Simple Stochastic Games

Wiesław Zielonka
www.irif.fr/~zielonka

November 20, 2018

1 Remarks

A good comprehension of the definitions is sufficient to work out most of the exercises given in the text. Thus even if the proofs given in these notes may seem challenging, with just elementary knowledge of probability you can still try to do exercises. In most (all?) cases mere intuition is largely sufficient. Some minimal experience with discrete Markov chains may help but is not really necessary.

2 Perfect information stochastic games

A two player perfect information stochastic games is played by two players on an arena consisting of finite set of states S partitioned into three sets :

- S_{Max} the states controlled by player Max,
- S_{Min} the states controlled by player Min,
- S_{Nat} the states controlled by the nature,

For each state $s \in S_{\text{Max}} \cup S_{\text{Min}}$, $\text{suc}(s) \subset S$ is a non-empty set of successor states.

For each state $s \in S_{\text{Nat}}$ controlled by the nature there is a fixed probability distribution $p(\cdot \mid s)$, where $p(u \mid s)$ is the probability to go in one step from s to u. We define $\text{suc}(s) = \{u \in S \mid p(u \mid s) > 0\}$ to be the set of successor states of $s \in S_{\text{Nat}}$.

We assume that for each state s the set $\text{suc}(s)$ is non-empty.

Max and Min play the following infinite. If the current state is $s \in S_{\text{Max}}$ then player Max chooses a successor state $s' \in \text{suc}(s)$ and the game moves to s'. If the current state is $s \in S_{\text{Min}}$ then player Min chooses a successor state $s' \in \text{suc}(s)$ and the game moves to s'. If the current state $s \in S_{\text{Nat}}$ is controlled by the nature then the game moves to a state s' with probability $p(s' \mid s)$.

2.1 Plays and histories

A play of is an infinite sequence

$$\omega = s_1, s_2, s_3, \ldots$$

of states such that $s_{i+1} \in \text{suc}(s_i)$. The set of all plays is denoted Ω.

A history is a finite sequence

$$h = s_1, s_2, s_3, \ldots, s_n$$

The set of histories is denoted H.

2.2 Strategies

A strategy of player Max is a mapping σ from histories to states such that, for each history h terminating in a state s controlled by Max,

$$\sigma(h) \in \text{suc}(s).$$

Strategies for player Min are defined in a similar way.

A pair (σ, τ) is a *strategy profile* if σ is a strategy of player Max and τ a strategy of Min.

2.3 Memoryless strategies

A selector for player Max is a mapping $\sigma' : S_{\text{Max}} \to S$ such that $\sigma'(s) \in \text{suc}(s)$ for each $s \in S_{\text{Max}}$.

A strategy σ is memoryless if there exists a selector σ' such that for each history h terminating in a state s controlled by Max, $\sigma(h) = \sigma'(s)$.

In the sequel we identify memoryless strategies with the corresponding selectors.

2.4 Probability

Let $h \in H$ be a history. By h^+ we denote the cylinder generated by h:

$$h^+ = \{\omega \in \Omega \mid h < \omega\},$$

where $h < \omega$ means that h is a prefix of ω.

Given a strategy profile (σ, τ), a history $h = s_1 s_2 \ldots s_i$ and state s_{i+1} we define the probability $p^{\sigma,\tau}(s_{i+1} \mid s_1 s_2 \ldots s_i)$ to move to s_{i+1} given h

$$p^{\sigma,\tau}(s_{i+1} \mid s_1 s_2 \ldots s_i) = \begin{cases} p(s_{i+1} \mid s_i) & \text{if } s_i \in S_{\text{Nat}}, \\ 1 & \text{if } s_i \in S_{\text{Max}} \text{ and } \sigma(s_1 \ldots s_i) = s_{i+1} \\ 1 & \text{if } s_i \in S_{\text{Min}} \text{ and } \tau(s_1 \ldots s_i) = s_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

And now, given a strategy profile (σ, τ) and an initial state s, we can define the probability of cylinders :

$$\text{for } h = s_1, s_2, s_3, \ldots, s_n \in H,$$

$$\mathbb{P}_s^{\sigma,\tau}(h^+) = \begin{cases} \prod_{t=1}^{n-1} p^{\sigma,\tau}(s_{t+1}|s_1 \ldots s_t) & \text{if } s = s_1, \\ 0 & \text{if } s \neq s_1. \end{cases}$$

$\mathbb{P}_s^{\sigma,\tau}$ extends in a unique way to a probability over the σ-algebra generated by cylinders.

We denote $\mathbb{E}_s^{\sigma,\tau}$ the corresponding expectation.

We assume that there is a measurable payoff mapping $\phi : \Omega \to \mathbb{R}$ from the set of all plays to real numbers. After an infinite play $\omega \in \Omega$ player Max receives from player Min the payoff $\phi(\omega)$. The aim of player Max is to maximize his expected gain and the aim of Min is to minimize his loss.

3 Qualitative analysis of perfect information stochastic games

There are two directions in the study of stochastic games. The qualitative analysis tries just to find the set of states where one player wins almost surely or the sets of states where he wins with a positive probability without trying to find this probability.

For perfect information stochastic games the qualitative analysis is quite simple and we propose some study it in a series of exercises.

Let $T \subseteq S$ be a set of states. In the reachability game player Max wins (he has payoff 1) if the play hits T at some moment.

In the exercises below we assume that the players play a perfect information stochastic game with a finite number of states.

Exercise 1. Give a polynomial time algorithm that finds the set of states R such that player Max has a strategy to hit T almost surely if the game starts in R. Find the corresponding strategy of Max.

Find a strategy of Min on the complement of R such that when Min plays this strategy then the probability to hit R is strictly smaller than 1.

Exercise 2. Give a polynomial algorithm that finds the set of states U such that player Max has a strategy to hit T with a positive probability. Find the corresponding strategy of Max. In the complement of U find the strategy of Min such that when Min plays this strategy then T is never visited.

Recall that the Büchi game is the game where Max wins if the set T is visited infinitely often.

Exercise 3. Give a polynomial time algorithm that finds the set of states R such that player Max has a strategy to hit T almost surely infinitely often for the game starting in R. Find the corresponding strategy of Max.

Find a strategy of Min on the complement of R such that when Min plays this strategy then the probability to hit R infinitely often is strictly smaller than 1.

Exercise 4. Give a polynomial algorithm that finds the set of states U such that player Max has a strategy to hit T infinitely often with a positive probability. Find the corresponding strategy of Max. In the complement of U find the strategy of Min such that when Min plays this strategy then almost surely T is visited finitely many times.

Exercise 5. Suppose that the value of the reachability game is strictly positive for all states (player Max has a strategy to hit R with a positive probability).

Is it possible in this case that some states have value smaller than 1? Justify you response either by showing an example or by proving that if the value of all states is positive then in fact this value is 1 for all states.

Exercise 6. Suppose that the value of the Büchi game is strictly positive for all states (player Max has a strategy to hit R infinitely often with a positive probability).

Is it possible in this case that some states have value smaller than 1?

Justify you response either by showing an example or by proving that if the value of all states is positive then in fact this value is 1 for all states.

4 Simple stochastic games

A *simple stochastic game* is a game where the is a fixed set $T \subset S$ of terminal states and the reward mapping $r : T\mathbb{R}_+$ from T to non-negative real numbers. We assume that each terminal state $s \in T$ is *absorbing* which means that $\text{suc}(s) = \{s\}$, i.e. once the game hits such a state it remains there forever.

The payoff mapping is defined in the following way:

- for the plays that never hit T the payoff is 0,
- for plays that hit a terminal state $s \in T$ the payoff is equal to the reward $r(s)$.

We denote this payoff mapping ϕ_r.

We can always normalize r and assume that $r(s) \leq 1$ for all states $s \in T$ (just set $r'(s) = r(s)/M$ where $M = \max_{s \in T} r(s)$.

Exercise 7. Show that we can modify the simple stochastic game G with the reward r in the interval $[0, 1]$ for all states of T to obtain an equivalent simple stochastic game G' with terminal states T' and the reward mapping r' and such that

- r' takes only two values, 0 and 1, $r'(s) \in \{0, 1\}$ for each state $s \in T'$,
- for all non-terminal states $\mathbb{E}_s^{\sigma,\tau}(\phi_r) = \mathbb{E}_s^{\sigma',\tau'}(\phi_{r'})$, i.e. playing in G or in G' we have the same payoff.

In G' we can add new terminal states and change transition probabilities.

This exercise shows that simple stochastic games an be transformed to reachability games.

4.1 One-player simple stochastic games (with Min as the unique player)

The aim of this section is to solve the simple stochastic game when the set S_{Max} of states of player Max empty.

Since player Max is absent instead of speaking about the payoff obtained by Max it is more natural to speak about the loss of player Min.

Thus the loss of Min is 0 is the terminal states are never visited and it is $r(s)$ if a terminal state $s \in T$ is visited at some moment.

We start by finding the set R_T of states such that, whatever the strategy of player Min, when the game starts in a state $u \in R_T$ then the game hits T with a strictly positive probability.

To this end we calculate the increasing sequence $R_0, R_1, R_2, \ldots$ of sets such that starting in R_i the game hits T with a positive probability in at most i steps.

We initialize $R_0 := T$ since the game starting in T hits T in 0 steps.

Suppose that R_i is already known. We initialize $R_{i+1} = R_i$ and for each state $s \in S \setminus R_i$:

1. if $s \in S_{\mathrm{Nat}}$ and $p(s' \mid s) > 0$ for some $s' \in R_i$ then $R_{i+1} := R_i \cup \{s'\}$,
2. if $s \in S_{\mathrm{Min}}$ and for all $s' \in \mathrm{suc}(s), s' \in R_i$ then $R_{i+1} := R_i \cup \{s\}$.

At some moment k we obtain $R_k = R_{k+1}$ and then stop our algorithm and we set $R_T = R_k$. Clearly, from the construction it follows that whatever the strategy of player Min if the game starts in R_T then with a positive probability it hits T at the stage $i \leq k$.

On the other hand, from the complement

$$\mathrm{Safe} := S \setminus R_T$$

satisfies the following condition:

(a) if $s \in \mathrm{Safe}$ then there exists $s' \in \mathrm{suc}(s)$ such that $s' in \, \mathrm{Safe}$ and

(b) if $s \in \mathrm{Safe}$ then, for all $s' \in \mathrm{suc}(s)$, $s' in \, \mathrm{Safe}$.

Thus for player Min it suffices to take in Safe the moves described in (a) and the game will never leave Safe which ensures for player Min the minimal loss equal to 0.

For the states $s \in T$ the game value is obviously $r(t)$.

It remains to find out the optimal strategy of Min and the values in the states

$$R_T^+ := R_T \setminus T$$

Lemma 1. *(a) There is a positive constant $c > 1$ such that for each strategy of player* Min *if the game starts in a state $s \in R_T$ then it hits T with a probability greater or equal to c in at most n steps where n is the number of states in R_T.*

(b) For each strategy of player Min *if the game starts in a state $s \in R_T$ then almost surely the game hits at some moment the set $T \cup \mathrm{Safe}$.*

(c) *For each strategy of* Min *almost surely for each play $s_1, s_2, s_3, \dots$ there exists k such that either all states s_i for $i \geq k$ belong to T or all states s_i for $i \geq k$ belong to* Safe.

Proof. Let
$$\beta = \min\{p(s' \mid s) \mid s' \in S, s \in S_{\text{Nat}} \text{ and } p(s' \mid s) > 0 \}$$
be the minimal positive transition probability from the states of the nature.

Then from the states of R_1 the game hits T with probability at least β in one step, from the states of $R_2 \setminus R_1$ the game hits T with probability at least β^2 in two steps, and, in general, from the states of $R_i \setminus R_{i-1}$ the game hits T with probability at least β^i in i steps.

We can see that with probability at least β^k the game hits T in at most k steps if the starting state is in $R_k = R_T$ which proves (a) for $c := \beta^k$.

Now note that the probability *to not to hit* $T \cup \text{Safe}$ in the first k steps is at most $1 - \beta^k$.

But if after k steps the game is still in R_T^+ then the above reasoning applies again.

We deduce that the probability that the game does not hit $T \cup \text{Safe}$ in mk steps is at most $(1 - \beta^k)^m$ and it tends to O as m increases.

Let X be the set of plays $s_1 s_2 s_3 \dots$ such that $s_k \in R_T^+$ for infinitely many k. The aim is to prove that the probability of X is equal to 0.

Suppose that X has measure > 0. Then with positive probability, for some i the set R_i is visited infinitely often while R_{i-1} is visited finitely often (R_i are the sets defined during the construction of R_T). Thus some state $s \in R_i$ is visited infinitely often with a positive probability. However, from the definition of R_i,

- either $s \in S_{\text{Min}}$ and then all successors of s are in R_{i-1} which implies that some state of R_{i-1} is visited infinitely often immediately after s or
- $s \in S_{\text{Nat}}$ and there exists $q \in R_{i-1}$ such that $p(q \mid s) > 0$. But if s if such s is visited infinitely often then the transition from s to q is selected infinitely often almost surely. Thus again there exists a state of S_{i-1} visited infinitely often.

We have proved that it is impossible to visit R_T^+ infinitely often without hitting T.

Therefore the probability of X is equal to 0.

□

The following lemma shows that the values of all states in the one-player game and the optimal strategy of Min can be obtained by solving a linear programming problem.

Lemma 2. *The value of the states R_T^+ can be obtained by solving the following linear programming problem:*

$$\max \sum_{s \in S} x_s$$

subject to

$$x_s = 0 \qquad \textit{for } s \in \text{Safe} \tag{1}$$
$$x_s = r(s) \qquad \textit{for } s \in T \tag{2}$$
$$x_s \leq x_q \qquad \textit{for all } s \in S_{\text{Min}} \cap R_T^+ \textit{ and all } q \in \text{suc}(s) \tag{3}$$
$$x_s \leq \sum_q p(q|s) \cdot x_q \qquad \textit{for } s \in S_{\text{Nat}} \cap R_T^+ \tag{4}$$

Moreover,

(a) the solution vector $(x^\star_s)$ *of the LP problem satisfies condition* (4) *with the equality,*

(b) for each state $s \in R^+_T \cap S_{\text{Min}}$ *there exist* $t \in \operatorname{suc}(s)$ *such that* $x^\star_s = x^\star_t$ *and the optimal strategy of* Min *in* s *is to move to such a state* t*.*

Proof. Let us recall that a *feasible solution* of an LP problem is any vector $(x_s)_{s\in S}$ satisfying the constraints.

Now let us note that $x_s = 0$ for $s \in S \setminus T$ and $x_t = r(t)$ for $t \in T$ satisfies all constraints (1)-(4) proving that the set of feasible solutions is non-empty.

We will show that for each strategy τ of player Min and each feasible solution (x_s), player's Min loss is greater or equal to x_s for any game starting in state $s \in R^+_T$.

Since the aim of Min is to minimize his loss we can assume that once the game is in Safe then player Min plays in such a way that the game remains forever in Safe and then his loss takes the minimal value 0.

First we show that all feasible solutions are bounded from above by the maximal reward $\max\{r(t) \mid t \in T\}$.

Indeed let (x_s) be a feasible solution and let

$$M = \max_{q \in R^+_T} x_q \tag{5}$$

the maximum of x_s over the states of R^+_T. If $M \le \max\{r(t) \mid t \in T\}$ then our claim is proved.

Suppose that

$$M > \max\{r(t) \mid t \in T\} \tag{6}$$

and let

$$S_M = \{s \in R^+_T \in \mid x_s = M\}$$

be the states of R^+_T where (x_s) attains the maximum. Let us look at the sets R_i constructed when we calculated R_T. For any state of $s \in R_1 \setminus T$, the inequalities (3) and (4) and (6) imply that x_s should be strictly smaller than M since such a state has at least one successor in T. By the same token, each state of $s \in R_2 \setminus R_1$ has at least one successor in R_1 and therefore x_s is strictly smaller than M. We continue in this way for all R_i (induction) to show that the elements of $R_i \setminus R_{i-1}$ have x_s strictly smaller than M. But if this is true for all R_i then in general all states in R_T have x_s strictly smaller than M, i.e. S_M is empty.

But is the set of feasible solutions is non-empty and a bounded from above then the LP problem has a solution.

Let $s_i, i = 1, 2, \ldots$ be discrete stochastic process such that s_i is the state visited at stage i of the game. In particular, $s_1 = s$ as the game starts at stage 1 in state $s \in R^+_T$.

For a given feasible solution (x_s) we define the mapping $x : S \to \mathbb{R}_+$ such that

$$x(s) = x_s \quad \text{for all } s \in S.$$

Thus $x(s)$ and x_s denote the same real number but this double notation will be useful since it allows to avoid a cascade of indices.

Let us consider the the expected value $\mathbb{E}_s^\tau(x(s_i))$, i.e. the expected value of the mapping x at stage i. By definition of the expectation we have

$$\mathbb{E}_s^\tau(x_{s_i}) = \sum_{q \in S} \mathbb{P}_s^\tau(s_i = q) \cdot x(q),$$

where $\mathbb{P}_s^\tau(s_i = q)$ is the probability that the game is in state q at stage i.

From (3) and (4) this expected value is non-decreasing at each stage (and in each state), i.e.

$$\mathbb{E}_s^\tau(x(s_i)) \leq \mathbb{E}_s^\tau(x(s_{i+1}))$$

for all i. But at the first stage the game is in state $s_1 = s$ so that $\mathbb{E}_s^\tau(x(s_1)) = x(s)$. We conclude that

$$x(s) \leq \mathbb{E}_s^\tau(x(s_i))$$

for all i.

But $\mathbb{E}_s^\tau(x(s_i))$ can be partitioned into three sums

$$\mathbb{E}_s^\tau(x(s_i)) = \sum_{q \in T} x(q) \cdot \mathbb{P}_s^\tau(s_i = q) + \sum_{q \in \text{Safe}} x(q) \cdot \mathbb{P}_s^\tau(s_i = q) + \sum_{q \in R_T^+} x(q) \cdot \mathbb{P}_s^\tau(s_i = q). \qquad (7)$$

However, $x(q) = 0$ for $q \in$ Safe, $x(q) = r(q)$ for $q \in T$ and

$$\sum_{q \in R_T^+} x(q) \cdot \mathbb{P}_s^\tau(s_i = q) \leq M \cdot \mathbb{P}_s^\tau(s_i \in R_T^+).$$

By Lemma 1, $\mathbb{P}_s^\tau(s_i \in R_T^+)$ tends to 0 when i goes to infinity.

Therefore the limit of right-hand side of (7) is equal to the expected loss of player Min.

To conclude, we have established that for each strategy τ of Min, each feasible solution (x_s) of the LP problem, and each initial state $s \in R_T^+$ the loss incurred by player Min is not smaller than $x(s)$.

Thus the same holds for the maximal feasible solution, for the solution $(x_s^\star)$ of the LP problem we have

$$x_s^\star \leq \mathbb{E}_s^\tau(\phi_r).$$

It remains to show that with an appropriate strategy τ player Min can limit his loss to $x_s^\star$ for the game starting at s.

First note that the solution $(x_s^\star)$ of the LP problem satisfies (3) and (4) with the equality in place of $\leq$.

Indeed if in some state $s \in R_T^+$ either the inequality (3) or (4) is strict then we can increase in this state the value of $x_s^\star$ until the equality is obtained and this would not violate the constraints of the LP problem.

This increase of $x_s^\star$ would increase the objective of the LP problem contradicting the optimality of $(x_s^\star)$.

Let us consider the following memoryless strategy $\tau^\star$ of player Min:

- in states of Safe he chooses always a move which leads to another state of Safe so that the game never leaves Safe and never hits T,

- in a state $s \in R_T^+$ he chooses a successor $q \in \mathrm{suc}(s)$ such that $x^\star_s = x^\star_q$. (As we have just shown such a successor always exists.)

With this strategy the expected value of the mapping $x^\star : S \to \mathbb{R}_+$, such that $x^\star(s) = x^\star_s$ remains constant at each stage

$$\mathbb{E}_s^{\tau^\star}(x^\star(s_i)) = \mathbb{E}_s^{\tau^\star}(x^\star(s_{i+1})$$

where, as previously, s_i is the state visited at stage i. Thus we have $x^\star_s = \mathbb{E}_s^{\tau^\star}(x^\star(s_1)) = \mathbb{E}_s^{\tau^\star}(x^\star(s_i))$ for all i.

And again consider the partition of the expectation as in (7) (with $x^\star$ replacing x and $\tau^\star$ replacing τ). As noted previously, the limit of (7) as i tends to infinity is equal to the expected loss $\mathbb{E}_s^{\tau^\star}(\phi_r)$ of player Min, i.e. if he plays $\tau^\star$ his (expected) loss is $x^\star_s$.

□

4.2 One-player simple stochastic games with Max as the unique player

Suppose that we have simple stochastic games with the set S_{Min} of states of Min player empty.

Can we solve such games by linear programming?

Exercise 8. Try to reformulate the LP problem in order to solve one-player games with Max and Nat states.

How you will redefine the set Safe?

4.3 The Hoffman-Karp algorithm

The Hoffman-Karp algorithm solves two-player simple stochastic games using as a sub-procedure the algorithm for one-player simple stochastic games developed in Section 4.1.

Let σ be a memoryless strategy of player Max. Suppose that we restrain the moves of Max by leaving at each state $s \in S_{\mathrm{Max}}$ just one successor : $\sigma(s)$. Such a game will be noted $G[\sigma]$ and it can be seen as a one-player game since it is not important who controls the states of S_{Max} in $G[\sigma_i]$ because there is no choice left in these states. The one-player game $G[\sigma]$ can be solved by the algorithm described in the preceding section which gives an optimal strategy τ of player Min when he plays against σ (the best response of Min against σ).

The Hoffman-Karp algorithm

(1) We start with any memoryless strategy σ_0 for Max. Fixing σ_0 find the optimal memoryless strategy of player Min in the game $G[\sigma_0]$ where player Max is forced to play σ_0.

(2) Set $i = 0$.

(3) For each state s, let $v_i(s) = \mathbb{E}_s^{\sigma_i,\tau_i}(\phi_r)$ be the payoff obtained by Max when the players play according to σ_i and τ_i respectively.

Let

$$S_i^+ = \{s \in S_{\text{Max}} \mid v_i(s) < \max_{q \in \text{suc}(s)} v_i(q)\}. \tag{8}$$

the set of *improvement states* at stage i.

If $S_i^+ = \emptyset$ then the strategies σ_i, τ_i are optimal. Return (σ_i, τ_i, v_i).

(4) Otherwise, if $S_i^+ \neq \emptyset$ then

(a) define a new memoryless strategy σ_{i+1} for player Max: for each state s controlled by Max,

$$\sigma_{i+1}(s) = \begin{cases} \arg\max_{q \in \text{suc}(s)} v_i(q) & \text{if } s \in S_i^+ \\ \sigma(s) & \text{otherwise,} \end{cases}$$

where $\arg\max_{q \in \text{suc}(s)} v_i(q)$ is a successor of s maximizing v_i.

(b) Given the strategy σ_{i+1} solve the one-player game $G[\sigma_{i+1}]$ where Max is constrained to play σ_{i+1}. Let τ_{i+1} be the optimal memoryless strategy of Min in the game $G[\sigma_{i+1}]$.

(c) Set $i = i + 1$. Jump to (3).

The mappings v_i defined in the algorithm are called valuations.

Roughly speaking, while the set of improvement states is non-empty the Hoffman-Karp algorithm tries to improve the strategy of Max in improvement states.

Let us note that it is crucial that the strategy of Max does not change in states that are not improvement states.

Exercise 9. Let us modify the step (5a) of the algorithm in the following way:

$$\sigma_{i+1}(s) = \arg\max_{q \in \text{suc}(s)} v_i(q)$$

i.e. we allows to modify the strategy of Max also in states that are not improvement states. Intuitively this means that if $s \in S_{\text{Max}}$ is not an improvement state then we are allowed now to change the strategy in s of Max in s by choosing another successor state. The only constraint is that $q = \sigma_{i+1}(s)$ should maximize v_i taken over all successors of s.

Show that such an algorithm may not converge.

Lemma 3. *The valuations calculated by the Hoffman-Karp algorithm satisfy the following conditions: for each state s,*

- $v_{i+1}(s) \geq v_i(s)$ *and*
- $v_{i+1}(s) \geq \max_{q \in \text{suc}(s)} v_i(q) > v_i(s)$ *for each state* $s \in S_i^+$.

Proof. Let us consider the situation when player Max plays according to strategy σ_{i+1} while player Min plays any strategy τ (not necessarily memoryless).

Let $s_1, s_2, s_3, \ldots$ be an infinite sequence of visited states (formally, $s_k, k = 1, 2, \ldots$ is the stochastic process such that s_k is the state visited at time k).

We examine how the real valued sequence $v_i(s_1), v_i(s_2), v_i(s_3), \ldots$ evolves in time (note that the valuation v_i is fixed and this the valuation from the stage i of the algorithm but player Max plays the strategy σ_{i+1} obtained at stage $i+1$. Player Min is supposed to play any strategy τ).

We have the following cases:

1. if $s_k \in S_i^+$ then, by definition of σ_{i+1}, player Max will select a successor state s_{k+1} maximizing v_i, thus $v_i(s_k) < v_i(s_{k+1})$.

2. if $s_k \in S_{\text{Max}} \setminus S_i^+$ then, by definition of σ_{i+1}, $\sigma_{i+1}(s_k) = \sigma_i(s_k)$ but such a move does not change the value of v_i, i.e. $v_i(s_k) = v_i(s_{k+1})$,

3. if $s_k \in S_{\text{Min}}$ then player Min has no move that can decrease the value v_i in the one-player game $G[\sigma_i]$. Therefore whatever move player by Min we have always $v_i(s_k) \leq v_i(s_{k+1})$.

4. if $s_k \in S_{\text{Nat}}$ then $v_i(s_k) = \sum_{q \in \text{suc}(s_k)} v_i(q) \cdot p(q \mid s_k)$. Again this follows from the fact that v_i is the value mapping in the one-player game $G[\sigma_i]$ and, as noted in Lemma 2(a), the moves of the nature neither increase nor decrease the game value.

Thus the stochastic process $v_i(s_k), k = 1, 2, 3, \ldots$ does not decrease its value, in the probability jargon, this process is a submartingale. Again in the probability this is noted as

$$\mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_{k+1}) \mid s_1, \ldots, s_k) \geq v_i(s_k)$$

which reads as "the expected value of v_i at stage $k+1$ given $s_1, \ldots, s_k$ is not smaller than the value $v_i(s_k)$ at stage k".

v_i is bounded from above by $\max_{t \in T} r(t)$ and the standard fact in probability is that a submartingale bounded from above converges almost surely[1].

This means that for almost all infinite plays $s_1, s_2, s_3, \ldots$ the real valued sequence $v_i(s_1), v_i(s_2), v_u(s_3), \ldots$ converges.

However, there is only a finite number of states thus a finite number of different values in the sequence $v_i(s_1), v_i(s_2), v_u(s_3), \ldots$. Such a sequence can converge only if starting at some moment all values are equal, i.e. there exists m (depending on the sequence) such that $v_i(s_l) = v_i(s_m)$ for all $l \geq m$. This can happen in two ways. Either the play $s_1, s_2, s_3, \ldots$ hits at some moment T which implies that there exists $t \in T$ and a moment m such that $s_k = t$ for all $k \geq m$.

The other possibility is that the play $s_1, s_2, s_3, \ldots$ never hits T but starting from some moment it visits only the states with the same value x of v_i.

Let $x \in \mathbb{R}_+$ and let X be the set of plays consisting of all plays $s_1 s_2 s_3 \ldots$ such that

- all states s_k belong to $S \setminus T$ and

- there exists k such that $v_i(s_l) = x$ for all $l \geq k$.

[1] This is Doob's martingale convergence theorem that can be found in any decent probability book, for example in `Williams, Probability with Martingales, Cambridge University Press`. In our setting this can be proved from scratch but why to do it from scratch if a more general proof is simple and avoids hand-waving.

Suppose that X has a positive probability, $\mathbb{P}_s^{\sigma_{i+1}\tau}(X) > 0$.

Since $G[\sigma_{i+1}]$ is essentially a one-player game with the unique player Min Lemma 1 applies. And this lemma stipulates that almost surely all plays either hit T or starting from some moment they do not leave the set $\mathrm{Safe} = \{s \in S \setminus T \mid v_i(s) = 0\}$.

We conclude that $\mathbb{P}_s^{\sigma_{i+1}\tau}(X)$ can only be positive for $x = 0$.

Therefore we have proved that if player Max plays using strategy σ_{i+1} then almost surely either the play hits T or the play from some moment on remains forever in the set $\{s \in S \mid v_i(s) = 0\}$ of states with v_i equal to 0.

Let

$$A_k = \{s_k \in T\} \tag{9}$$

be the set of plays that at time k are in T (and therefore they remain in T in the future as the states of T are absorbing).

Let

$$B_k = \{\text{for all } l \geq k, v_i(s_l) = 0\}$$

be the set of plays that at time k and at all subsequent moments are in the states having v_i equal to 0. Clearly the sets A_k and B_k form an increasing sequence, $A_k \subseteq A_{k+1}$, $B_k \subseteq B_{k+1}$.

Let

$$A = \bigcup_{k=1}^{\infty} A_k = \{\exists k, s_k \in T\}$$

be the set of plays that hit T at some moment and let

$$B = \bigcup_{k=1}^{\infty} B_k = \{\exists k, \forall l \geq k, v_i(s_l) = 0\}$$

to be the set of plays that from some moment on remain forever in the states having $v_i = 0$. We have proved above that

$$\mathbb{P}_s^{\sigma_{i+1},\tau}(A \cup B) = 1.$$

Thus for each $\varepsilon > 0$ there exists k such that

$$\mathbb{P}_s^{\sigma_{i+1},\tau}(A_k \cup B_k) > 1 - \varepsilon. \tag{10}$$

Since the expectation on v_i is non-decreasing at each step[2] the expectation of v_i at time k is greater or equal to the value of v_i in the initial state s:

$$v(s) = v_i(s_1) \leq \mathbb{E}_{s_1}^{\sigma_{i+1},\tau}(v_i(s_k)).$$

To estimate the expectation on the right-hand side we evaluate the expectation separately on three disjoint sets: A_k, B_k and $\overline{A_k \cup B_k} = \Omega \setminus (A_k \cup B_k)$:

$$\mathbb{E}_{s_1}^{\sigma_{i+1},\tau}(v_i(s_k)) = \mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_k); A_k) + \mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_k); B_k) + \mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_k); \overline{A_k \cup B_k}). \tag{11}$$

However, for the terminal states of T v_i is equal to the reward r, i.e. it is equal to the value of the payoff ϕ_r

$$\mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_k); A_k) = \mathbb{E}_s^{\sigma_{i+1},\tau}(\phi_r; A_k).$$

[2] More precisely the process $v_i(s_k), k = 1, 2, \ldots$ is a submartingame.

By definition, the value of $v_i(s_k) = 0$ on B_k,

$$\mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_k); B_k) = 0.$$

By (10), $\mathbb{P}_s^{\sigma_{i+1},\tau}(\overline{A_k \cup B_k}) < \varepsilon$ implying

$$\mathbb{E}_s^{\sigma_{i+1},\tau}(v_i(s_k); \overline{A_k \cup B_k}) < M\varepsilon,$$

where $M = \max_{q \in T} r(q) = \max\{v_i(q) \mid q \in S\}$. Therefore,

$$v_i(s) < \mathbb{E}_s^{\sigma_{i+1},\tau}(\phi_r; A_k) + M\varepsilon \leq \mathbb{E}_s^{\sigma_{i+1},\tau}(\phi_r) + M\varepsilon.$$

The last inequality holds for each $\varepsilon > 0$ we obtain

$$v_i(s) \leq \mathbb{E}_s^{\sigma_{i+1},\tau}(\phi_r).$$

However as this inequality holds for each strategy τ of Min it holds also for τ_{i+1} thus $v_i(s) \leq \mathbb{E}_s^{\sigma_{i+1},\tau_{i+1}}(\phi_r) = v_{i+1}(s)$.

The second assertion of Lemma 3 follows from the fact that, by definition of σ_{i+1}, when the game is in $s \in S_i^+$ then player Max moves in one step to the state $q = \arg\max_{z \in \mathrm{suc}(s)} v_i(z)$. But we have just proved that once in q using σ_{i+1} he wins at least $v_i(q)$. □

Corollary 4. *The Hoffman-Karp algorithm terminates after a finite number of steps and returns optimal strategies for both players.*

Proof. Let us compare the strategies obtained in iterations i and $i+1$ of the algorithm. Note that, for each state s, $v_{i+1}(s)$ is the minimal expected payoff that player Max obtains if he plays according to strategy σ_{i+1} against the best response τ_{i+1} of player Min. By Lemma 3 there exist a state s such that $v_{i+1}(s) > v_i(s)$ and for all states q, $v_{i+1}(s) \geq v_i(s)$. This means that playing σ_{i+1} player Max obtains the expected payoff not worse than when he plays σ_i, and at least in one state his payoff is strictly better, in both cases he plays against the best response of his adversary.

This implies strategies $\sigma_i, i = 1, 2, \ldots, k$ calculated by the Hoffman-Karp algorithm are pairwise different. But there is only a finite number of memoryless strategies thus the algorithm will stop at some moment.

Suppose that the algorithm stops at stage i and (σ_i, τ_i) are the strategies computed by the algorithm. From the fact that τ_i is the best response of Min against σ_i it follows that player Min loses $v_i(s)$ if the game starts at s and he cannot lose less when he plays against σ_i.

For player Max the situation is less clear, can Max win more than $v_i(s)$ for the game starting in s?

Suppose that Max plays a strategy σ (not necessarily memoryless) against τ_i. Then at each step the expectation of v_i cannot increase[3].

The proof goes now in a similar way as in Lemma 3 but all inequalities are in the reverse direction. In fact the situation is much more simpler, we do not even need to bother if the

[3] More exactly the stochastic process $v_i(s_1), v_i(s_2), v_i(s_3), \ldots$ is now a supermartingale, i.e. $\mathbb{E}_s^{\sigma,\tau_i}(v_i(s_{k+1}) \mid s_1, \ldots, s_k) \leq v_i(s_k)$.

process $v_i(s_k), k = 1, 2, 3, \dots$ converges[4]. This is due to the fact that for player Max it is prejudicial to stay forever outside of T as in this case his payoff takes the minimal value 0.

Thus we have for each k,

$$v_i(s) = v_i(s_1) \geq \mathbb{E}_s^{\sigma,\tau_i}(v_i(s_k)) = \mathbb{E}_s^{\sigma,\tau_i}(v_i(s_k); A_k) + \mathbb{E}_s^{\sigma,\tau_i}(v_i(s_k); \overline{A_k}) \geq \mathbb{E}_s^{\sigma,\tau_i}(v_i(s_k); A_k),$$

where A_k are defined as in (9). But as noted previously, for plays in A_k we have $v_i(s_k) = \phi_r$ thus

$$v_i(s) \geq \mathbb{E}_s^{\sigma,\tau_i}(\phi_r; A_k).$$

A_k is an increasing sequence of sets with the limit A consisting of all plays that hit T at some moment thus for each $\varepsilon > 0$ there exists k such that $\mathbb{P}_s^{\sigma,\tau_i}(A_k) + \varepsilon \geq \mathbb{P}_s^{\sigma,\tau_i}(A)$ and therefore

$$v_i(s) \geq \mathbb{E}_s^{\sigma,\tau_i}(\phi_r; A) - M\varepsilon$$

where M is the maximum of v_i which is the same as the maximal reward $\max_s r(s)$. Since the last inequality holds for all ε we obtain $v_i(s) \geq \mathbb{E}_s^{\sigma,\tau_i}(\phi_r; A)$. But the payoff ϕ_r is equal to 0 for the plays in the complement of A thus $v_i(s) \geq \mathbb{E}_s^{\sigma,\tau_i}(\phi_r)$ and we can see that whatever the strategy of player Max he cannot win more than $v_i(s)$ for the games starting in s when he plays against τ_i.

□

5 Concurrent stochastic games

A concurrent stochastic game is a zero-sum game played by Max and Min. In each state s player Max has a non-empty finite set of action $A(s)$ and player Min has a finite non-empty set of actions $B(s)$. Player Max chooses and action $a \in A(s)$, player Min chooses independently and simultaneously an action $b \in B(s)$ and the game moves to a new state s' with the probability $p(s' \mid s, a, b)$.

We assume that these transition probabilities sum up to 1, i.e. $\sum_{s' \in S} p(s' \mid s, a, b) = 1$ for each triple (s, a, b).

A history is a finite sequence $h = s_1, (a_1, b_1), s_2, (a_2, b_2), s_3(a_3, b_3), \dots, s_n$ alternating states and pairs of actions and such that $(a_i, b_i) \in A(s_i) \times B(s_i)$ for all $i < n$.

For each finite set X by $\Delta(X)$ we denote the set of probability distributions on X. Thus an element of $\Delta(X)$ is any mapping $\delta : X \to [0, 1]$ such that $\sum_{x \in X} \delta(x) = 1$.

A strategy of player Max is mapping σ such that for each history $h = s_1, (a_1, b_1), s_2, (a_2, b_2), s_3, (a_3, b_3), \dots, s_n$, $\sigma(h) \in \Delta(A(s_n))$. Thus the strategy of Max gives for each history h and each action $a \in A(s_n)$ the probability $\sigma(h)(a)$ of plying a after this history.

Strategies for the player Min are defined in a similar way.

Given history h and strategies σ, τ of both players, $\sigma(h)(a) \cdot \tau(h)(b)$ is the probability that the players will play the pair (a, b) of actions.

The probability of h^+ - the cylinder generated[5] by h, is defined inductively,

[4] But it converges almost surely since this is a supermartingale bounded from below.

[5] h^+ is the set of infinite histories with prefix h.

$\mathbb{P}_s^{\sigma,\tau}(s_1^+) = 1$ is $s = s_1$ and 0 otherwise.

Suppose that $h_{n-1} = s_1, (a_1, b_1), s_2, (a_2, b_2), s_3, \ldots, s_{n-1}$, $h_n = h_{n-1}(a_{n-1}, b_{n-1}), s_n$ and that $\mathbb{P}_s^{\sigma,\tau}(h_{n-1}^+)$ is already defined.

Then

$$\mathbb{P}_s^{\sigma,\tau}(h_n^+) = \mathbb{P}_s^{\sigma,\tau}(h_{n-1}^+) \cdot \sigma(h_{n-1})(a_{n-1}) \cdot \tau(h_{n-1})(b_{n-1}) \cdot p(s_n \mid s_{n-1}, a, b).$$

The quantitative analysis of such games is difficult, optimal strategies may not exist[6].

However the qualitative analysis of these games is elementary.

5.1 Reachability games

A reachability game is the game when the aim of one of the players, say player Max, is to visit a fixed set $T \subseteq S$ of states. We can assume that the states of T are absorbing i.e. for $t \in T$ $p(t \mid t, a, b) = 1$ for all actions a, b.

Sure winning.

Sure winning is not a probability notion. It is rather a notion of related to non-determinism.

Let SURE be the set of infinite histories h consisting of all infinite histories such that for each finite prefix g of h, $\mathbb{P}_s^{\sigma,\tau}(g^+) > 0$. Thus $h \in$ SURE if all finite prefixes have positive probability.

Strategy σ of Max is surely winning if each infinite history $g \in$ SURE hits T at some moment.

Almost sure winning. Let HIT be the set of infinite histories that hit T at some moment.

Strategy σ of Max is almost surely winning if $\mathbb{P}_s^{\sigma,\tau}(\text{HIT}) = 1$.

A sure winning strategy is almost sure winning. The converse is not true. The inclusion HIT $\subset$ SURE can be strict but if the difference SURE $\setminus$ HIT has measure 0 then the strategy σ is almost surely winning but not surely winning.

For example if we have two states x and y. The target set $T = \{y\}$ and the game starts in x. Each player has only one action in x, they have no choice. The pair of actions played in x leads to y with probability 1/2 and with the same probability the game returns to x.

Almost surely the game terminates in y and is winning for Max.

However the infinite path $xxxx\ldots$ loops in x forever, each finite prefix has a positive probability thus $xxx\ldots$ belongs to $SURE$. But this infinite pat is not winning for Max. Thus Max does not win surely.

Limit winning.

Max wins at the limit if for each $\varepsilon > 0$ there exists a strategy σ such that for each strategy τ of Min $\mathbb{P}_s^{\sigma,\tau}(HIT) \geq 1 - \varepsilon$.

Again Max can win in the limit but not almost surely.

Exercise 10. Give an example of a reachability game where Max wins in the limit but not almost surely.

[6] For reachability games player Max has only ε-optimal strategy, player Min has optimal strategy. Both strategies are randomized. For general parity games both players have only ε-optimal strategies and such strategies depend on the past, i.e. are not memoryless.

Exercise 11. Give an algorithm that finds the set of states X such that Max wins surely if the game starts in X.

Exercise 12. Give an algorithm that finds the set of states X such that Max wins almost surely if the game starts in X.

Exercise 13. Give an algorithm that finds the set of states X such that Max wins in the limit if the game starts in X.

(The last exercise is difficult, the first two are much easier.)