

## Lecture 9 - part 2: Undiscounted stochastic games

Lecturer: Peter Bro Miltersen

Scribe: Kasper Damgaard

## 1 Undiscounted Stochastic games

We consider two-player zero-sum stochastic games in the undiscounted payoff model. In the undiscounted payoff model, the payoff to player I is given by the  $\liminf$ , as  $t$  approaches infinity, of the weighted average of the rewards up to iteration  $t$ :

$$\liminf_{t \rightarrow \infty} \frac{r_1 + r_2 + \dots + r_t}{t}$$

The following theorem is the analogue of the minimax theorem for this class of games. The sets  $S_1$  and  $S_2$  is the sets of unrestricted strategies of the two players. We make the *perfect monitoring* assumption meaning that a player is informed of all actions his opponent made in the past.

**Theorem 1 (Mertens and Neyman 1978)** *For a two-player undiscounted stochastic game,*

$$\sup_{x \in \tilde{S}_1} \inf_{y \in S_2} u_1(x, y) = \inf_{y \in S_2} \sup_{x \in S_1} u_1(x, y)$$

As usual, we refer to this as the value of the game.

The proof is quite complicated, but takes advantage of the fact that we understand the discounted version very well. But for undiscounted payoffs, it is *not* true in general that

$$\max_{x \in S_1} \min_{y \in S_2} u_1(x, y) = \min_{y \in S_2} \max_{x \in S_1} u_1(x, y)$$

In fact, these two expressions are not even well-defined. In particular, for some games, such as “The big match” below, for any strategy  $x$  of Player 1, we can find a strategy  $x'$  with a strictly better guarantee. Also, there are games where  $\sup_{x \in \tilde{S}_1} \inf_{y \in S_2} u_1(x, y) < \inf_{y \in \tilde{S}_2} \sup_{x \in S_1} u_1(x, y)$  where  $\tilde{S}_i$  is the set of behavior strategies of Player  $i$ . That is, the value of the game can not even be *approximated* using behavior strategies. Again, “The big match” below is an example of this.

### 1.1 The big match

The small and very unfair game Matching Pennies is the game where Player 2 hides a penny heads up or tails up and Player 1 guesses if it is head up or tails up. If he guesses correctly, he wins the penny. This game is a matrix game with the following payoff matrix, with the unique maximin and minimax strategies being the uniform distributions.

	H	T	
h	1	0	$\frac{1}{2}$ $\frac{1}{2}$
t	0	1	
	$\frac{1}{2}$	$\frac{1}{2}$	

H is hiding heads and T is hiding tails while h is guessing heads and t is guessing tails. This slightly boring game has a value of  $\frac{1}{2}$ , and is in itself not very interesting. We therefore expand the game into *The Big Match*, which can be described the following way:

	H	T
h	A	B
t	C	D

Where

A:= reward of 1 and repeat the game.

B:= reward of 0 and repeat the game.

C:= reward of 0 and repeat a sub-game yielding an infinite sequence of rewards of 0.

D:= reward of 1 and repeat a sub-game yielding an infinite sequence of rewards of 1.

As usual player I would like to maximize the expected payoff and player II would like to minimize it. If player II chooses uniformly at random, he is guaranteed an expected loss of at most  $1/2$ . Thus we get that  $\inf_y \sup_x u_1(x, y) \leq \frac{1}{2}$ .

Is it possible for player I to get  $\sup_x \inf_y \geq \frac{1}{2}$  with the sup being over the set of behavior strategies? The answer is no. Actually the best behavior strategy can only yield a guarantee on the expected payoff of 0, since there will always be a way for player II to counter the behavior efficiently (there are two cases: If Player 1 puts any positive probability mass on t, Player 2 can play H, and if Player 1 puts probability mass 0 on t, Player 2 can play T). Therefore player I has to think of a more general strategy, considering the history of the game and reacting on what player II did earlier. It can be shown that the following strategy achieves a guarantee of  $1/2 - \epsilon$ , where the number  $k_\epsilon$  is a (large!) positive integer depending on  $\epsilon > 0$ .

$$\text{Play t with probability} = \frac{1}{(\#heads_2 - \#tails_2 + k_\epsilon)^2}$$

$$\text{Play h with probability} = 1 - \frac{1}{(\#heads_2 - \#tails_2 + k_\epsilon)^2}$$

Here,  $\#heads_2/\#tails_2$  is the number of times player II chose heads/tails so far. We shall not prove that this strategy has a guarantee of  $1/2 - \epsilon$ , but we note that the strategy has a very intuitive explanation:

In the beginning, Player I will (probably) guess a long sequence of heads. If Player II hides strictly more heads than tails in “the long run”, Player I will (probably) *never* guess tails (for this, it is important that the expression in the denominator is squared!). On the other hand, if Player II hides at least as many tails as heads in “the long run”, Player I will at some distant point in time surprise Player II by suddenly guessing tails (thereby “freezing” the play of the game), and, since Player II is taken by surprise by this, it is roughly as likely that he was hiding tails as heads at this point in time.

In conclusion, the value of “the big match” is  $1/2$ , but Player I can not achieve this value exactly at all, he can only approximate it, and to approximate it, a behavior strategy is not sufficient.

Though some work has been done on computing strategies (and values) based on the description of general undiscounted stochastic games, no very good algorithm is known. It is challenging to even define which output representation for the strategy would be appropriate, since we would not be computing a behavior strategy. Therefore, we shall take a look at some special cases.

**Special case 1:** Undiscounted *perfect information* games. When we enter this world, we are granted the most desired property - namely that the minimax theorem holds (with positional

strategies). We are therefore in a situation where:

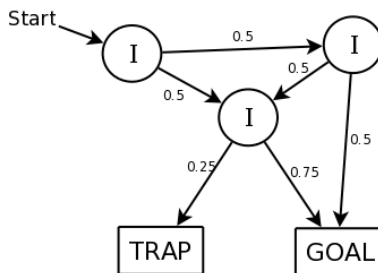
$$\max_{x \in \tilde{S}_1} \min_{y \in S_2} u_1(x, y) = \min_{y \in \tilde{S}_2} \max_{x \in S_1} u_1(x, y)$$

In this world we are able to slightly alter the STRATEGY ITERATION algorithm that we described for discounted games and make it work in the undiscounted case. For simplicity, we shall only give the algorithm and (a sketch of) the proof for the case of *all rewards being 0, except at absorbing states, that is, states where both player have one action only and the state is recurring with probability 1*). We recall that this is in fact a slight generalization of the *simple stochastic games* we defined earlier. Let us, by slight misuse of terminology, use the term simple stochastic games for the slightly more general class of undiscounted perfect information stochastic games with all rewards at non-absorbing states being 0.

Let us simply rewrite the strategy iteration algorithm so that it works for this undiscounted case. As a subroutine in Strategy iteration for a 2-player game, we recall that we must solve a 1-player game (when we compute a best reply of Player 2), so let us first look solving 1-player games. Why not use a strategy improvement algorithm for this? After all, we will define one for the 2-player case anyway, and the 1-player case should be simpler! In fact, it turns out that the correctness proof of Strategy iteration for the 1-player case (i.e, the subroutine) generalizes quite easily to a correctness proof of the entire algorithm.

The 1-player case of a simple stochastic game is also known as an *absorbing Markov decision process*. Thus, we first consider a strategy iteration algorithm for a solving absorbing Markov decision processes. But, starting to write down this algorithm, we realize of course that as a subroutine in *that* algorithm, we should know how to solve the 0-player case!

A 0-player simple stochastic game is also known as an *absorbing Markov process*, as illustrated in the following figure. In an absorbing Markov process, we place a pebble at the starting node



and move around within the given probabilities until we reach either a trap node or the goal. In our case, we collapse all trap-nodes into one single trap. We are also ensure by construction that there is a path that can be taken with positive probability from every node to both the trap-node and the goal (if for some node there are only positive probability paths to, say, the trap, we may collapse that node with the trap) . This ensures that  $\Pr[\text{Reach TRAP}] + \Pr[\text{Reach GOAL}] = 1$ , since infinite play occurs with probability 0. The probability of reaching goal rather than trap is the value of the “game”. This probability can be determined with simple linear algebra, as we shall see next.