Learning to precode in outage minimization games over MIMO interference channels

Elena Veronica Belmega, Hamidou Tembine, Samson Lasaulce

Learning to precode in outage minimization games over MIMO interference channels

Elena Veronica Belmega
Signals and Systems Laboratory
SUPÉLEC
Gif-sur-Yvette
France
Email: belmega@lss.supelec.fr

Hamidou Tembine
Department of Telecommunications
SUPÉLEC
Gif-sur-Yvette
France
Email: tembine@ieee.org

Samson Lasaulce
Signals and Systems Laboratory
SUPÉLEC
Gif-sur-Yvette
France
Email: lasaulce@lss.supelec.fr

***Abstract*—In this paper, we consider a network composed of several interfering transmitter-receiver pairs where all the terminals are equipped with multiple antennas. The problem of finding the precoding matrices minimizing the outage probabilities is analyzed using a game theoretical framework under the assumption of slow fading links and non-cooperative transmissions. An analytical solution of this game is very difficult to be found in general. Even in the most simple case of single-user, the problem remains an open issue. However, the existence of a pure-strategy Nash equilibrium solution is proven in the extreme SNR regimes. Furthermore, we exploit a simple reinforcement algorithm and show that, based only on the knowledge of one ACK/NACK bit, the users may converge to an Nash equilibrium solution of the game under investigation.**

I. Introduction

Game theory appears to be the unifying tool for studying resource allocations problems in interference channels. The wireless environment and the mutual interference between the simultaneous transmissions gives rise to the competition for common resources. This competition leads to strategic interaction amongst the users which is modeled as a non-cooperative game. The non-cooperative resource allocation game in multiple-input multiple-output (MIMO) interference channel (IC) has been extensively studied in the literature. The players, the transmitter-receiver pairs, are assumed to choose their best precoding matrices to maximize their Shannon achievable rates. In [1], [2], [3], [4], the particular case of parallel IC and, recently, in [5], the general MIMO IC was studied. In [5], the authors give sufficient conditions that ensure both the uniqueness of the NE and convergence of asynchronous iterative water-filling algorithms. In the vast majority of the papers treating the IC, the static channel model is assumed, i.e. the channel gains are deterministic and static over the whole transmission duration.

In this paper, we study a similar power allocation game in the MIMO IC. The players are the transmitter-receiver pairs. The main difference with the aforementioned works consists in the statistics of the channels. Here, we assume the channel gains to be slow fading, i.e. the realizations of random variables, known only at the receivers and static over the transmission duration. This difference implies important changes in the structure of the game under study. First of all, the Shannon achievable rates are no longer suited to measure the performance of the transmissions (they are strictly equal to zero). Therefore we assume that the users chose their best precoding matrices to minimize their individual outage probabilities [6]. The problem is very difficult in general. The main reason is that even in the single-user MIMO slow fading case and assuming i.i.d. standard Gaussian entries of the channel matrix gain the problem of finding the optimal transmit precoding matrix is an open issue. The result was conjectured by Telatar in [7] and was solved in some particular cases: i) multiple-input single-output (MISO) channel in [8]; ii) two-input single-output (TISO) channel in [9]; iii) MIMO channel assuming the high and low SNR regimes in [8]. Telatar's conjecture states that the optimal precoding matrix consists in uniformly spread all the available power over a subset of antennas. The number of active antennas depends on the system parameters (i.e., target rate, noise variance, available transmit power).

Second, motivated by this conjecture, we study the discrete game where the set of possible covariance matrices is reduced to the set of uniformly spreading the power over a subset of antennas. Because of the difficulties encountered when trying to find ordering relations between the users' payoffs, the existence of a pure-strategy Nash equilibrium stable solution [10] in the general case will be illustrated via numerical simulations alone. However, we exploit the results in [8] and prove mathematically the existence of at least one pure-strategy NE in the high or low SNR regimes at the receivers. The most important contribution of this paper is the study of a simple reinforcement learning technique, similar to [11], that allows the users to converges to the pure-strategy NE of the discrete game. Based only on the knowledge of their own action spaces and a single ACK/NACK bit at each iteration, the users apply simple updating rules completely ignorant of the structure of the game (i.e., other players, other players' actions and payoffs, their own payoff functions). Provided a pure-strategy NE exists, the algorithm converges to this optimal solution minimizing the individual outage probabilities of the users.

This paper is structured as follows. In Sec. II, we introduce the model and basic assumptions. The non-cooperative power

allocation game where the users maximize their own success probabilities is defined in Sec. III and the existence of the Nash equilibrium solution is proven in the extreme SNR regimes (see Subsec. III-A). In Sec. IV, we propose a simple reinforcement learning algorithm to converge to the Nash equilibrium in a distributed manner, having only the knowledge of one ACK/NACK bit. The single-user case is analyzed thoroughly in Subsec. IV-B. We illustrate the convergence results and the trade-off between the convergence time and the convergence to the optimum via numerical simulations in Sec. V. We conclude with several remarks and open issues.

II. System Model

We consider an interference channel (IC) composed of K transmitter-receiver pairs. The transmitters are assumed to send their private messages to the intended receivers. Transmitter $k \in \mathcal{K} \triangleq \{1, \ldots, K\}$ is equipped with $n_{t,k}$ antennas whereas the receiver k has $n_{r,k}$ antennas. The slow fading channel model is investigated where only the receivers are assumed to have perfect channel state information. The equivalent baseband signals write as:

$$\underline{Y}_k \quad = \quad \sum_{\ell=1}^{K} \mathbf{H}_{\ell k} \underline{X}_k + \underline{Z}_k,$$

where, for the sake of simplicity, the time index was ignored. The vector $\underline{X}_k$ represents the $n_{t,k}$-dimensional column vector of symbols transmitted by user k, $\mathbf{H}_{\ell k} \in \mathbb{C}^{n_{r,k} \times n_{t,\ell}}$ is the channel matrix (stationary and ergodic process) between transmitter ℓ and the receiver k and $\underline{Z}_k$ is the $n_{r,k}$-dimensional complex white Gaussian noise distributed as $\mathcal{N}(\underline{0}, \sigma_k^2 \mathbf{I}_{n_{r,k}})$, for all $k, \ell \in \mathcal{K}$. The channel matrices $\mathbf{H}_{\ell k}$ are assumed to contain i.i.d. standard complex Gaussian random entries.

In this context, the mutual information is a random variable, varying from block to block, and thus it is not possible (in general) to guarantee that it is always above a certain threshold. In this case, the achievable transmit rate in the sense of Shannon is zero. A suited performance metric is the probability of an outage for a fixed transmission rate [6]. This metric allows one to quantify the probability that the rate target is not reached by using a good channel coding scheme and is defined as:

$$\mathrm{P}_{\mathrm{out},k}(\mathbf{Q}_k, \mathbf{Q}_{-k}, R_k) = \Pr\left[\mu_k(\mathbf{Q}_k, \mathbf{Q}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-k,k}) < R_k\right],$$

where V_{-k} denotes the super-vector $(V_1, \ldots, V_{k-1}, V_{k+1}, \ldots, V_K)$ for any quantity V and μ_k denotes the instantaneous mutual information. The matrix $\mathbf{Q}_k = \mathbb{E}[\underline{X}_k \underline{X}_k^H]$ denotes the input precoding matrix of user k in the convex and compact set of positive definite matrices:

$$\mathcal{A}_k = \{\mathbf{Q} \in \mathbb{C}^{n_{t,k} \times n_{t,k}} : \mathbf{Q} \succeq \mathbf{0}, \mathrm{Tr}(\mathbf{Q}) \leq \overline{P}_k\}. \qquad (1)$$

Assuming that the interference is treated as noise at the receiver level, the instantaneous mutual information of user k writes as:

$$\begin{aligned} \mu_k(\mathbf{Q}_k, \mathbf{Q}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-kk}) \quad &= \quad \theta(\mathbf{Q}_k, \mathbf{Q}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-kk}) - \\ &\quad \eta_k(\mathbf{Q}_{-k}, \mathbf{H}_{-kk}), \end{aligned} \qquad (2)$$

$$\begin{aligned} \theta(\mathbf{Q}_k, \mathbf{Q}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-kk}) \quad &= \quad \log_2 \left| \mathbf{I}_{n_{r,k}} + \rho_k \sum_{\ell=1}^{K} \mathbf{H}_{\ell k} \mathbf{Q}_\ell \mathbf{H}_{\ell k}^H \right| \\ \eta_k(\mathbf{Q}_{-k}, \mathbf{H}_{-kk}) \quad &= \quad \log_2 \left| \mathbf{I}_{n_{r,k}} + \rho_k \sum_{\ell \neq k} \mathbf{H}_{\ell k} \mathbf{Q}_\ell \mathbf{H}_{\ell k}^H \right| \end{aligned} \qquad (3)$$

where $\rho_k = \frac{1}{\sigma_k^2}$.

At this point, an important observation has to be made. Having assumed that channels are i.i.d. complex Gaussian, the search for the optimal precoding matrices in $\mathcal{A}_k$ is reduced to its subset of diagonal matrices. The proof is based on Lemma 5 in [7] stating that the distribution of the channel matrix does not change when multiplied to the right and/or left by unitary matrices. Therefore, the search for the optimal precoding matrices is reduced to solving the power allocation problem over the available eigen-modes.

III. Non-cooperative power allocation game

In this section, we describe the non-cooperative power allocation game defined by the triplet $\mathcal{G} = \left(\mathcal{K}, \{\mathcal{D}_k\}_{k \in \mathcal{K}}, \{u_k\}_{k \in \mathcal{K}}\right)$. The game components are: i) *the players (in the set $\mathcal{K}$)*: the transmitter-receiver pairs assumed to be autonomous non-cooperative; ii) *the players' strategies* consisting of their power allocation policies $\underline{d}_k \in \mathcal{D}_k$; iii) *the players' payoff functions*: the success probabilities $u_k(\underline{d}_k, \underline{d}_{-k}) = 1 - \mathrm{P}_{\mathrm{out},k}(\mathbf{D}_k), \mathbf{D}_{-k}, R_k)$[1] Notice that the optimal precoding matrix and the optimal success probability of each user will depend implicitly on both target rates R_1 and R_2. In this paper, we assume that the action set of user k is a simple discrete version of $\mathcal{A}_k$:

$$\mathcal{D}_k = \left\{ \frac{\overline{P}_k}{\ell} \underline{e}_\ell \;\middle|\; \ell \in \{1, \ldots, n_t\}, \underline{e}_\ell \in \{0,1\}^{n_t}, \sum_{i=1}^{n_t} e_\ell(i) = \ell \right\}. \qquad (4)$$

$\mathcal{D}_k$ represents the set of power allocation vectors that consists in allocating uniform power over only a subset of ℓ eigen-modes. The choice of these sets is motivated by several reasons:

- As argued in the previous section, the search for the optimal precoding matrices in $\mathcal{A}_k$ is reduced to its subset of diagonal matrices.
- It can be proven that, saturating the available power, i.e. $\mathrm{Tr}(\mathbf{Q}_k) = \overline{P}_k$, is the dominant strategy for any user k.
- For the single-user case, Telatar [7] conjectured that the optimal covariance matrix is to uniformly allocate the power on a subset of antennas.

Let us index the elements of $\mathcal{D}_k$, i.e., $\mathcal{D}_K = \{\underline{d}_k^{(1)}, \ldots, \underline{d}_k^{(m_k)}\}$ with $m_k = \mathrm{Card}(\mathcal{D}_k)$ (i.e., the cardinal of $\mathcal{D}_k$). We denote by $\Delta(\mathcal{D}_k)$ the set of mixed-actions (i.e., discrete probability measures over $\mathcal{D}_k$) of user k. Thus, $\underline{p}_k \in \Delta(\mathcal{D}_k)$ denotes a mixed-strategy for user k and p_{kj_k} represents the probability of choosing the allocation vector $\underline{d}_k^{(j_k)}$.

[1] We will use the notation $\mathbf{D}_k \triangleq \mathbf{diag}(\underline{d}_k)$ throughout the rest of the paper.

A natural solution concept in non-cooperative games is the Nash equilibrium (i.e. a strategy profile from which no user can gain by unilateral deviation, see [12] [13] for a detailed discussion). We know from [10] that at least one mixed-strategy Nash equilibrium exists in any discrete finite game. However, the existence of a pure-strategy Nash equilibrium is not always guaranteed and it depends on the values of the payoffs and the ordering relations between them. Establishing these relations in general is a very difficult problem since closed-form expressions of the outage probability are not yet available. Notice that, for the particular case where the transmitters are equipped with single antennas (i.e., $n_{t,k} = 1$), the problem is trivial since the action sets reduce to singletons. This means that every user is allowed to transmit on the only antenna available. In what the MISO case is concerned, i.e. $n_{r,k} = 1$, the solution is far from being trivial and is left as an useful extension of this paper. The idea is to exploit the exact solution given for the single-user case in [8].

A. Extreme SNR regimes

In what follows, we will investigate the extreme SNR particular cases, i.e., $\rho_k \to 0$ or $\rho_k \to +\infty$ and prove that in these cases there is at least one NE.

Theorem 1: If, for all $k \in \mathcal{K}$, we have either $\rho_k \to 0$ or $\rho_k \to +\infty$, then the game $\mathcal{G}$ has at least a pure-strategy Nash equilibrium.

In the low SNR regime, $\rho_k \to 0$, we prove in Appendix A that, regardless of the strategy of the other user, beam-forming (BF) is the optimal strategy for user k, i.e., $\underline{d}_k^{\mathrm{BF}} \in \left\{\overline{P}_k \underline{e}_1\right\}$ is a dominating strategy for user k. On the other hand, in the high SNR regime, a dominant strategy for user k is the uniform power allocation policy (UPA) over all the antennas $\underline{d}_k^{\mathrm{UPA}} = \frac{\overline{P}_k}{n_{t,k}} \underline{e}_{n_{t,k}}$. For exemple, if $K = 2$, we have four different situations: i) $\rho_1 \to 0$ and $\rho_2 \to 0$, then $(\underline{d}_1^{\mathrm{BF}}, \underline{d}_2^{\mathrm{BF}})$ is NE; ii) $\rho_1 \to +\infty$ and $\rho_2 \to 0$, then $(\underline{d}_1^{\mathrm{UPA}}, \underline{d}_2^{\mathrm{BF}})$ is NE; iii) $\rho_1 \to 0$ and $\rho_2 \to +\infty$, then $(\underline{d}_1^{\mathrm{BF}}, \underline{d}_2^{\mathrm{UPA}})$ is NE; iv) $\rho_1 \to +\infty$ and $\rho_2 \to +\infty$, then $(\underline{d}_1^{\mathrm{UPA}}, \underline{d}_2^{\mathrm{UPA}})$ is a NE.

IV. Learning algorithms in games

In this section, we discuss a class of iterative algorithms that converge to a certain desirable state (e.g., the equilibrium points of the power allocation game described previously or a certain global optimum). The users are not assumed to be rational devices but simple automata that know only their own action sets. They start at a completely naive state choosing randomly their action (e.g., following the uniform distribution over their own action sets for example). After the play, each user obtains a certain feedback from the nature (e.g., the realization of a random variable, the value of its own instantaneous payoff).

We assume that the only feedback that user $k \in \mathcal{K}$ receives is an ACK/NACK signal. It receives the realization of the following random variable $S_k = 0$ if $\mu_k(\mathbf{D}_k, \mathbf{D}_{-k}, \mathbf{H}_{1k}, \mathbf{H}_{2k}) \leq R_k$ otherwise $S_k = 1$. If an outage has occurred at time t the receiver feedbacks $s_k^{[t]} = 0$ to the transmitter, otherwise it sends $s_k^{[t]} = 1$. Notice that the random variable S_k is a Bernoulli distributed with parameter $q_k = 1 - \mathrm{P}_{\mathrm{out},k}(\mathbf{D}_k, \mathbf{D}_{-k}, R_k)$ such that its expected value is equal to $1 - \mathrm{P}_{\mathrm{out},k}(\mathbf{D}_k, \mathbf{D}_{-k}, R_k)$. Thus if the instantaneous payoff is $s_k^{[t]}$ then the expected payoff of user k is exactly the success probability $1 - \mathrm{P}_{\mathrm{out},k}(\mathbf{D}_k, \mathbf{D}_{-k}, R_k)$.

Based only on this value, $s_k^{[t]}$, each user applies a simple updating rule over its own probability distribution or mixed strategy. It turns out that in the long run, the updating rules converge to some desirable system states (i.e., the NE of the game $\mathcal{G}$). Note that the rationality assumption is no longer needed. The transmitters don't even need to know the structure of the game or even that they play a game. The price to pay will be reflected in slower convergence time.

A. A reinforcement learning algorithm

Here, we consider a stochastic learning algorithm similarly to [11]. At step $n > 0$ of the iterative process, User k randomly chooses a certain action $\underline{d}_k^{[n]} \in \mathcal{D}_k$ based on the probability distribution $\underline{p}_k^{[n-1]}$ from the previous iteration. As a consequence, it obtains the realization of a random variable, which is, in our case, $s_k^{[n]} = s_k\left(\underline{d}_k^{[n]}, \underline{d}_{-k}^{[n]}\right)$. Based on this value, Player k updates its own probability distribution as follows:

$$p_{k,j_k}^{[n]} = p_{k,j_k}^{[n-1]} - \gamma^{[n]} s_k^{[n]} p_{k,j_k}^{[n-1]} + \gamma^{[n]} s_k^{[n]} \mathbb{1}_{\left(\underline{d}_k^{[n]} = \underline{d}_k^{(j_k)}\right)}, \quad (5)$$

where $0 < \gamma^{[n]} < 1$ is the quantization or learning step and $p_{k,j_k}^{[n]}$ represents the probability that user k chooses action $\underline{d}_k^{(j_k)}$ at iteration n. We denote by $\underline{p}^{[n]}$ the super-vector containing the mixed strategies of all users.

Using the results from the stochastic approximation theory (see [14], Chapter 8 in [15], Chapter 2 in [16]), the sequence $\underline{p}^{[n]}$ can be approximated in the asymptotic regime ($n \to +\infty$) with the solution of the deterministic ordinary differential equation (ODE):

$$\frac{\mathrm{d}p_{kj_k}}{\mathrm{d}t} = p_{kj_k} \sum_{i_k=1}^{m_k} p_{ki_k} [h_{kj_k}(\underline{p}_{-k}) - h_{ki_k}(\underline{p}_{-k})], \quad (6)$$

where

$$h_{kj_k}(\underline{p}_{-k}) = \sum_{i_{-k}} u_k\left(\underline{d}_k^{(j_k)}, \underline{d}_{-k}^{(i_{-k})}\right) \prod_{\ell \neq k} p_{\ell i_\ell}$$

However, these convergence results are proven in a probabilistic manner: i) for constant step-size $\gamma^{[n]} = \gamma \to 0$ the convergence is proven in distribution (see Chapter 8 in [15]); ii) for diminishing step-size $\gamma^{[n]}$ verifying certain conditions (see Chapter 2 [16]), the convergence is proven almost surely.

This means that, in order to study the stochastic process $p_{k,j}^{[n]}$ in the asymptotic regime, we can focus on the study of the deterministic ODE that captures its average behavior. Notice that the ODE (6) is similarly to the replicator dynamics. The mixed and pure-strategy NE are rest points of this dynamics. However, all the pure-strategy profiles, even those which are not NE are also rest points.

Notice that this ODE is just an approximation that allows us to explain the asymptotic behaviour of the discrete process given in (5). One main difference is that only pure strategies can be stationary points of the discrete process, while this is not generally true in the continuous-time dynamics given in (6).

B. The single-user particular case

An interesting particular case that can be solved analytically and thus allowing us to gain insight on the general problem is the single-user case. The game is reduced to an optimization problem where, let's say, user 1 has to choose his best precoding matrix to maximize his success probability $u_1(\underline{d}_1) = 1 - P_{\text{out}}(\mathbf{D}_1, R_1)$. One nice property of the discrete finite ($\underline{d}_1 \in \mathcal{D}_1$) optimization problem is that there always exists a solution:

$$\mathcal{S}_P = \left\{ j \in \{1, \ldots, m_1\} \middle| j \in \arg \max_{i \in \{1,\ldots,m_1\}} u_1(\underline{d}_1^{(i)}) \right\} \quad (7)$$

the set of pure-strategy solutions and

$$\mathcal{S}_M = \left\{ \underline{p} \in \Delta(\mathcal{D}_1) \middle| \underline{p} = \sum_{i \in \mathcal{S}_p} \alpha_i \underline{e}_i, \ \forall j \in \mathcal{S}_p : \alpha_j \geq 0, \ \sum_{i \in \mathcal{S}_p} \alpha_i = 1 \right\} \quad (8)$$

the convex set of mixed-strategy solutions where $\underline{e}_i \in \{0,1\}^{m_1}$ corresponds to the canonical vector taking value one on the $i-$ th position.

The updating rule is identical to (5). The only difference consists in the random payoff which depends only on $\underline{d}_1$ $s_1^{[n]} = s_1\left(\underline{d}_1^{[n]}\right)$ which is equal to zero if an outage has occurred, i.e., $\log_2 \left| \mathbf{I}_{n_{r,1}} + \rho \mathbf{H}_{11} \mathbf{D}_1 \mathbf{H}_{11}^H \right| < R_1$ or equal to one otherwise.

The deterministic mean ODE in (6) becomes:

$$\frac{\mathrm{d}p_{1,j}}{\mathrm{d}t} = p_{1,j} \left[u_1(\underline{d}_1^{(j)}) - \sum_{i=1}^{m_1} p_i u_1(\underline{d}_1^{(i)}) \right], \quad (9)$$

for all $j \in \{1, \ldots, m_1\}$. In this particular case, the exact solution of the ODE can be found (see [17]) depending on the initial condition $\underline{p}_1(0) \in \Delta(\mathcal{D}_1)$ and is given by:

$$p_{1,j}(t) = \frac{p_{1,j}(0) \mathrm{e}^{t u_1(\underline{d}_1^{(j)})}}{\sum_{i=1}^{m_1} p_{1,i}(0) \mathrm{e}^{t u_1(\underline{d}_1^{(i)})}}, \quad (10)$$

for all $j \in \{1, \ldots, m_1\}$ and $t > 0$. Observe that, if $\underline{p}_1(0)$ is a degenerate probability distribution corresponding to a pure-strategy, then $\underline{p}_1(t) = \underline{p}_1(0)$ for all $t > 0$ (i.e. all pure-strategies are stationary points of the dynamics in (9)). Now, if $\underline{p}_1(0)$ lies in the relative interior of $\Delta(\mathcal{D}_1)$, we can find the convergence point of the trajectories of the ODE by taking the limit when $t \to +\infty$ in (10) and obtain:

$$\lim_{t \to +\infty} p_{1,j}(t) = \begin{cases} \frac{p_j(0)}{\sum_{i \in \mathcal{S}_p} p_i(0)} & \text{if} \quad j \in \mathcal{S}_p, \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

The solution is similar if the initial distribution lies on the border of the simplex $\Delta(\mathcal{D}_k)$ by taking into account the fact that the border is an invariant set of the ODE.

Notice that if the set $\mathcal{S}_p$ is a singleton, then the trajectories of the ODE convergent to this point. Otherwise, the trajectories of the continuous-time ODE convergent to one of the solutions in $\mathcal{S}_m$ depending on the initial point $\underline{p}_1(0)$. However, we will see in the numerical simulations that the discrete process converges to one of the pure-strategy solutions in $\mathcal{S}_p$. Notice that, the function

$$V(\underline{p}_1) = u_{\max} - \sum_{i=1}^{m_1} p_{1,i} u_1(\underline{d}_1^{(i)}), \quad (12)$$

where $u_{\max} = \max_j u_1(\underline{d}_1^{(j)})$ is a Lyapunov function for all the distributions in $\mathcal{S}_m$ and thus they are stable points of the dynamics (9).

In conclusion, using the simple adaptive rule in (5) a transmitter is able to learn the optimal precoding matrix which minimizes the outage probability. This is an important result since optimizing the outage probability in the single-user scenario is still an open issue [18]. Furthermore, numerical methods based on Monte-Carlo simulations and exhaustive search are very expensive in terms of computational cost. We will see in Sec. V, that using learning algorithms that require only one bit of feedback, the optimal precoding matrix can be computed in a more efficient way.

V. Numerical simulations

Single-user particular case. Consider the scenario where $n_t = n_r = 2$, $R_1 = 1 bpcu$, $\overline{P}_1 = 1$ W, $\sigma_1^2 = 1$ W. In this case, the user can choose between beam-forming and the uniform power allocation. The success probability is given by $u_1(\underline{d}_1^{\text{BF}}) = 0.7359$, $u_1(\underline{d}_1^{\text{UPA}}) = 0.8841$. These values were calculated using 10^6 Monte-Carlo iterations. Because the channel matrix is i.i.d. Gaussian, the position of active antennas does not matter only the number of active modes has an influence on the success probability. The choice of the initial distribution is the uniform one.

Fixed learning step-size. In Fig. 1, we trace the expected payoff $\sum_{j=1}^{m_1} p_{1,j}^{[n]} u_1(\underline{d}_1^{(j)})$. Notice that, for $\gamma^{[n]} = \gamma = 0.01$ (constant step-size), the user converges to the optimal solution in 2554 iterations. However, the performance of the algorithm depends on the choice of the learning parameter. The larger γ, the smaller the convergence time. The problem when choosing large steps is that the algorithm may converge to a corner of the simplex which is not a maximizer of the success probability. In Tab. I, we summarize the results obtained after 1000 experiments in terms of average number of iterations and convergence to the maximum point. We observe that there is a trade-off between the convergence time and the convergence to the optimal point which can be controlled by tuning the learning step. *Variable learning step-size.* For the same scenario, consider the case where the step-size is variable: $\gamma^{[n]} = \frac{\alpha_1}{(n+\alpha_2)^{\alpha_3}}$ for $n \geq 1$ such that $0 < \alpha_1 \leq 1$, $\alpha_2 \geq 0$,

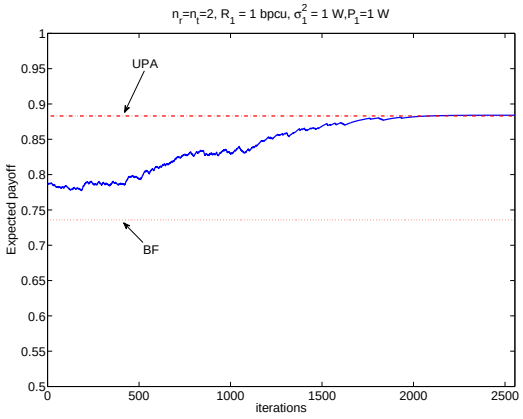


Fig. 1. Average payoff vs. number of iterations.

TABLE I
TRADE-OFF BETWEEN THE CONVERGENCE TIME AND THE CONVERGENCE TO THE OPTIMAL POINT (CONSTANT STEP-SIZE)

γ	Time [nb. iterations]	Convergence to optimum [%]
0.001	3755	100
0.1	261	71
0.5	27	45
0.9	9	39

$0.5 < \alpha_3 \leq 1$ which ensure the asymptotic convergence in probability (see condition (A2) in Chapter 2 [16]) of the discrete learning process to the solution of the mean ODE. It turns out that a careful choice of these parameters is needed to ensure good performances of the algorithm. For example, consider the case where $\gamma^{[1]} = 0$, $\gamma^{[n]} = \frac{1}{n}$ for $n > 1$ ($\alpha_1 = 1$, $\alpha_2 = 0$, $\alpha_3 = 1$). Assume that the initial distribution is uniform one and that the chosen strategy is w.l.o.g. $\underline{d}_1^{[1]} = \underline{d}_1^{(j)}$ with $j \in \{1, \ldots, 3\}$. If an outage does not occur at the first iteration, i.e., $s^{[1]} = 1$, then we see that $p_j^{[1]} = 1$ and $p_i^{[1]} = 0$ for all $i \neq j$ and the algorithm stops. Therefore, if an outage hasn't occurred at the first iteration, the first strategy chosen (which is any strategy in $\mathcal{D}_1$ with equal probability) will be the rest point of the algorithm. We see that there is only a 30% probability that the algorithm stops at the optimal point which is very different w.r.t. the theoretical analysis (telling us that almost surely, the algorithm converges to the optimal point). Now, let us consider $\alpha_1 = 1$, $\alpha_3 = 0.55$ and focus on the impact of parameter α_2. In Tab. II, we summarize the results obtained after 1000 experiments (the convergence time is longer for the variable step-size) in terms of average number of iterations and convergence to the maximum point. Here as well there is a trade-off between the convergence time and the convergence to the optimal point. Even though theoretically the variable step-size algorithm performs better in terms of convergence to the optimal point, simulations show that the performance of the algorithm depends on a very careful choice of the learning step. Furthermore, there is a trade-off between the convergence time and the convergence to the optimum.

Two-user case. Now we assume the $K = 2$ scenario where

TABLE II
TRADE-OFF BETWEEN THE CONVERGENCE TIME AND THE CONVERGENCE TO THE OPTIMAL POINT (VARIABLE STEP-SIZE)

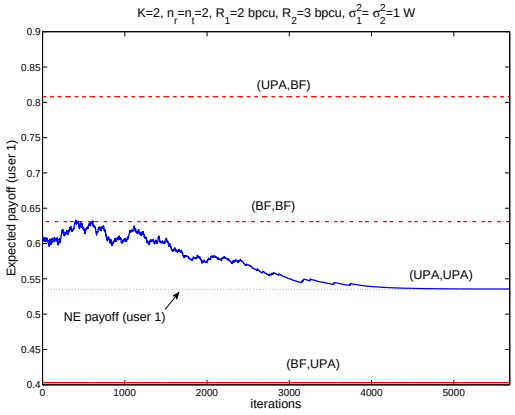
α_2	Time [nb. iterations]	Convergence to optimum [%]
1	34	43
10	435	71
100	1354	91
1000	2533	100

$n_r = n_t = 2$, $\sigma_1^2 = \sigma_2^2 = 1$ W, $\overline{P}_1 = \overline{P}_2 = 10$ W, the transmission rates $R_1 = 2$bpcu, $R_2 = 3$bpcu. The actions that the users can take are $\underline{d}_k^{(1)} = \overline{P}_k(0,1)$, $\underline{d}_k^{(2)} = \overline{P}_k(1,0)$, $\underline{d}_k^{(3)} = \frac{\overline{P}_k}{2}(1,1)$. Since the channels are i.i.d. Gaussian, the beam-forming strategies are identical in terms of payoff and the users can be considered as having two strategies: beam-forming (BF) (either $\underline{d}_k^{(1)}$ or $\underline{d}_k^{(2)}$) and uniform power allocation (UPA) ($\underline{d}_k^{(3)}$). The payoff matrix for user 1 is given by the success probability:

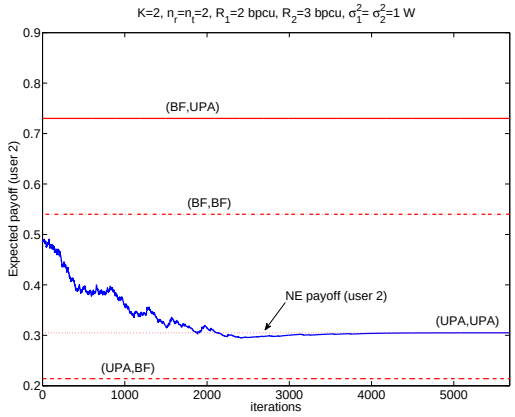
$\mathbf{U}_1 = \begin{pmatrix} 0.631 & 0.402 \\ 0.801 & 0.535 \end{pmatrix}$ $\mathbf{U}_2 = \begin{pmatrix} 0.540 & 0.731 \\ 0.214 & 0.305 \end{pmatrix}$ where $\mathbf{U}_k(1,1)$ corresponds to the case where both users apply BF, $\mathbf{U}_k(1,2)$ user 1 applies BF while the other one UPA, $\mathbf{U}_k(2,1)$ user 1 applies UPA while the other BF, $\mathbf{U}_k(2,2)$ both users apply UPA. We observe that the unique NE is given by the UPA for both users. Furthermore, we observe that the system optimal state w.r.t. the average of the success probabilities is the state where both players use BF and that the NE is the worse state w.r.t. this measure. We apply the reinforcement algorithm proposed in the previous section. In Fig. 2, we plot the expected payoff depending on the probability distribution over the action sets at every iteration for User 1 in Fig. 2(a) and for User 2 in Fig. 2(b) assuming $\overline{P}_1 = \overline{P}_2 = 5$ W. We observe that the users converge to the Nash equilibrium after approximatively 6000 iterations.

VI. CONCLUSION

The non-cooperative power allocation game in the slow-fading MIMO interference channels where the users wish to minimize their outage probabilities was studied. Analytical solutions to the general problem is very hard to be obtained. It turns out that a simple reinforcement algorithm may allow the transmitters to learn their best precoding matrix with respect to their individual outage probabilities. This algorithm has several appealing features. It is adaptive, of low complexity and requires only the knowledge of one bit of feedback from the environment and no rationality assumption. However, all these benefits come at the cost of long convergence time. Moreover, the algorithms are stochastic in nature and only asymptotic convergence in probability can be ensured. In practice, this translates the fact that a very careful choice of the learning step has to be made to ensure a good performance of the algorithms. We have seen that there is a trade-off between the probability (frequency) of convergence and the convergence time. Interesting extensions of this work could be: to prove the existence of the NE for the MISO case (exploiting the



(a) User 1.



(b) User 2.

Fig. 2. Expected payoff vs. iteration number for $K = 2$ users.

solution for the single-user case); to use reinforcement learning allowing the users to converge to other system optimal points than the Nash equilibrium.

APPENDIX A
EXTREME SNR REGIMES

We will exploit the results available for the single-user MIMO channel in [8]: a) in the low SNR regime, the outage probability is a Schur-concave function w.r.t. the power allocation vector and BF is the optimal power allocation policy; b) in the high SNR regime, the outage probability is a Schur-convex function w.r.t. the power allocation vector and UPA is the optimal power allocation policy.

Let us prove that, when $\rho_k \to 0$ then $u_k(\underline{d}_k, \underline{d}_{-k})$ is Schur-convex w.r.t. $\underline{d}_k$.

The proof follows from the following steps:

- We assume that $\underline{d}_\ell \in \mathcal{C}_\ell \triangleq \left\{ \underline{v} \in \mathbb{R}_+^{n_{t,k}} \left| \sum_{i=1}^{n_{t,k}} v(i) = \overline{P}_k \right. \right\}$ for all $\ell \in \mathcal{K}$.
- Assuming that $\rho_k \to 0$ then we prove that $\Pr[\theta(\mathbf{D}_k, \mathbf{D}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-kk}) < \tilde{R}]$ is Schur-concave w.r.t. $(\underline{d}_k, \underline{d}_{-k}) \in \prod_{\ell=1}^{K} \mathcal{C}_\ell$. Indeed, by denoting $\widetilde{\mathbf{D}} = \mathbf{diag}(\underline{d}_k, \underline{d}_{-k})$ and $\widetilde{\mathbf{H}} = [\mathbf{H}_{kk} \mathbf{H}_{-\mathbf{k},\mathbf{k}}]$, then the results for the single-user MIMO channel in [8] apply directly.
- It is easy to prove that, for arbitrary $\underline{d}_{-k} \in \prod_{\ell \neq k} \mathcal{C}_\ell$, the function $\Pr[\theta(\mathbf{D}_k, \mathbf{D}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-k,k}) < \tilde{R}]$ is Schur-concave w.r.t. $\underline{d}_k$.
- Since the previous result holds for any rate $\tilde{R} > 0$, by choosing $\tilde{R} = R_k + \eta_k(\mathbf{D}_{-k}, \mathbf{H}_{-k,k})$, we obtain that $\Pr[\theta(\mathbf{D}_k, \mathbf{D}_{-k}, \mathbf{H}_{kk}, \mathbf{H}_{-k,k}) - \eta_k(\mathbf{D}_{-k}, \mathbf{H}_{-k,k}) < R_k]$ is Schur-concave w.r.t. $\underline{d}_k$.
- This implies that $u_k(\underline{d}_k, \underline{d}_{-k})$ is Schur-convex w.r.t. $\underline{d}_k \in \mathcal{C}_k$ for any $\underline{d}_{-k} \in \mathcal{C}_{-k}$ and that beam-forming is an optimal strategy.
- Since $\underline{d}_k^{\mathrm{BF}} \in \mathcal{D}_k$, then it follows that, for any $\underline{d}_{-k}$, it is an optimal strategy for user k.

In the high SNR regime, when $\rho_k \to +\infty$, we have that $u_k(\underline{d}_k, \underline{d}_{-k})$ is Schur-concave w.r.t. $\underline{d}_k$ for all $\underline{d}_{-k}$. The proof follows similarly and will be omitted.

REFERENCES

[1] W. Yu, G. Ginis, and J. M. Cioffi, "Distributed multiuser power control for digital subscriber lines," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 1105–1115, Jun. 2002.

[2] S. T. Chung, S. J. Kim, J. Lee, and J. M. Cioffi, "A game theoretic approach to power allocation in frequency-selective gaussian interference channels," in *Proc. IEEE Intl. Symposium on Information Theory (ISIT)*, Pacifico Yokohama, Kanagawa, Japan, Jun./Jul. 2003, pp. 316–316.

[3] G. Scutari, D. P. Palomar, and S. Barbarossa, "Optimal linear precoding strategies for wideband non-cooperative systems based on game theory-Part I: Nash equilibria," *IEEE Trans. Signal Process.*, vol. 56, pp. 1230–1249, Mar. 2008.

[4] ——, "Competitive design of multiuser MIMO systems based on game theory: A unified view," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 1089–1103, Aug. 2008.

[5] ——, "The MIMO iterative waterfilling algorithm," *IEEE Trans. Signal Process.*, vol. 57, pp. 1917–1935, May 2009.

[6] L. H. Ozarow, S. S. (Shitz), and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 10, pp. 359–378, May 1994.

[7] E. Telatar, "Capacity of multi-antenna gaussian channels," *AT&T Bell Labs, Technical Report*, 1995.

[8] E. A. Jorswieck and H. Boche, "Outage probability in multiple antenna systems," *European Transactions on Telecommunications*, vol. 18, pp. 217–233, 2006.

[9] M. Katz and S. Shamai, "On the outage probability of a multiple-input single-output communication link," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 4120–4128, Nov. 2007.

[10] J. F. Nash, "Equilibrium points in n-points games," *Proc. of the Nat. Academy of Science*, vol. 36, no. 1, pp. 48–49, Jan. 1950.

[11] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thatchar, "Decentralized learning of nash equilibria in multi-person stochastic games with incomplete information," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, pp. 769–777, May 1994.

[12] D. Fudenberg and J. Tirole, "Game theory," *The MIT Press*, 1991.

[13] M. J. Osborne, *An introduction to game theory*. Oxford University Press, 2003.

[14] M. Benaïm, "Dynamics of stochastic approximation algorithms," *Séminaire de probabilités (Strasbourg)*, vol. 3, pp. 1–68, 1999.

[15] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*. Springer-Verlag New York, 1997.

[16] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Hindustan Book Agency (Cambridge University Press), 2008.

[17] J. Hofbauer and K. Sigmund, "Evolutionary game dynamics," *Bulletin of the American Mathematical Society*, vol. 40, pp. 479–519, Jul. 2003.

[18] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Europ. Trans. Telecommunications, ETT*, vol. 10, no. 6, pp. 585–596, Nov. 1999.