*, and by nature. We also fix a *local reward* function $r : E \to \mathbb{R}$, and probabilities $p(v, u)$ for all arcs (v, u) going out of $v \in V_R$. Vertices $v \in V$ and arcs $e \in E$ are called *positions* and *moves*, respectively. In a personal position $v \in V_W$ or $v \in V_B$ the corresponding player WHITE or BLACK selects an arc (v, u), while in a random position $v \in V_R$ a move (v, u) is chosen with the given probability $p(v, u)$.

From a given initial position $v_0 \in V$ the game produces an infinite walk (called a *play*). WHITE's objective is to maximize the *limiting mean payoff*

$$c = \liminf_{n\to\infty} \frac{\sum_{i=0}^{n} b_i}{n+1}, \tag{1}$$

where b_i is the expected reward incurred at step i of the play, while the objective of BLACK is the opposite, that is, to minimize c.

For this class of *BWR-games*, it is known that a *saddle point* exists in *pure positional uniformly optimal* strategies. Here "pure" means that the choice of a move (v, u) in a personal position $v \in V_B \cup V_W$ is deterministic; "positional" means that this choice depends solely on v, not on previous positions or moves; finally, "uniformly optimal" means that it does not depend on the initial position v_0, either. This fact was proved by Gillette [Gil57] and Liggett and Lippman [LL69] by considering the *discounted* version, in which the payoff is discounted by a factor β^i at step i, giving the effective payoff:

$$a_\beta = (1 - \beta) \sum_{i=0}^{\infty} \beta^i b_i, \tag{2}$$

and then proceeding to the limit as the *discount factor* $\beta \in [0, 1)$ goes to 1.

The important special case of BWR-games without random vertices, i.e., $V_R = \emptyset$, is known as *cyclic* or *mean payoff* games [Mou76b, Mou76a, EM79, GKK88]. A BWR-game is reduced to a *minimum mean cycle problem* in case $V_W = V_R = \emptyset$ or $V_B = V_R = \emptyset$, which can be solved in polynomial time [Kar78]. If one of the sets V_B or V_W is empty, we obtain a *Markov decision process (MDP)* for which polynomial-time algorithms are also known [MO70]. Finally, if both sets are empty $V_B = V_W = \emptyset$, we get a *weighted Markov chain*.

In the special case of a BWR-game, when all rewards are zero except at a single node t (called the terminal), which has a self-loop with reward 1, we obtain the so-called *simple*

stochastic games (SSGs), introduced by Condon [Con92, Con93] and considered in several papers (e.g. [GH08, Hal07]). In these games, the objective of WHITE is to maximize the probability of reaching the terminal, while BLACK wants to minimize this probability. Recently, it was shown that Gillette games (and hence BWR-games by [BEGM09]) are equivalent to SSGs under polynomial-time reductions [AM09]. At the heart of these reductions is the fact, established in [AM09], that it is enough to take

$$\beta = 1 - [(N!)^2 2^{2N+3} D^{2N^2} R]^{-1} \tag{3}$$

to guarantee that an optimal pair of strategies in the discounted game remains optimal in the undiscounted one. Here, and throughout the paper, $N = |V|$ denotes the total number of nodes, R the maximum absolute value of a reward (assuming integral rewards), and D the common denominator of the transition probabilities (assuming rational transition probabilities).

While there are numerous pseudo-polynomial algorithms known for BW-games (the case when there are no random nodes) [GKK88, Pis99, ZP96], no such algorithm is known for the BWR-case, even if we restrict the number of random vertices. For *ergodic* BWR-games (in which the equilibrium values do not depend on the initial position) with constant number of random nodes, a pseudo-polynomial algorithm was given in [BEGM10], but a similar result in the non-ergodic case was left open.

One approach towards this end is to consider the β-discounted game, which can be solved in time polynomial in the input size and $\frac{1}{1-\beta}$, and then set β sufficiently close to 1. In the absence of random positions, such approach yields indeed a pseudo-polynomial algorithm: to get the exact solution of an undiscounted BW-game with N positions and maximum absolute reward R, it is enough to solve the corresponding β-discounted game with any $\beta > 1 - 1/(4N^3R)$ [ZP96]. However, such approach requires exponential time in the general BWR-case, since one must choose $\beta > 1 - \varepsilon/2^N$ to approximate the value of the game with accuracy ε, as follows follow from an example in [Con92], with only random nodes (that is a weighted Markov chain). We note, however, that the number of random nodes in this example is N, and thus a question that naturally arises is whether one can get a bound similar to (3) in which the exponent $O(N^2)$ is replaced by some function of $|V_R|$ only. If this was the case, it would imply a pseudo-polynomial algorithm for BWR-games with $|V_R| = O(1)$. In this short note, we rule-out this possibility by showing that, in general, the discount factor may need to be chosen exponentially close to 1, even for games with a single player and a single random node, that is, for MDP's where random choices are allowed only for a single action in a single state.

Theorem 1. *There exists a WR-game $\mathcal{G}$ with one random node, $D = O(N)$ and $R = O(N^{20})$, such that solving $\mathcal{G}$ to optimality using discounts requires a discount factor of at least $1 - O(\frac{N^{1/12}}{2^{N/3}})$, where N the total number of nodes.*

Similar lower bounds for more general classes of stochastic games were considered by different authors; see, e.g., the recent survey by Miltersen [Mil11]. However, our bound seems to be the smallest such example, in terms of "the amount of randomness", in the sense that it show that a single random position in an MDP makes an algorithm based on discounts exponentially slower than the one needed for deterministic MDP's.

In contrast, for some special classes of games, β does not have to be exponentially close to 1. In particular, in Section 6, we consider the so-called *irreducible games* defined by Hoffman and Karp [HK66] as follows. For every pair of strategies of the two players, the resulting Markov chain consists of a single absorbing class. As an example, consider a BWR-game in which the graph induced by the set V_R is strongly connected; for every node in V_W (resp., V_B) there is at least one incoming arc from V_R; and there are no arcs between V_W and V_B.

We show in the case of irreducible BWR-games with perfect information and a bounded number of random nodes that it is enough to choose β only polynomially close to 1. In the rest of the paper we denote by $K = |V_R|$ the number of random nodes.

Theorem 2. *Let $\mathcal{G}$ be an irreducible BWR-game. Then for $\beta_* := 1 - [8(NKD)^{2K+4}R]^{-1}$, any uniformly optimal strategy in β_*-discounted game is also uniformly optimal in $\mathcal{G}$.*

As a corollary, we obtain a pseudo-polynomial algorithm for irreducible games, based on solving the corresponding discounted game. We remark that a similar result was obtained in [BEGM10], by a different technique (that does not use discounts) for a more general class of games (the so-called *ergodic* games) which includes irreducible games.

2 A basic lemma

Let n be a positive integer, and P be a set of primes p such that $n \leq p \leq n^2$ and $|P| = 2n$. By Chebyshev's prime number theorem we know that the number $\pi(X)$ of primes not larger than X satisfies the inequalities

$$\frac{7}{8}\frac{X}{\ln X} \;\leq\; \pi(X) \;\leq\; \frac{9}{8}\frac{X}{\ln X}$$

if X is large enough, and thus there are more than $2n$ primes between n and n^2 for all large enough integers n.

For a positive integer k let us denote by $\binom{P}{k}$ the family of k element subsets of P. For a subset $I \subseteq P$ we define

$$r(I) = \sum_{p \in I} \frac{1}{p} \quad \text{and} \quad s(I) = \sum_{p \in I} p. \tag{4}$$

Lemma 1. *Let $n > 0$ be a large enough integer. There exist subsets (possibly multisets) of integers $I, J \subseteq \mathbb{Z}_+$ such that $I \cap J = \emptyset$, $|I| = |J| \leq n + 2$, and such that the following inequalities are satisfied:*

$$0 \ < \ r(J) - r(I) \ \leq \ \frac{\sqrt{n}}{2^{2n-1}}, \tag{5}$$

$$s(I) - s(J) \geq 1, \quad \textit{and} \tag{6}$$

$$s(I) \leq 2n^3. \tag{7}$$

Proof. For the family $\mathcal{F} = \binom{P}{n}$, by Stirling's approximation, we obtain

$$|\mathcal{F}| \geq \frac{2^{2n-1}}{\sqrt{n}}. \tag{8}$$

Let us observe next that for all subsets $I \in \mathcal{F}$ we have $n^{\frac{1}{n^2}} \leq r(I) \leq n^{\frac{1}{n}}$, and thus in particular

$$0 \leq r(I) \leq 1. \tag{9}$$

We claim that if $I, J \in \mathcal{F}$, $I \neq J$, then $r(I) \neq r(J)$. To this end, let us define $D = \prod_{p \in I \cup J} p$ and let $q \in I \setminus J$. Then we have

$$(Dr(I) \mod q) = \left(\sum_{p \in I} \frac{D}{p} \mod q \right) = \left(\frac{D}{q} \mod q \right) \neq 0$$

since $\frac{D}{q}$ is a product of primes different from q. On the other hand, for J we have

$$(Dr(J) \mod q) = \left(\sum_{p \in J} \frac{D}{p} \mod q \right) = 0$$

since $q \notin J$. Thus, by the above claim the reals $r(I)$ for $I \in \mathcal{F}$ are pairwise distinct, and all belong to the unit interval $[0, 1]$ by (9). Therefore, by (8), we must have two subsets, say $I', J' \in \mathcal{F}$ satisfying

$$0 < r(J') - r(I') \leq \frac{\sqrt{n}}{2^{2n-1}}.$$

Then, the sets $\widetilde{I} = I' \setminus J'$ and $\widetilde{J} = J' \setminus I'$ satisfy (5).

If they also satisfy (6), then setting $I = \widetilde{I}$ and $J = \widetilde{J}$ completes our proof, since (7) follows simply by our choice of P.

Otherwise, let us note that we must have

$$s(\widetilde{I}) - s(\widetilde{J}) \geq |\widetilde{I}|n - |\widetilde{J}|n^2 \geq n^2 - n^3, \tag{10}$$

since we have $|\widetilde{I}| = |\widetilde{J}| \leq n$.

Let us then choose an integer a such that

$$n^3 \geq 2a^2 \geq 2(a-1)^2 \geq n^3 - n^2 + 1, \tag{11}$$

and define $I = \widetilde{I} \cup \{a, a(2a-1)\}$ and $J = \widetilde{J} \cup \{2a-1, 2a-1\}$. Note that J became a multiset now in which $2a-1$ has multiplicity 2. With this, we still have $|I| = |J|$.

Furthermore, since $\frac{1}{a} + \frac{1}{a(2a-1)} = \frac{1}{2a-1} + \frac{1}{2a-1}$ we also have $r(J) - r(I) = r(\widetilde{J}) - r(\widetilde{I}) = r(J') - r(I')$, and hence I and J satisfy (5). Finally, we have

$$s(I) - s(J) = s(\widetilde{I}) + a + a(2a-1) - \Big(s(\widetilde{J}) + 2(2a-1)\Big) = s(\widetilde{I}) - s(\widetilde{J}) + 2(a-1)^2 \geq 1,$$

where the last inequality follows by the lower bounds in (10) and (11).

To see (7), let us note that $s(\widetilde{I}) \leq |\widetilde{I}|n^2 \leq n^3$ and $a + a(2a-1) = 2a^2 \leq n^3$ by (11).

Thus, the sets I and J satisfy all claimed inequalities, completing our proof. □

3 Construction

Let us choose two subsets $I, J \subseteq \mathbb{Z}_+$ of integers as in Lemma 1. Let $|I| = |J| = k$, and denote $I = \{p_1, p_2, \cdots, p_k\}$ and $J = \{q_1, q_2, \cdots, q_k\}$. Set $N = 4 + \sum_{j=1}^{k}(p_j + q_j)$, and note that by Lemma 1 we have $N = O(n^3)$.

Let us next associate to this input a WR-game on a directed graph $G = (V, E)$ having N vertices, defined as follows:

$$V = \{w_0, w_1, w_2, w_3\} \cup \left(\bigcup_{j=1}^{k}\{u_0^j, u_1^j, \cdots, u_{p_j-1}^j\}\right) \cup \left(\bigcup_{j=1}^{k}\{v_0^j, v_1^j, \cdots, v_{q_j-1}^j\}\right).$$

Let us define cycles $C_j^u = \{u_0^j, u_1^j), (u_1^j, u_2^j), \cdots, (u_{p_j-1}^j, u_0^j)\}$ and $C_j^v = \{(v_0^j, v_1^j), (v_1^j, v_2^j), \cdots, (v_{q_j-1}^j, v_0^j)\}$ and set the arc set as

$$E = \{(w_0, w_1), (w_0, w_2), (w_2, w_2), (w_3, w_3)\} \cup \bigcup_{j=1}^{k}\{(w_1, u_0^j), (w_1, v_0^j)\} \cup \bigcup_{j=1}^{k}(C_j^u \cup C_j^v).$$

Let the initial node w_0 has three outgoing arcs, left, top and right. The first two arcs have 0 as local rewards, while the third arc (w_0, w_3) has reward 1. The left and right neighbors w_2 and w_3 have single loop arcs with local rewards 0 and $-\frac{(1-\beta_0)}{\beta_0}$, respectively, where $\beta_0 := 1 - \frac{1}{36n^{20}}$. The top neighbor of w_0 is w_1, a random node, with $2k$ outgoing arcs to the nodes u_0^j and v_0^j for $j = 1, ..., k$. All these arcs have rewards 0 and have transition probabilities $\frac{1}{2k}$. The local rewards on the arcs of the cycles C_j^u and C_j^v, $j = 1, ..., k$, are all 0, except the first arcs of the cycles, where we have

$$r(u_0^j, u_1^j) = 1 \quad \text{and} \quad r(v_0^j, v_1^j) = -1 \quad \text{for all } j = 1, ..., k.$$

All the nodes, except w_1, are be controlled by one player, say the maximizer (WHITE), i.e., the game is actually a Markov decision process.

For an illustration see Figure 1. Note that the number of vertices N satisfies: $n \leq N \leq 4(n^3+1)$, and the common denominator of all probabilities $D = 2k \leq 2(n+2)$. Furthermore, by multiplying all the rewards by $36n^{20}-1$ we obtain an equivalent game with integral rewards whose maximum absolute value is $R = 36n^{20}-1$.

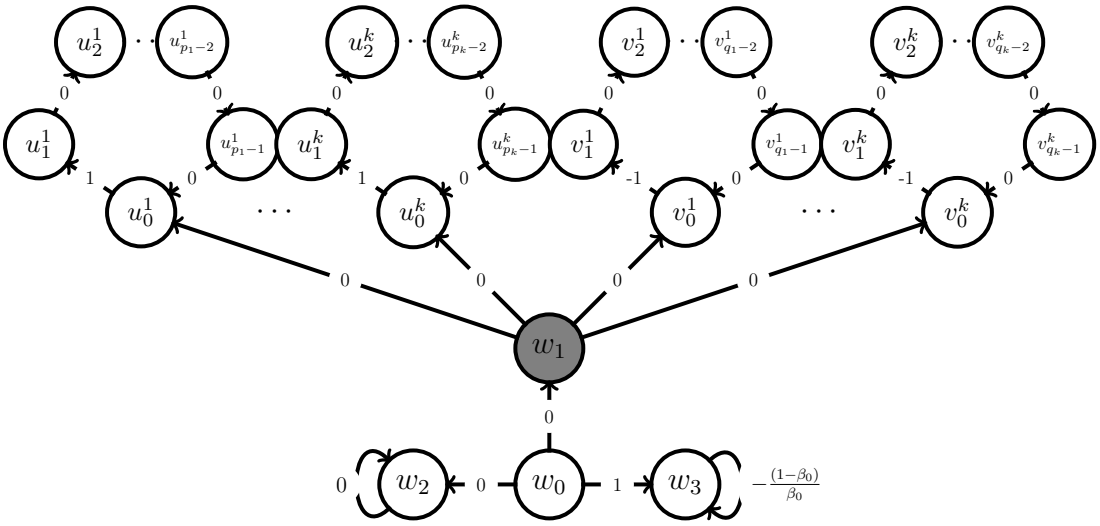

Figure 1: The graph $G = (V, E)$ with a single RANDOM node w_1. The probabilities on the arcs leaving w_1 are all equal to $\frac{1}{2k}$. All other nodes are controlled by WHITE .

4 Game values

Let us denote by $\mu(v)$ the undiscounted game value form initial point v, and denote by $\mu^\beta(v)$ the value from the same initial point of the discounted game with discount factor $0 < \beta < 1$.

Lemma 2. *We have* $\mu(w_2) = \mu^\beta(w_2) = 0$ *and* $\mu(w_3) = \mu^\beta(w_3) = -\frac{(1-\beta_0)}{\beta_0}$, *for all* $0 < \beta < 1$. *Furthermore we have*

$$\mu(w_1) \;=\; \frac{1}{2k}\left(\sum_{p\in I}\frac{1}{p} \;-\; \sum_{q\in J}\frac{1}{q}\right) \;=\; \frac{r(I)-r(J)}{2k}, \quad \textit{and} \tag{12}$$

$$\mu^\beta(w_1) \;=\; \frac{1}{2k}\left(\sum_{p\in I}\frac{(1-\beta)\beta}{1-\beta^p} \;-\; \sum_{q\in J}\frac{(1-\beta)\beta}{1-\beta^q}\right). \tag{13}$$

If WHITE chooses the left strategy at w_0, then her undiscounted and discounted values are $\mu(w_0) = \mu^\beta(w_2) = 0$; if she chooses the top strategy, her values are $\mu(w_0) = \mu(w_1)$ and $\mu^\beta(w_0) = \beta\mu^\beta(w_1)$; and if she chooses the right strategy, her values are $\mu(w_0) = \mu(w_3)$ and $\mu^\beta(w_0) = (1-\beta) + \beta\mu^\beta(w_3)$.

Proof. It is straightforward from the definitions. □

5 Proof of Theorem 1

We show that in the discounted game WHITE can guarantee a positive value by either moving from w_0 to w_3, when $\beta < \beta_0$, or to w_1, when $\beta_0 \le \beta < \beta_1 := 1 - \frac{6n^{1/4}}{2^n}$. This will prove Theorem 1 since the corresponding strategies provide WHITE with a negative value in the undiscounted game, while his optimum is 0, attained by moving to w_2.

To see the first half of the above claim, suppose that WHITE chooses the arc (w_0, w_3). Then $\mu^\beta(w_0) > 0$ for all $\beta \in [0, \beta_0)$ follows by Lemma 2 implying $\mu^\beta(w_0) = 1 - \frac{\beta}{\beta_0}$. In the undiscounted game we have $\mu(w_0) = 1 - \frac{1}{\beta_0} < 0$ for the corresponding strategy. For $\beta > \beta_0$ the discounted value is negative if WHITE moves to w_3.

To complete the proof of the theorem, we consider next the second part. We will use the following Lemma.

Lemma 3. *For any positive integer p we have*

$$\frac{1}{1+\beta+\beta^2+\cdots+\beta^{p-1}} - \frac{1}{p} \;=\; (1-\beta)\frac{p-1}{2p} + (1-\beta)^2\frac{p^2-1}{12p} + (1-\beta)^3 R(p) \qquad (14)$$

where $R(p) = O(p^6)$.

Proof. It follows from Taylor expansion around $\beta = 1$ and routine calculations. Let

$$f(\beta) = \left(\sum_{i=0}^{p-1} \beta^i\right)^{-1} - \frac{1}{p}.$$

Then

$$\begin{aligned}
f'(\beta) &= -\left(\sum_{i=0}^{p-1}\beta^i\right)^{-2}\left(\sum_{i=0}^{p-1} i\beta^{i-1}\right)\\
f''(\beta) &= -\left(\sum_{i=0}^{p-1}\beta^i\right)^{-2}\left(\sum_{i=0}^{p-1} i(i-1)\beta^{i-2}\right)+2\left(\sum_{i=0}^{p-1}\beta^i\right)^{-3}\left(\sum_{i=0}^{p-1} i\beta^{i-1}\right)^2\\
f'''(\beta) &= -\left(\sum_{i=0}^{p-1}\beta^i\right)^{-2}\left(\sum_{i=0}^{p-1} i(i-1)(i-2)\beta^{i-3}\right)+2\left(\sum_{i=0}^{p-1}\beta^i\right)^{-3}\left(\sum_{i=0}^{p-1} i\beta^{i-1}\right)\left(\sum_{i=0}^{p-1} i(i-1)\beta^{i-2}\right)+\\
&\quad 4\left(\sum_{i=0}^{p-1}\beta^i\right)^{-3}\left(\sum_{i=0}^{p-1} i\beta^{i-1}\right)\left(\sum_{i=0}^{p-1} i(i-1)\beta^{i-2}\right)-6\left(\sum_{i=0}^{p-1}\beta^i\right)^{-4}\left(\sum_{i=0}^{p-1} i\beta^{i-1}\right)^3.
\end{aligned}$$

In particular, $f(1)=0$, $f'(1)=-\frac{p-1}{2p}$, $f''(1)=\frac{p^2-1}{6p}$ and for any $0\le\xi\le 1$, $-f'''(\xi)\le 6\left[\binom{p}{4}+\binom{p}{2}^3\right]$. By Taylor expansion of $f(\beta)$ around $\beta=1$,

$$f(p)=f(1)-f'(1)(1-\beta)+\frac{f''(1)}{2}(1-\beta)^2-\frac{f'''(\xi)}{6}(1-\beta)^3,$$

for some $\xi\in[\beta,1]$. Thus, we obtain (14). □

Lemma 4. *Assume that* WHITE *chooses* (w_0,w_1) *as his first move. Then, we have* $\mu(w_0)=\mu(w_1)<0$. *Furthermore, for all discount factors satisfying* $\beta_0\le\beta<\beta_1$, *we have* $\mu^\beta(w_0)=\beta\mu^\beta(w_1)>0$.

Proof. Since k and β here are positive constants, clearly an equivalent statement is that

$$A:=2k\mu(w_1) \;=\; r(I)-r(J) \;<\; 0$$

while

$$B:=\frac{2k}{\beta}\mu^\beta(w_1) \;=\; \sum_{p\in I}\frac{(1-\beta)}{1-\beta^p} \;-\; \sum_{q\in J}\frac{(1-\beta)}{1-\beta^q} \;>\; 0.$$

The first claim follows immediately from (5), so it remains to prove the second claim.

To this end let us note that for a positive integer p we have

$$\frac{(1-\beta)}{1-\beta^p}=\frac{1}{1+\beta+\beta^2+\cdots+\beta^{p-1}}.$$

To continue with the proof of the lemma, let us write

$$X \;:=\; \sum_{p\in I}\frac{p-1}{2p}-\sum_{q\in J}\frac{q-1}{2q} \;=\; \frac{r(J)-r(I)}{2}$$

since we have $|I| = |J|$;

$$Y \;:=\; \sum_{p\in I} \frac{p^2-1}{12p} - \sum_{q\in J} \frac{q^2-1}{12q} \;=\; \frac{1}{12}\left[(s(I)-s(J)) + (r(J)-r(I))\right];$$

and

$$Z \;:=\; \sum_{p\in I} R(p) - \sum_{q\in J} R(q), \quad |Z| \;=\; O(n^{19})$$

since $p \le n^3$ for $p \in I \cup J$, and $|I| = |J| = k \le n+2$.

The above and Lemma 1 imply that

$$X \ge 0, \quad Y \ge \frac{1}{12}, \quad \text{and} \quad Z \ge -Cn^{19},$$

for a constant C. Thus, we get by Lemma 3

$$\begin{aligned} B &= A + (1-\beta)X + (1-\beta^2)Y + (1-\beta^3)Z \\ &\ge A + \tfrac{1}{12}(1-\beta)^2 - (1-\beta)^3 Cn^{19} \\ &\ge A + (1-\beta)^2 \tfrac{1}{18} \end{aligned}$$

for all discount factors $\beta \ge 1 - \frac{1}{36Cn^{19}}$. Consequently, for all discount factors satisfying

$$1 - \frac{6n^{1/4}}{2^n} > \beta \ge 1 - \frac{1}{36Cn^{19}}$$

we have $B > 0$, proving the lemma. □

6 A bound on the discount factor for the irreducible case

In the following we let $\mathcal{G}$ be an irreducible BWR-game with integral rewards of maximum absolute value R and rational transition probabilities with common denominator at most D. For $\beta \in (0,1]$, we will denote by $\mathcal{G}^\beta$ the corresponding β-discounted game. For a situation (that is, a pair of pure stationary strategies) $s = (s_W, s_B)$, we will denote by $\mu_s^\beta \in \mathbb{R}^V$ the value vector of the β-discounted Markov chain defined by s, and by $\bar{r}_s \in \mathbb{R}^V$ the vector with components $\bar{r}_s(v) = \sum_{u:(v,u)\in E} P_s(v,u) r(v,u)$, for $v \in V$, where $P_s \in [0,1]^{N\times N}$ is the stochastic matrix defined as

$$\begin{aligned} P_s(v,u) &= p(v,u) \quad \text{if } v \in V_R; \\ P_s(v,u) &= 1 \quad \text{if } v \in V_W \text{ and } u = s_W(v) \text{ or } v \in V_B \text{ and } u = s_B(v); \\ P_s(v,u) &= 0 \quad \text{if } v \in V_W \text{ and } u \ne s_W(v) \text{ or } v \in V_B \text{ and } u \ne s_B(v). \end{aligned}$$

(Note that $\bar{r}_s(v)$ is the expected payoff of the next move at v.) By (2), the vector $\mu_s^\beta = (\mu_s^\beta(v) : v \in V)$ of values, obtained if we fix situation s, is given by

$$\mu_s^\beta = (1-\beta)\sum_{i=0}^{\infty} \beta^i P_s^i \bar{r}_s = (1-\beta)(I-\beta P_s)^{-1}\bar{r}_s. \tag{15}$$

To prove Theorem 2, we use essentially the same argument as in [AM09], but involving a more careful analysis[1]. In the following, we assume without loss of generality that $K \geq 1$.

Lemma 5. *For any situation s in $\mathcal{G}^\beta$, each component of μ_s^β is a rational polynomial $\frac{q(\beta)}{p(\beta)}$ in β, with the following properties: $p(\beta)$ is of degree at most $(N-K+1)K$ and integer coefficients bounded by $(K!)N^K D^K$ in absolute value; and $q(\beta)$ is of degree at most $(N-K+1)K+N-K$ and integer coefficients bounded by $2[(K!)N^K D^K]R$, in absolute value.*

Proof. Fix a situation s in $\mathcal{G}^\beta$, and for simplicity, write $P = P_s$, $\bar{r}_s = \bar{r}$, and $\mu = \mu_s^\beta$. Then (15) can be rewritten as:

$$(I-\beta P)\mu = (1-\beta)\bar{r}. \tag{16}$$

Since the game is irreducible, unless it is a BW-game, for each deterministic node $u \in V_W \cup V_B$ there is a path of deterministic nodes $u = u_0, u_1, \ldots, u_{\ell(u)} = \bar{u}$, leading to a (unique) random node $\bar{u} \in V_R$. Then (16) allows us to write $\mu(u)$ as a function of $\mu(\bar{u})$:

$$\mu(u) = \beta^\ell \mu(\bar{u}) + (1-\beta)\sum_{i=0}^{\ell-1} \beta^i r(u_{i-1}, u_i) = \beta^\ell \mu(\bar{u}) + f_u(\beta, r), \tag{17}$$

where $\ell = \ell(u)$ and $f_u(\beta, r)$ is a polynomial with degree at most ℓ in β each coefficient of which is of the form $r(e_1) \pm r(e_2)$, for some arcs $e_1, e_2 \in E$.

Substituting μ_u, for $u \in V_W \cup V_B$, from (17) in (16), we get, for $v \in V_R$,

$$\mu(v) = \beta \sum_{u \in V_W \cup V_B} p(v,u)(\beta^{\ell(u)}\mu(\bar{u}) + f_u(\beta, r)) + \beta \sum_{u \in V_R} p(v,u)\mu(u) + (1-\beta)\bar{r}(v).$$

Rearranging terms, we end-up with a *full-rank* system of equations $A(\beta)x = b(\beta)$ in the vector of values $x = \mu[V_R]$ of the random nodes, where

(*i*) each entry a_{ij} of $A(\beta)$ is a polynomial in β with degree at most $N-K+1$ and rational coefficients of the form $\frac{s}{D}$, where $s \in \mathbb{Z}$, $|s| \leq ND$; and

(*ii*) each entry b_i of $b(\beta)$ is a polynomial in β with degree at most $N-K+1$ and rational coefficients of the form $\frac{s}{D}$, where $s \in \mathbb{Z}$ and $|s| \leq 2NRD$.

[1] We emphasize that we made no attempt at optimizing the exponents in the bounds.

By (i), the determinant of $A(\beta)$ is a polynomial in β with degree at most $(N-K+1)K$ and rational coefficients of the form $\frac{s}{D^K}$, where $s \in \mathbb{Z}$ and $|s| \leq (K!)(ND)^K$. By this, (ii), and Cramer's rule, each component of the solution vector x is of the form $\frac{q(\beta)}{p(\beta)}$, where $p(\beta)$ and $q(\beta)$ satisfy the properties stated in the claim. This establishes the claim of the lemma for $v \in V_R$. The corresponding claim for $v \in V_W \cup V_B$, can be seen to hold if we substitute the values for the random nodes into (17). □

Lemma 6. *Let $f(\beta) = \sum_{i=0}^{d} a_i \beta^i$ be a polynomial in β with degree $d \geq 1$ and integer coefficients a_i, each bounded from above by α in absolute value, and assume that $f(1) \neq 0$. Let $\beta_* = 1 - \frac{1}{4d^2\alpha}$. Then $f(\beta)$ has no root in the interval $[\beta_*, 1]$ (and hence $f(\beta)$ does not change its sign in this interval).*

Proof. Let us do the substitution $\beta := 1 - \gamma$, to get a polynomial $g(\gamma) = \sum_{j=0}^{d} b_j \gamma^j$. It can be easily seen that $b_j = (-1)^j \sum_{i=j}^{d} \binom{i}{j} a_i$, and hence, $|b_j| \leq \alpha d^{j+1}$. It follows that, for any $\gamma \in [0, 1-\beta_*]$,

$$|\sum_{j=1}^{d} b_j \gamma^j| \leq \alpha \sum_{j=1}^{d} d^{j+1}\gamma^j \leq \alpha \sum_{j=1}^{d} \frac{1}{d^{j-1}(4\alpha)^j} \leq \alpha \sum_{j=1}^{d} \frac{1}{(4\alpha)^j} < \frac{\alpha}{4\alpha - 1} < 1. \tag{18}$$

By assumption $f(1) = g(0) = b_0 \neq 0$. If $f(\beta)$ has a root in $[\beta_*, 1]$ then $g(\gamma)$ has root in $[0, 1-\beta_*]$. But this is not possible since (18) implies that $|g(\gamma)| = |b_0 + \sum_{j=1}^{d} b_j \gamma^j| \geq |b_0| - |\sum_{j=1}^{d} b_j \gamma^j| > 0$, for any $\gamma \in [0, 1-\beta_*]$. The statement follows. □

Proof of Theorem 2. Let s_* be a uniformly optimal situation in $\mathcal{G}^{\beta_*}$. To show the (uniform) optimality of s^* in $\mathcal{G}^{\beta}$ for any $\beta \in [\beta_*, 1]$, it is enough to show that, for any $v \in V$ and any arbitrary situation s, the difference $\mu^{\beta}_{s_*}(v) - \mu^{\beta}_{s}(v)$ does not change sign in the interval $[\beta_*, 1]$. By Lemma 5, we can write $\mu^{\beta}_{s_*}(v))$ and $\mu^{\beta}_{s}(v)$ respectively as rational polynomials $\frac{q_1(\beta)}{p_1(\beta)}$ and $\frac{q_2(\beta)}{p_2(\beta)}$ in β, with the numerators and denominators satisfying the properties stated in the lemma. Then the difference can be written as $\frac{q_1(\beta)p_2(\beta) - p_1(\beta)q_2(\beta)}{p_1(\beta)p_2(\beta)} = \frac{(1-\beta)^i q_3(\beta)}{p_3(\beta)}$, where $i \in \mathbb{Z}$, $p_3(\beta)$ and $q_3(\beta)$ are polynomials with degree at most $d = (NK)^2$, and integer coefficients bounded by $\alpha = 2[(K!)N^K D^K]^2 R$, and such that $p_3(1) \neq 0$ and $q_3(1) \neq 0$. Now we invoke Lemma 6 with $\alpha = 2[(K!)N^K D^K]^2 R$ and $d = (NK)^2$ to get the conclusion of the theorem. □

We will show now that Theorem 2 implies that a discounting algorithm can be used to solve any undiscuounted irreducible BWR-game.

Lemma 7. *Let $\mathcal{G}$ be an irreducible BWR-game and $\beta = 1 - \frac{1}{B} \in [0, 1)$ be a rational number, with $B \in \mathbb{Z}_+$. Then for any v, $\mu^{\beta}(v)$ is a rational number of the form $\frac{s}{t}$, where $s, t \in \mathbb{Z}$, and $t \leq t_0(N, K, D, B) := 2(K+1)!N^{K+1}D^K B^{2(NK)}$.*

Proof. This is immediate from Lemma 5, since $\mu^\beta(v)$ can be written as $\frac{B^{d'-d}\sum_{i=0}^{d} a_i B^{d-i}(B-1)^i}{\sum_{i=0}^{d'} b_i B^{d'-i}(B-1)^i}$, where $d, d' \leq 2(NK)$, $|a_i| \leq 2[(K!)N^K D^K]R$, and $|b_i| \leq [(K!)N^K D^K]$. □

Corollary 1. *For $K = O(1)$, there is a pseudo-polynomial algorithm that solves, in uniformly optimal strategies, any irreducible BWR-game $\mathcal{G}$ with integral rewards and rational transition probabilities, in time $O((NKD)^{2K+6}|E|R\log R)$.*

Proof. It is known (see e.g. [BEGM09]) that for $\beta \in (0,1)$, the values of any discounted BWR-game can be approximated within an absolute error of ε in $\frac{\log R - \log \varepsilon}{1-\log(1+\beta)} \sim 2\ln 2 \frac{(\log R - \log \varepsilon)}{(1-\beta)}$ iterations, each taking $O(|E|)$ time. By Theorem 2 and Lemma 7, it is enough to take $\beta = \beta_*$, and $\varepsilon = \frac{1}{t_0^2}$, where $t_0 = t_0(N, D, K, B)$ and $\beta_* = 1 - 1/B$. Plugging in the value for $B = 8(NKD)^{2K+4}R$, we arrive at the bound stated in the theorem. □

References

[AM09] D. Andersson and P. B. Miltersen. The complexity of solving stochastic games on graphs. In *Proc. 20th ISAAC*, volume 5878 of *LNCS*, pages 112–121, 2009.

[BEGM09] E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. Every stochastic game with perfect information admits a canonical form. RRR-09-2009, RUTCOR, Rutgers University, 2009.

[BEGM10] 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*, pages 341–354, 2010.

[Con92] A. Condon. The complexity of stochastic games. *Information and Computation*, 96:203–224, 1992.

[Con93] A. Condon. An algorithm for simple stochastic games. In *Advances in computational complexity theory, volume 13 of DIMACS series in discrete mathematics and theoretical computer science*, 1993.

[EM79] A. Eherenfeucht and J. Mycielski. Positional strategies for mean payoff games. *International Journal of Game Theory*, 8:109–113, 1979.

[GH08] H. Gimbert and F. Horn. Simple stochastic games with few random vertices are easy to solve. In *Proc. 11th FoSSaCS*, volume 4962 of *LNCS*, pages 5–19, 2008.

[Gil57] 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.

[GKK88] 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.

[Hal07] N. Halman. Simple stochastic games, parity games, mean payoff games and discounted payoff games are all LP-type problems. *Algorithmica*, 49(1):37–50, 2007.

[HK66] A. J. Hoffman and R. M. Karp. On non-terminating stochastic games. *Management Science*, 12:359–370, 1966.

[Kar78] R. M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Math.*, 23:309–311, 1978.

[LL69] T. M. Liggett and S. A. Lippman. Stochastic games with perfect information and time-average payoff. *SIAM Review*, 4:604–607, 1969.

[Mil11] P. B. Miltersen. Discounted stochastic games poorly approximate undiscounted ones, manuscript. Technical report, 2011.

[MO70] H. Mine and S. Osaki. *Markovian decision process*. American Elsevier Publishing Co., New York, 1970.

[Mou76a] H. Moulin. Extension of two person zero sum games. *Journal of Mathematical Analysis and Application*, 5(2):490–507, 1976.

[Mou76b] H. Moulin. Prolongement des jeux à deux joueurs de somme nulle. *Bull. Soc. Math. France, Memoire*, 45, 1976.

[Pis99] N. N. Pisaruk. Mean cost cyclical games. *Mathematics of Operations Research*, 24(4):817–828, 1999.

[ZP96] U. Zwick and M. Paterson. The complexity of mean payoff games on graphs. *Theoret. Comput. Sci.*, 158(1-2):343–359, 1996.