Game Theory, Optimization, Reinforcement Learning **NTUA, Spring 2021**

Meeting 1: Shapley's Stochastic Games

Advisors: ***Ioannis Panageas, Emmanouil Vlatakis*** **Student:** ***Fivos Kalogiannis***

1.1 Introduction

Here we will discuss the kind of games that take place in multiple steps while also the available moves of each player are subject to change and depend on the state the game occupies at each time-step. Specifically, we will be concerned with somewhat of an extension to 2-person zero-sum games. In a 2-person, zero-sum game each player is given m and n available moves accordingly, they make their choices in secret and then simultaneously reveal their moves; what player 1 wins, player 2 will lose and both player have knowledge of how each pair of two opposing moves affects the utility, the payoff matrix is known to both players. Hereafter, we will be referring to conventional single round N-person games as *matrix games*.

Since the original paper was published in 1953, it does not adhere to Markov Decision Processes standard notations but we will try to bridge the gap between the notation of contemporary literature and the original work.

1.1.1 Formal Setting

Let us now define our setting a bit more concretely. The *stochastic game* we are concerned with is played between 2 players. They still play their moves in the fashion of a conventional 2-person matrix game, i.e. they will simultaneously reveal their choice of move while they hold knowledge of the corresponding payoff matrix of the state they are in.

Definition 1 (Stochastic Game)**.** *A stochastic game* Γ *consists of:*

- 2 *opposing players*
- $N+1$ *states (to which we will refer to with an integer* k *or* $l \in \{1,\dots,N\}$*). N of them stand for states of an ongoing game (non-terminal states) and* 1 *of them stands for the end of the game (terminal state). The former are all potentially connected to one another while the game-over state is a total sink (i.e. we can access it from every state, but we can access no other state once we are there and we declare the game finished).*
- *a collection payoff matrices* $\boldsymbol{A}^k$*. This means the number of payoff different payoff matrices is N and each one of them has a dimension* $m^k \times n^k$ *for a state k.*
- *a collection of transition matrices* $\boldsymbol{P}^{kl}$*. Since every state is possibly accessible from any other, every state k has corresponding* 1 *matrix for the terminal state and N matrices for every other state. This is sums to* N^2+N *such matrices in total.*

If we define Γ^k *as the game starting at state k, then the game* Γ *is the collection* $\Gamma = \{\Gamma^k | k = 1,2,\dots\}$*. Which means a game* Γ *is the set of games that start at any non-terminal state.*

Let's elaborate a bit on the matter on notation; the matrix $\boldsymbol{P}^{kl}$ holds entries p_{ij}^{kl} which denote the probability of transition from state k to state l when the players draw the moves i and j accordingly. Respectively, s_{ij}^k denotes the probability the game finishes when the two players draw the latter moves. We also note that s_{ij}^k is set to be $s_{ij}^k > 0$, meaning there is always a chance the game will be over at the corresponding next state at any given state, with any given tuple of actions i, j. It may have become apparent that:

$$\sum_{l \in 1,\dots,N} p_{ij}^{kl} + s_{ij}^k = 1$$

In 1.1 we try to illustrate the latter relations since for someone familiar with Markov Processes the notation can feel a bit awkward.

As far as payoffs are concerned, at every state k, player 1 will be winning a_{ij}^k over player 2 if they play moves i and j accordingly.

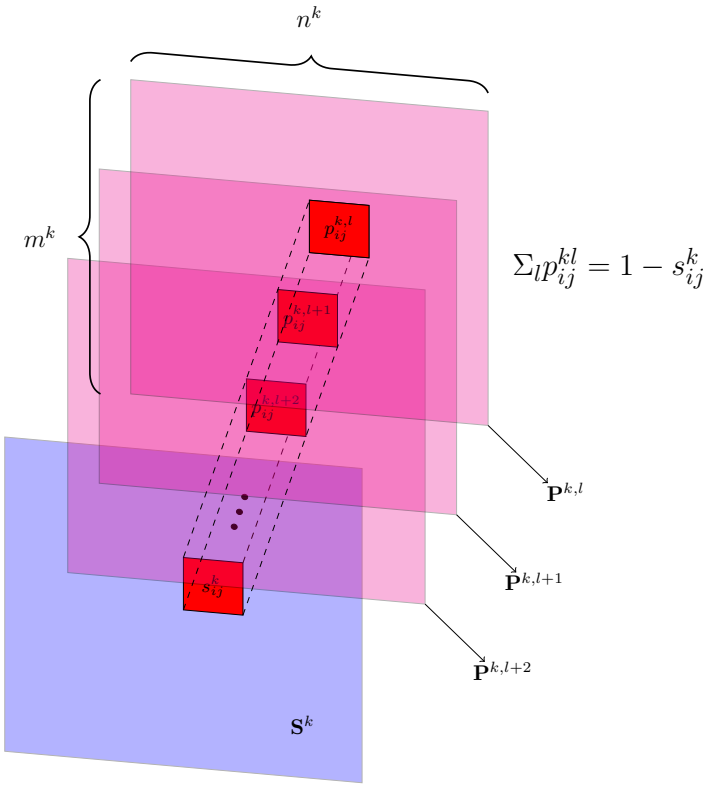

Figure 1.1: A set of transition matrices for state k. The values of the red squares sum to 1.

We move on to make some auxiliary conventions on notation and some useful observations; value s stands for the minimum probability of moving to the terminal state that is related to any possible tuple of actions i, j and state k:

$$s = \min_{i,j,k} s_{ij}^k$$

In the same spirit, we define M as the maximum absolute value of any possible payoff; that is, for any tuple of actions i, j on state any state k:

$$M = \max_{i,j,k} \left| a_{ij}^k \right|$$

We could now observe that the expected total gain/loss is bounded by:

$$M + (1-s)M + (1-s)^2 M + \ldots = \frac{M}{s}$$

or

$$\mathbb{E}[\text{total gain/loss}] \leq \frac{M}{s}$$

And conclusively, the following, more or less obvious, inequalities will hold:

$$p_{ij}^{kl} \geq 0$$
$$\left| a_{ij}^k \right| \leq M$$
$$p_{ij}^{kl} = 1 - s_{ij}^k \leq 1 - s < 1$$

We could say that time is implied to be discrete but still the full sets of pure or mixed strategies for each player will enormous and rather redundant. We restrict our concerns on *stationary strategies*, namely a rule of betting that will take into account solely a player's position at the current moment. The latter translates to playing the same strategy every time one reaches a certain state regardless of the route followed to get there. The set of these strategies is represented by tuples $\boldsymbol{x}$ and $\boldsymbol{y}$ each holding N vector entries of mixed strategies one for every non-terminal state:

$$\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N)$$

and

$$\boldsymbol{y} = (\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^N)$$

. Every one of the latter vectors has a length dependent on the given state:

$$\boldsymbol{x}^k = (x_1^k, \ldots, x_{m^k}^k)$$

and

$$\boldsymbol{y}^k = (y_1^k, \ldots, y_{n^k}^k)$$

1.2 Existence of a Solution

As we seek to prove the existence of a solution to any such game, we will have to make some preliminary observations. We will have to momentarily revert to simple, matrix games to introduce mapping $\mathrm{val}[\cdot] \doteq \min_{\boldsymbol{y}} \max_{\boldsymbol{x}}\{\cdot\}$ and observe that for any two matrices of the same shape $\boldsymbol{B}$ and $\boldsymbol{C}$:

$$|\mathrm{val}[\boldsymbol{B}] - \mathrm{val}[\boldsymbol{C}]| \leq \max_{i,j} |b_{ij} - c_{ij}| \tag{1.1}$$

While for the purposes of proving the existence of a solution for stochastic games, we are introducing the matrices $\boldsymbol{A}^k(\boldsymbol{\delta}_t)$. Each $\boldsymbol{A}^k(\boldsymbol{\delta}_t)$ holds entries:

$$a_{ij}^k + \sum_l p_{ij}^{kl} \delta_t^l$$

We observe matrix $\boldsymbol{A}^k(\boldsymbol{\delta}_t)$ to be a sum of the plain matrix $\boldsymbol{A}$ and a matrix holding a weighted sum of the entries of $\boldsymbol{\delta}_t$ the coefficients of which depend on the actions chosen the running state and the state each entry of $\boldsymbol{\delta}_t$ corresponds to. Vector $\boldsymbol{\delta}_t$ is dependent on time and will hold N numerical entries. After initializing $\boldsymbol{\delta}_0$ to whichever vector we may, we can recursively induce the value of $\boldsymbol{\delta}_t$ as:

$$\delta_t^k = \mathrm{val}[\boldsymbol{A}^k(\boldsymbol{\delta}_{t-1})]$$

It is notable that if every entry δ_0^k is initialized to $\mathrm{val}\, \boldsymbol{A}^k$, then δ_t^k will be the value of the game that starts at state k and, if it lasts, it is cut off after t time steps.

We can now state the first theorem on the value of a stochastic game Γ: The value of the stochastic game Γ is the unique solution $\boldsymbol{\delta}^*$ of the system:

$$\delta^{k^*} = \mathrm{val}[\boldsymbol{A}^k(\boldsymbol{\delta}^*)]$$

For the needs of the proof in question, we need to prove that for $t \to \infty$, $\boldsymbol{\delta}_t$ is independent of the initial vector $\boldsymbol{\delta}_0$ and the components of the resulting vector are the values of the infinite game Γ^k.

Proof. **Existence of $\boldsymbol{\delta}^*$:** For the needs of this proof we will be using the l_∞ and will be shortly referring to simply by $\|\cdot\|$. (i.e. $\|\boldsymbol{a}\| = \max|a^k|$)

Let $\boldsymbol{T}$ be a transform $\boldsymbol{T} : \mathbb{R}^N \to \mathbb{R}^N$ defined as $\boldsymbol{T}(\boldsymbol{\delta}) = \boldsymbol{\xi} = (\xi^1, \ldots, \xi^k, \ldots, \xi^N)^T$ where $\xi^k = \mathrm{val}[\boldsymbol{A}^k(\boldsymbol{\delta})]$.

In order to prove that there does exist a fixed point under $\boldsymbol{T}$ we write the following:

$$\|\boldsymbol{T}\boldsymbol{\xi} - \boldsymbol{T}\boldsymbol{\delta}\| \doteq \max_k \left|\mathrm{val}\, \boldsymbol{A}^k(\boldsymbol{\xi}) - \mathrm{val}\, \boldsymbol{A}^k(\boldsymbol{\delta})\right|$$

$$\overset{1.1}{\leq} \max\left|(a_{ij}^k + \sum_l p^{kl}\xi^l) - (a_{ij}^k + \sum_l p^{kl}\delta^l)\right|$$
$$= \max\left|\sum_l p^{kl}\xi^l - \sum_l p^{kl}\delta^l\right| = \max\left|\sum_l p^{kl}(\xi^l - \delta^l)\right|$$
$$\leq \max\left|\sum_l p^{kl}\right| \max_l|\xi^l - \delta^l|$$
$$= (1 - \min_{i,j,k} s_{ij}^k)\|\boldsymbol{\xi} - \boldsymbol{\delta}\|$$
$$= (1-s)\|\boldsymbol{\xi} - \boldsymbol{\delta}\| \Rightarrow$$

$$\|\boldsymbol{T\xi} - \boldsymbol{T\delta}\| \leq (1-s)\|\boldsymbol{\xi} - \boldsymbol{\delta}\| \tag{1.2}$$

Since, $\boldsymbol{T\xi} = \boldsymbol{T}(\boldsymbol{T\delta}) = \boldsymbol{T}^2\boldsymbol{\delta}$ and as just shown $\|\boldsymbol{T\xi} - \boldsymbol{T\delta}\| = \|\boldsymbol{T}(\boldsymbol{T\delta} - \boldsymbol{\delta})\| \leq (1-s)\|\boldsymbol{T\delta} - \boldsymbol{\delta}\|, s > 0$ we can conclude that the sequence $\boldsymbol{T}^n\boldsymbol{\delta}$ converges and as $n \to \infty$, $\boldsymbol{T}^n\boldsymbol{\delta}$ converges to a fixed point $\boldsymbol{\delta}^*$.

Uniqueness of $\boldsymbol{\delta}^*$: Furthermore, we will demonstrate that the latter is unique. Let $\boldsymbol{\phi}, \boldsymbol{\psi}$ be two fixed points, meaning $\boldsymbol{\phi} = \boldsymbol{T\phi}, \boldsymbol{\psi} = \boldsymbol{T\psi}$. Then for the norm of their difference we show that:

$$\|\boldsymbol{\psi} - \boldsymbol{\phi}\| = \|\boldsymbol{T\psi} - \boldsymbol{T\phi}\| \overset{1.2}{\leq} (1-s)\|\boldsymbol{\phi} - \boldsymbol{\psi}\|$$

This inevitably leads to $\|\boldsymbol{\phi} - \boldsymbol{\psi}\| = 0$ and $\boldsymbol{\phi} = \boldsymbol{\psi}$, since something non-negative is equal to itself multiplied by something other than 1.

We have now seen that there does exist a fixed point under $\boldsymbol{T}$ and that it also is unique and independent of an initial point $\boldsymbol{\delta}_0$.

$\boldsymbol{\delta}^*$'s entries hold the values of Γ^k: In order to show that entry $\delta^{k^\star}$ holds the value of a game that starts at state k, namely Γ^k, we note the following; let λ_t^k be the value of the game Γ^k if an optimal strategy is followed for the t initial steps. If the player were to play ad infinitum following random choices, the expected value of the game λ_∞^k would be:

$$\lambda_t^k - (1-s)^t M - (1-s)^{t+1}M - \ldots \leq \lambda_\infty^k \leq \lambda_t^k + (1-s)^t M + (1-s)^{t+1}M + \ldots \Rightarrow$$
$$\lambda_t^k - (1-s)^t (M\sum_{i=0}^{\infty}(1-s)^i) \leq \lambda_\infty^k \leq \lambda_t^k + (1-s)^t (M\sum_{i=0}^{\infty}(1-s)^i) \Rightarrow$$
$$\lambda_t^k - (1-s)^t \frac{M}{s} \leq \lambda_\infty^k \leq \lambda_t^k + (1-s)^t \frac{M}{s} \Rightarrow$$

$$-(1-s)^t \frac{M}{s} \leq \lambda_\infty^k - \lambda_t^k \leq (1-s)^t \frac{M}{s}$$

We define $e^t = (1-s)^t \frac{M}{s}$ and observe that while $t \to \infty$, $e^t \to 0$. Hence, $\boldsymbol{\delta}^*$ indeed holds the values of every game Γ^k.

□

Next, we will be concerned with whether a set *optimal stationary strategies* exists or not. And such a set does exist, but in order to prove so we will slightly modify the rules of the game. Instead of playing a game of indefinite time steps, the game will be agreed to stop at step t, and at this last step, both players will receive a surplus to the payoff dictated by the payoff matrix corresponding to the their current state. The surplus for the state h, given the pair of actions i, j will be $\sum_l p_{ij}^{hl}$. Let's now state the promised theorem before going on with its proof. The stationary strategies $\boldsymbol{x}^*, \boldsymbol{y}^*$ (whose l-th vector-elements $\boldsymbol{x}^l, \boldsymbol{y}^l$ belong to the sets of optimal mixed strategies of the matrix game $A^l(\boldsymbol{\delta}^*)$ for every state $l = 1, \ldots, N$ and for player 1 and player 2 accordingly) are optimal for each player respectively in every game $\Gamma^k \in \Gamma$

Proof. As we started discussing, we have changed the rules a bit. The game will end at step t when the players are occupying state h of the game and the actions of their choice are i, j. They will receive (or lose)

at the final step:

$$a_{ij}^h + \sum_l p_{ij}^{hl} \delta^{h^*}$$

While at every previous step at state q, the were receiving/losing the familiar:

$$a_{ij}^q$$

This ensures that by following such an optimal stationary strategy, at the end of the game that started on state k every player gets (or loses) δ^{k^*}.

Returning to the initial setting, after step t, player 1 will have won the amount δ_t^k:

$$\begin{aligned}
\delta_t^k &\geq \delta^{k^*} - (1-s)^{t-1} \max_{i,j,h} \sum_l p_i^{hl} j \delta^{k^*} \\
&\geq \delta^{k^*} - (1-s)^{t-1} \max_{i,j,h} \sum_l p_i^{hl} j \max_l \delta^{l^*} \\
&\geq \delta^{k^*} - (1-s)^{t-1} \max_{i,j,h} \sum_l p_i^{hl} j \max_l \delta^{l^*} \\
&\geq \delta^{k^*} - (1-s)^{t-1} (1-s) \max_l \delta^{l^*} \\
&\geq \delta^{k^*} - (1-s)^t \max_l \delta^{l^*}
\end{aligned}$$

So for, δ_∞^k we can conclude the following inequality:

$$\delta_\infty^k \geq \delta^{k^*} - (1-s)^t \max_l \delta^{l^*} - (1-s)^t \frac{M}{s}$$

Then, letting t go to infinity, and since δ_∞^k is also bounded from above, playing on stationary strategies from the set of optimal solutions to every instantaneous matrix game, does provide an optimal strategy for the overall stochastic game. □

References

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences* **39** 1095–1100.