Learning through Gradient-Type Reinforcement for N-person Repeated Constrained Games: Counter-Coalition Space Approach

Alexander S. Poznyak[a], Member, IEEE, Martín Godoy-Alcántar[b] and Eduardo Gómez-Ramírez[c]

[a] CINVESTAV-IPN, Departamento de Control Automático, A.P. 14-740, CP 07000 México D.F., México, e-mail: apoznyak@ctrl.cinvestav.mx.

[b] CINVESTAV-IPN, Departamento de Control Automático, A.P. 14-740, CP 07000 México D.F., México, and Instituto Mexicano del Petróleo, e-mail: mgodoy@ctrl.cinvestav.mx

[c] Laboratorio de Investigación y Desarrollo de Tecnología Avanzada, Universidad La Salle, Benjamín Franklin No. 47 Col. Condesa, CP 06140, México, D. F., México, e-mail: egomez@ci.ulsa.mx

Abstract— **The paper tackles the design and analysis of a learning gradient-type strategy for N-person averaged constrained game with incomplete information. Each player is modelled by a stochastic variable-structure learning automaton (a simplest single state Markov Chain). Using the "joint payoff function", the considered game problem is formulated in terms of, so-called, counter-coalition variables. A special δ-regularization is introduced. Such approach does not require the "diagonal concavity" conditions to guarantee the uniqueness of the Nash equilibrium. The asymptotic convergence of the suggested learning procedure is analyzed.**

Keywords— **Reinforcement learning, Repeated game, Nash equilibrium**

I. Introduction

Games theory dealing with data of stochastic nature (or, with imperfect information) has seen a tremendous growth in the last decades. Two models of players are more frequently considered in the literature: *dynamic and static models.* The first ones are given by a Markov chain characterized by sets (may be, finite or infinite) of states and actions (several actions are allowed to be applied at each state) with the corresponding transition (in general, probabilistic) rules. The second ones (*static models)* represent the simplest case of a finite Markov chain with a single state, that is, there are not any internal dynamic effects and the only actions selected by the participants are responsible for the further behavior of each player; last models are considered within the *Learning Automata paradigm* [8], [6], [3] and [4]. In this paper the simplest static player models will be considered.

A. Motivation

A few papers and rigorous results, dealing with learning (for static or dynamic players model) in constrained repeated stochastic games, are actually known. The fundamental paper of Rosen [9] concerns the uniqueness problem for Nash equilibrium but with no stochastic in the model. Several games with learning were considered by Lakshmivarahan [2] and by Narendra and Thathachar [6], but there were no studies of constraint game situations. The spreading of the Rosen's ideas to the class of strictly diagonal concave games with reinforcement automata learning was realized by Nazin and Poznyak in [7] (see, also [8]). The "*strict convexity condition*" (see also [9]) is also required that restricts significantly the areas of possible applications. Recently, the results dealing with learning to such sort of game have been published by Poznyak and Najim [5] where the Bush-Mosteller reinforcement technique was applied. But this paper also requires the restricting "*strict convexity condition*".

The rigorous description of the class of the repeated games having non unique equilibrium points as well as the analysis of the convergence and rate of learning for different reinforcement schemes still practically stay open that serves a lot for the motivation of this study.

B. Discussion of basic assumptions and restrictions

The most important assumptions (and, as the result, restrictions) concerning the considered game are as follows:

- this is the multi-person repeated game where the behavior of each player is the static model given by the simplest single-state Markov Chain, or, a *finite state automaton* with "time-variable structure" (at each stage of the game the distribution of the selected control actions can vary according to a reinforcement procedure);
- if some controls are realized by players at a current stage, then each participant immediately get the information on the realized values of his payoff and constraint functions which are individual for each player. The necessary information (realizations of both payoffs and constraints) is obtained during the course of the game. These realized payoffs and constraints (below called "*realizations*") are random variables (non obligatory binary) having constant first (conditional) moments which are assumed to be *a* priori unknown as well as uniformly bounded second conditional moment. In other words, the game is one of incomplete information.
- each player has *his own set of constraints* (constraint functions) which may be dependent on the selected control actions of other participants (the strategic interdepen-

dence);
• all participants accept some sort of agreement to use the same reinforcement learning procedure (in this case the stochastic pseudo-gradient scheme) and not to change it to another one during all long-range time of the game. Any other "cooperations" during the game are prohibited. Only the parameters of this procedure can be modified by player during the game, but not the *fixed structure* of the reinforcement procedure.

C. Main Contribution

Briefly, the basic contributions of this paper can be summarized as follows:
• using the "*joint payoff function*", the given game problem is formulated in terms of, so-called, "*counter-coalition variables*" that significantly simplifies the problem description making it linear with respect to these variables;
• for each player the suggested pseudo-gradient reinforcement procedure, based on Lagrange multipliers and an appropriate regularization, is shown to be oriented to obtain the *optimal response* (in average sense) within the given constraints;
• this learning tactics is proven to *lead to the unique Nash equilibrium point* corresponding to the "*extended adjusted vector*" with a minimal norm;
• any "*Strictly Diagonal Conditions*", commonly supposed in the finite game theory, do not required;
• based on the stochastic approximation technique (the Robbins- Siegmund theorem), the *convergence* (with probability one and in mean square sense) of the considered learning procedure to the Nash equilibrium is stated and the *rate of learning* is also estimated.

II. N-person Repeated Constraint Game

Let $(\ ,F,P)$ be a probability space, where $\ $ is the sample space, F a minimal σ-algebra of subsets of $\ $, and P a probability measure on $(\ ,F)$. The symbol ω denotes the canonical point (event) in $\ $. All subsequent random variables will be defined in this space.

A. Player model as Learning Automaton

Each player is modelled by a *stochastic variable-structure learning automaton* which consists of a simple Markov chain containing only one state (memoryless or static system) [3] and [4]. A stochastic automaton operating in a random environment (medium) is an adaptive discrete machine described by a collection $\left\{\Xi, U^k, \left\{\xi_n^k\right\}, \left\{u_n^k\right\}, \left\{p_n^k\right\}\right\}$ where Ξ is the automaton input bounded set, U^k denotes the set $\left\{u^k(1), u^k(2), ..., u^k(N_k)\right\}$ $\left(k=\overline{1,N}\right)$ of actions of the player (automaton), $\left\{\xi_n^k\right\}, \left\{\eta_n^{k,l_k}\right\}$ is a sequence of the k-player (automaton) inputs (payoffs and constrains $\left(\xi_n^k ; \eta_n^{k,l_k}\right) \in \Xi$, $l_k = \overline{1,M_k}$) provided by the game mechanism in a binary (P-model environment) or continuous (S-model environment) form, $\left\{u_n^k\right\}$ is a sequence of the k-player (automaton) outputs (actions) and $p_n^k = \left[p_n^k(1), p_n^k(2), ..., p_n^k(N_k)\right]^\intercal$ is the conditional probability distribution at time n, that is, $p_n^k(i) = P\left\{\omega \in \ : u_n^k = u^k(i) \ / \ F_{n-1}\right\}$, $\sum_{i=1}^{N_k} p_n^k(i) = 1$ Here $F_n = \sigma(u_1^k, p_1^k, \xi_1^k, \eta_1^{k,l_k};\ ...;\ u_n^k, p_n^k, \xi_n^k, \eta_n^{k,l_k})$ is the minimal σ-algebra generated by the corresponding events ($F_n \subseteq F$).

B. How game is played

This game is played in the following way. According to the probability distributions (mixed strategy) p_n^k, at each stage n, simultaneously and independently (we consider a non-cooperative game in which each player has only its own payoff realization $\xi_n^k = \xi_n^k\ (\omega)$ and the realizations $\eta_n^{k,l_k} = \eta_n^{k,l_k}(\omega)$ of the constraints), each player chooses randomly an action, for instance $u^k(i_k)$ $\left(i_k = \overline{1,N_k}, k = \overline{1,N}\right)$:

$$\begin{cases} \xi_n^k = \xi^k\left(\omega; u_n^k = u^k(i_k)\right), \ k = \overline{1,N} \\ \eta_n^k = \left(\eta_n^{k,1}, ..., \eta_n^{k,M_k}\right)^\intercal, \eta_n^{k,l_k} = \eta^{k,l_k}\left(\omega; u_n^k = u^k(i_k)\right) \end{cases}$$

The k^{th} random component ξ_n^k represents the payoff (Borel function) earned by the k^{th} player, and η_n^{k,l_k} is the random realization of the l_k-th constraint for this player. To obtain reasonable generality without excessive complexity, the following assumptions will be in force throughout this paper:
(**A1**) The conditional expectations of ξ_n^k and η_n^k are independent of the history of the game (the game rules are assumed to be unchangeable):

$$\begin{aligned} E\left\{\xi_n^k \mid u_n^k = u^k(i_k) \wedge F_{n-1} \ \ k = \overline{1,N}\right\} &\overset{a.s.}{=} v_{i_1,...,i_N}^k \\ E\left\{\eta_n^{k,l_k} \mid u_n^k = u^k(i_k) \wedge F_{n-1} \ \ k = \overline{1,N}\right\} &\overset{a.s.}{=} w_{i_1,...,i_N}^{k,l_k} \end{aligned} \tag{1}$$

(**A2**) For any realized actions u_n^k at current stage n of the play, the second absolute moments of ξ_n^k and η_n^{k,l_k} are uniformly (on ω and n) bounded with probability one, that is,

$$\begin{aligned} E\left\{\left|\xi^k\left(\omega; u_n^k = u^k(i_k), k=\overline{1,N}\right)\right|^2 \mid F_{n-1},\right. \\ \left. k = \overline{1,N}\right\} \overset{a.s.}{\leq} \sigma_{\xi,k}^+ < \infty \\ E\left\{\left|\eta^{k,l_k}\left(\omega; u_n^k = u^k(i_k), k=\overline{1,N}\right)\right|^2 \mid F_{n-1},\right. \\ \left. k = \overline{1,N}\right\} \overset{a.s.}{\leq} \sigma_{\eta,k,l_k}^+ < \infty \end{aligned} \tag{2}$$

where $|\cdot|$ stands for the absolute value. For each player the average payoff and constrains form the collection of M_k+1 tensors

$$\begin{cases} V^k = \left[v_{i_1,...,i_N}^k\right] \ \left(i_k = \overline{1,N_k}, k = \overline{1,N}\right) \\ W^{k,l_k} = \left[w_{i_1,...,i_N}^{k,l_k}\right] \ \left(l_k = \overline{1,M_k}\right) \end{cases}$$

which are assumed to be *a priori* unknown. Then each player changes his mixed strategy according to the accepted reinforcement $T_n^k : \left(p_n^k \overset{T_n^k}{\to} p_{n+1}^k\right)$ trying for large n to maximize his individual averaged payoff $\liminf_{n\to\infty} \Phi_n^k$ where $\Phi_n^k := n^{-1} \sum_{t=1}^{n} \xi_t^k$ maintaining (with probability one) the constraints $\liminf_{n\to\infty} \Psi_n^{k,l_k} \leq b^{k,l_k}$, $\Psi_n^{k,l_k} := n^{-1} \sum_{t=1}^{n} \eta_t^{k,l_k}$.

Definition 1: At stage n, for the k^{th} player of the considered game, and for any F_{n-1}-measurable conditional probability distribution $(p_n^1, ..., p_n^N)$, the Borel function

$$V^k\left(p_n^1, ., p_n^N\right) := \sum_{i_1=1}^{N_1} \cdot \sum_{i_N=1}^{N_N} v^k_{i_1,...,i_N} \prod_{s=1}^{N} p_n^s\left(i_s\right)$$

are said to be the **expected payoff**, and

$$W^{k,l_k}\left(p_n^1, ., p_n^N\right) := \sum_{i_1=1}^{N_1} \cdot \sum_{i_N=1}^{N_N} w^{k,l_k}_{i_1,...,i_N} \prod_{s=1}^{N} p_n^s\left(i_s\right)$$

corresponds to the **expected constraints**.
Define also the *averaged expected payoff* and *constraints* as follows:

$$V_n^k := n^{-1} \sum_{t=1}^{n} V^k\left(p_t^1, ., p_t^N\right)$$
$$W_n^{k,l_k} := n^{-1} \sum_{t=1}^{n} W^{k,l_k}\left(p_t^1, ., p_t^N\right)$$

. The following lemma (see [5]) states the asymptotic equivalence between $(\Phi_n^k, \Psi_n^{l_k})$ and (V_n^k, W_n^{k,l_k}).

Lemma 1: Under the condition **A1** for any conditional distributions $\left\{p_n^1, ..., p_n^N\right\}$ it follows

$$\Phi_n^k \overset{a.s.}{=} V_n^k + o_\omega\left(n^{-\frac{1}{2}}\right), \Psi_n^k \overset{a.s.}{=} W_n^{k,l} + o_\omega\left(n^{-\frac{1}{2}}\right)$$

III. Mixed Strategies and Nash Equilibrium

A. Randomized strategies

The *randomized (or mixed) strategy* of the k^{th} player is any sequence D^k of vectors $\left\{p_n^k\right\}\ \left(k = \overline{1, N}\right)$ with F_{n-1}-measurable components belonging to the simplex $S^{N_k}_{\varepsilon=0}$, that is,

$$p_n^k \in S^{N_k}_{\varepsilon^k=0} := \left\{p_n^k \in R^{N_k} : p_n^k\left(i\right) \geq \varepsilon^k \geq 0, \sum_{i=1}^{N_k} p_n^k\left(i\right) = 1\right\}$$

Definition 2: The strategies $\overline{D}^1, ..., \overline{D}^N$ are said to be the **non-cooperative equilibrium strategies** (in the Nash sense) if

1) they are admissible, that is, $\min\limits_{p_1^s, s=\overline{1,N}} \liminf\limits_{n\to\infty} W_n^{k,l_k} := \widetilde{W}\left(\overline{D}^1, ..., \overline{D}^N\right) \leq b^{k,l_k}$ for any $k = \overline{1, N}$, $l_k = \overline{1, M_k}$, where $b^{k,l_k} \in R$ are *a priori* given;

2) for any integer k and any admissible strategy D^k

$$\begin{aligned}\widetilde{V}^k\left(\overline{D}^1, ..., \overline{D}^N\right) &:= \min_{p_1^s, s=\overline{1,N}} \liminf_{n\to\infty} V_n^k \\ &\overset{a.s.}{\geq} \widetilde{V}\left(\overline{D}^1, ..., \overline{D}^{k-1}, D^k, \overline{D}^{k+1}, ..., \overline{D}^N\right)\end{aligned} \tag{3}$$

where the minimization is done over all the initial probability distributions $p_1^s, s = \overline{1, N}$.

B. Equilibrium within the subclass of stationary mixed strategies

Consider now the subclass of stationary mixed strategies $D^k = \left\{p^k\right\}\ \left(k = \overline{1, N}\right)$.

Definition 3: The point $\left(\bar{p}^1, ..., \bar{p}^N\right)$ is said to be **an equilibrium point** [10] of the given N-person game within the class of stationary strategies if

$$V^k\left(\bar{p}^1, ..., \bar{p}^N\right) = \max_{p^k \in R^k} V^k\left(\bar{p}^1, ..., p^k, ..., \bar{p}^N\right)$$

$$R^k := S_0^{N_k} \bigcap_{l=1}^{M_k} \left\{p^k : W^{k,l_k}\left(\bar{p}^1, ..., \bar{p}^k, ..., \bar{p}^N\right) \leq b^{k,l_k}\right\}$$

for each $k = \overline{1, N}$
At this point no player can increase his payoff by a unilateral change in his strategy.

Remark 1: The set of all equilibrium strategies $\overline{D}^k\ \left(k = \overline{1, N}\right)$ contains (see [5]) the subset of **admissible stationary strategies** $\left\{\overline{p}^k\right\}$ realizing the inequalities $V^k\left(\overline{p}^1, ..., \overline{p}^N\right) \geq V^k\left(\overline{p}^1, ..., \overline{p}^{k-1}, p^k, \overline{p}^{k+1}, ..., \overline{p}^N\right)$, $W^{k,l_k}\left(\overline{p}^1, ..., \overline{p}^N\right) \leq b^{k,l_k}\ \left(k = \overline{1, N}, l_k = \overline{1, M_k}\right)$ for any $\overline{p}^k \in S^{N_k}_{\varepsilon^k=0}$.

IV. Game as Reinforcement: Problem Setting

Now we are ready to formulate the N-person repeated constrained game problem with *a priori* unknown average payoffs and constrains: *based on current information, generate randomized (mixed) admissible strategies* $\left\{p_n^k\right\}\ \left(k = \overline{1, N}\right)$ *in order to achieve a Nash equilibrium realizable within the subclass of stationary strategies.*
To achieve this objective, first emphasize the following fact. According to Nash theorem [10], the set of admissible stationary distributions $\left(\overline{p}^1, ..., \overline{p}^N\right)$, may contain more than one element. As it is shown in [5] that if an admissible mixed strategy $D^k\ \left(k = \overline{1, N}\right)$ converges to such probability distribution $\left(\overline{p}^1, ..., \overline{p}^N\right)$, that is, $\limsup\limits_{n\to\infty} n^\tau E\left\{\sum\limits_{k=1}^{N} \left\|p_n^k - \overline{p}^k\right\|^2\right\} < \infty$ for some $\tau > 0$, or, in another words, the strategy $D^k\ \left(k = \overline{1, N}\right)$ is asymptotically stationary realizing a Nash equilibrium, then the associated random functions Φ_n^k and Ψ_n^{k,l_k} also converge (with the same rate) to the corresponding average values $V^k\left(\overline{p}^1, ..., \overline{p}^N\right)$ and $W^{k,l_k}\left(\overline{p}^1, ..., \overline{p}^N\right)$ respectively, that is, $\limsup\limits_{n\to\infty} n^\tau E\left\{\sum\limits_{k=1}^{N} \left|\Phi_n^k - V^k\left(\overline{p}^1, ..., \overline{p}^N\right)\right|^2\right\} < \infty$ and $\limsup\limits_{n\to\infty} n^\tau E\left\{\sum\limits_{l_k=1}^{M_k} \left|\Psi_n^{k,l_k} - W^{k,l_k}\left(\overline{p}^1, ..., \overline{p}^N\right)\right|^2\right\} < \infty$.
So, if we construct an asymptotically stationary mixed strategy $\left\{p_n^k\right\}$ converging to a stationary distribution $\left(\overline{p}^1, ..., \overline{p}^N\right)$ realizing a Nash equilibrium, we will be able to attain the main aim of the game. But to do it in a rigorous manner, first, the questions related to the *existence* and the *uniqueness* of the Nash equilibrium (within stationary strategies) should be resolved. Talking on the uniqueness

of the finite matrix games, one can consider the uniqueness of the optimal policy for each player, and the uniqueness of the Nash equilibrium point.
As it is shown by Rosen [9], the non-uniqueness of the Nash equilibrium points cannot be offset by a small regularization (perturbation) term. The condition for the uniqueness of equilibria are known as "*strict diagonal concavity*". As it was shown in [8] and [5], if the given (nonregularized) matrix game is "*diagonal concave*", then the corresponding regularized game turns out to be strictly diagonal concave that, by Rosen´s theorem (the theorem 2 in [9]), implies the uniqueness of the equilibria policy. But in this paper we suggest another approach which does not requires the "*diagonal concavity condition*", but demands the representation of the initial problem in new (counter-coalition) variables with some additional regularization.

V. Equivalent counter-coalition variables representation

A. Joint Payoff Function and Nash Equilibrium

Following the approach presented in [9], let us introduce the joint payoff function $\rho(p,q)$ defined as

$$\rho(p,q) := \sum_{k=1}^{N} V^k\left(p^1,...,p^{k-1},q^k,p^{k+1},...,p^N\right) = \sum_{k=1}^{N} V^k\left(p^{\hat{k}},q^k\right) \tag{4}$$

for any $(p,q)\in(S\times Q)\times(S\times Q)$ where:

$$V^k\left(p^{\hat{k}},q^k\right) = \sum_{j_1=1}^{N_1}\sum_{j_2=1}^{N_2}\cdot\cdot\sum_{j_N=1}^{N_N} V^k_{j_1,...,j_N}\, p^{\hat{k}}_{(\cdot)} q^k_{j_k}$$

with

$$p^{\hat{k}}_{(\cdot)} = p^{\hat{k}}_{j_1,...,j_{k-1},j_{k+1},...,j_N} = p^{\hat{k}}_{j_{\hat{k}}} := \prod_{s=1,s\neq k}^{N} p^s_{j_s} \in S^{N_{\hat{k}}}$$

where $N_{\hat{k}} := \prod_{s=1,s\neq k}^{N} N_s$ being referred to as the *extended counter coalition vector*.

Example 1: For a three player game with two action for each one we have for player one

$$p^{\hat{1}} := \left(p^2_1p^3_1, p^2_1p^3_2, p^2_2p^3_1, p^2_2p^3_2\right)^{\intercal}$$

Using Kakutani's fixed point theorem, Rosen proved the following important result [9], that any N-person matrix game with constraints has an equilibrium point $p^*\in R^{\overline{N}}\left(R^{\overline{N}} := R^{N_1}\times R^{N_2}\times\cdots\times R^{N_N}\right)$, in Nash sense (see Definition 3), satisfying

$$\rho(p^*,p^*) = \max_{q\in S\times Q}\rho(p^*,q) \tag{5}$$

where $Q\subset R^{\overline{N}}$ defined as

$$\begin{aligned} Q &:= Q^1\times\cdots\times Q^N;\ Q^k := Q^{k,1}\times\cdots\times Q^{k,M_k}\\ Q^{k,l_k} &:= \Big\{p\in R^{\overline{N}} \mid k=\overline{1,N},\quad l_k=\overline{1,M_k},\ W^{k,l_k}(p)\\ &:= \sum_{j_1=1}^{N_1}\cdots\sum_{j_N=1}^{N_N} W^{k,l_k}_{j_1,...,j_N}\prod_{i=1}^{N} p^i_{j_i} \le b^{k,l_k}\Big\} \end{aligned} \tag{6}$$

The constraints (6) can be written as

$$p^{\hat{k},\intercal}\overline{W}^{k,l_k}p^k - b^{k,l_k} \le 0, (k=\overline{1,N},\ l_k=\overline{1,M_k})$$

where the matrix $\overline{W}^{k,l_k}$ is given by $\overline{W}^{k,l_k} = \left[w^{k,l_k}_{j_{\hat{k}},j_k}\right]$. The index j_k of the actions of player k grows in the columns and the combination of indexes $j_{\hat{k}}$ of the counter coalition grows in the rows of $\overline{W}^{k,l_k}$, or, in an equivalent form,

$$\widetilde{W}^{k,l_k}\left(p^k,p^{\hat{k}}\right) = p^{k,\intercal}\left(\left(\overline{W}^{k,l_k}\right)^{\intercal} - b^{k,l_k}\mathbf{1^k}\right)p^{\hat{k}}\left(p^k\right) \le 0$$

where $\mathbf{1^k}$ is a matrix of ones of size N_k by $N_{\hat{k}}$ where $N_{\hat{k}} := \prod_{i=1,i\neq k}^{N} N_i$.

Assumption A3: We will assume that the Slater's condition hold, that is, there exist pairs $\left(\hat{p}^k,\hat{p}^{\hat{k}}\right)\in S^{N_k}\times S^{N_{\hat{k}}}$ such that

$$\widetilde{W}^{k,l_k}\left(\hat{p}^k,\hat{p}^{\hat{k}}\right) < 0$$

for all $k=\overline{1,N}$.

B. Criterion of Nash-equilibrium using counter-coalition variables

Define the matrices $\overline{V}^k$ given by $\overline{V}^k = \left[V^k_{j_{\hat{k}},j_k}\right]$ where the index j_k of the actions of player k grows in the columns and the combination of indexes $j_{\hat{k}}$ of the counter coalition grows in the rows of $\overline{V}^k$ and

$$\widetilde{W}^k\left(\beta^k\right) = \sum_{l_k=1}^{M_k}\beta^k_{l_k}\widetilde{W}^{k,l_k};\ \widetilde{W}^{k,l_k} = \overline{W}^{k,l_k} - b^{k,l_k}\mathbf{1^{k,\intercal}}$$

The following theorem represents the criterion of a Nash-equilibrium (see [1]).

Theorem 1: A necessary and sufficient condition that $\widehat{p}^*$ be a Nash equilibrium point is that the collection $(\widehat{p}^*,\alpha^*,\beta^*,t^*)$ is the solution of the following bilinear programing problem

$$\overline{F}(t,\alpha) = \sum_{k=1}^{N} t^k - \sum_{k=1}^{N}\alpha^k \to \max_{p^{\hat{k}}\in R^{N_{\hat{k}}},\alpha^k,t^k\in R,\beta^k\in R^{M_k}}$$

subject to

$$\begin{aligned} \left(\overline{V}^{k,\intercal} - \widetilde{W}^{k,\intercal}\left(\beta^k\right)\right)p^{\hat{k}} - \alpha^k e^k \le 0\\ -\left(\overline{V}^{k,\intercal} - \widetilde{W}^{k,\intercal}\left(\beta^k\right)\right)p^{\hat{k}} + t^k e^k \le 0\\ \underline{M^{k,\hat{s}}\ p^{\hat{s}} = M^{k,\hat{r}}\ p^{\hat{r}};\ p^{\hat{k}}} \in S^{N_{\hat{k}}}\\ k=\overline{1,N},\ l_k=\overline{1,M_k},\quad s=\overline{1,N},\ s\neq k, s\neq r \end{aligned} \tag{7}$$

The following property holds: $\overline{F}(t^*,\alpha^*) = 0$.

VI. Learning Algorithm

A. Regularization, Lagrange multipliers and projection gradient procedure: the complete information case

Using the Lagrange approach define the δ-regularized Lagrangian as:

$$
\begin{array}{l}
L_{\delta}\left(\widehat{p},\alpha,t,\beta,\lambda_{\alpha},\lambda_{t},\lambda_{\widehat{p}}\right):=\sum\limits_{k=1}^{N}t^{k}\text{-}\sum\limits_{k=1}^{N}\alpha^{k}\\
\text{-}\sum\limits_{k=1}^{N}\lambda_{\alpha}^{k,T}\left[\left(\overline{V}^{k,\intercal}\text{-}\widetilde{W}^{k,\intercal}\left(\beta^{k}\right)\right)p^{\widehat{k}}\text{-}\alpha^{k}e^{k}\right]\\
\text{-}\sum\limits_{k=1}^{N}\lambda_{t}^{k,T}\left[\text{-}\left(\overline{V}^{k,\intercal}\text{-}\widetilde{W}^{k,\intercal}\left(\beta^{k}\right)\right)p^{\widehat{k}}+t^{k}e^{k}\right]\\
\text{-}\sum\limits_{k=1}^{N}\sum\limits_{r=1,r\neq k}^{N}\sum\limits_{s=1,s\neq r,k}^{N}\lambda_{\widehat{p},\widehat{r},\widehat{s}}^{k,T}\left[M^{k,\widehat{s}}\;p^{\widehat{s}}\text{-}M^{k,\widehat{r}}\;p^{\widehat{r}}\right]\\
+\frac{\delta}{2}\sum\limits_{k=1}^{N}\left(\text{-}\left\|p^{\widehat{k}}\right\|^{2}\text{-}\left(\alpha^{k}\right)^{2}\text{-}\left(t^{k}\right)^{2}\text{-}\left\|\beta^{k}\right\|^{2}+\right.\\
\left.\left\|\lambda_{\alpha}^{k}\right\|^{2}+\left\|\lambda_{t}^{k}\right\|^{2}+\left\|\lambda_{\widehat{p}}^{k}\right\|^{2}\right)\\
\rightarrow\max\limits_{p^{\widehat{k}}\in S_{\varepsilon_{n}^{k}}^{N_{\widehat{k}}},\alpha^{k},t^{k}\in R,\beta^{k}\in R^{M_{k}}}\;\min\limits_{\lambda_{\alpha}^{k},\lambda_{t}^{k},\lambda_{\widehat{p},\widehat{r},\widehat{s}}^{k}\in R^{N_{k}}}
\end{array}
\tag{8}
$$

In the case when the complete information on the expected payoffs and constraints is available, then the *gradient-like* technique may be applied to attain the equilibrium point:

$$
\begin{array}{l}
p_{n+1}^{\widehat{k}}=\pi_{S_{\varepsilon_{n}^{k}}^{N_{\widehat{k}}}}\left\{p_{n}^{\widehat{k}}+\gamma_{n}^{\widehat{p}}\nabla_{p^{\widehat{k}}}L_{\delta_{n}}^{k}\left(\cdot\right)\right\}\\
\alpha_{n+1}^{k}=\alpha_{n}^{k}+\gamma_{n}^{\alpha}\nabla_{\alpha^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\\
t_{n+1}^{k}=t_{n}^{k}+\gamma_{n}^{t}\nabla_{t^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\\
\beta_{n+1}^{k}=\left[\beta_{n}^{k}+\gamma_{n}^{\beta}\nabla_{\beta^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\right]_{+}\\
\lambda_{\alpha,n+1}^{k}=\left[\lambda_{\alpha,n}^{k}-\gamma_{n}^{\lambda_{\alpha}}\nabla_{\lambda_{\alpha}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\right]_{+}\\
\lambda_{t,n+1}^{k}=\left[\lambda_{t,n}^{k}-\gamma_{n}^{\lambda_{t}}\nabla_{\lambda_{t}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\right]_{+}\\
\lambda_{\widehat{p},\widehat{r},\widehat{s}}^{k}(n+1)=\left[\lambda_{\widehat{p},\widehat{r},\widehat{s}}^{k}(n)-\gamma_{n}^{\lambda_{\widehat{p}}}\nabla_{\lambda_{\widehat{p}}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\right]_{+}\\
k=\overline{1,N},\;n=1,2,...
\end{array}
\tag{9}
$$

where

$$
L_{\delta_{n}}^{k}\left(\cdot\right)=L_{\delta_{n}}^{k}(\widehat{p}_{n},\alpha_{n},t_{n},\beta_{n},\lambda_{\alpha,n},\lambda_{t,n},\lambda_{\widehat{p},n})
$$

and $\pi_{S_{\varepsilon_{n}^{k}}^{N_{\widehat{k}}}}\{\cdot\}$ is the projection operator to the simplex set $S_{\varepsilon_{n}^{k}}^{N_{\widehat{k}}}$ and $[\cdot]_{+}$ is the "positive part" operator. If the parameters of this procedure satisfy

$$
0<\gamma_{n}^{\widehat{p}},\delta_{n},\varepsilon_{n}^{k}\rightarrow 0
$$

$$
\sum_{n=1}^{\infty}\gamma_{n}^{\widehat{p}}\delta_{n}=\infty,\;\sum_{n=1}^{\infty}\kappa_{n}<\infty
$$

(κ_{n} is defined by (12)), it provides the convergence of the estimates $p_{n}^{\widehat{k}}$ to the corresponding unique equilibrium strategy $\overline{p}^{\widehat{1}}\left(\vartheta\right),...,\overline{p}^{\widehat{N}}\left(\vartheta\right)$ having the minimal norm over all possible equilibrium points.

B. Stochastic approximation approach and the "measure keeping problem": incomplete information case

When we deal with incomplete information case, that is, the only current realizations of payoff ξ_{n}^{k} and constraints $\eta_{n}^{k,l_{k}}$ are available, the "stochastic approximation" version of the procedure (9) can be applied (see, for example, [4]) where instead of $\nabla_{p^{\widehat{k}}}L_{\delta_{n}}^{k}\left(\cdot\right)$, $\nabla_{\alpha_{n}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)$, $\nabla_{t^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)$, $\nabla_{\beta^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)$, $\nabla_{\lambda_{\alpha}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)$, $\nabla_{\lambda_{t}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)$, and $\nabla_{\lambda_{\widehat{p}}^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)$ their estimates using the current realizations are implemented. Such procedure is known as the "*reinforcement learning algorithm*" and can be realized by different ways (with different estimates in the use). Below, we present the concrete learning procedure which is based on Learning Automata concept [3], [4].

C. Pseudo-gradient reinforcement with Lagrange multipliers adjustment

The pseudo-gradient reinforcement scheme presented in will be applied hereafter to design a new learning algorithm for N-person constrained repeated games in terms of the complemented variables with unknown expected payoff and constraints. In fact, we assume that after each stage, the payoff to each player as well as the constraints are random variables. No information concerning the distribution of the payoff and constraints is *a priori* available. The necessary information is obtained during the course of the game.

Learning control is an iteration process involving an adaptation at each stage (time step). We now present the "four-step" recursive algorithm.

Step 1. On the basis of the available data

$$
\begin{array}{rcl}
u_{n}^{k} & = & u^{k}\left(i_{k}\right),\xi_{n}^{k},\eta_{n}^{k,l_{k}},p_{n}^{\widehat{k}}(p_{n}^{k}\left(i_{k}\right)\geq\varepsilon_{n}^{k}>0)\\
 & & \lambda_{t,n}^{k},\lambda_{\alpha,n}^{k},\lambda_{\widehat{p},\widehat{r},\widehat{s},n}^{k}
\end{array}
$$

built the following functions for $k=\overline{1,N}$, $n=1,2,...$

$$
\begin{array}{l}
\mathrm{R}\left(\nabla_{p^{\widehat{k}}}L_{\delta_{n}}^{k}\left(\cdot\right)\right):=\dfrac{\left(\xi_{n}^{k}\text{-}\sum\limits_{l_{k}=1}^{M_{k}}\beta_{l_{k}}^{k}\eta_{n}^{k,l_{k}}\right)}{\prod\limits_{s=1,s\neq k}^{N}p^{\widehat{s}}\left(u_{n}^{\widehat{s}}\right)}I^{\widehat{k},k}\left(u_{n}^{\widehat{k}},u_{n}^{k}\right)\times\\
\left(\lambda_{t}^{k}\text{-}\lambda_{\alpha}^{k}\right)+\sum\limits_{i=1,i\neq k}^{N}\sum\limits_{r=1,r\neq i}^{N}\sum\limits_{s=1,s\neq r,i}^{N}\left(M^{i,\widehat{r}}\right)^{T}\lambda_{\widehat{p},\widehat{r},\widehat{s}}^{i}\text{-}\delta_{n}p_{n}^{\widehat{k}}\\
\mathrm{R}\left(\nabla_{\alpha^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\right):=-1+\lambda_{\alpha,n}^{k,T}e^{k}-\delta_{n}\alpha_{n}^{k}\\
\mathrm{R}\left(\nabla_{\beta^{k}}L_{\delta_{n}}^{k}\left(\cdot\right)\right)_{l_{k}}:=\left(\lambda_{t}^{k}-\lambda_{\alpha}^{k}\right)^{T}\\
\times\dfrac{\eta_{n}^{k,l_{k}}}{\prod\limits_{s=1,s\neq k}^{N}p^{\widehat{s}}\left(u_{n}^{\widehat{s}}\right)}\left(I^{\widehat{k},k}\left(u_{n}^{\widehat{k}},u_{n}^{k}\right)\right)^{T}p_{n}^{\widehat{k}}\text{-}\delta_{n}\beta_{n}^{k}
\end{array}
$$

$$
\begin{array}{l}
\mathrm{R}\left(\nabla_{\lambda_\alpha^k} L_{\delta_n}^k(\cdot)\right) := -\dfrac{\left(\xi_n^k - \sum\limits_{l_k=1}^{M_k} \beta_{l_k}^k \eta_n^{k,l_k}\right)}{\prod\limits_{s=1, s\neq k}^{N} p^{\widehat{s}}\left(u_n^{\widehat{s}}\right)} \times \\
\qquad \left(I^{\widehat{k},k}\left(u_n^{\widehat{k}}, u_n^k\right)\right)^T p_n^{\widehat{k}} + \alpha_n^k e^k + \delta_n \lambda_{\alpha,n}^{k,T} \\
\mathrm{R}\left(\nabla_{t^k} L_{\delta_n}^k(\cdot)\right) := 1 - \lambda_{t,n}^{k,T} e^k - \delta_n t_n^k \\
\mathrm{R}\left(\nabla_{\lambda_t^k} L_{\delta_n}^k(\cdot)\right) := \dfrac{\left(\xi_n^k - \sum\limits_{l_k=1}^{M_k} \beta_{l_k}^k \eta_n^{k,l_k}\right)}{\prod\limits_{s=1, s\neq k}^{N} p^{\widehat{s}}\left(u_n^{\widehat{s}}\right)} \times \\
\qquad \left(I^{\widehat{k},k}\left(u_n^{\widehat{k}}, u_n^k\right)\right)^T p_n^{\widehat{k}} - t_n^k e^k + \delta_n \lambda_{\alpha,n}^{k,T} \\
\mathrm{R}\left(\nabla_{\lambda_{\widehat{p},\widehat{r},\widehat{s}}^k} L_{\delta_n}^k(\cdot)\right) := M^{k,\widehat{s}}\, p^{\widehat{s}} - M^{k,\widehat{r}}\, p^{\widehat{r}} + \delta_n \lambda_{\widehat{p},\widehat{r},\widehat{s}}^k
\end{array}
\tag{10}
$$

Step 2. Update the probability distributions p_{n+1}^k and the Lagrange multipliers λ_{n+1} using the following iterative schemes:

$$
\begin{array}{l}
p_{n+1}^{\widehat{k}} = \pi_{S_{\varepsilon_n^k}^{N_{\widehat{k}}}}\left\{p_n^{\widehat{k}} + \gamma_n^{\widehat{p}} \mathrm{R}\left(\nabla_{p^{\widehat{k}}} L_{\delta_n}^k(\cdot)\right)\right\} \\
\alpha_{n+1}^k = \alpha_n^k + \gamma_n^\alpha \mathrm{R}\left(\nabla_{\alpha^k} L_{\delta_n}^k(\cdot)\right) \\
t_{n+1}^k = t_n^k + \gamma_n^t \mathrm{R}\left(\nabla_{t^k} L_{\delta_n}^k(\cdot)\right) \\
\beta_{n+1}^k = \beta_n^k + \gamma_n^\beta \mathrm{R}\left(\nabla_{\beta^k} L_{\delta_n}^k(\cdot)\right) \\
\lambda_\alpha^k(n+1) = \left[\lambda_\alpha^k(n) - \gamma_n^{\lambda_\alpha} \mathrm{R}\left(\nabla_{\lambda_\alpha^k} L_{\delta_n}^k(\cdot)\right)\right]_0^{\lambda_{\alpha,n+1}^+} \\
\lambda_t^k(n+1) = \left[\lambda_t^k(n) - \gamma_n^{\lambda_t} \mathrm{R}\left(\nabla_{\lambda_t^k} L_{\delta_n}^k(\cdot)\right)\right]_0^{\lambda_{t,n+1}^+} \\
\lambda_{\widehat{p},\widehat{r},\widehat{s}}^k(n+1) = \left[\lambda_{\widehat{p},\widehat{r},\widehat{s}}^k(n) - \gamma_n^{\lambda_{\widehat{p}}} \mathrm{R}\left(\nabla_{\lambda_{\widehat{p},\widehat{r},\widehat{s}}^k} L_{\delta_n}^k(\cdot)\right)\right]_0^{\lambda_{\widehat{p},\widehat{r},\widehat{s},n+1}^+} \\
k = \overline{1,N},\ n = 1,2,...
\end{array}
\tag{11}
$$

where if $u_n^k = u^k(j_k)$ and $u_n^{\widehat{k}} = u^{\widehat{k}}\left(j_{\widehat{k}}\right)$ then $I^{\widehat{k},k}\left(u_n^{\widehat{k}}, u_n^k\right) = \left[\delta_{(\widehat{i}_k, i_k)}\right] \in R^{N_{\widehat{k}} \times N_k}$ with $\delta_{(\widehat{i}_k, i_k)} = 1$ if $\widehat{i}_k = j_{\widehat{k}}$ and $i_k = j_k$ and $\delta_{(\widehat{i}_k, i_k)} = 0$ otherwise. The operator $[y]_0^{\lambda_{n+1}^+}$ is defined as follows:

$$
[y]_0^{\lambda_{n+1}^+} = \begin{cases} y \text{ if } y \in [0, \lambda_{n+1}^+] \\ \lambda_{n+1}^+ \text{ if } y > \lambda_{n+1}^+ \\ 0 \text{ if } y < 0 \end{cases}
$$

Step 3. According to $\Pr\left\{u_{n+1}^k = u^k(i) \mid F_n\right\} = p_{n+1}^k(i)$ generate randomly (for each player) new discrete random variables u_{n+1}^k as in learning stochastic automata implementation, and get a new observations (realizations) ξ_{n+1}^k and η_{n+1}^{k,l_k} that corresponds to the environment vector-reactions.

Step 4. Return to Step 1.

VII. Convergence Analysis

A. Main theorem on the convergence with probability one

Theorem 2 (convergence with probab one) Suppose that assumptions **A1- A2** hold for the learning reinforcement procedure (10)-(11) and the given game is diagonal convex. In addition, assume that:

1. there exist the following nonnegative sequences $\left\{\gamma_n^{\widehat{p}}\right\}$, $\left\{\gamma_n^\alpha\right\}, \left\{\gamma_n^t\right\}, \left\{\gamma_n^\beta\right\}, \left\{\gamma_n^{\lambda_\alpha}\right\}, \left\{\gamma_n^{\lambda_t}\right\}, \left\{\gamma_n^{\lambda_{\widehat{p}}}\right\}, \{\delta_n\}$ and $\left\{\varepsilon_n^k\right\}$ such that

$$
\begin{array}{c}
\gamma_n^{\widehat{p}}, \gamma_n^\alpha, \gamma_n^t, \gamma_n^\beta, \gamma_n^{\lambda_\alpha}, \gamma_n^{\lambda_t}, \gamma_n^{\lambda_{\widehat{p}}} \downarrow 0,\ \delta_n \in \left(0, \delta^+\right),\ \delta_n \downarrow 0 \\
\gamma_n^{\widehat{p}} = c_\alpha \gamma_n^\alpha = c_t \gamma_n^t = c_\beta \gamma_n^\beta = c_{\lambda_\alpha} \gamma_n^{\lambda_\alpha} = c_{\lambda_t} \gamma_n^{\lambda_t} = c_{\lambda_{\widehat{p}}} \gamma_n^{\lambda_{\widehat{p}}} \\
\varepsilon_n^k \in \left(0, \frac{1}{N_{\widehat{k}}}\right),\ \varepsilon_n^k \downarrow 0,\ \limsup_n \frac{\varepsilon_n^k}{\delta_n} < \infty
\end{array}
$$

2. the updating factors $\gamma_n^{\widehat{p}}$, γ_n^α, γ_n^t, γ_n^β, $\gamma_n^{\lambda_\alpha}$, $\gamma_n^{\lambda_t}$, $\gamma_n^{\lambda_{\widehat{p}}}$ satisfy $\sum\limits_{n=1}^{\infty} \gamma_n^{\widehat{p}} \delta_n = \infty$ and $\sum\limits_{n=1}^{\infty} \left[\phi_n + \kappa_n^2 \left(\gamma_n^{\widehat{p}} \delta_n\right)^{-1}\right] < \infty$ where

$$
\begin{array}{l}
\kappa_n := C_1 \left\|\varepsilon_{n+1}^k - \varepsilon_n^k\right\| + C_2 \left|\delta_{n+1} - \delta_n\right| \\
\qquad + C_3 \left\|\varepsilon_{n+1}^k \delta_{n+1}^{-1} - \varepsilon_n^k \delta_n\right\|
\end{array}
\tag{12}
$$

$$
\begin{array}{l}
\phi_n := \kappa_n^2 + C_A^2 \sum\limits_{k=1}^{N} \left(\gamma_n^{\widehat{p}}\right)^2 + 2 C_A \kappa_n \sum\limits_{k=1}^{N} \gamma_n^{\widehat{p}} \\
\qquad + 2M \left(\gamma_n^{\widehat{p}}\right)^2 \left[\left(\delta_n \lambda_n^+\right)^2 + \left(\sigma_\eta^+\right)^2\right] \\
\qquad + 2\sqrt{M} \gamma_n^{\widehat{p}} \kappa_n \left(\delta_n \lambda_n^+ + \sigma_\eta^+\right)
\end{array}
$$

Then the mixed strategies of the players ensure the convergence of the game to the equilibrium point with probability one, that is,

$$
\begin{array}{l}
\sum\limits_{k=1}^{N} r_k \left(\left\|p_{n+1}^{\widehat{k}} - \overline{p}^{\widehat{k}}\left(\varepsilon_{n+1}^k, \delta_{n+1}\right)\right\|^2 \right. \\
\qquad \left. + \left\|\lambda_{n+1} - \lambda\left(\varepsilon_{n+1}, \delta_{n+1}\right)\right\|^2\right) \overset{a.s.}{\to} 0
\end{array}
$$

VIII. Conclusion

This paper was concerned with the development of a new learning algorithm for N-person constrained repeated game with unknown expected payoff and constraints. Based on Lyapunov approach and martingale theory, the convergence with probability one has been stated. Simulations results justifying this approach are available.

References

[1] M. Godoy-Alcántar, E. Gómez-Ramírez, A. Poznyak and K. Najim, Alternate Linear Programming Approach for Noncooperative Constrained Finite Games, *International Journal Systems Science*, submitted.

[2] S. Lakshmivarahan, Learning Algorithms Theory and Applications, *Springer-Verlag*, Berlin, 1981.

[3] K. Najim and A. S. Poznyak, *Learning Automata: Theory and Applications*. Oxford: Pergamon Press, 1994.

[4] A. S. Poznyak and K. Najim, *Learning Automata and Stochastic Optimization*. Berlin:Springer-Verlag, 1997.

[5] A. Poznyak and K.Najim, "Learning through reinforcement for N-person repeated constraint games", *IEEE Trans. on Systems, Man, and Cybernetics: Part B Cybernetis*, vol. 32, No.6, 2002.

[6] K. S. Narendra and M. A. L. Thathachar, Learning Automata: an Introduction, *Prentice-Hall*, Englewood Cliffs, New Jersey, 1989.

[7] A. V. Nazin and A. S. Poznyak, "Matrix N-person game with incomplete information," Economics and Mathematical Methods, vol. 14, pp. 958-968, 1978.

[8] Nazin, A. V. and A. S. Poznyak, *Adaptive Choice of Variants* (in Russian). Moscow: Nauka, 1986.

[9] J. B. Rosen, Existence and uniqueness of equilibrium points for concave N-persons games, *Econometrica*, vol. 33, pp. 520-534, 1965.

[10] J. Nash, "Equilibrium points in n-person games," Proc. Nat. Acad. U.S.A., vol.36, pp. 48-49, 1950.