Adaptive Spectrum Management in MIMO-OFDM Cognitive Radio: An Exponential Learning Approach*

Panayotis Mertikopoulos
French National Center for Scientific Research (CNRS) and Laboratoire d'Informatique de Grenoble (LIG), Grenoble, France
panayotis.mertikopoulos@imag.fr

E. Veronica Belmega
ETIS/ENSEA – Université de Cergy-Pontoise – CNRS
Cergy-Pontoise, France
belmega@ensea.fr

ABSTRACT

In this paper, we examine cognitive radio systems that evolve dynamically over time as a function of changing user and environmental conditions. To take into account the advantages of orthogonal frequency division multiplexing (OFDM) and recent advances in multiple antenna (MIMO) technologies, we consider a full MIMO-OFDM Gaussian cognitive radio system where users with several antennas communicate over multiple non-interfering frequency bands. In this dynamic context, the objective of the network's secondary users (SUs) is to stay as close as possible to their optimum power allocation and signal covariance profile as it evolves over time, with only local channel state information at their disposal. To that end, we derive an adaptive spectrum management policy based on the method of matrix exponential learning, and we show that it leads to *no regret* (i.e. it performs asymptotically as well as any fixed signal distribution, no matter how the system evolves over time). As it turns out, this online learning policy is closely aligned to the direction of change of the users' data rate function, so the system's SUs are able to track their individual optimum signal profile even under rapidly changing conditions.

1. INTRODUCTION

As a result of the explosive spread of Internet-enabled mobile devices, the radio spectrum has become a scarce resource which, if not properly managed, will be unable to accommodate the soaring demand for wireless broadband and the ever-growing volume of data traffic and cellphone calls. Exacerbating this issue, studies by the US Federal Communications Commission (FCC) and the National Telecommunications and Information Administration (NTIA) have shown that this vital commodity is effectively squandered through underutilization and inefficient use: only 15% to 85% of the licensed radio spectrum is used on average, leaving ample spectral voids that could be exploited for opportunistic radio access [8, 25].

In view of the above, the emerging paradigm of cognitive radio (CR) has attracted considerable interest as a promising counter to spectrum scarcity [11, 12, 20, 34]. At its core, this paradigm is simply a two-level hierarchy between communicating users induced by spectrum licensing: on the one hand stand the network's primary users (PUs) who have purchased spectrum rights but allow others to access it (provided that the resulting interference remains below a certain threshold); on the other hand, the network's secondary users (SUs) are free-riding on the licensed part of the spectrum, and they try to communicate under the constraints imposed by the PUs (but with no quality of service (QoS) guarantees). In this way, by opening up the unfilled "white spaces" of the licensed spectrum to opportunistic access, the overall spectrum utilization can be greatly increased without compromising the performance guarantees that the network's licensed users have already paid for.

Orthogonally to the above, the seminal prediction that the use of multiple-input and multiple-output (MIMO) technologies can lead to substantial gains in information throughput [9, 31] opens up another way for overcoming spectrum scarcity. In particular, by employing multiple antennas for communication, it is possible to exploit spatial degrees of freedom in the transmission and reception of radio signals, the only physical limit being the number of antennas that can be deployed on a portable device. As a result, the existing wireless medium can accommodate greater volumes of data traffic without requiring the reallocation (and subsequent re-regulation) of additional frequency bands.

In this paper, we combine these two approaches and focus on a dynamic MIMO cognitive radio system comprised of several wireless users (primary and secondary alike) who communicate over multiple non-interfering channels, and are going online or offline based on their individual needs. In this unregulated (and evolving) context, the intended receiver of a message has to cope with unwarranted interference from a large number of transmitters, a factor which severely limits the capacity of the wireless system in question. On that account, and given that the theoretical performance limits of MIMO systems still elude us (even in basic network models such as the interference channel), a widespread approach is to treat the interference from other users as additive colored noise, and to use the mutual information for Gaussian input and noise as a unilateral performance metric [31]. In a similar vein, users cannot be

*This work was supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (contract no. 318306) and by ENSEA, Cergy-Pontoise, France.

Valuetools'13, December 10–12, 2013, Turin, Italy.

VALUETOOLS 2013, December 10-12, Torino, Italy

DOI 10.4108/icst.valuetools.2013.254385

assumed to have full information on the wireless system as it evolves over time (due e.g. to the arrival of new users, fluctuations in the PUs' demand, congestion, etc.), so they will have to optimize their signal characteristics "on the fly", based only on locally available information. Accordingly, our aim will be to derive a distributed learning scheme that allows the system's SUs to adapt to changes in the wireless medium and track their individual optimum signal profile using only local channel state information (CSI).

Of course, the setting above is fairly general in scope as it allows the network's SUs significant control over both spatial and spectral degrees of freedom: in the spatial component, the users control the covariance of their transmit directions (essentially the spread of their symbols over the transmitting antennas), whereas in the frequency domain, they control the allocation of their transmit power over the different channels at their disposal. To wit, when users are only equipped with a single antenna and are not allowed to split power across subcarriers, the problem boils down to deriving an efficient online channel selection policy as in [2, 10, 17, 21]. Alternatively, in the static, single-channel regime where the system's SUs only react to each other and the PUs' spectrum utilization is fixed, the main objective is to optimize the users' spectrum sharing policy [33] and/or to identify the Nash equilibria of the resulting game [26, 32]. That said, since the PUs' changing behavior is not affected (and cannot be predicted) by the system's SUs, our dynamic environment may no longer be modeled as a game, so static solution concepts (such as that of Nash equilibrium) are no longer meaningful. Thus, merging the analysis of [2] for dynamic channel selection with that of [26, 33] for static MIMO systems, we will focus on adaptive full spectrum management policies that lead to *no regret* in the sense that they perform asymptotically as well as *any* fixed policy, irrespective of the system's evolution over time [6, 27]. Intuitively, this means that the proposed scheme performs at least as well as the best fixed transmission strategy, even though the latter cannot be anticipated by the transmitter in our dynamic scenario with only local information.

Motivated by the no-regret properties of the exponential weight algorithm [6, 13, 28], our approach will be based on the distributed optimization method of *matrix exponential learning* that was recently introduced in [18]. More precisely, by decomposing our online rate maximization problem into a signal covariance and a power allocation component (corresponding to spatial and frequency degrees of freedom respectively), we derive an augmented exponential learning policy for adaptive spectrum management in dynamic MIMO CR environments. Then, by studying the evolution of the so-called "free energy" of the users' transmit profile, we show that this learning policy leads to no regret (Theorem 3.5); in fact, our exponential learning scheme turns out to be closely aligned to the direction of change of the users' rate function, so the system's SUs are able to track their individual optimum signal profile as it evolves over time, even under rapidly changing channel conditions.

2. SYSTEM MODEL

The cognitive radio system that we will focus on consists of a set of non-cooperative wireless MIMO users (primary and secondary alike), all communicating over several non-interfering channels by means of an orthogonal frequency division multiplexing (OFDM) scheme [3, 16]. Specifically, let $\mathcal{Q} = \mathcal{P} \cup \mathcal{S}$ denote the set of the system's users, with $\mathcal{P}$ (resp. $\mathcal{S}$) representing the system's primary (resp. secondary) users; assume further that each user $q \in \mathcal{Q}$ is equipped with m_q transmit antennas, and that the radio spectrum is partitioned into a set $\mathcal{K} = \{1, \ldots, K\}$ of K orthogonal frequency bands [3]. Then, the aggregate signal $\mathbf{y}_k^s \in \mathbb{C}^{n_s}$ received on the k-th frequency subcarrier at the intended destination of the secondary user $s \in \mathcal{S}$ (assumed equipped with n_s receive antennas) will be:

$$\mathbf{y}_k^s = \mathbf{H}_k^{ss}\mathbf{x}_k^s + \sum\nolimits_{p\in\mathcal{P}} \mathbf{H}_k^{ps}\mathbf{x}_k^p + \sum\nolimits_{r\in\mathcal{S},r\neq s} \mathbf{H}_k^{rs}\mathbf{x}_k^r + \mathbf{z}_k^s, \quad (1)$$

where $\mathbf{x}_k^q \in \mathbb{C}^{m_q}$ is the transmitted message of user $q \in \mathcal{Q}$ (primary or secondary) over the k-th subcarrier, $\mathbf{H}_k^{qs}$ is the corresponding channel matrix between the q-th transmitter and the intended receiver of user s, and $\mathbf{z}_k^s \in \mathbb{C}^{n_s}$ is the noise in the channel, including thermal, atmospheric and other peripheral interference effects (and modeled as a zero-mean circularly symmetric complex Gaussian random vector with non-singular covariance). Accordingly, if we focus on a particular secondary user and drop the index $s \in \mathcal{S}$ in (1) for simplicity, we obtain the unilateral signal model:

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{w}_k, \quad (2)$$

where $\mathbf{w}_k$ now denotes the multi-user interference-plus-noise over the frequency subcarrier $k \in \mathcal{K}$ at the receiver end.

The covariance of $\mathbf{w}_k$ in (2) obviously evolves over time e.g. due to users going offline for a period of time, or of modulating their transmit profiles to achieve better transmission rates.[1] In this evolving, decentralized context, employing sophisticated interference cancellation techniques at the receiver is all but impossible, especially with regards to the system's unregulated secondary users; as such, we will assume that interference by other users (primary and secondary alike) at the receiver is treated as additive, colored noise. In this single user decoding (SUD) regime, the transmission rate of a user under the signal model (2) will then be given by the familiar expression [3, 31]:

$$\Psi(\mathbf{Q},\mathbf{p}) = \sum\nolimits_k \Big[\log\det\big(\mathbf{W}_k + p_k\mathbf{H}_k\mathbf{Q}_k\mathbf{H}_k^\dagger\big) - \log\det\mathbf{W}_k\Big], \quad (3)$$

where:

1. $\mathbf{W}_k = \mathbb{E}\big[\mathbf{w}_k\mathbf{w}_k^\dagger\big]$ is the multi-user interference-plus-noise covariance matrix on subcarrier k at the receiver.
2. $p_k = \mathbb{E}\big[\mathbf{x}_k^\dagger\mathbf{x}_k\big] = \mathbb{E}\big[\|\mathbf{x}_k\|^2\big]$ is the user's transmit power over subcarrier k, and $\mathbf{p} = (p_1, \ldots, p_K)$ is the overall power allocation vector.
3. $\mathbf{Q}_k = \mathbb{E}\big[\mathbf{x}_k\mathbf{x}_k^\dagger\big] / \mathbb{E}\big[\mathbf{x}_k^\dagger\mathbf{x}_k\big]$ is the *normalized* covariance matrix of the user's transmitted signal, and $\mathbf{Q} = \bigoplus_{k=1}^K \mathbf{Q}_k = \operatorname{diag}(\mathbf{Q}_1, \ldots, \mathbf{Q}_K)$ denotes the aggregate covariance profile over all subcarriers.[2]

Thus, given that $\mathbf{W}_k$ might change over time as a result of evolving user conditions, we obtain the *time-dependent* objective:

$$\Psi(\mathbf{P};t) = \sum\nolimits_k \log\det\big[\mathbf{I} + \widetilde{\mathbf{H}}_k(t)\,\mathbf{P}_k\,\widetilde{\mathbf{H}}_k^\dagger(t)\big], \quad (4)$$

[1] That said, we will be assuming that such changes occur at a sufficiently slow rate relative to the coherence time of the channel so that the standard results of information theory continue to hold [31].

[2] Throughout this paper, $\bigoplus_{k=1}^K \mathbf{A}_k \equiv \operatorname{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_K)$ will denote the block-diagonal (direct) sum of the matrices $\mathbf{A}_k$.

where

$$\mathbf{P}_k = \mathbb{E}[\mathbf{x}_k \mathbf{x}_k^\dagger] = p_k \mathbf{Q}_k, \quad \mathbf{P} = \operatorname{diag}(\mathbf{P}_1, \dots, \mathbf{P}_K), \tag{5}$$

denotes the unnormalized covariance matrix of the transmitter's signal on the k-th subcarrier, and the effective channel matrices $\widetilde{\mathbf{H}}_k$ are given by

$$\widetilde{\mathbf{H}}_k(t) = \mathbf{W}_k(t)^{-1/2} \mathbf{H}_k. \tag{6}$$

Motivated by the "white-space filling" paradigm advocated (e.g. by the FCC) as a means to minimize interference by unlicensed users in MIMO CR networks by sensing spatial and/or spectral voids in the wireless medium [8, 14, 25, 26, 29], we will consider the following constraints for the SUs' transmit policies:

1. Constrained total power:[3]
$$\operatorname{tr}(\mathbf{P}) = \textstyle\sum_k p_k = P. \tag{7a}$$
2. Constrained transmit power per subcarrier:
$$\operatorname{tr}(\mathbf{P}_k) = p_k \leq P_k. \tag{7b}$$
3. Null-shaping constraints:
$$\mathbf{U}_k^\dagger \mathbf{P}_k = 0, \tag{7c}$$
for some tall complex matrix $\mathbf{U}_k$ with full column rank.

Of the constraints above, (7a) is a physical constraint on the user's total transmit power, (7b) imposes a limit on the interference level that can be tolerated on a given subcarrier, and (7c) is a "hard", spatial version of (7b) which guarantees that certain spatial dimensions per subcarrier (the columns of $\mathbf{U}_k$) will only be open to licensed, primary users (see also [16] for a more detailed discussion). After a suitable change of basis, the set of admissible signal shaping policies for the rate function (4) will thus be:

$$\mathcal{X} = \big\{(\mathbf{P}_1, \dots, \mathbf{P}_K) : \ \mathbf{P}_k \in \mathbb{C}^{m_k \times m_k}, \mathbf{P}_k \succcurlyeq 0, \\ 0 \leq \operatorname{tr}(\mathbf{P}_k) \leq P_k \text{ and } \textstyle\sum_k \operatorname{tr}(\mathbf{P}_k) = P\big\}, \tag{8}$$

where $m_k = \text{nullity}(\mathbf{U}_k)$ is the number of spatial dimensions that are open to SUs on subcarrier k. Accordingly, writing $\mathbf{P}_k$ in the decoupled form $\mathbf{P}_k = p_k \mathbf{Q}_k$ as before, we obtain the decomposition $\mathcal{X} = \prod_k \mathcal{D}_k \times \mathcal{X}_0$ where

$$\mathcal{D}_k = \big\{\mathbf{Q}_k \in \mathbb{C}^{m_k \times m_k} : \ \mathbf{Q}_k \succcurlyeq 0, \operatorname{tr}(\mathbf{Q}_k) = 1\big\} \tag{9}$$

is the "spectrahedron" of admissible normalized covariance matrices for subcarrier k and

$$\mathcal{X}_0 = \big\{\mathbf{p} \in \mathbb{R}^K : \ 0 \leq p_k \leq P_k, \textstyle\sum_k p_k = P\big\} \tag{10}$$

denotes the set of admissible power allocation vectors.

In view of the above, the unilateral objective of each SU at time t will be given by the online rate maximization problem:

$$\begin{aligned} &\text{maximize} \ \ \Psi(\mathbf{P}; t), \\ &\text{subject to} \ \ \mathbf{P} = \textstyle\bigoplus_{k=1}^K p_k \mathbf{Q}_k, \\ &\qquad\qquad (p_1, \dots, p_K) \in \mathcal{X}_0, \ \mathbf{Q}_k \in \mathcal{D}_k. \end{aligned} \tag{RM}$$

[3] If users are energy-aware, we should consider a more general total power constraint of the form $\sum_k p_k \leq P$ and incorporate a cost of power consumption in the user's objective. In our current setting however, the users have nothing to gain by not transmitting at full power, so there is no need to consider a softer constraint for the total transmit power.

Clearly, if the behavior of the network's PUs were known (or could otherwise be predicted) ahead of time, every SU would only have to react to each other's transmit policy, thus allowing us to model the situation as a non-cooperative game – see e.g. [26, 32] where the authors study the existence of a unique equilibrium in the "low-interference" regime of a stationary, single-carrier network. In our setting however, the PUs' transmit profiles are not influenced by the choices of the SUs, but instead change arbitrarily over time as a function of their individual needs; as a result, we obtain an evolving "game against nature" where static solution concepts (such as Nash equilibria) do not apply.

The most prominent solution concept in this online optimization context is that of a *no-regret* learning policy [6, 27], i.e. a dynamic transmit strategy $\mathbf{P}(t)$, $t \geq 0$, which performs asymptotically as well as *any* fixed profile $\mathbf{P}_0 \in \mathcal{X}$, and *for all* possible evolutions of the objective (4) over time. More precisely, the *regret* of a dynamic transmit policy $\mathbf{P}(t) \in \mathcal{X}$ with respect to $\mathbf{P}_0 \in \mathcal{X}$ is defined as:

$$\operatorname{Reg}(\mathbf{P}_0, t) = \frac{1}{t} \int_0^t \big[\Psi(\mathbf{P}_0; s) - \Psi(\mathbf{P}(s); s)\big] \, ds, \tag{11}$$

i.e. $\operatorname{Reg}(\mathbf{P}_0, t)$ simply measures the average transmission rate difference between $\mathbf{P}_0$ and $\mathbf{P}(t)$ up to time t. Obviously, large positive values of $\operatorname{Reg}(\mathbf{P}_0, t)$ indicate that the user could have achieved a higher transmission rate in the past by employing $\mathbf{P}_0$ instead of $\mathbf{P}(t)$, making him "regret" his policy choice. We will thus say that the policy $\mathbf{P}(t)$ *leads to no regret* if

$$\limsup_{t \to \infty} \operatorname{Reg}(\mathbf{P}_0, t) \leq 0 \tag{12}$$

for all $\mathbf{P}_0 \in \mathcal{X}$, and no matter how the objective $\Psi(\cdot; t)$ of (RM) evolves over time as a function of the effective channel matrices $\widetilde{\mathbf{H}}_k(t)$, $t \geq 0$. Alternatively, if we interpret $\lim_{t\to\infty} \int_0^t \Psi(\mathbf{p}_0; s) \, ds$ as the long-term average rate associated to $\mathbf{P}_0$, then (12) simply means that the average data rate of the dynamic policy $\mathbf{P}(t)$ is at least as good as that of any $\mathbf{P}_0 \in \mathcal{X}$, irrespective of how $\Psi(\cdot\,; t)$ evolves over time.

The notion of regret will be central in our analysis, so a few remarks are in order:

Remark 1. Obviously, if the optimum transmit policy $\mathbf{P}^*(t)$ which maximizes (RM) could be predicted at every $t \geq 0$ by some oracle-like device, we would have $\operatorname{Reg}(\mathbf{P}_0, t) \leq 0$ in (11) for all $\mathbf{P}_0 \in \mathcal{X}$, so the no-regret property (12) would be trivially satisfied. Equation (12) is thus a fundamental requirement for performance evaluation in the context of online programming, and negative regret is a key indicator of tracking the maximum of (RM) over time.

Remark 2. When the network's PUs induce changes to the wireless medium (e.g. with respect to network topology, channel conditions, congestion, etc.), the notion of regret becomes especially relevant because the PUs' behavior is not affected (at least, ideally) and cannot be predicted by the network's SUs. In this context, uniform power allocation is generally considered to be a realistic simple policy which is also robust against channel uncertainty: in particular, if the channel matrices are drawn at each realization from an isotropic distribution, then spreading power uniformly across carriers and antennas is the optimum policy in the worst-case scenario where nature (including the network's PUs) is actively choosing the worst possible channel realization for the transmitter [22]. A no-regret policy extends

this "min-max" concept by ensuring that no matter how the channels evolve over time (isotropically or otherwise), the policy's achieved transmission rate will be asymptotically as good as that of *any* transmit policy, including the uniform one (as a special case where nature is playing against the transmitter).

3. ONLINE SPECTRUM MANAGEMENT DYNAMICS

Even though there exists an extensive literature on no-regret learning policies for problems with discrete state spaces (see e.g. [6] for a panoramic survey), the situation is significantly more complicated in the case of online optimization programs with continuous action spaces and implicit constraints – such as the semidefiniteness constraints of (RM). To simplify matters, we will thus take a step-by-step approach consisting of: *a*) deriving a no-regret policy $\mathbf{Q}(t) \in \mathcal{X}_+ \equiv \prod_k \mathcal{D}_k$ for the covariance component of (RM) assuming a fixed power allocation profile $\mathbf{p} \in \mathcal{X}_0$; *b*) deriving a no-regret policy $\mathbf{p}(t) \in \mathcal{X}_0$ for the power allocation component of (RM) assuming a fixed covariance profile $\mathbf{Q} \in \mathcal{X}_+$; and *c*) combining the two policies in a joint no-regret spectrum management scheme for (RM) over the entire state space $\mathcal{X} = \mathcal{X}_+ \times \mathcal{X}_0$.

3.1 Online signal covariance optimization

We begin by analyzing the online rate maximization problem (RM) for a fixed power allocation profile $\mathbf{p} = (p_1, \ldots, p_K) \in \mathcal{X}_0$. To that end, if we fix $\mathbf{p} \in \mathcal{X}_0$, the derivation of a no-regret policy for the objective (4) over $\mathcal{X}_+ \equiv \prod_k \mathcal{D}_k$ boils down to the online semidefinite program:

$$\begin{array}{ll} \text{maximize} & \sum_k \log\det\big[\mathbf{I} + p_k \widetilde{\mathbf{H}}_k(t) \mathbf{Q}_k \widetilde{\mathbf{H}}_k^\dagger(t)\big], \\ \text{subject to} & \mathbf{Q}_k \succcurlyeq 0,\ \operatorname{tr}(\mathbf{Q}_k) = 1 \quad (k = 1, \ldots, K). \end{array} \tag{RM$_+$}$$

Motivated by the method of matrix exponential learning that was introduced in [18] for *static* optimization problems with constraints of this kind, we will consider the dynamics

$$\begin{aligned} \dot{\mathbf{Y}}_k &= \mathbf{V}_k, \\ \mathbf{Q}_k &= \frac{\exp(\mathbf{Y}_k)}{\operatorname{tr}[\exp(\mathbf{Y}_k)]}, \end{aligned} \tag{XL}$$

where $\mathbf{Y}_k \in \mathbb{C}^{m_k \times m_k}$ is an auxiliary "scoring" matrix and

$$\mathbf{V}_k = \frac{\partial \Psi}{\partial \mathbf{Q}_k^*} = p_k \widetilde{\mathbf{H}}_k^\dagger \big[\mathbf{I} + p_k \widetilde{\mathbf{H}}_k \mathbf{Q}_k \widetilde{\mathbf{H}}_k^\dagger\big]^{-1} \widetilde{\mathbf{H}}_k \tag{13}$$

is the (conjugate) gradient of the objective of (RM$_+$) with respect to $\mathbf{Q}_k \in \mathcal{D}_k$.

There are two reasons behind this choice of learning dynamics: first, in the static regime where the objective function does not change over time (i.e. $\widetilde{\mathbf{H}}_k(t) = \widetilde{\mathbf{H}}_k$ for all $t \geq 0$), the analysis of [18] shows that (XL) converges to the maximum of Ψ, so the no-regret property is satisfied trivially in that case. Secondly, if $\mathbf{Q}_k$ and $\widetilde{\mathbf{H}}_k^\dagger \widetilde{\mathbf{H}}_k$ are simultaneously diagonalizable, the dynamics (XL) reduce to the continuous-time exponential weight algorithm [6, 28]:

$$\dot{Y}_{k\alpha} = V_{k\alpha} \tag{14}$$

$$q_{k\alpha} = \frac{\exp(Y_{k\alpha})}{\sum_{\beta=1}^{m_k} \exp(Y_{k\beta})}, \tag{15}$$

where $Y_{k\alpha}$ and $V_{k\alpha}$, $\alpha = 1, \ldots, m_k$, denote the diagonal elements of $\mathbf{Y}_k$ and $\mathbf{V}_k$ respectively. For linear objectives of the form $\Psi(q; t) = \sum_{k=1}^K \sum_{\alpha=1}^{m_k} V_{k\alpha}(t)\, q_{k\alpha}$, it was shown in [28] that (14) is a no-regret policy for any (continuous) "payoff stream" $V_{k\alpha}(t)$; hence, in conjunction with our previous observation, one would hope that this stays true for the matrix-valued extension (XL) of (14), and for any stream of nonlinear objective functions $\Phi(\cdot; t)$ of the form (4). In the rest of this section, we will show that this is indeed the case, extending in this way the results of [28] to a semidefinite setting with nonlinear objectives.

Our analysis will hinge on a matrix variant of the so-called (Helmholtz) *free energy* [15] defined as:

$$A(\mathbf{Y}, \mathbf{Q}) = \operatorname{tr}[\mathbf{Y}\mathbf{Q}] - h(\mathbf{Q}), \tag{16}$$

where $\mathbf{Y}$ is Hermitian, $\mathbf{Q}$ is positive–semidefinite, and $h(\mathbf{Q})$ is (minus)[4] the von Neumann (quantum) entropy

$$h(\mathbf{Q}) = \operatorname{tr}(\mathbf{Q} \log \mathbf{Q}). \tag{17}$$

As we show in the following proposition, the key property of the free energy function A is that the normalized matrix exponential in (XL) is the unique solution of the positive-definite Legendre–Fenchel problem [23]:

$$\begin{array}{ll} \text{maximize} & A(\mathbf{Y}, \mathbf{Q}), \\ \text{subject to} & \mathbf{Q} \succcurlyeq 0,\ \operatorname{tr}(\mathbf{Q}) = 1. \end{array} \tag{LF$_+$}$$

More precisely, we have:

PROPOSITION 3.1. *Let $\mathbf{Y}$ be an $m \times m$ Hermitian matrix. Then, the unique solution to the Legendre–Fenchel problem (LF$_+$) is $\mathbf{Q}_\mathbf{Y} = \exp(\mathbf{Y}) / \operatorname{tr}[\exp(\mathbf{Y})]$, and the corresponding maximum value will be:*

$$h^*(\mathbf{Y}) \equiv A(\mathbf{Y}, \mathbf{Q}_\mathbf{Y}) = \log \operatorname{tr}[\exp(\mathbf{Y})]. \tag{18}$$

In view of Proposition 3.1 (which we prove in Appendix A), the so-called *convex conjugate* $h^*(\mathbf{Y}) = \log \operatorname{tr}[\exp(\mathbf{Y})]$ of h [23] will be particularly important because it represents the maximum value of the "average" $\langle \mathbf{Y} \rangle = \operatorname{tr}[\mathbf{Y}\mathbf{Q}]$ of $\mathbf{Y}$ adjusted for the "information cost" $h(\mathbf{Q})$ of $\mathbf{Q}$. To wit, by studying the evolution of h^* over time (see Appendix A for the details), we may show that the dynamics of matrix exponential learning provide a no-regret policy for the online rate maximization problem (RM$_+$); more precisely, we have:

THEOREM 3.2. *Let $\Psi(\cdot\,; t)$ be a continuous stream of rate functions of the form (4) for some evolving configuration of effective channel matrices $\widetilde{\mathbf{H}}_k(t)$, $t \geq 0$. Then, for every $\mathbf{Q}_0 \in \mathcal{X}_+$, the signal covariance policy (XL) satisfies*

$$\operatorname{Reg}(\mathbf{Q}_0, t) \leq \frac{h_+(\mathbf{Q}_0) + h_+^*(\mathbf{Y}_0)}{t} \tag{19}$$

where $\mathbf{Y}_0 \equiv \mathbf{Y}(0)$ is the initialization of $\mathbf{Y}$ in (XL) and h_+, h_+^* are the analogues of the von Neumann entropy (17) and its convex conjugate (18) over $\mathcal{X}_+$:

$$h_+(\mathbf{Q}) = \textstyle\sum_k \operatorname{tr}(\mathbf{Q}_k \log \mathbf{Q}_k), \tag{20a}$$

$$h_+^*(\mathbf{Y}) = \textstyle\sum_k \log \operatorname{tr}[\exp(\mathbf{Y}_k)]. \tag{20b}$$

In particular, (XL) is a no-regret policy for the online covariance optimization problem (RM$_+$).

[4] We are omitting the traditional minus sign for notational convenience – so that h be strictly convex instead of strictly concave.

REMARK. By taking the uniform initialization $\mathbf{Y}_0 = 0$ and noting that $h_+(\mathbf{Q}) \leq 0$ for all $\mathbf{Q} \in \mathcal{X}_+$, the bound (19) readily yields

$$\mathrm{Reg}(\mathbf{Q}_0, t) \leq t^{-1} \textstyle\sum_{k=1}^K \log m_k. \tag{21}$$

We thus see that a user's regret grows at most linearly in the number of available subcarriers and at most logarithmically in the number of transmit dimensions m_k per subcarrier, a fact with important consequences as far as the policy's "curse of dimensionality" is concerned [6].

3.2 Dynamic power allocation

Having established a no-regret learning policy for the covariance component (RM_+) of (RM), we now turn to the orthogonal problem of optimizing the user's power allocation for a fixed covariance profile $\mathbf{Q} \in \mathcal{X}_+$. More specifically, our aim in this section will be to devise a no-regret policy for the online power allocation problem:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_k \log\det\big[\mathbf{I} + p_k \mathbf{S}_k(t)\big], \\ \text{subject to} \quad & 0 \leq p_k \leq P_k,\ \textstyle\sum_k p_k = P, \end{aligned} \tag{RM_0}$$

where $\mathbf{Q}_k \in \mathcal{D}_k$ is now kept fixed and, for notational convenience, we have set $\mathbf{S}_k(t) = \widetilde{\mathbf{H}}_k(t)\, \mathbf{Q}_k\, \widetilde{\mathbf{H}}_k^\dagger(t)$.

Motivated by the analysis of the previous section, our approach will consist of the following steps:

1. Define an entropy-like function on the state space $\mathcal{X}_0 = \{\mathbf{p} \in \mathbb{R}^K : 0 \leq p_k \leq P_k,\ \sum_k p_k = P\}$ of the online power allocation problem (RM_0).
2. Introduce a "score vector" $\mathbf{y} = (y_1, \dots, y_K)$ to track the performance of each subcarrier based on the gradient component $v_k = \frac{\partial \Psi}{\partial p_k}$ of the objective of (RM_0).
3. Map these "scores" to a power allocation vector $\mathbf{p} \in \mathcal{X}_0$ by solving a Legendre–Fenchel problem as in Proposition 3.1 (which characterizes the exponential choice map of (XL)).

This approach is driven in no small part by the proof of Theorem 3.2 which relies on the convex conjugate of the von Neumann entropy (17).[5] Our first step will thus be to define a strictly convex entropy function $h_0 \colon \mathcal{X}_0 \to \mathbb{R}$ which becomes infinitely steep near the boundary $\mathrm{bd}(\mathcal{X}_0)$ of $\mathcal{X}_0$.

Inspired by a construction of [1], let

$$h_0(\mathbf{p}) = \sum\nolimits_{k=1}^K \big[\theta(p_k) + \theta(P_k - p_k)\big], \tag{22}$$

where

$$\theta(x) = x \log x \tag{23}$$

is the single-dimensional analogue of the von Neumann entropy (17). It is then easy to see that *a)* h_0 is strictly convex over $\mathcal{X}_0$; and *b)* it becomes infinitely steep at the boundary $\mathrm{bd}(\mathcal{X}_0)$ of $\mathcal{X}_0$. As a result, the associated Legendre–Fenchel problem

$$\begin{aligned} \text{maximize} \quad & \mathbf{y} \cdot \mathbf{p} - h_0(\mathbf{p}), \\ \text{subject to} \quad & 0 \leq p_k \leq P_k,\ \textstyle\sum_k p_k = P, \end{aligned} \tag{LF_0}$$

will admit a unique interior solution $\mathbf{p}^*(\mathbf{y}) \in \mathcal{X}_0$ for every $\mathbf{y} \in \mathbb{R}^K$ (see e.g. Chapter 25 in [23]). Hence, in view of Proposition 3.1, and echoing the definition (XL) of the dynamics of matrix exponential learning, we will consider the power allocation dynamics:

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{v}, \\ \mathbf{p} &= \mathbf{p}^*(\mathbf{y}), \end{aligned} \tag{24}$$

where, as noted before, the auxiliary vector $\mathbf{y}$ "scores" the performance of each subcarrier by tracking the gradient vector $\mathbf{v} = (v_1, \dots, v_K)$ with components given by

$$v_k = \frac{\partial \Psi}{\partial p_k} = \mathrm{tr}\left[\big(\mathbf{I} + p_k \mathbf{S}_k\big)^{-1} \mathbf{S}_k\right] = p_k^{-1}\, \mathrm{tr}[\mathbf{V}_k \mathbf{Q}_k]. \tag{25}$$

To implement the power allocation scheme (24), we need to have an explicit expression for the solution mapping $\mathbf{p}^*(\mathbf{y})$ of (LF_0) – just as Proposition 3.1 provided the exponential expression $\exp(\mathbf{Y})/\,\mathrm{tr}[\exp(\mathbf{Y})]$ for the positive-definite problem (LF_+). To that end, the Karush–Kuhn–Tucker (KKT) conditions for (LF_0) give:

$$y_k - \theta'(p_k) + \theta'(P_k - p_k) = \lambda, \tag{26}$$

where λ is the Lagrange multiplier for the total power constraint $\sum_k p_k = P$. With $\theta'(x) = 1 + \log x$, we then obtain $\log(P_k - p_k) - \log p_k = \lambda - y_k$, or, after solving for p_k:

$$p_k = \frac{P_k}{1 + \exp(\lambda - y_k)}. \tag{27}$$

We are thus left to calculate the Lagrange multiplier λ; to that end, summing (27) over $k \in \mathcal{K}$ yields:

$$P = \sum\nolimits_{k=1}^K \frac{P_k}{1 + \exp(\lambda - y_k)}. \tag{28}$$

Since the RHS of this last equation is bounded below and (strictly) decreasing in λ, it is straightforward to calculate λ numerically – e.g. by performing a simple line search for e^λ [5].

On the other hand, by carrying out the summation at the RHS of (28), it is easy to see that deriving a closed-form expression for e^λ involves solving a polynomial equation of degree K – which is impossible for $K \geq 5$ (and leads to fairly complicated expressions, even for $K = 2$). Instead, an explicit analytic expression for the power allocation dynamics (24) is provided by the following proposition (see Appendix A for a proof):

PROPOSITION 3.3. *The solution orbits of the dynamics (24) satisfy the adjusted replicator dynamics:*

$$\dot{p}_k = \rho_k \big[v_k - \rho^{-1} \textstyle\sum_\ell \rho_\ell v_\ell\big], \tag{aRD}$$

where $\rho_k = p_k(P_k - p_k)/P_k$, $\rho = \sum_k \rho_k$, and the gradient components v_k are given by (25).

REMARK 1. The terminology "adjusted replicator equation" is due to the similarity between (aRD) and the replicator dynamics of evolutionary game theory [24, 30]. In our setting, this last equation would take the form [19]:

$$\dot{p}_k = p_k \big[v_k - P^{-1} \textstyle\sum_\ell p_\ell v_\ell\big], \tag{RD}$$

with the transmit powers p_k assumed to satisfy the total power constraint (7a), but not necessarily the "low-interference" constraints (7b).

[5] See also [27] for links with the discrete-time learning method of online mirror descent, [1] for a closely related entropy-driven approach to static optimization problems, and [7] for applications to learning in games).

REMARK 2. To gain some intuition about the form of the adjusted variables ρ_k, it is instructive to note that (aRD) satisfies the constraints (7a) and (7b) for all $t \geq 0$. Indeed, $\rho_k = 0$ if and only if $p_k \in \{0, P_k\}$, so we will have $0 \leq p_k(t) \leq P_k$ for all $t \geq 0$; the second term of (aRD) then ensures that $\sum_k \dot{p}_k = 0$, so the total transmit power $\sum_k p_k$ is conserved along (aRD).[6]

A fundamental property of the replicator dynamics is that they lead to no regret [13]; as it turns out (see Appendix A for the proof), this property also holds for (aRD) as well:

THEOREM 3.4. *Let $\Psi(\cdot\,;t)$ be a continuous stream of rate functions of the form (4) for some evolving configuration of effective channel matrices $\widetilde{\mathbf{H}}_k(t)$, $t \geq 0$. Then, for every $\mathbf{p}_0 \in \mathcal{X}_0$, the power allocation policy (aRD) satisfies:*

$$\mathrm{Reg}(\mathbf{p}_0, t) \leq C_0/t, \tag{29}$$

for some positive constant C_0 which depends only on $\mathbf{p}_0$ and the initialization of (aRD). In particular, (aRD) is a no-regret policy for the online power allocation problem (RM_0).

3.3 Joint covariance-and-power management

In view of the analysis of the previous sections, our joint signal covariance and power management scheme will be a combination of matrix exponential learning (for the covariance component) and the adjusted replicator dynamics (for the power allocation component). More precisely, by combining (XL) and (aRD), we obtain the *augmented* dynamics:

$$\begin{aligned} \dot{p}_k &= \rho_k \left[v_k - \rho^{-1} \textstyle\sum_{\ell=1}^K \rho_\ell v_\ell \right], \\ \dot{\mathbf{Y}}_k &= \mathbf{V}_k, \\ \mathbf{Q}_k &= \frac{\exp(\mathbf{Y}_k)}{\mathrm{tr}[\exp(\mathbf{Y}_k)]}, \end{aligned} \tag{30}$$

where, as before:

$$\begin{aligned} \mathbf{V}_k &= \frac{\partial \Psi}{\partial \mathbf{Q}_k^*} = p_k \widetilde{\mathbf{H}}_k^\dagger \left[\mathbf{I} + \widetilde{\mathbf{H}}_k \mathbf{Q}_k \widetilde{\mathbf{H}}_k^\dagger \right]^{-1} \widetilde{\mathbf{H}}_k, \\ v_k &= \frac{\partial \Psi}{\partial p_k} = p_k^{-1} \,\mathrm{tr}\left[\mathbf{V}_k \mathbf{Q}_k \right], \\ \rho_k &= p_k(P_k - p_k)/P_k, \quad \rho = \textstyle\sum_{\ell=1}^K \rho_\ell. \end{aligned} \tag{31}$$

By combining Theorems 3.2 and 3.4, we may then show that the dynamics (30) lead to no regret in the online rate maximization problem (RM) (see Appendix A for the details):

THEOREM 3.5. *Let $\Phi(\cdot\,;t)$ be a continuous stream of rate functions of the form (4) for some evolving configuration of effective channel matrices $\widetilde{\mathbf{H}}_k(t)$, $t \geq 0$. Then, the augmented dynamics (30) satisfy:*

$$\mathrm{Reg}(\mathbf{P}_0, t) \leq C/t, \tag{32}$$

for some positive constant C which depends only on $\mathbf{P}_0$ and the initialization of (30). In particular, (30) leads to no regret in the online rate maximization problem (RM).

[6]Note also that in the limit $P_k \to \infty$ where the constraints (7b) become redundant, we also get $\rho_k \to p_k$.

4. NUMERICAL RESULTS

In this section, our aim will be to evaluate the performance of the online spectrum management dynamics (30) via numerical simulations. We begin by providing an algorithmic version of (30) that could be employed by every SU in the network in a fully distributed fashion:

Algorithm 1 Augmented Exponential Learning (AXL)

$n \leftarrow 0$;
Choose $\eta > 0$;
foreach subcarrier $k \in \mathcal{K}$ **do**
  Initialize Hermitian score matrix $\mathbf{Y}_k \in \mathbb{C}^{m_k \times m_k}$ and normalized signal covariance $\mathbf{Q}_k \in \mathcal{D}_k$;
  Initialize subcarrier score $y_k \in \mathbb{R}$ and transmit power p_k.
Repeat
  $n \leftarrow n + 1$;
  foreach subcarrier $k \in \mathcal{K}$ **do**
    Measure effective channel matrix $\widetilde{\mathbf{H}}_k$;
    Set $\mathbf{V}_k \leftarrow p_k \widetilde{\mathbf{H}}_k^\dagger \left[\mathbf{I} + \widetilde{\mathbf{H}}_k \mathbf{Q}_k \widetilde{\mathbf{H}}_k^\dagger \right]^{-1} \widetilde{\mathbf{H}}_k$;
    Update covariance score $\mathbf{Y}_k \leftarrow \mathbf{Y}_k + \eta \mathbf{V}_k$;
    Update subcarrier score $y_k \leftarrow y_k + \eta\, p_k^{-1} \,\mathrm{tr}[\mathbf{V}_k \mathbf{Q}_k]$ and calculate λ from (28);
    Set signal covariance $\mathbf{Q}_k \leftarrow \exp(\mathbf{Y}_k) \big/ \mathrm{tr}[\exp(\mathbf{Y}_k)]$;
    Set transmit power $p_k \leftarrow P_k/(1 + \exp(\lambda - y_k))$.

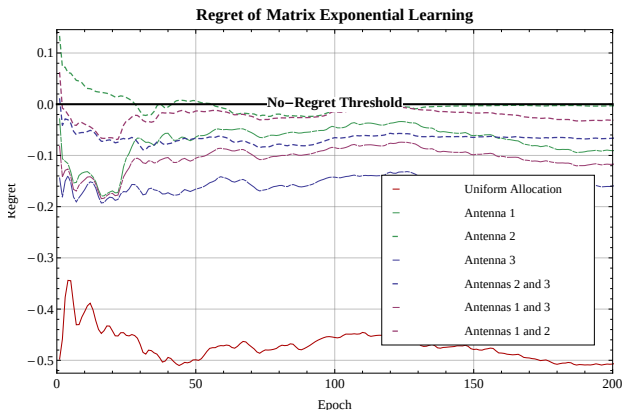
The augmented exponential learning (AXL) algorithm above will be the main focus of this section, so a few remarks are in order:

REMARK 1. From an implementation point of view, AXL has the following desirable properties:

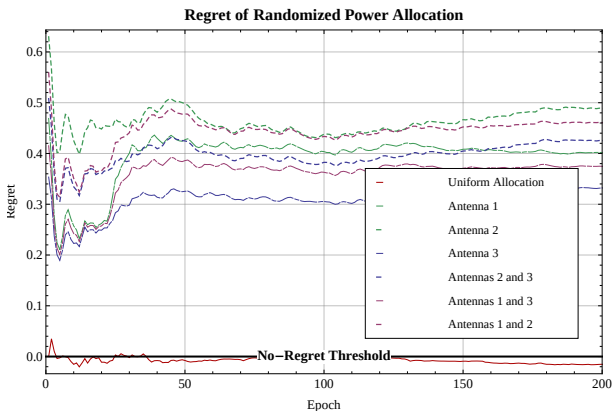
(P1) It is *distributed*: users only need to update their individual signal characteristics using local channel state information (the channel matrices $\widetilde{\mathbf{H}}_k$).
(P2) It is *asynchronous*: there is no need for a global update timer requiring user synchronization and coordination.
(P3) It is *stateless*: users do not need to know the state of the system and they may be completely oblivious to each other's actions.
(P4) It is *reinforcing*: users tend to increase their unilateral transmission rates.

REMARK 2. Essentially, AXL may be viewed as a discretization of the spectrum management dynamics (30): given that the adjusted replicator equation (aRD) is equivalent to the dynamic power allocation scheme (24), the score-updating step of AXL is just an Euler discretization of (XL) and (24) with constant step size η. The reason for discretizing (24) instead of (aRD) is that we do not need to impose bounds on the step-size η to ensure that the algorithm will always remain in the state space $\mathcal{X}$ of (RM).

Of course, this approach carries the cost of having to calculate the Lagrange multiplier λ numerically; however, given that (28) is bounded and decreasing in λ, this calculation can be performed in an efficient and stable way (e.g. by a low-complexity line search [5]), so this implicit numerical step does not have a perceptible impact on the performance of AXL. If required, this step could be avoided altogether



(a) Long-term regret of augmented exponential learning.



(b) Long-term regret of randomized power allocation.

Figure 1: The long-term regret induced by augmented exponential learning (Fig. 1(a)) and a random sampling power allocation policy (Fig. 1(b)) with respect to different signal covariance profiles (see text for details). In tune with Theorem 3.5, augmented exponential learning quickly gets below the no-regret threshold whereas the randomized policy leads to positive regret in 6 out of the 7 benchmark signal profiles.

by introducing an explicit Euler discretization of (aRD) and taking η sufficiently small; in our numerical simulations however, we never needed to do so (and CPU monitoring did not show a perceptible difference between the two approaches).

Now, to validate the predictions of Section 3 for the AXL algorithm, we simulated in Fig. 1 a network consisting of 4 PUs and 8 SUs, all equipped with $m_k = 3$ transmit/receive antennas, and communicating over $K = 6$ orthogonal subcarriers with a base frequency of $\nu = 2\,\text{GHz}$. Both the PUs and the SUs were assumed to be mobile with an average speed of 5 km/h (pedestrian movement), and the channel matrices $\mathbf{H}_k^{qs}$ of (1) were modeled after the well-known Jakes model for Rayleigh fading [4]. For simplicity, we assumed that the PUs were going online and offline following a Poisson process but were not otherwise modulating their transmit signal characteristics; on the other hand, the simulated SUs tried to optimize their spectrum exploitation by using the AXL algorithm with $\eta = 1$ and an update epoch of $\delta = 5\,\text{ms}$.

We then picked a sample secondary user to focus on, and we calculated the regret induced by the AXL policy with respect to 7 different fixed signal profiles: the uniform one (where power is spread equally across antennas and frequency bands), and all possible combinations of spreading power uniformly across subcarriers while keeping one or two transmit dimensions closed (the legend of Fig. 1 indicates the antennas that were not employed in each benchmark policy).[7] The results of these simulations were plotted in Fig. 1(a): as predicted by Theorem 3.5, AXL leads to no regret in the long term; in fact, AXL falls below the no-regret threshold within a few epochs, indicating that it is performing strictly better on average than any of the benchmark signal shaping profiles.

For comparison purposes, we also simulated the same scenario, but with the SUs employing a randomized signal shaping policy. In particular, motivated by the analysis of [22], we simulated the randomized sampling scheme:

$$\begin{aligned} \mathbf{Q}_k(n+1) &= (1-r)\mathbf{Q}_k(n) + r\mathbf{R}_k(n), \\ \mathbf{Q}_k(0) &= m_k^{-1}\mathbf{I}, \end{aligned} \tag{33}$$

where the matrix $\mathbf{R}_k(n)$ is drawn uniformly from the spectrahedron $\mathcal{D}_k$ of $m_k \times m_k$ positive-definite matrices with unit trace, and $r \in [0,1]$ is a discount parameter interpolating between the uniform distribution $\mathbf{Q}_k \propto \mathbf{I}$ for $r = 0$ and the completely random policy $\mathbf{R}_k$ for $r = 1$ (we took $r = 0.9$ in our simulations).[8] Even though this online spectrum management policy is sampling the state space essentially uniformly for large values of r, Fig. 1(b) shows that it leads to positive regret in 6 out of the 7 benchmark policies.[9] In other words, the no regret property of AXL is not a spurious artifact of the exploration aspect of the entropy-driven dynamics (30), but a direct consequence of the underlying learning process.

The negative regret results of Fig. 1 also suggest that the transmission rate achieved by the focal SU is close to the user's (evolving) maximum possible rate. To test this hypothesis, we plotted in Fig. 2 the achieved data rate of a SU employing the AXL algorithm versus the user's maximum possible unilateral data rate and the transmission rates achieved by uniform power allocation and the random sampling policy (33). We see there that AXL adapts to the changing channel conditions and tracks the maximum achievable rate remarkably well, staying within 10% of the maximum rate for almost all epochs (in stark contrast to both fixed and randomized signal shaping policies).

5. CONCLUSIONS

[7] We chose these benchmark profiles so as to sample the covariance component $\mathcal{X}_+$ of the problem's state space as uniformly as possible.

[8] In practice, it is not possible to switch to very different covariance profiles when the update epoch is very short; we introduced the averaging parameter r in order to smooth out the random process somewhat.

[9] The random sampling policy (33) leads to no regret against the uniform power allocation policy because the average of (33) is just the uniform policy.

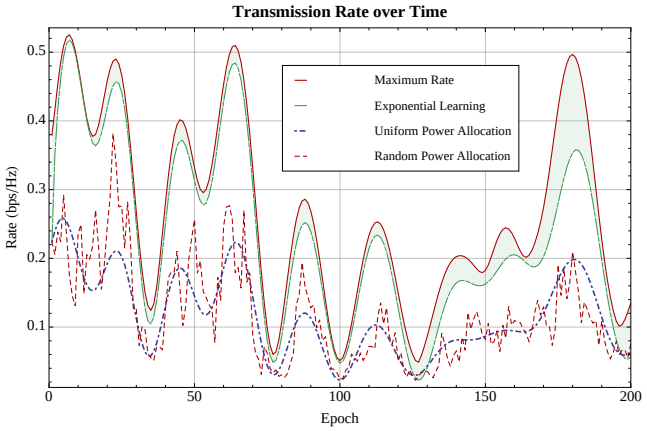


Figure 2: Data rates achieved by AXL in a changing environment: by following the AXL algorithm, users are able to track the (evolving) transmission profile that maximizes their unilateral data rates.

In this paper, we introduced an adaptive spectrum management policy for MIMO-OFDM cognitive radio systems that evolve dynamically over time as a function of changing user and environmental conditions. By drawing on the method of matrix exponential learning and decomposing the users' online rate maximization into a signal covariance and a power allocation component, we derived an augmented exponential learning scheme which leads to *no regret*: for every SU, the proposed dynamic learning policy performs asymptotically as well as any fixed signal distribution, irrespective of how the system evolves over time. In fact, this learning scheme is closely aligned to the direction of change of the users' data rate function, so the system's SUs are able to track their individual optimum signal distribution even under rapidly changing conditions.

To a large extent, our learning scheme owes its no-regret properties to the associated entropy-like function defined on the problem's state space (for instance, the von Neumann quantum entropy for the problem's signal covariance component). As a result, by properly extending the driving entropy function, our method can lead to no-regret strategies in significantly more general situations – including for example pricing and/or energy-awareness constraints.

APPENDIX

A. TECHNICAL PROOFS

We begin by showing that the normalized exponential of (XL) solves the positive-definite problem (LF_+):

Proof of Proposition 3.1. Our proof strategy will be to first show that the problem (LF_+) admits a unique solution in the (relative) interior $\mathcal{D}^\circ$ of the spectrahedron $\mathcal{D} = \{\mathbf{Q} \in \mathbb{C}^{m\times m} : \mathbf{Q} \succcurlyeq 0, \mathrm{tr}(\mathbf{Q}) = 1\}$, and then use the KKT conditions to establish our claim. To that end, we will first need to evaluate the directional derivative $\nabla_{\mathbf{Z}} A(\mathbf{Y},\mathbf{Q}) = \frac{d}{dt}\big|_{t=0} A(\mathbf{Y},\mathbf{Q}+t\mathbf{Z})$ of the free energy (16) on the tangent directions $\mathbf{Z} \in T_{\mathbf{Q}}\mathcal{D}^\circ = \{\mathbf{A} \in \mathbb{C}^{m\times m} : \mathbf{A}^\dagger = \mathbf{A}, \mathrm{tr}(\mathbf{A}) = 0\}$ to $\mathcal{D}^\circ$ at $\mathbf{Q}$. Accordingly, if $\{q_j, \mathbf{u}_j\}_{j=1}^m$ is an eigen-decomposition of $\mathbf{Q}+t\mathbf{Z}$, we readily obtain $A(\mathbf{Y},\mathbf{Q}+t\mathbf{Z}) = \mathrm{tr}[\mathbf{Y}\mathbf{Q}] + \mathrm{tr}[\mathbf{Y}\mathbf{Z}]\,t - \sum_j q_j \log q_j$, and hence:

$$\nabla_{\mathbf{Z}} A(\mathbf{Y},\mathbf{Q}) = \tfrac{d}{dt}\big|_{t=0} A(\mathbf{Y},\mathbf{Q}+t\mathbf{Z}) = \mathrm{tr}[\mathbf{Y}\mathbf{Z}] - \textstyle\sum_j \dot{q}_j \log q_j, \tag{34}$$

where we have used the fact that $\sum_j \dot{q}_j = 0$ (recall that $\sum_j q_j = \mathrm{tr}(\mathbf{Q}+t\mathbf{Z}) = 1$ for all t such that $\mathbf{Q}+t\mathbf{Z} \in \mathcal{D}^\circ$). Moreover, differentiating the defining relation $(\mathbf{Q}+t\mathbf{Z})\mathbf{u}_j = q_j\mathbf{u}_j$ yields $\mathbf{Z}\mathbf{u}_j + (\mathbf{Q}+t\mathbf{Z})\dot{\mathbf{u}}_j = \dot{q}_j\mathbf{u}_j + q_j\dot{\mathbf{u}}_j$, so, after multiplying from the left by $\mathbf{u}_j^\dagger$, we get:

$$\dot{q}_j = \mathbf{u}_j^\dagger \mathbf{Z}\mathbf{u}_j + \mathbf{u}_j^\dagger(\mathbf{Q}+t\mathbf{Z})\dot{\mathbf{u}}_j - q_j\mathbf{u}_j^\dagger\dot{\mathbf{u}}_j = \mathbf{u}_j^\dagger\mathbf{Z}\mathbf{u}_j. \tag{35}$$

Summing over j gives $\sum_j \dot{q}_j \log q_j = \sum_j \mathbf{u}_j^\dagger\mathbf{Z}\mathbf{u}_j \log q_j = \mathrm{tr}[\mathbf{Z}\log\mathbf{Q}]$; then, by substituting in (34), we get:

$$\nabla_{\mathbf{Z}} A(\mathbf{Y},\mathbf{Q}) = \mathrm{tr}[\mathbf{Z}(\mathbf{Y} - \log\mathbf{Q})]. \tag{36}$$

With this expression at hand, it is easy to see that $A(\mathbf{Y},\cdot)$ becomes infinitely steep at the boundary $\mathrm{bd}(\mathcal{D})$ of $\mathcal{D}$, i.e. $|\nabla_{\mathbf{Z}} A(\mathbf{Y},\mathbf{Q}_n)| \to \infty$ whenever $\mathbf{Q}_n \to \mathrm{bd}(\mathcal{D})$. Since h is strictly convex, it follows that A will be of Legendre type [1, 23], so (LF_+) will admit a unique solution $\mathbf{Q}_{\mathbf{Y}}$ at the interior $\mathcal{D}^\circ$ of $\mathcal{D}$ – see e.g. Chap. 26 in [23]. The KKT conditions for (LF_+) thus yield $\nabla_{\mathbf{Z}} A(\mathbf{Y},\mathbf{Q}_{\mathbf{Y}}) = 0$ for all tangent directions $\mathbf{Z}$ to $\mathcal{D}^\circ$ at $\mathbf{Q}_{\mathbf{Y}}$, i.e. $\mathrm{tr}[\mathbf{Z}(\mathbf{Y}-\log\mathbf{Q}_{\mathbf{Y}})] = 0$ for all Hermitian $\mathbf{Z}$ such that $\mathrm{tr}(\mathbf{Z}) = 0$. From this last condition, we immediately deduce $\mathbf{Y} - \log\mathbf{Q}_{\mathbf{Y}} \propto \mathbf{I}$, and with $\mathrm{tr}(\mathbf{Q}_{\mathbf{Y}}) = 1$, we obtain $\mathbf{Q}_{\mathbf{Y}} = \exp(\mathbf{Y})/\mathrm{tr}[\exp(\mathbf{Y})]$; the expression for the convex conjugate $h^*(\mathbf{Y})$ of h then follows by substituting $\mathbf{Q}_{\mathbf{Y}}$ in (16). □

Proof of Theorem 3.2. Let $h_k(\mathbf{Q}_k) = \mathrm{tr}(\mathbf{Q}_k\log\mathbf{Q}_k)$, $\mathbf{Q}_k \in \mathcal{D}_k$, so $h_k^*(\mathbf{Y}_k) = \log\mathrm{tr}[\exp(\mathbf{Y}_k)]$ by Proposition 3.1. Letting $\mathbf{Y} = \bigoplus_k \mathbf{Y}_k = \mathrm{diag}(\mathbf{Y}_1,\dots,\mathbf{Y}_K)$ and noting that $h_+^*(\mathbf{Y}) = \sum_k h_k^*(\mathbf{Y}_k)$, we obtain:

$$\begin{aligned}\frac{dh_+^*}{dt} &= \sum\nolimits_k \mathrm{tr}[\exp(\mathbf{Y}_k)]^{-1}\frac{d}{dt}\mathrm{tr}[\exp(\mathbf{Y}_k)]\\ &= \sum\nolimits_k \mathrm{tr}[\exp(\mathbf{Y}_k)]^{-1}\mathrm{tr}[\dot{\mathbf{Y}}_k\exp(\mathbf{Y}_k)] = \mathrm{tr}[\mathbf{V}\mathbf{Q}],\end{aligned} \tag{37}$$

where, in the last step, we used the fact that $\dot{\mathbf{Y}}_k = \mathbf{V}_k$ on account of (XL). A simple integration then gives:

$$\begin{aligned}\int_0^t \mathrm{tr}[\mathbf{Q}(s)\mathbf{V}(s)]\,ds &= \sum\nolimits_k \big[h_k^*(\mathbf{Y}_k(t)) - h_k^*(\mathbf{Y}_k(0))\big]\\ &\geq \mathrm{tr}[\mathbf{Y}(t)\mathbf{Q}_0] - h_+(\mathbf{Q}_0) - h_+^*(\mathbf{Y}_0),\end{aligned} \tag{38}$$

where the last inequality follows from the fact that $h_k^*(\mathbf{Y}_k)$ is the maximum value of $\mathrm{tr}[\mathbf{Y}_k\mathbf{Q}_k] - h_k(\mathbf{Q}_k)$ over $\mathcal{D}_k$ (Proposition 3.1). Recalling that $\mathbf{Y}(t) = \int_0^t \mathbf{V}(s)\,ds$ from the definition of the dynamics (XL) and rearranging then yields:

$$\int_0^t \mathrm{tr}[\mathbf{Q}_0\mathbf{V}(s)] - \mathrm{tr}[\mathbf{Q}(s)\mathbf{V}(s)]\,ds \leq h^*(\mathbf{Y}_0) + h(\mathbf{Q}_0). \tag{39}$$

However, with $\Psi(\cdot;t)$ (strictly) concave over $\mathcal{X}_+$ and $\mathbf{V} = \nabla\Psi$, we will also have $\mathrm{tr}[(\mathbf{Q}_0-\mathbf{Q})\mathbf{V}] = \mathrm{tr}[(\mathbf{Q}_0-\mathbf{Q})\nabla\Psi(\mathbf{Q})] \geq \Psi(\mathbf{Q}_0;s) - \Psi(\mathbf{Q};s)$ for all $\mathbf{Q} \in \mathcal{X}_+$ by monotonicity (we are suppressing the fixed power allocation argument $\mathbf{p} \in \mathcal{X}_0$ for simplicity). The bound (19) then follows directly from (39), and the no-regret property of (XL) is obtained by dividing (19) by t and taking the lim sup as $t \to \infty$. □

We now move on to the power allocation component (RM_0) of the online rate maximization problem (RM). We begin by establishing the equivalence between (24) and (aRD):

PROOF OF PROPOSITION 3.3. Let $p_k^* \equiv p_k^*(\mathbf{y})$, $k \in \mathcal{K}$, denote the components of the solution $\mathbf{p}^*(\mathbf{y})$ of the Legendre–Fenchel problem (LF$_0$). By differentiating the power update step of (24), we get

$$\dot{p}_k = \sum\nolimits_\ell \frac{\partial p_k^*}{\partial y_\ell} \dot{y}_\ell = \sum\nolimits_\ell \frac{\partial p_k^*}{\partial y_\ell} v_\ell, \tag{40}$$

so, to establish the equivalence between (24) and (aRD), it suffices to show that

$$\frac{\partial p_k^*}{\partial y_\ell} = \rho_k(\delta_{k\ell} - \rho_\ell/\rho), \tag{41}$$

where $\rho_k = p_k(P_k - p_k)/P_k$, $\rho = \sum_k \rho_k$.

To that end, differentiating (26) w.r.t. y_ℓ readily yields:

$$\delta_{k\ell} - \frac{P_k}{p_k(P_k - p_k)} \frac{\partial p_k^*}{\partial y_\ell} = \frac{\partial \lambda}{\partial y_\ell}, \tag{42}$$

with λ being the Lagrange multiplier for the total power constraint $\sum_k p_k = P$. Solving for $\partial p_k^*/\partial y_\ell$ then gives

$$\frac{\partial p_k^*}{\partial y_\ell} = \rho_k \left(\delta_{k\ell} - \frac{\partial \lambda}{\partial y_\ell} \right), \tag{43}$$

so it suffices to show that $\partial\lambda/\partial y_\ell = \rho_\ell/\rho$. However, by differentiating (28) and using (27), we get:

$$\begin{aligned} 0 &= \sum\nolimits_{k=1}^K \frac{P_k \exp(\lambda - y_k)}{(1 + \exp(\lambda - y_k))^2} \left(\delta_{k\ell} - \frac{\partial \lambda}{\partial y_\ell} \right) \\ &= \sum\nolimits_{k=1}^K p_k(P_k - p_k)/P_k \left(\delta_{k\ell} - \frac{\partial \lambda}{\partial y_\ell} \right) = \rho_\ell - \rho \frac{\partial \lambda}{\partial y_\ell}, \end{aligned} \tag{44}$$

and our claim follows. □

PROOF OF THEOREM 3.4. Shadowing the proof of Theorem 3.2, let

$$h_0^*(\mathbf{y}) = \max_{\mathbf{p} \in \mathcal{X}_0} \{\mathbf{p} \cdot \mathbf{y} - h_0(\mathbf{p})\} = \mathbf{y} \cdot \mathbf{p}^*(\mathbf{y}) - h_0(\mathbf{p}^*(\mathbf{y})) \tag{45}$$

be the convex conjugate of $h_0(\mathbf{p})$. We will then have:

$$\begin{aligned} \frac{\partial h_0^*}{\partial y_k} &= p_k^* + \sum\nolimits_{\ell=1}^K y_\ell \frac{\partial p_\ell^*}{\partial y_k} - \sum\nolimits_{\ell=1}^K \frac{\partial h_0}{\partial p_\ell} \frac{\partial p_\ell^*}{\partial y_k} \\ &= p_k^* + \lambda \sum\nolimits_{\ell=1}^K \frac{\partial p_\ell^*}{\partial y_k} = p_k^*, \end{aligned} \tag{46}$$

where we have used the KKT condition (26) in the second equality and the fact that $\sum_{\ell=1}^K p_\ell^* = P$ in the last one. The dynamics (24) then yield

$$\frac{dh_0^*}{dt} = \sum\nolimits_k \frac{\partial h_0^*}{\partial y_k} \dot{y}_k = \sum\nolimits_k p_k v_k, \tag{47}$$

so, after integrating, we obtain

$$\begin{aligned} \int_0^t \sum\nolimits_k p_k(s) v_k(s)\, ds &= h_0^*(\mathbf{y}(t)) - h_0^*(\mathbf{y}_0) \\ &\geq \mathbf{p}_0 \cdot \mathbf{y}(t) - h_0(\mathbf{p}_0) - h_0^*(\mathbf{y}_0), \end{aligned} \tag{48}$$

where $\mathbf{y}_0$ is the initial condition of the dynamics (24) and we have used the fact that $h_0^*(\mathbf{y}) \geq \mathbf{p} \cdot \mathbf{y} - h(\mathbf{p})$ for all $\mathbf{p} \in \mathcal{X}_0$. Thus, noting that $\mathbf{y}(t) = \int_0^t \mathbf{v}(s)\, ds$ where $\mathbf{v} = \nabla_{\mathbf{p}} \Psi$ is the gradient (25) of Ψ w.r.t. $\mathbf{p}$, we obtain:

$$\int_0^t \big[\mathbf{p}_0 - \mathbf{p}(s)\big] \cdot \mathbf{v}(s)\, ds \leq h_0^*(\mathbf{y}_0) + h_0(\mathbf{p}_0). \tag{49}$$

Our claim then follows by recalling that $\Psi(\cdot\,; t)$ is (strictly) concave, so $\Psi(\mathbf{p}_0; s) - \Psi(\mathbf{p}; s) \leq (\mathbf{p}_0 - \mathbf{p}) \cdot \mathbf{v}$ for all $\mathbf{p} \in \mathcal{X}_0$ and for all $s \geq 0$ (simply take $C = h_0^*(\mathbf{y}_0) + h_0(\mathbf{p}_0)$). □

PROOF OF THEOREM 3.5. Let $H(\mathbf{Q}, \mathbf{p}) = h_+(\mathbf{Q}) + h_0(\mathbf{p}) = \sum_k \big[\operatorname{tr}(\mathbf{Q}_k \log \mathbf{Q}_k) + \theta(p_k) + \theta(P_k - p_k)\big]$ denote the aggregate entropy over the state space $\mathcal{X} = \mathcal{X}_+ \times \mathcal{X}_0$ of (RM), and consider the associated Legendre–Fenchel problem:

$$\begin{aligned} \text{maximize} \quad & \operatorname{tr}[\mathbf{Y}\mathbf{Q}] + \mathbf{y} \cdot \mathbf{p} - H(\mathbf{Q}, \mathbf{p}), \\ \text{subject to} \quad & \mathbf{Q} \in \mathcal{X}_+, \mathbf{p} \in \mathcal{X}_0. \end{aligned} \tag{LF}$$

Clearly, (LF) may be decomposed as a sum of (LF$_0$) and K copies of the positive-definite problem (LF$_+$) (one for each component $\mathcal{D}_k$ in $\mathcal{X}_+$ – i.e. one copy per subcarrier $k \in \mathcal{K}$), so the convex conjugate of H will be:

$$H^*(\mathbf{Y}, \mathbf{y}) = h_+^*(\mathbf{Y}) + h_0^*(\mathbf{y}), \tag{50}$$

with h_+^* and h_0^* given by (20b) and (45) respectively. Accordingly, following the same steps as in the proof of Theorems 3.2 and 3.4, we obtain:

$$\frac{dH^*}{dt} = \operatorname{tr}[\mathbf{Q}\mathbf{V}] + \mathbf{p} \cdot \mathbf{v}, \tag{51}$$

and hence, combining (39) and (49):

$$\begin{aligned} \int_0^t \operatorname{tr}[(\mathbf{Q}_0 - \mathbf{Q}(s))\mathbf{V}(s)]\, ds + \int_0^t \big[\mathbf{p}_0 - \mathbf{p}(s)\big] \cdot \mathbf{v}(s)\, ds \\ \leq H^*(\mathbf{Y}_0, \mathbf{y}_0) + H(\mathbf{Q}_0, \mathbf{p}_0), \end{aligned} \tag{52}$$

where $\mathbf{Q}_0 \in \mathcal{X}_+$ and $\mathbf{p}_0 \in \mathcal{X}_0$ are the covariance and power allocation components of $\mathbf{P}_0 \in \mathcal{X}$ respectively, and $(\mathbf{Y}_0, \mathbf{y}_0)$ are the initial conditions of (30). Thus, with $\mathbf{V}$ and $\mathbf{v}$ being the gradient of $\Psi(\mathbf{Q}, \mathbf{p}; t)$ along $\mathbf{Q}$ and $\mathbf{p}$ respectively, and with $\Psi(\cdot\,; t)$ concave, our claim follows as in the last part of the proof of Theorems 3.2 and 3.4. □

References

[1] F. Alvarez, J. Bolte, and O. Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004.

[2] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE J. Sel. Areas Commun.*, 29(4):731–745, April 2011.

[3] H. Bölcskei, D. Gesbert, and A. J. Paulraj. On the capacity of OFDM-based spatial multiplexing systems. *IEEE Trans. Commun.*, 50(2):225–234, February 2002.

[4] G. Calcev, D. Chizhik, B. Göransson, S. Howard, H. Huang, A. Kogiantis, A. F. Molisch, A. L. Moustakas, D. Reed, and H. Xu. A wideband spatial channel model for system-wide simulations. *IEEE Trans. Veh. Technol.*, 56(2):389, March 2007.

[5] C. D. Cantrell. *Modern mathematical methods for physicists and engineers*. Cambridge University Press, Cambridge, UK, 2000.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[7] P. Coucheney, B. Gaujal, and P. Mertikopoulos. Entropy-driven dynamics and robust learning procedures in games. http://arxiv.org/abs/1303.2270.

[8] FCC Spectrum Policy Task Force. Report of the spectrum efficiency working group. Technical report, Federal Communications Comission, November 2002.

[9] G. J. Foschini and M. J. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6:311–335, 1998.

[10] Y. Gai, B. Krishnamachari, and R. Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *DySPAN '10: Proceedings of the 2010 IEEE Symposium on Dynamic Spectrum Access Networks*, 2010.

[11] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa. Breaking spectrum gridlock with cognitive radios: An information theoretic perspective. *Proc. IEEE*, 97(5):894–914, 2009.

[12] S. Haykin. Cognitive radio: Brain-empowered wireless communications. *IEEE J. Sel. Areas Commun.*, 23(2):201–220, February 2005.

[13] J. Hofbauer, S. Sorin, and Y. Viossat. Time average replicator and best reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, May 2009.

[14] J. Huang and Z. Han. *Cognitive Radio Networks: Architectures, Protocols, and Standards*, chapter Game theory for spectrum sharing. Auerbach Publications, CRC Press, 2010.

[15] L. D. Landau and E. M. Lifshitz. Statistical physics. In *Course of Theoretical Physics*, volume 5. Pergamon Press, Oxford, 1976.

[16] K. B. Letaief and Y. J. A. Zhang. Dynamic multiuser resource allocation and adaptation for wireless systems. *Wireless Communications, IEEE*, 13(4):38–47, August 2006.

[17] H. Li. Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case. In *SMC '09: Proceedings of the 2009 International Conference on Systems, Man and Cybernetics*, pages 1893–1898, 2009.

[18] P. Mertikopoulos, E. V. Belmega, and A. L. Moustakas. Matrix exponential learning: Distributed optimization in MIMO systems. In *ISIT '12: Proceedings of the 2012 IEEE International Symposium on Information Theory*, pages 3028–3032, 2012.

[19] P. Mertikopoulos, E. V. Belmega, A. L. Moustakas, and S. Lasaulce. Dynamic power allocation games in parallel multiple access channels. In *ValueTools '11: Proceedings of the 5th International Conference on Performance Evaluation Methodologies and Tools*, 2011.

[20] J. Mitola III and G. Q. Maquire Jr. Cognitive radio for flexible mobile multimedia communication. *IEEE Personal Commun. Mag.*, 6(4):13–18, aug 1999.

[21] N. Nie and C. Comaniciu. Adaptive channel allocation spectrum etiquette for cognitive radio networks. In *DySPAN '05: Proceedings of the 2005 IEEE Symposium on Dynamic Spectrum Access Networks*, pages 269–278, 2005.

[22] D. P. Palomar, J. M. Cioffi, and M. Lagunas. Uniform power allocation in MIMO channels: a game-theoretic approach. *IEEE Trans. Inf. Theory*, 49(7):1707, July 2003.

[23] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[24] W. H. Sandholm. *Population Games and Evolutionary Dynamics*. Economic learning and social evolution. MIT Press, Cambridge, MA, 2010.

[25] K. V. Schinasi. Spectrum management: Better knowledge needed to take advantage of technologies that may improve spectrum efficiency. Technical report, United States General Accounting Office, May 2004.

[26] G. Scutari and D. P. Palomar. MIMO cognitive radio: A game theoretical approach. *IEEE Trans. Signal Process.*, 58(2):761–780, February 2010.

[27] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[28] S. Sorin. Exponential weight algorithm in continuous time. *Mathematical Programming*, 116(1):513–528, 2009.

[29] C. R. Stevenson, G. Chouinard, Z. Lei, W. Hu, and S. J. Shellhammer. IEEE 802.22: The first cognitive radio wireless regional area network standard. *IEEE Commun. Mag.*, 47(1):130–138, jan 2009.

[30] P. D. Taylor and L. B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

[31] I. E. Telatar. Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications and Related Technologies*, 10(6):585–596, 1999.

[32] J. Wang, G. Scutari, and D. P. Palomar. Robust MIMO cognitive radio via game theory. *IEEE Trans. Signal Process.*, 59(3):1183–1201, March 2011.

[33] Y. J. A. Zhang and M.-C. A. So. Optimal spectrum sharing in MIMO cognitive radio networks via semidefinite programming. *IEEE J. Sel. Areas Commun.*, 29(2):362–373, 2011.

[34] Q. Zhao and B. M. Sadler. A survey of dynamic spectrum access. *IEEE Signal Process. Mag.*, 24(3):79–89, May 2007.