Learning in User-Centric IPTV Services Selection in Heterogeneous Wireless Networks

Manzoor Ahmed Khan*
DAI-Labor, Technical University Berlin, Germany
Email: manzoor-ahmed.khan@dai-labor.de

Hamidou Tembine
SUPELEC, Paris, France
Email: tembine@supelec.fr

Stefen Marx
DAI-Labor, Technical University Berlin, Germany
Email: stefan.marx@dai-labor.de

***Abstract*—It is envisioned that in IPTV services will be run on the top a heterogeneous mix of network infrastructure. The existence of multi-mode terminals enable users get associated to the best available networks according to user preferences and application specific requirements. In this paper we study the user-centric service selection for IPTV services. We propose the user satisfaction function for video services and validate this function against the objective measurement results. The users select the service based on his satisfaction function. We model the service selection problem using dynamic games. We also introduce novel scheme called *cost of learning* that incorporates the cost to switch to an alternate IPTV service provider. Using the *Evalutionary game dynamics*, we study the convergence and stability properties in user-centric IPTV service selection problem. We propose a trusted third party based architecture solution to realize the user-centric network selection. For the proof of concept simulations were run in OPNET and Mathematica.**

I. Introduction

After the initial monopoly-like era in telecommunication era, an increasing number of (real and virtual) network operators and service providers have been observed on the market in most countries. The disintegration of service providers, content providers, network infrastructure providers etc. have introduced new dynamics in the telecommunication market. Service consumers benefit from resulting competition among all the stake holders by having much wider spectrum for more competitive service quality, price etc. This in turn results in recently most addressed concept of user Quality of Experience (QoE). Given such a user-centric network selection approach, the focus of service providers/operators shift from the only throughput maximization to a more subjective objective function of increasing satisfied user pool. Concentrating on IPTV, such an objective function of operators or service providers necessitates a framework that estimates the user satisfaction for video applications in terms of i) network technology technical indices (which in turn requires mapping of QoS over QoE) and ii) economical preferences of users such as service pricing etc. Game-theory proves to be a natural candidate to address the problems of network selection in the disintegrated telecommunication paradigm. We further consider the dynamic games to address the problem network(service) selection for extension of IPTV services and introduce the novel concept of learning in this regard. We consider more realistic, disintegrated, and dynamic communication scenario.

One of the reasons that motivated authors to consider dynamic scenarios in evolving networks is that they seem to show up in reality more often. In the recent research literature game theory finds its application in the network traffic, routing, congestion games, security. However, in most of the studies a static network model is considered which includes a game which is framed over static network, static user demand and a fixed iterative learning scheme. We can not neglect the fact that with increase in the complexity of existing system we cannot assume the environment to be constant. Thus we need to study and explore the dynamic behavior of such systems which involve not only the time dependencies and the state of the environment but also the variability of the demands, the uncertainty of the system parameters, the random activity of the users, the time delays, error and noise in the measurement over long-run interactions, etc. [9] shows that algorithms that do not require information about other stake-holders' actions or payoffs can not cause the user period-by-period behavior to converge to Nash equilibrium for a large scale class of games. Hence, most of the time, there is no guarantee that the behaviors of fully distributed learning algorithms and dynamics will come close to Nash equilibrium. By introducing public signals (but irrelevant-payoff signals) into the interaction, each user (player) can choose his/her action according to her observation of the value of the signal. Then, a strategy assigns an action to every possible observation a user can make. If no user would want to deviate from the recommended strategy (assuming the others don't deviate), the distribution is called a correlated equilibrium. The works in [5], [8] showed that *regret-minimizing procedures* can cause the empirical frequency distribution of play to converge to the set of correlated equilibria. Note that the set of correlated equilibria is convex and includes the convex hull of the set of Nash equilibria.

Another aspect in networking and communication games is the *uncertainty*. A category of games with uncertainty is known as *Robust games*, users in such games need to make their decisions using algorithms that accommodate limitations in information gathering and processing. This disqualifies some of the well known decision making models (such as *fictitious play*, best reply, gradient descent, model-based algorithms etc.)

Recently, distributed learning algorithms and feedback based update rules have been extensively developed in networking and communication systems. Particular cases of Bush-Mosteller [3] with slight changes have been examined in [13], [15]. Xing & Chandramoulli [15] have studied stochastic learning algorithm in a distributed discrete power control problem. The authors in [15] have investigated in detail the convergence and the divergence issues for the two-user two-action case. However, a payoff-reinforcement learning (Q-value learning) is not examined in their models.

The authors in [6] proposed Q-learning algorithms for non-zero-sum finite stochastic games in wireless networks. However a general convergence result of such algorithms remain a challenging open problem. We give a convergence result for such games with uncontrolled and ergodic state transitions. In [4], the authors analyzed the robustness of the dynamics when users join and leave the network. However, the case where the users have different behavior (different learning patterns and different speed of learning) is not examined in [4]. As we will see in this paper, these two parameters are very important in terms of convergence time of the combined learning in a dynamic unknown environment.

Different from *distributed learning optimization*, we use the term *strategic learning* [16].

978-1-4244-9920-5/11/$26.00 ©2011 IEEE

Only few convergence results are known in strategic learning. These are obtained for a particular structure of the payoffs and action spaces: $[R1]$ Lyapunov expected games (any finite robust game in which the expected payoff leads to an hybrid dynamics which has a Lyapunov function). Particular classes of these games are potential games, common interest games, dummy games, congestion games etc under specific dynamics. $[R2]$ Two-user-two-action games for well-chosen learning patterns and generic payoffs, $[R3]$ Particular class of games with monotone expected payoffs, $[R4]$ Particular classes of supermodular games, submodular games in low dimension (2 or 3), $[R5]$ Dominant solvable games (games with a dominant strategy).

Detailed analysis of these results can be found in [12], [1].

All the above convergence results [R1-R5] for specific classes of games can be extended into the class of robust games.

Once we move from these specific classes of games, the convergence of learning schemes must be proven. Cases of non-convergence under homogeneous learning including *cycling games* which leads to limit cycles and oscillating behaviors may occur. Using specific learning rates and by carefully choosing the learning scheme, the multiple-scale learning is known to be convergent in specific classes of games that generalize Shapley's games, Jordan games, matching pennies, variations of Rock-Scissor-Paper games etc. The generalization uses the Dulac's theorem and Poincaré - Bendixson theorem (see [7]) which states that for planar systems if the w-limit set is non-empty and if the trace of the Jacobian of the system (the divergence) is of constant sign for all pair of the variables, then the system is convergent. Note that these results are limited to planar systems i.e they can be used only for two-action games or at most three-actions symmetric games. Using the multiple time-scale stochastic approximations developed in [2], we study various combined learning algorithms for stochastic games with particular state transition structures.

A. Contribution

In this paper, we focus on hybrid and combined strategic learning in future user-centric IPTV service provider selection. The hybrid learning focuses on general-sum stochastic dynamic games with incomplete information and action-independent state transition with the following novelties: i) In contrast to the standard learning approaches widely studied in the literature [17], [8] where the users follow the same predetermined scheme, here we relax this assumption and the users do not need to follow the same learning patterns. We propose different learning schemes that the users can adopt. This leads to *heterogeneous learning*. Our motivation for heterogeneous learning in user-centric IPTV service selection follows from the observation that, in mentioned problem of service selection, the users may not see the environment in the same way, they may have different capabilities and different adaptation degrees (in this case options of IPTV service providers, terminal capabilities etc.). Thus, it is important to take into consideration these differences when analyzing the behavior of the wireless medium characteristics and varying service provider offers. *As we will see the heterogeneity in the learning is crucial in term of convergence of certain systems.* ii) Each user does not need to update his strategy at each iteration. The updating times are random and unknown by the users. Usually, in the iterative learning schemes the time slots during which the user updates are fixed. Here we do not restrict to fixed updating time. This is because some users come in or exit temporarily (specifically when it comes to IPTV like services. Owing to this dynamic behavior, IPTV Service Provider (SP) vary the service cost offers dynamically), and it may be costly to update or for some other reasons, the users may prefer to update their strategies at another time. One may think that if some of the user does not update often, the strategic learning process will be slower in terms of convergence time; this statement is less clear because the off-line users may indirectly help the online users to converge and, when they wake-up they respond to an already converged system, and so on. iii) We propose a *cost of learning* CODIPAS-RL which takes into consideration the cost of switching the IPTV SPs. (applicable in the context of technology selection as well) iv) Our theoretical findings are illustrated numerically in heterogeneous wireless networks with multiple classes of users and multiple technologies: wireless local area networks (WLAN) and long term evolution (LTE) using Mathematica and OPNET Simulation.

To the best of the authors' knowledge, this is the first paper analyzing (i) the cost of learning in an heterogeneous and unknown environment (ii) convergence results for hybrid learning schemes, (iii) mean field learning in games subject to uncertainty and their connection to evolutionary game dynamics, (iv) combining theoretical results with the experimental learning scenarios using OPNET simulator.

We summarize some of the notations in Table I.

TABLE I
SUMMARY OF NOTATIONS

Symbol	Meaning
$\mathbb{R}^k$	$k-$dimensional Euclidean space
$\mathcal{W} \subseteq \mathbb{R}^k$	state space
$\mathcal{N}$	set of potential users (finite or infinite)
$\mathcal{B}^n(t)$	random set of active users at time t.
$\mathcal{A}_j$	set of actions of user j
$s_j \in \mathcal{A}_j$	a generic element of $\mathcal{A}_j$
$\mathcal{X}_j := \Delta(\mathcal{A}_j)$	set of probability distributions over $\mathcal{A}_j$
$a_{j,t} \in \mathcal{A}_j$	action of the user j at time t
$\mathbf{x}_{j,t} \in \mathcal{X}_j$	strategy of the user j at t
$u_{j,t}$	perceived payoff by user j at t
$\hat{\mathbf{u}}_{j,t} \in \mathbb{R}^{\|\mathcal{A}_j\|}$	estimated payoff vector of user j at t
l^2	space of sequences $\{\lambda_t\}_{t\geq 0}$, $\sum_{t\in\mathbb{N}} \|\lambda_t\|^2 < +\infty$
l^1	space of sequences $\{\lambda_t\}_{t\geq 0}$, $\sum_{t\in\mathbb{N}} \|\lambda_t\| < +\infty$
$(\lambda_{j,t}, \nu_{j,t})$	learning rates of user j at t
$m_t^p(.)$	Mean field limit at time t
$\zeta_{k,c}(\hat{l})$	user sensitivity towards values of $\hat{l}$.

II. The setting

A. Description of the IPTV service providers dynamic environment

We examine a system with a finite number of *potential users*. The set of users is denoted by $\mathcal{N} = \{1, 2, \dots, n\}$, $n = |\mathcal{N}|$. The number n can be 10, 10^4 or 10^6. Each user has a finite number of actions denoted by $\mathcal{A}_j$ (which can be arbitrary large, a union of IPTV SPs and Infrastructure providers). Time is discrete and the space of time is $\mathbb{N} = \{0, 1, 2, \dots\}$. A user does not necessarily interact at all the time steps. Each user can be in one of the two modes: *active mode* or *sleep mode*. The set of users interacting at the current time is the set of active users $\mathcal{B}^n(t) \subseteq \mathcal{N}$. This time-varying set is unknown to the users. When an user is in active mode, he does an experiment, and gets a measurement or a reaction to his decision (of selecting any SP or operator), denoted $u_{j,t} \in \mathbb{R}$ (this may be delayed as we will see). Let $\mathcal{X}_j := \Delta(\mathcal{A}_j)$ be the set of probability distributions over $\mathcal{A}_j$ i.e the simplex of $\mathbb{R}^{|\mathcal{A}_j|}$. The number $u_{j,t} \in \mathbb{R}$ is the realization of a random variable $\tilde{U}_{j,t}$ which depends on the state of nature $\mathbf{w}_t \in \mathcal{W}$ and the action of the users where the set $\mathcal{W}$ is a subset of a finite dimensional Euclidean space. Each *active user* j updates his current strategy $\mathbf{x}_{j,t+1} \in \Delta(\mathcal{A}_j)$ based on his experiment and its prediction for his future interaction via the payoff estimation $\hat{\mathbf{u}}_{j,t+1} \in \mathbb{R}^{|\mathcal{A}_j|}$.(where payoff is defined as the user satisfaction, detailed in the later sections)

This leads into the class of dynamic games with unknown payoff function and with imperfect monitoring (the last decisions of the other users are not observed). A payoff in the long-run interaction is the average payoff which we assume to have a limit. In that case, under the stationary strategies, the limiting of the average payoff can be expressed as an expected game i.e the game with payoff $v_j : \prod_{j' \in \mathcal{N}} \mathcal{X}_{j'} \longrightarrow \mathbb{R}$, $v_j(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n} \left(\mathbb{E}\tilde{U}_j\right)$

Assumptions on user's information: The only information assumed is that each user is able to observe or to measure a noisy value of its payoff when he is active and update its strategy based on this measurement, which is realistic.

Note that the users do not need to know their own action space in advance. Each user can learn his action space (using for example exploration techniques). In that case, we need to add an exploration phase or a progressive exploration during the dynamic game.

In order to define rigourously the dynamic robust game, we need some preliminaries. Next, we introduce the notions of histories, strategies and payoffs (performance metrics). The payoff is associated to a (behavioral) strategy profile which is a collection of mapping from the set of histories to the available actions at the current time.

Histories An user's information consists of his (own) past activities, own-actions and measured own-payoffs. A private history up to t for user j is a collection $h_{j,t} = (b_{j,0}, a_{j,0}, u_{j,0}, b_{j,1}, a_{j,1}, u_{j,1}, \ldots, b_{j,t-1}, a_{j,t-1}, u_{j,t-1})$ in the set $H_{j,t} := (\{0,1\} \times \mathcal{A}_j \times \mathbb{R})^t$. where $b_{j,t} = \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}}$ which is 1 if j is active at time t and 0 otherwise.

Behavioral Strategy A behavioral strategy for user j is a mapping $\tilde{\tau}_j : \bigcup_{t \geq 0} H_{j,t} \longrightarrow \mathcal{X}_j$. We denote by Σ_j the set of behavioral strategies of user j.

The set of complete histories of the dynamic robust game after t stages is $H_t = (2^{\mathcal{N}} \times \mathcal{W} \times \prod_{j \in \mathcal{N}} \mathcal{A}_j \ \times \mathbb{R}^n)^t$, it describes the set of active users, the states, the chosen actions and the received payoffs for all the users at all past stages before t. The set $2^{\mathcal{N}}$ denotes the set of all the subsets of $\mathcal{N}$ (except the empty set). A behavioral strategy profile $\tilde{\tau} = (\tilde{\tau}_j)_{j \in \mathcal{N}} \in \prod_j \Sigma_j$ and a initial state $\mathbf{w}$ induce a probability distribution $P_{\mathbf{w},\tilde{\tau}}$ on the set of plays $H_\infty = (\mathcal{W} \times \prod_j \mathcal{A}_j \ \times \mathbb{R}^n)^{\mathbb{N}}$.

Payoffs Assume that $\mathbf{w}, \mathcal{B}^n$ are independent and independent of the strategy profiles. For a given $\mathbf{w}, \mathcal{B}^n$, we denote $U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{x}) := \mathbb{E}_{(\mathbf{x}_k)_{k \in \mathcal{B}^n}} \tilde{U}_j^{\mathcal{B}^n}(\mathbf{w}, (a_k)_{k \in \mathcal{B}^n})$. Let $\mathbb{E}_{\mathbf{w},\mathcal{B}^n}$ be the mathematical expectation relatively to the measure generated by the random variables $\mathbf{w}, \mathcal{B}^n$. Then, the expected payoff can be written as $\mathbb{E}_{\mathbf{w},\mathcal{B}^n} \tilde{U}_j^{\mathcal{B}^n}(.,.)$.

We focus on the limiting of the average payoff i.e $F_{j,T} = \frac{1}{T}\sum_{t=1}^{T} u_{j,t}$. The long-term payoff reduces to $\frac{1}{\sum_{t=1}^{T} \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}}} \sum_{t=1}^{T} \ u_{j,t} \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}}$, when considering only the activity of user j. We assume that we do not have short-term users or equivalently the probability for an user j to be active is strictly positive. Given a initial state $\mathbf{w}$ and a strategy profile $\tilde{\tau}$, the payoff of user j is the superior limiting of the Cesaro-mean payoff $\mathbb{E}_{\mathbf{w},\tilde{\tau},\mathcal{B}^n} F_{j,T}$. We assume that $\mathbb{E}_{\mathbf{w},\tilde{\tau},\mathcal{B}^n} F_{j,T}$ has a limit. Then, the expected payoff of an active user j is denoted by $v_j(e_{s_j}, \mathbf{x}_{-j}) = \mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, e_{s_j}, \mathbf{x}_{-j})$ where e_{s_j} is the vector unit with 1 at the position of s_j and zero otherwise.

Definition 1 (Expected robust game)**.** *We define the expected robust game as* $\left(\mathcal{N}, (\mathcal{X}_j)_{j \in \mathcal{N}}, \mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, .)\right)$.

Definition 2. *A strategy profile* $(\mathbf{x}_j)_{j \in \mathcal{N}} \in \prod_{j=1}^{n} \mathcal{X}_j$ *is a (mixed) state-independent equilibrium for the expected robust game if and only if* $\forall j \in \mathcal{N}$, $\forall \mathbf{y}_j \in \mathcal{X}_j$,

$$\mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{y}_j, \mathbf{x}_{-j}) \leq \mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{x}_j, \mathbf{x}_{-j}), \tag{1}$$

The existence of solution of Equation (1) is equivalent to the existence of solution of the following *variational inequality problem*: find $\mathbf{x}$ such that $\langle \mathbf{x} - \mathbf{y}, V(\mathbf{x}) \rangle \geq 0$, $\forall \mathbf{y} \in \prod_j \mathcal{X}_j$ where $\langle ., . \rangle$ is the inner product, $V(\mathbf{x}) = [V_1(\mathbf{x}), \ldots, V_n(\mathbf{x})]$, $V_j(\mathbf{x}) = [\mathbb{E}_{\mathbf{w},\mathcal{B}} U_j^{\mathcal{B}}(\mathbf{w}, e_{s_j}, \mathbf{x}_{-j})]_{s_j \in \mathcal{A}_j}$.

Lemma 1. *Assume that $\mathcal{W}$ is compact. Then, The expected robust game with unknown state and variable number of interacting users has at least one (state-independent) equilibrium.*

The existence of such equilibrium points is guaranteed since the mappings $v_j : (\mathbf{x}_j, \mathbf{x}_{-j}) \longmapsto \mathbb{E}_{\mathbf{w},\mathcal{B}} U_j^{\mathcal{B}}(\mathbf{w}, \mathbf{x}_j, \mathbf{x}_{-j})$ is jointly continuous, quasi-concave in $\mathbf{x}_j$, the spaces $\mathcal{X}_j$, are non-empty, convex and compact. Then, the result follows by using Kakutani fixed point theorem or by applying Nash theorem to the expected robust game.

Since we have existence of state-independent equilibrium under suitable conditions, we seek for heterogeneous and combined algorithms to locate the equilibria.

III. CODIPAS-RL

We propose an hybrid, delayed, COmbined fully DIstributed PAyoff and Strategy Reinforcement Learning in the following form: (hybrid-delayed-CODIPAS-RL)

$$\left\{\begin{array}{c}
\mathbf{x}_{j,t+1}(s_j) - \mathbf{x}_{j,t}(s_j) = \\
\mathbb{1}_{\{j \in \mathcal{B}^n(t)\}} \sum_{l \in \mathcal{L}} \mathbb{1}_{\{l_{j,t}=l\}} K_{j,s_j}^{1,(l)}(\lambda_{j,\theta_j(t)}, a_{j,t}, u_{j,t-\tau_j}, \hat{\mathbf{u}}_{j,t}, \mathbf{x}_{j,t}), \\
\hat{\mathbf{u}}_{j,t+1}(s_j) - \hat{\mathbf{u}}_{j,t}(s_j) = \\
\mathbb{1}_{\{j \in \mathcal{B}^n(t)\}} K_{j,s_j}^{2}(\nu_{j,\theta_j(t)}, a_{j,t}, u_{j,t-\tau_j}, \hat{\mathbf{u}}_{j,t}, \mathbf{x}_{j,t}), \\
j \in \mathcal{N}, t \geq 0, a_{j,t} \in \mathcal{A}_j, s_j \in \mathcal{A}_j, \\
\theta_j(t+1) = \theta_j(t) + \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}}, \\
t \geq 0, \ \mathcal{B}^n(t) \subseteq \mathcal{N}, \\
\mathbf{x}_{j,0} \in \mathcal{X}_j, \hat{\mathbf{u}}_{j,0} \in \mathbb{R}^{|\mathcal{A}_j|}.
\end{array}\right.$$

where $\hat{\mathbf{u}}_{j,t} = (\hat{u}_{j,t}(s_j))_{s_j \in \mathcal{A}_j} \in \mathbb{R}^{|\mathcal{A}_j|}$ is a vector payoff estimation of user j at time t. Note that when user j uses $a_{j,t} = s_j$, he observes only his measurement corresponding to that action but not those of the other actions $s'_j \neq s_j$. Hence he needs to estimate/predict them via the vector $\hat{\mathbf{u}}_{j,t+1}$. The functions K^1 and λ are based on estimated payoffs and perceived measured payoff (delayed and noisy) such that the invariance of simplex is preserved almost surely. The function K_j^1 defines the strategy learning pattern of user j and $\lambda_{j,\theta_j(t)}$ is its strategy learning rate. If at least two of the functions K_j are different then we refer to *heterogeneous learning* in the sense that the learning schemes of the users are different. If all the K_j^1 are identical but the learning rates λ_j are different, we refer to *learning with different speed*: slow learners, medium or fast learners. Note that the term $\lambda_{j,\theta_j(t)}$ is used instead of $\lambda_{j,t}$ because the global clock $[t]$ is not known by user j (he knows only how many times he has been active, the activity of others is not known by j). $\theta_j(t)$ is a random variable that determines the local clock of j. Thus, the updates are asynchronous. The functions K_j^2, and ν_j are well-chosen in order to have a good estimation of the payoffs. τ_j is a time delay associated to user j in its payoff measurement. The payoff $u_{j,t-\tau_j}$ at $t - \tau_j$ is perceived at time t. We examine the case where the users can choose different CODIPAS-RL patterns during the dynamic game. They can select among a set of CODIPAS-RLs denoted by $\mathcal{L}_1, \ldots, \mathcal{L}_m, m \geq 1$. The resulting learning scheme is called *hybrid CODIPAS-RL*. The term $l_{j,t}$ is the CODIPAS-RL pattern chosen by user j at time t.

A. CODIPAS-RL patterns

We examine the above dynamic game in which each user learns according to a specific CODIPAS-RL scheme.

1) Bush-Mosteller based CODIPAS-RL: $\mathcal{L}_1$: The learning pattern $\mathcal{L}_1$ is given by

$$
\begin{aligned}
x_{j,t+1}(s_j) - x_{j,t}(s_j) = \lambda_{\theta_j(t)} 1\!\!1_{\{j\in\mathcal{B}^n(t)\}} \times \\
\frac{u_{j,t} - \Gamma_j}{\sup_{\mathbf{a},w} |U_j(w,a) - \Gamma_j|} \left(1\!\!1_{\{a_{j,t}=s_j\}} - \mathbf{x}_{j,t}(s_j)\right), \qquad (2)\\
\hat{u}_{j,t+1}(s_j) - \hat{u}_{j,t}(s_j) = \\
\nu_{\theta_j(t)} 1\!\!1_{\{a_{j,t}=s_j, j\in\mathcal{B}^n(t)\}} \left(u_{j,t} - \hat{u}_{j,t}(s_j)\right) \qquad (3)\\
\theta_j(t+1) = \theta_j(t) + 1\!\!1_{\{j\in\mathcal{B}^n(t)\}} \qquad (4)
\end{aligned}
$$

where Γ_j is a reference level of j. The first equation of $\mathcal{L}_1$ is widely studied in machine learning and have been initially proposed by Bush & Mosteller in 1949-55 [3]. The second equation of $\mathcal{L}_1$ is a payoff estimation for the experimented action by the users. Combined together one gets a specific combined fully distributed payoff and strategy reinforcement learning based on Bush-Mosteller reinforcement learning.

B. Main results

We introduce the following assumptions. $[H2]$, $\forall j \in \mathcal{N},\ \liminf_{t\longrightarrow\infty} \frac{\theta_j(t)}{t} > 0$

$[H3]$ $\lambda_t \geq 0$, $\lambda \in l^2\backslash l^1$, $\mathbb{E}\left(M_{j,t+1} \mid \mathcal{F}_t\right) = 0$, $\forall\ j$, $\mathbb{E}\left(\| M_{j,t+1} \|^2\right) \leq c_1 \left[1 + \sup_{t'\leq t} \| \mathbf{x}_{t'} \|^2\right]$ where $c_1 > 0$ is a constant.

It is important to mention that these assumptions H2-H3 are standard assumptions in stochastic approximations for almost sure convergence. However the vanishing learning rate can be time-consuming. In order to design fast convergent learning algorithms, *constant* learning rate ($\lambda_t = \lambda$) can be used as well, and convergence in law can be proved under suitable conditions. In this case the expectation of the gap between the solution of differential equations and the stochastic process is in order of the constant learning rate i.e $O(\lambda)$. In particular, if $\lambda \longrightarrow 0$ one has a weak convergence. Below the give the main results for time-varying learning rate under H2-H3.

Consequence for IPTV service selection games Under suitable conditions of the learning rate, the above learning schemes can be studied by their differential equation counterparts, and the result applies directly to autonomous self-organizing networks with randomly changing IPTV service offers, variable number of interacting users that impact on the video service quality and in turn influences the user QoE, and random updating time slots. Next, we provide our second main result which establishes heterogeneous learning convergence and capture the impact of different behavior of the users.

Theorem 1 (heterogenous rates)**.** *Assume H2-H3 and Assume that the payoff-learning rates are faster than strategy learning rates i.e* $[H4]$ $\lambda_t \geq 0, \nu_t \geq 0$, $(\lambda,\nu) \in (l^2\backslash l^1)^2$, $\frac{\lambda_t}{\nu_t} \longrightarrow 0$. *Then, hybrid-delayed-CODIPAS-RL scheme with variable number of players has the asymptotic pseudo trajectory of the following non-autonomous system:*

$$
\left\{
\begin{array}{c}
\dot{x}_{j,t}(s_j) = g_{j,t} \sum_{l\in\mathcal{L}} p_{j,t,l} f^{(l)}_{j,s_j}(\mathbf{x}_{j,t}, \mathbb{E}_{\mathbf{w},\mathcal{B}} U^{\mathcal{B}}_j(\mathbf{w}, ., \mathbf{x}_{-j,t})) \\
x_j(s_j) > 0 \Longrightarrow \hat{\mathbf{u}}_{j,t}(s_j) \longrightarrow \mathbb{E}_{\mathbf{w},\mathcal{B}} U^{\mathcal{B}}_j(\mathbf{w}, \mathbf{e}_{s_j}, \mathbf{x}_{-j})
\end{array}
\right.
$$

We define two properties:

- NS: Nash stationary property refers to the configuration in which the set of Nash equilibria of the expected game coincide with the rest points (stationary points) of the resulting hybrid dynamics.
- PC: Positive Correlation property refers to the configuration where the covariance between the strategies generated by the dynamics and the payoff is positive. i.e $F(x) \neq 0 \Longrightarrow \sum_{j,s_j} u_j(e_{s_j}, \mathbf{x}_{-j}) F_{j,s_j}(\mathbf{x}) > 0$ where F is the drift of the dynamics. We say that the expected robust game is a potential game if there exists a regular function W such that $u_j(e_{s_j}, \mathbf{x}_{-j}) = \frac{\partial}{\partial x_j(s_j)} W(\mathbf{x})$.

The proof is omitted due to the space limitations.

Impact of these results for IPTV service selection games In many cases, the games have specific structures such as *aggregative games, potential games, supermodular games*. This result gives the convergence of heterogeneous learning to equilibria in dynamic robust potential games but also in dynamic monotone games.

IV. Mean field hybrid learning

In this subsection we show how to extend the learning framework to large number of players called *mean field learning*.

A. Learning under noisy strategy

Following the above lines, one can generalize the CODIPAS-RL in the context of Itô's stochastic differential equation (SDE). Typically, the case where the strategy learning has the following form: $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda_t(f(\mathbf{x}_t, \hat{\mathbf{u}}_t) + M_{t+1}) + \sqrt{\lambda_t}\sigma(\mathbf{x}_t, \hat{\mathbf{u}}_t)$, can be seen as an Euler scheme of the Itô's SDE: $d\mathbf{x}_{j,t} = f_j(\mathbf{x}_t, \hat{\mathbf{u}}_t)dt + \sigma_j(\mathbf{x}_t, \hat{\mathbf{u}}_t)d\mathbb{B}_{j,t}$, where $\mathbb{B}_{j,t}$ is a standard Brownian motion in $\mathbb{R}^{|\mathcal{A}_j|}$. This leads stochastic evolutionary game dynamics where the stochastic stability of equilibria can be used to find robustness of the system under stochastic fluctuations. Note that the distribution the noisy strategy-learning or equivalently the mean field learning can be characterized by a solution of the following partial differential equation called Fokker-Planck-Kolmogorov equation $\partial_t m_{j,t}(x) + div(f_j m_t) - \frac{1}{2} trace(\sigma\sigma^t \partial^2_{xx} m_t) = 0$, where div is the divergence operator and ∂^2_{xx} is the matrix of second derivatives of $m_t(.)$ with the respect to x. Particular case of this class of dynamics are *evolutionary game dynamics with diffusion terms*. We refer to [11] for the derivation of these equations which require the theory of distribution and integration by parts.

B. Cost of learning CODIPAS-RL

In this subsection we introduce a novel way of learning under switching cost called *Cost-To-Learn CODIPAS-RL*. Usually in learning in games , the cost of switching between the actions, the cost of experimenting with another option are not taken into consideration. In this section we take these issues into account and study their effects in the learning outcome. In our scenario, the learning cost can arise in three different situations: (i) service provider switch, (ii) infrastructure switchover (codec-switchover, handover), (iii) joint service provider-switch-handover-and-codec switch-over. In a more general setting, one can think about a cost to have a new technology or a cost to produce a specific product. The reason for this cost of learning approach is that, in many situations, changing, improving the performance, the quality of experience of a user, guaranteeing to a quality of service etc has cost. At a given time t, if user j changed its selection (SP switchover, codec switchover, handover etc) i.e., if user j moves, its objective function is translated form the standard utility plus an additional cost for moving from the old configuration to the new one. Then, there is no additional cost to learn if the action remains the same.

V. User Satisfaction Function

When it comes to the measurement of user satisfaction for IPTV services, one is faced with few natural questions. i) How to quantify the perceived user satisfaction with respect to video quality? ii) Does the service cost affect user satisfaction? if yes, how to quantify the user satisfaction with respect to service costs? iii) How to capture

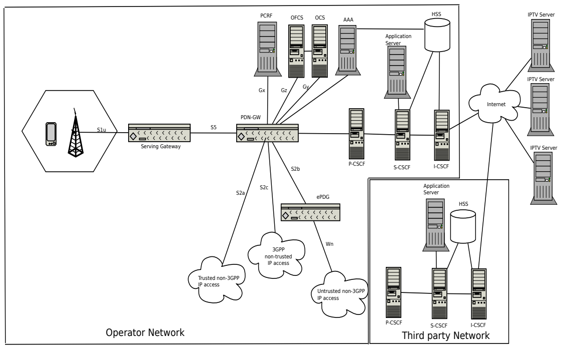

Fig. 1. Proposed IPTV Architecture

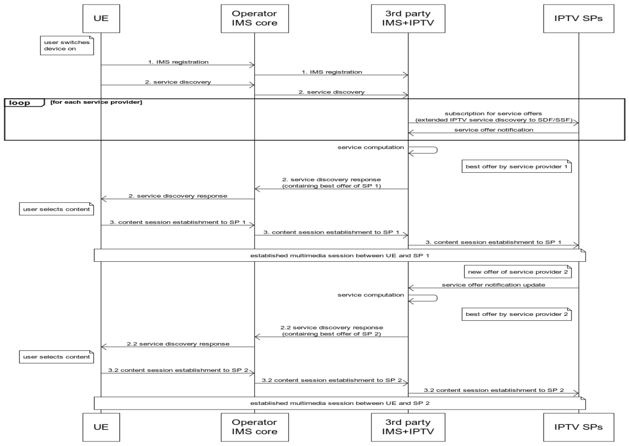
UE
Operator IMS core
3rd party IMS+IPTV
IPTV SPs
user switches device on
1. IMS registration
1. IMS registration
2. service discovery
2. service discovery
loop
[for each service provider]
subscription for service offers (extended IPTV service discovery to SDF/SSF)
service offer notification
service computation
best offer by service provider 1
2. service discovery response (containing best offer of SP 1)
2. service discovery response
user selects content
3. content session establishment to SP 1
3. content session establishment to SP 1
3. content session establishment to SP 1
established multimedia session between UE and SP 1
new offer of service provider 2
service offer notification update
service computation
best offer by service provider 2
2.2 service discovery response (containing best offer of SP 2)
2.2 service discovery response
user selects content
3.2 content session establishment to SP 2
3.2 content session establishment to SP 2
3.2 content session establishment to SP 2
established multimedia session between UE and SP 2
UE
Operator IMS core
3rd party IMS+IPTV
IPTV SPs

Fig. 2. Procedures overview

user satisfaction with respect to service contents, security etc. To answer the mentioned questions, we come to the conclusion that there is a need to model the user satisfaction function that can capture the quantitatively capture user satisfaction. In this connection, we use our previous work on user QoE [10].

VI. User-centric IMS based IPTV architecture

We assume that operators are integrated through a neutral and trusted third party (as shown in Figure-1) which is responsible for Accounting, Authorization, Authentication etc. In the proposed architecture each user maintains a separate business relationship with SP he is subscribed to, thus complete user profile should be maintained by HSS of trusted third party and each SP should receive only SP specific user data from HSS. It should further be noted that our approach addresses the fully disintegrated scenario where the SP or content providers are owned by the network infrastructure providers. However, the proposed analytical framework is rich enough to be applicable to any third party entity. The function components of third party entity include the standardized IPTV components such as Service Discovery Function (SDF), Service Control Function (SCF), Service Selection Function (SSF), User Profile Server Function (UPSF), and Decision Maker (DM).

Decision Maker (DM) entity on third party collects IPTV service data such as Video on Demand (VoD) catalogue, Electronic Program Guide (EPG), and TV channels (from SSF and SDF components) as offers from the SP. DM makes the decision over the best available SP, the decision related information is then delivered to users in the service attachment and selection process of UE. Owing to the fact that users do not have a contract(s) with IPTV service providers, SCF has to bridge to IPTV service providers by acting as Back-to-Back User Agent (B2BUA) in case of session establishment. On the other hand IPTV SPs also consists of standardized IPTV components, their SDFs and SSFs are requested for service data by third party functional entity. As the third party requires service related information (i.e., service costs and provided QoS) for service selection decision making. Such process of acquiring the mentioned information can be realized by extending the common standardized service descriptions. In the proposed architecture, IPTV SPs implement the standardized functionalities for authentication, service authorization, and session establishment.

Generally the standardized IPTV procedures over UE start up and service consumption consists of following steps [14]: i) Network attachment, ii) IMS registration, iii) IPTV service attachment, iv) IPTV service selection, and v) IPTV session establishment. We now briefly discuss the important interaction details.

A. IPTV Service Discovery: After the UE is switched ON, the network attachment and IMS registration procedures trigger and UE attaches to the network of the visited operator network. Since UE is not a subscriber of the operator, therefore the procedures of roaming case are applied. The IMS registration from the P-CSCF of the operator network are forwarded from the operator IMS network to the third party IMS. After being successfully authenticated in the third party IMS, the UE is available for communication over a secured connection.

B. IPTV Service discovery: The service discovery process starts after UE's successful IMS registration. This process is divided in service attachment and selection. The consequence of this process is the service data allocation from different IPTV SPs and triggering decision making process. Given our basic assumption of existence of reliable third party, it is realistic to assume that the trusted third party owns a subscription with each SP and UE is successfully registered and authenticated with each SP prior to the following steps (however UE by default is authenticated to the default infrastructure provider): After the trusted third party SDF retrieves service discovery request from the UE it requests the DM for service data that is to be deliver to the user. The DM then carry out the SP selection decision for the UE generate service request, DM evaluates the SPs' offered data from SDF and SSF functionalities of the IPTV SPs (which are extended by auction bidding like capabilities to provide enriched service information). In fact the decision making process starts after getting service data from all IPTV SPs. As the result all service data of the winning IPTV service provider(s) is aggregated in service description documents provided by the trusted third party SSF(s) to the UE in the service discovery response. This dictates that SSF can make use of an additional filter process that matches the user preferences and UE capabilities. As depicted in Figure-2 IPTV SPs can send updated service data offers within the subscriptions from the trusted third party. New offers may trigger a new DM process and result in updated service data notification to the UE, thus enabling UEs to successively select and consume the media of different service providers following the proposed service selection approach.

C. IPTV session establishment: The signalling for media session initiation to the SP network pass through IMS networks as well as the SCF of the trusted third party network. This follows that fact that only the trusted third party is a registered user of the service provider. Therefore, the SCF acts as a B2BUA and intercepts a user initiated session, creates a new session to the IPTV service providers and stores an association between both for a later intermediation of session signalling (session tear-down).

Note: Figure 2 depicts the details of signalling for overall procedures.

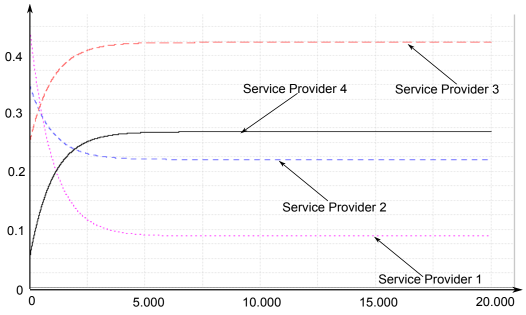

Fig. 3. Strategy convergence curves of IPTV consumer in fair service settings

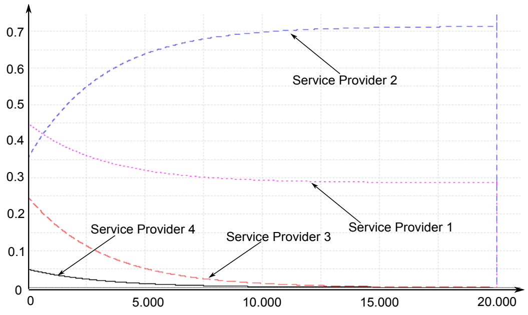

Fig. 4. Strategy convergence curves of IPTV consumer in excellent service settings

VII. Numerical Analysis

In order to demonstrate the user-centric IPTV service selection, and demonstrate the effect of learning in such a telecommunication landscape, we run extensive rounds of simulation runs. The simulation scenario dictates that IPTV consumers are under the coverage of heterogeneous technologies owned by different infrastructure providers. We consider the Long Term Evolution (LTE) and WLAN access network technologies. We extensively implement the integration of these two technologies following 3GPP standards for intra-operator heterogeneous technologies integration. Intra-operator mobility management is carried out using Mobile IPv6. Furthermore in total there are four IPTV service providers, who are considered as potential candidate service providers (competitors) to extend IPTV services to the consumers. Service requests of different quality classes, content types are generated by consumers. The arrival of requests is modeled by Poisson process, and the service quality class is chosen randomly. In order to capture the different consumer preferences we assume that the sizes of different quality class requests are assumed to be static and are 200kbps, 500kbps, and 800kbps for low, medium, and high video quality respectively. The capacities of LTE and WLAN network technologies are 32Mbps (Downlink)/ 8Mbps (Uplink), 8Mbps respectively. As the network technologies are owned by two different operators, the technical configuration of the technologies owned by both the operators are very similar. However the service pricing scheme is operator specific, which influences the user-centric service selection decision.

Within the simulation settings, we configure that all the users in the system have the same initial probability list i.e., 0.45,0.35,0.25,0.05 for SP-1, SP-2, SP-3 and SP-4 respectively. We also configure that SP-1 and SP-2 offer higher service costs when compared with the SP-3 and SP-4. To capture the system behavior for users preferences over the service costs, we further configure two simulation settings namely i) excellent service settings and ii) fair service settings. In the early settings a IPTV users prefer quality over the service costs, whereas the later case is converse to the earlier. Figure-4 depicts the results of user strategy convergence in excellent service settings. As can be seen that user initial probability converges such that SP-1 and SP-2 are assigned the equilibrium probabilities of 0.3 and 0.7 respectively, whereas the low quality offering service providers are assigned $zero$ probabilities. On the other hand the initial strategies of user in Figure-3 converge such that user prefers SP-3 and SP-4 more as compared to the other relatively more expensive service providers. However it should be noted that the decision of service provider selection is based on the user satisfaction function and not only on the cost preferences of the users.

The configuration of the technical indices are the same for all the underlying technologies, thus the operators offer of technical parameters are influenced by the congestion, available bandwidth, wireless medium characteristics etc. The simulation was run for number of iterations and the convergence of user probabilities of network selection was observed for variable learning schemes. Thus on the basis of the presented results we can confidently claim that the proposed learning scheme fits well to the future user-centric IPTV service selection paradigm.

VIII. Conclusion

In this paper, we have presented hybrid and heterogeneous strategic learning schemes in IPTV service selection in heterogeneous 4G networks. We have illustrated how important these learning schemes are in dynamic service selection scenario, where the measurement can be imperfect, noisy and delayed and the environment random and changing. We proposed user satisfaction function. Our contributions are validated through extensively simulating the realistic scenario using Mathematica numerical examples and OPNET simulations. We considered and simulated the LTE, and WLAN technologies taking into consideration the effect of switching costs in the payoff function. We illustrated the proposed cost of learning CODIPAS-RL scheme to find the corresponding solution in an iterative fashion.

References

[1] Markos P. Anastasopoulos, Dionysia K. Petraki, Rajgopal Kannan, and Athanasios V. Vasilakos. Tcp throughput adaptation in wimax networks using replicator dynamics. *IEEE TSMC partB*, June 2010.

[2] V. S. Borkar. Stochastic approximation: a dynamical systems viewpoint. 2008.

[3] R. Bush and F. Mosteller. Stochastic models of learning. *Wiley Sons, New York.*, 1955.

[4] Shah Devavrat and Shin Jinwoo. Dynamics in congestion games. *Sigmetrics*, 2010.

[5] D. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

[6] F. Fu and M. van der Schaar. Learning to compete for resources in wireless stochastic games. *IEEE Trans. Veh. Tech.*, 58(4):1904–1919, May 2009.

[7] J. Guckenheimer and P. Holmes. Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. *Springer-Verlag, New York*, 1983.

[8] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[9] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *Amer. Econ. Rev.*, 93, 2003.

[10] Manzoor Ahmed Khan and Umar Toseef. User utility function as quality of experience (qoe). In *Proceedings of the ICN'11*, pages 99–104, 2011.

[11] J.M. Lasry and P.L. Lions. Mean field games. *Japan. J. Math.*, 2:229–260, 2007.

[12] H. Tembine. Distributed strategic learning for wireless engineers. *Lecture notes, Supelec, 200 pages*, 2010.

[13] M.A.L. Thathachar, P.S. Sastry, and V.V. Phansalkar. Decentralized learning of nash equilibria in multiperson stochastic games with incomplete information. *IEEE transactions on system, man, and cybernetics*, 24(5), 1994.

[14] ETSI TISPAN. ETSI TS 182 027 V2.4.1 (2009-07); Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IPTV Architecture; IPTV functions supported by the IMS subsystem, 2009.

[15] Y. Xing and R. Chandramouli. Stochastic learning solution for distributed discrete power control game in wireless data networks. *IEEE/ACM Transactions on Networking*, 16(4):932–944, August 2008.

[16] H. P. Young. Strategic learning and its limits. *Oxford University Press*, 2004.

[17] H. P. Young. Learning by trial and error. *Games and Economic Behavior, Elsevier*, 65:626–643, March 2009.