

Discounted stochastic games poorly approximate undiscounted ones

Peter Bro Miltersen*

March 30, 2011

The purpose of this note is to summarize recent results [6, 4, 5] on stochastic games by the author and his collaborators. These results have appeared or will appear in proceedings of computer science conferences. The intended reader of this note has an interest in finite stochastic games, but no particular interest in computation, that is, the computational aspects of the results obtained have been weeded out.

We consider two-player zero-sum finite (but infinite duration) stochastic games G with N positions and at most m actions available to each of the two players in each position. The reward to Player I when Player I plays i and Player II plays j in position k is denoted a_{ij}^k . Transition probabilities are denoted p_{ij}^{kl} . We assume stopping probabilities are 0, i.e., for all k, i, j we have $\sum_l p_{ij}^{kl} = 1$. To be able to state our results as simply as possible, we shall also assume throughout that for all k, l, i, j , we have $a_{ij}^k, p_{ij}^{kl} \in \{0, 1\}$. In particular, we assume deterministic dynamics of nature and non-negative payoffs.

By G we denote the game with limiting average (undiscounted) payoffs, i.e., payoff $\liminf_{t \rightarrow \infty} (\sum_{i=0}^{t-1} r_i)/t$ to Player I, where r_i is the reward collected by Player I at stage i . By G_λ we denote the game with payoffs discounted with a discount factor of $1 - \lambda$, i.e., with payoff $\lambda \sum_{i=0}^{\infty} (1 - \lambda)^i r_i$ to Player I. By G_T we denote the finite game with T stages and payoff $(\sum_{i=0}^{T-1} r_i)/T$ to Player I. We shall be interested in the special case where G is a recursive game in the sense of Everett [3]. In a recursive game, all non-zero rewards occur at absorbing states with only one action available to each player ("terminal states"). When G is recursive, we let G'_T denote the finite game with payoff r_{T-1} to Player I (i.e., the payoff is the reward Player I collects at the last stage of play). Note that when rewards are non-negative, $\text{val}(G_T) \leq \text{val}(G'_T)$.

The seminal result of Mertens and Neyman [7] states:

Theorem 1 (Mertens and Neyman).

$$\text{val}(G) = \lim_{T \rightarrow \infty} \text{val}(G_T) = \lim_{\lambda \rightarrow 0^+} \text{val}(G_\lambda).$$

In the works reported here, we ask: *How good are the approximations of Theorem 1 as a function of the combinatorial parameters of the game, i.e., N and m ?*

We provide lower bounds as well as upper bounds on the badness of the approximations. For the lower bounds, we exhibit for any $N \geq 1, m \geq 2$, a particular recursive game $P(N, m)$ (named "Generalized Purgatory" [4]), where the approximations are quite bad. We describe this game as being played between Dante (Player I) and Lucifer (Player II) as follows. *Lucifer repeatedly selects integers between 1 and m , hidden from Dante's view. Dante has to try to guess each of them. If he guesses correctly N times in a row, he wins the game (payoff 1). If he ever overshoots Lucifer's number he loses the game (payoff 0).*¹ It can be shown that the value of $P(N, m)$ is 1 (and Section 2 of this paper contains a proof for the case of $m = 2$). In Hansen, Koucky and Miltersen [6] and Hansen, Ibsen-Jensen and Miltersen [4], the following bounds are shown.

Theorem 2. *Let $\epsilon > 0$ be fixed. For fixed $m \geq 2$ and as a function of N , $\text{val}(G'_T) = o(1)$ and $\text{val}(G_\lambda) = o(1)$ for $G = P(N, m)$, $T = 2^{m^{(1-\epsilon)N}}$ and $\lambda = 2^{-m^{(1-\epsilon)N}}$.*

*Aarhus University, Computer Science Department. Work supported by Center for Algorithmic Game Theory, funded by the Carlsberg Foundation. Work also supported by the Sino-Danish Center for the Theory of Interactive Computation, funded by the Danish National Research Foundation and the National Science Foundation of China (under the grant 61061130540).

¹note that N is the number of non-absorbing states of $P(N, m)$; the total number of states is $N + 2$.

As $\text{val}(P(N, m)) = 1$ and all rewards are 0 or 1, the theorem suggests that essentially no meaningful approximation is obtained from the time-bounded and discounted versions of the game unless the time bound (resp., one minus the discount factor) is doubly exponentially large (resp., small) in the number of states of the game. The proof is elementary, but somewhat involved in the general case. In Section 2, we give a complete proof of a concrete (non-asymptotic) version of the theorem for the case $m = 2$ ("Purgatory", [6]).

To get upper bounds, the "big gun" of semi-algebraic geometry [1], a.k.a., the model theory of the first order theory of the real numbers, is applied, much in line with the classical works on stochastic games [2, 7]. The best bounds are obtained for recursive games. In Hansen *et al.* [5] we apply the *sampling theorem* [1, Theorem 13.11] to formalizations of statements of Everett in the first order theory of real numbers and get a lower bound on the *patience* of one ϵ -optimal stationary strategy of a given recursive games. In Hansen, Ibsen-Jensen and Miltersen [4] we show how lower bounds on patience of near-optimal strategies of a game yield bounds on the value of its time bounded version, which immediately imply bounds on the value of its discounted version. Overall, we obtain:

Theorem 3 (Hansen et al., [5, 4]). *There is a constant c so that the following holds. For all $N \geq 1, m \geq 2$ and all recursive games G with N states and m actions for each player in each state, 0-1 rewards, and deterministic dynamics of nature, we have for all $0 < \epsilon \leq \frac{1}{2}$:*

$$T \geq (1/\epsilon)^{m^{cN}} \Rightarrow |\text{val}(G) - \text{val}(G_T)| \leq \epsilon.$$

$$\lambda \leq \epsilon^{m^{cN}} \Rightarrow |\text{val}(G) - \text{val}(G_\lambda)| \leq \epsilon.$$

Note that up to the constant c , the dependence on m and N is tight, by Theorem 2. Unfortunately, it is hard to extract an explicit value for c from the literature on semi-algebraic geometry (in particular, we do not know a version of the sampling theorem without unspecified constants). Interestingly, if one does not mind getting m in the upper exponent, the sampling theorem can be avoided, and the explicit bounds $T = (1/\epsilon)^{2^{31mN}}$, $\lambda = \epsilon^{2^{31mN}}$ for $N \geq 10$ can be obtained [6].

Tightening the upper and lower bounds for the case of recursive games is currently work in progress. We conjecture that in fact the example of $P(N, m)$ is *extremal* among recursive games with 0-1 rewards and deterministic dynamics of nature, i.e., that it exhibits the largest possible difference in value between the infinite and the time bounded versions of the game among all games of parameters N and m . In fact, work in progress by Ibsen-Jensen using combinatorial techniques rather than semi-algebraic geometry seems to show that this conjecture can be established to be true for games of value 1. If true in general, it would mean that the constant c in Theorem 3 can be made approximately 1.

For the general case of stochastic games we do not know how to do quite as well as for recursive games. On the other hand, we essentially get a "free lunch": The fact that the theorem of Mertens and Neyman is true, an inspection of its formalization in the first order theory of the real numbers, and standard theorems of semi-algebraic geometry, including the sampling theorem and the quantifier elimination theorem [1, Theorem 14.16] together imply a bound on the accuracy of the approximation as a function of the combinatorial parameters of the game:

Theorem 4 (Hansen et al., [5]). *There is a constant c' so that the following holds. For all $N \geq 1, m \geq 2$ and all stochastic games G with N states and m actions for each player in each state, 0-1 rewards, and deterministic dynamics of nature, we have for all $0 < \epsilon \leq \frac{1}{2}$:*

$$T \geq (1/\epsilon)^{m^{c'N^2}} \Rightarrow |\text{val}(G) - \text{val}(G_T)| \leq \epsilon.$$

$$\lambda \leq \epsilon^{m^{c'N^2}} \Rightarrow |\text{val}(G) - \text{val}(G_\lambda)| \leq \epsilon.$$

We do not know examples of families of stochastic games that suggest that the quadratic dependence on N in the upper exponent is necessary and consider it an interesting problem to remove this dependence.

1 Analysis of $P(N, 2)$

In this section we present a self-contained analysis of the value of $P(N, 2)$ and the value of its time bounded version $P(N, 2)'_T$, optimizing simplicity. More precise bounds can be found in [6, 4]. The game $P(N, 2)$, called Purgatory in [6], can be conveniently reformulated as a repeated matching pennies game: *Lucifer repeatedly hides a penny. Dante has to guess if it is heads up or tails up. If he guesses correctly N times in a row, he wins the game (payoff 1, Paradise). If he ever incorrectly guesses “tails” he loses the game (payoff 0, Hell).* We call the state of the game where Dante already guessed correctly i times in a row for *terrace i* . Thus, terraces are numbered $0, 1, \dots, N - 1$.

1.1 The value of $P(N, 2)$ is 1

First we show that the value of $P(N, 2)$ is 1. This is an induction in N . Specifically, we will assume that we for any $\epsilon > 0$ can construct a stationary strategy for Dante guaranteed to win $P(N - 1, 2)$ with probability at least $1 - \epsilon$ and use this as an induction hypothesis to show the same statement for $P(N, 2)$. By convention, $P(0, 2)$ is the game that Dante wins immediately, so the induction basis is trivial.

In the induction step, we construct a stationary strategy that guarantees Dante a win with probability at least $1 - \epsilon^4$ in $P(N - 1, 2)$. In $P(N, 2)$, we use this strategy for terraces $0, 1, \dots, N - 2$ and must specify how Dante should play on terrace $N - 1$: He guesses “heads” with probability $1 - \epsilon^2$ and “tails” with probability ϵ^2 .

To show that Dante playing by the constructed strategy is guaranteed to win $P(N, 2)$ with probability at least $1 - \epsilon$, we assume that Dante commits to playing this strategy and that Lucifer plays a best reply. With Dante’s strategy frozen, Lucifer’s best reply is essentially in a one-person stochastic game (a.k.a. a Markov decision process) and it can therefore be assumed to be pure as well as stationary. There are two cases to check.

- Lucifer hides the penny heads up at terrace $N - 1$. Then, by induction, Dante reaches terrace $N - 1$ with probability $1 - \epsilon^4$ and then (correctly) guesses heads with probability $1 - \epsilon^2$. Thus, he wins the game with probability more than $1 - \epsilon^4 - \epsilon^2$.
- Lucifer hides the penny tails up at terrace $N - 1$. Then, the resulting dynamics may be represented by a Markov process with two transient states A and B and two absorbing states P (aradise) and H (ell), with state A representing terraces $0, 1, 2, \dots, N - 2$ and state B representing terrace $N - 1$. The probability of going from B to P is ϵ^2 , while the probability of going from B to A is $1 - \epsilon^2$. The probability of going from A to H is (at most) ϵ^4 , while the probability of going from A to B is (at least) $1 - \epsilon^4$. As play starts in state A , we may analyze the probability of ending in P as follows: The probability of getting immediately from A to B is at least $1 - \epsilon^2$. In plays starting in B , the probability of eventually ending in H rather than P is less than $\epsilon^4/\epsilon^2 = \epsilon^2$. Thus, the probability of ending in H when starting in A is less than $\epsilon^4 + \epsilon^2 < \epsilon$.

1.2 The value of $P(N, 2)'_T$

We show that for $N \geq 4$ and $T = 2^{2^{N-1}}$, we have $\text{val}(P(N, 2)'_T) < 0.68$. In words, no strategy of Dante guarantees him a win with probability 0.68 using less than $2^{2^{N-1}}$ guesses. We note that if there are $N = 7$ terraces of Purgatory and Dante makes one guess per second, $T = 2^{2^{N-1}}$ is more than 500 billion years.

We show the bound by exhibiting a strategy of Lucifer preventing Dante to win with probability more than 0.68: *Lucifer hides the penny heads up with probability $2^{-2^{N-1-j}}$ at terrace j .* We may assume that Dante plays a pure (but not necessarily stationary) reply to this strategy.

In $P(N, 2)'_T$, Dante has a budget of T guesses. For the analysis, we make the game slightly better for him by allowing him T epochs rather than T guesses, each epoch beginning when Dante finds himself at terrace 0, i.e., an epoch begins at the first stage of the game and each time Dante has incorrectly guessed “heads”.

We can assume that Dante at the beginning of each epoch plans ahead what he will guess at each stage of the epoch (optimistically assuming that he guessed correctly at all previous stages making the epoch continue to the stage under consideration). Assuming this, we divide epochs into two kinds: *Chicken epochs*,

where Dante plans to guess “heads” at every stage, and *risky epochs*, where Dante plans to guess “tails” at some stage. We will show that the probability that Dante wins in some chicken epoch is at most $1/128$ and the probability that he wins in some risky epoch is at most $2/3$. Since $1/128 + 2/3 < 0.68$, our claim follows.

First, we show that the probability that Dante wins in some chicken epoch is at most $1/128$. Indeed, fix a chicken epoch, and let W be the event that Dante wins during that epoch. Then, by the definition of Lucifer’s strategy, we have

$$\Pr[W] = 2^{-2^{N-1}} 2^{-2^{N-2}} \dots 2^{-2^0} = 2^{-2^N+1}.$$

Since there are only $2^{2^{N-1}}$ epochs altogether, the probability of a win in a chicken epoch is by the union bound at most $2^{2^{N-1}} \cdot 2^{-2^N+1} = 2^{-2^{N-1}+1} \leq 1/128$ since $N \geq 4$.

Next, we show that the probability that Dante wins in some risky period is at most $2/3$. Indeed, fix a risky epoch and let W denote the probability of winning during this epoch. Let L denote the probability of losing (the entire game) during this epoch. Let j be the smallest j so that Dante plans to guess “tails” at terrace $N - 1 - j$. Let J be the event of reaching terrace $N - 1 - j$ during the epoch. We note that if we can show $\Pr[W]/\Pr[L] \leq 2$, we are done (since the risky epoch fixed is arbitrary). But to show $\Pr[W]/\Pr[L] \leq 2$, it is enough to show $\Pr[W|J]/\Pr[L|J] \leq 2$, since W is a subevent of J . But by definition of Lucifer’s strategy, we have:

$$\Pr(L|J) = 2^{-2^j}$$

and

$$\Pr(W|J) = (1 - 2^{-2^j}) 2^{-2^{j-1}} 2^{-2^{j-2}} \dots 2^{-2^0} \leq 2^{-2^j+1}$$

and we are done.

References

- [1] S. Basu, R. Pollack, and M. Roy. *Algorithms in Real Algebraic Geometry*. Springer, 2nd edition, 2006.
- [2] Truman Bewley and Elon Kohlberg. The asymptotic theory of stochastic games. *Mathematics of Operations Research*, 1:197–208, 1976.
- [3] H. Everett. Recursive games. In *Contributions to the Theory of Games Vol. III*, volume 39 of *Ann. Math. Studies*, pages 67–78. Princeton University Press, 1957.
- [4] K.A. Hansen, R. Ibsen-Jensen, and P.B. Miltersen. The complexity of solving reachability games using value and strategy iteration. In *6th Int. Comp. Sci. Symp. in Russia, CSR*, LNCS. Springer, 2011.
- [5] K.A. Hansen, M. Koucký, N. Lauritzen, P.B. Miltersen, and E.P. Tsigaridas. Exact algorithms for solving stochastic games. In *Proc. 43rd Annual ACM Symp. Theory of Computing (STOC)*, 2011. (To appear).
- [6] K.A. Hansen, M. Koucký, and P.B. Miltersen. Winning concurrent reachability games requires doubly exponential patience. In *Proc. of IEEE Symp. on Logic in Comp. Sci., LICS*, pages 332–341, 2009.
- [7] J.F. Mertens and A. Neyman. Stochastic games. *Int. J. of Game Theory*, pages 53–66, 1981.