Transmit without Regrets: Online Optimization in MIMO–OFDM Cognitive Radio Systems

Panayotis Mertikopoulos, *Member, IEEE*, and E. Veronica Belmega

***Abstract*—In this paper, we examine cognitive radio systems that evolve dynamically over time due to changing user and environmental conditions. To combine the advantages of orthogonal frequency division multiplexing (OFDM) and multiple-input, multiple-output (MIMO) technologies, we consider a MIMO–OFDM cognitive radio network where wireless users with multiple antennas communicate over several non-interfering frequency bands. As the network's primary users (PUs) come and go in the system, the communication environment changes constantly (and, in many cases, randomly); accordingly, the network's unlicensed, secondary users (SUs) must adapt their transmit profiles "on the fly" in order to maximize their data rate in a rapidly evolving environment over which they have no control. In this dynamic setting, static solution concepts (such as Nash equilibrium) are no longer relevant, so we focus on dynamic transmit policies that lead to *no regret*, i.e. that perform at least as well as (and typically outperform) even the best fixed transmit profile throughout the entire transmission horizon, and irrespective of the systems' evolution over time. Drawing on the method of matrix exponential learning, we derive a no-regret transmit policy for the system's SUs which relies only on local channel state information (CSI); as a result, the system's SUs are able to track their individually optimum transmit profiles as they evolve over time remarkably well, even under rapidly (and randomly) changing conditions. Importantly, the proposed augmented exponential learning (AXL) policy retains its no-regret properties even if the SUs' channel measurements are subject to arbitrarily large observation errors (the imperfect CSI case), thus ensuring the method's robustness in the presence of uncertainties.**

***Index Terms*—Cognitive radio; exponential learning; MIMO; OFDM; regret minimization; online optimization.**

I. Introduction

THE explosive spread of Internet-enabled mobile devices has turned the radio spectrum into a scarce resource which, if not managed properly, may soon 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 [1, 2].

In view of the above, the emerging paradigm of cognitive radio (CR) has attracted considerable interest as a promising counter to spectrum scarcity [3–6]. 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 try to communicate under the constraints imposed by the PUs (the downside being that the SUs have no quality of service (QoS) guarantees since the PUs' protection is paramount). In this way, by opening up the unfilled "white spaces" of the licensed spectrum to opportunistic radio access, the overall utilization of the wireless medium 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 [7, 8] opens up additional ways 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 per Hertz without requiring the reallocation (and subsequent re-regulation) of additional frequency bands.

In this paper, we combine these two approaches and focus on dynamic MIMO cognitive radio systems comprising several wireless users (primary and secondary alike) who communicate over multiple non-interfering channels. In this evolving (and unregulated) 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. As a result, given that the system's SUs cannot rely on contractual QoS guarantees to achieve their desired throughput levels, the maximization of their achievable transmission rates under the operational constraints imposed by the network's PUs becomes a critical issue.

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 [8]. However, since users cannot be 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, etc.),

Part of this work was presented at the 7th International Conference on Performance Evaluation and Tools (ValueTools 2013), Turin, Italy, Dec. 2013.

P. Mertikopoulos is with the French National Center for Scientific Research (CNRS) and the Laboratoire d'Informatique de Grenoble, Grenoble, France; E.V. Belmega is with ETIS/ENSEA – Université de Cergy-Pontoise – CNRS, Cergy-Pontoise, France.

they must optimize their signal characteristics "on the fly", based only on locally available information. Our overall aim will thus be to derive a dynamic transmit policy that allows the system's SUs to adapt to changes in the wireless medium and to track their individually optimum transmission profiles using only local (and possibly imperfect) channel state information (CSI).

This setting is fairly general in scope as it involves cognitive SUs with significant control over both spatial and spectral degrees of freedom: in the spatial (MIMO) component, the users can control the covariance of their transmit directions (essentially the spread of their symbols over the transmitting antennas), whereas in the frequency domain (the OFDM component), they control the allocation of their transmit power over the different channels at their disposal. To the best of our knowledge, only special cases of this problem have been considered in a CR setting: for instance, [9–11] analyzed the case where there is only one channel and the environment is *static* (i.e. the system's SUs only react to each other and the PUs' spectrum utilization is fixed); in this context, [9] characterized the best spatial covariance profile for the interacting SUs whereas [10, 11] described how to reach a Nash equilibrium in the resulting non-cooperative game. On the other hand, in *dynamic* environments where the PUs' evolving behavior cannot be anticipated by the system's SUs, [12–15] proposed different learning schemes for optimal channel selection, but they only considered the case where the SUs are equipped with a single antenna and cannot split power across subcarriers.

Extending the above considerations, our goal in this paper will be to derive an adaptive transmit policy for SU rate optimization in dynamically evolving MIMO–OFDM cognitive radio networks. In this online optimization framework, the most widely used performance criterion is that of *regret minimization*, a concept which was first introduced by Hannan [16] and which has since given rise to a vigorous literature at the interface of optimization, statistics, game theory, and machine learning – see e.g. [17, 18] for a comprehensive survey. Specifically, in the language of game theory, the notion of regret compares the cumulative payoff obtained by an agent who changes actions based on how his environment evolves over time to the cumulative payoff that he would have obtained by constantly playing the same action. Accordingly, the purpose of regret minimization is to devise learning policies that lead to vanishingly small regret against *any* fixed action and *irrespective* of how the agent's environment evolves over time.

We will thus focus on *no-regret* policies that perform at least as well as the asymptotically best fixed policy in terms of each user's achievable transmission rate – despite the fact that the latter cannot be determined by the SUs when they have no means to anticipate the PUs' behavior. Motivated by the no-regret properties of the so-called exponential weight (EW) algorithm for problems with discrete action sets [17, 19–21], we propose an augmented exponential learning (AXL) approach that can be applied to the continuous regret minimization problem at hand with minimal information requirements. A key challenge here is that any learning algorithm must respect the problem's semidefiniteness constraints; as such, an important component of the AXL algorithm is the continuous-time technique of *matrix exponential learning* that was recently introduced for ordinary (as opposed to online) rate optimization problems in MIMO multiple access channels (MACs) [22] – and which is in turn related to the matrix regularization techniques of [23].

Of course, since the SUs' optimal transmit profile varies over time, the notions of convergence and/or convergence speed are no longer applicable; instead, the figure of merit will be the rate at which the SUs attain a no-regret state. In that respect, AXL guarantees a worst-case average regret of $\mathcal{O}(T^{-1/2})$ after T epochs, a bound which is well known to be optimal in machine learning [17]. Additionally, AXL retains its no-regret properties even if the SUs' channel measurements are subject to arbitrarily large observation errors (the imperfect CSI case), thus providing significant performance improvements over more traditional water-filling methods that are critically sensitive to perfect CSI. As a result, the system's SUs are able to track their individually optimum transmit profile as it evolves over time remarkably well, even under rapidly (and randomly) changing conditions.

Paper Outline and Summary of Results

The breakdown of our paper is as follows: in Section II, we introduce our MIMO–OFDM cognitive radio network model and the notion of a no-regret transmission policy in the context of optimizing the SUs' individual transmission rates. In Section III, we decompose this online rate optimization problem in two components, and we propose a no-regret algorithm for each one: specifically, in Section III-A we propose an adaptive power allocation policy for the problem's OFDM component, whereas in Section III-B, we derive a dynamic signal covariance policy based on matrix exponential learning for the problem's MIMO component. These components are merged in Section IV where we present our augmented exponential learning method for the general MIMO-OFDM setting and we show that it leads to no regret (Theorem 1). Importantly, we also show that the AXL algorithm retains its no-regret properties even when the user only has imperfect CSI at his disposal (Theorem 2). This theoretical analysis is validated and supplemented by numerical simulations in Section V where we also examine the users' ability to track their individually optimum transmit characteristics. To facilitate presentation, proofs and technical details have been delegated to a series of appendices at the end of the paper.

II. System Model

A. The Network Model

The cognitive radio system that we will focus on consists of a set of non-cooperative wireless MIMO users (primary and secondary alike) that communicate over several non-interfering channels (frequency bands) by means of an OFDM scheme [24, 25]. 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 [24]. Then, the aggregate signal $\mathbf{y}_k^s \in \mathbb{C}^{n_s}$ on the k-th subcarrier at the intended receiver 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 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 non-singular, zero-mean Gaussian vector). Accordingly, if we focus for simplicity on a specific SU and drop the user index $s \in \mathcal{S}$ in (1), we obtain the signal model

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

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

The covariance of $\mathbf{w}_k$ in (2) obviously changes over time e.g. due to modulations in the PUs' behavior.[1] In this setting, employing sophisticated successive interference cancellation (SIC) techniques at the receiver is highly nontrivial, especially with regards to the system's unregulated secondary users; as such, we will assume that interference by other users (primary and secondary alike) 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 be given by the familiar expression [8, 24]:

$$\Phi(\mathbf{P}) = \sum\nolimits_k \left[\log\det\left(\mathbf{W}_k + \mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\dagger\right) - \log\det\mathbf{W}_k\right], \quad (3)$$

where:

1) $\mathbf{W}_k = \mathbb{E}\left[\mathbf{w}_k\mathbf{w}_k^\dagger\right]$ is the multi-user interference-plus-noise covariance matrix over subcarrier k.
2) $\mathbf{P}_k = \mathbb{E}[\mathbf{x}_k\mathbf{x}_k^\dagger]$ is the covariance matrix of the user's transmitted signal on subcarrier k and $\mathbf{P} = \mathrm{diag}(\mathbf{P}_1, \dots, \mathbf{P}_K)$ denotes the user's transmit profile over all subcarriers.[2] In particular, we will write for convenience:

$$\mathbf{P}_k = p_k\mathbf{Q}_k, \quad (4)$$

where $p_k = \mathbb{E}[\mathbf{x}_k^\dagger\mathbf{x}_k]$ denotes the user's *transmit power* over subcarrier k and $\mathbf{Q}_k = \mathbb{E}\left[\mathbf{x}_k\mathbf{x}_k^\dagger\right]/\,\mathbb{E}\left[\mathbf{x}_k^\dagger\mathbf{x}_k\right]$ is the corresponding *normalized* signal covariance matrix.

Hence, given that $\mathbf{W}_k$ might change over time due to evolving user conditions, we obtain the *time-dependent* objective:

$$\Phi(\mathbf{P};t) = \sum\nolimits_k \log\det\left[\mathbf{I} + \widetilde{\mathbf{H}}_k(t)\,\mathbf{P}_k\,\widetilde{\mathbf{H}}_k^\dagger(t)\right], \quad (5)$$

where the *effective channel matrices* $\widetilde{\mathbf{H}}_k$ are given by

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

and the time variable $t = 1, 2, \dots$ will be assumed discrete (for instance, corresponding to the epochs of a time-slotted system). Obviously, since we are putting no constraints on the behavior of the system's users (who may connect and disconnect from the system and/or otherwise modulate their transmit profiles based on arbitrary criteria), the evolution of the effective channel matrices $\widetilde{\mathbf{H}}_k(t)$ over time can be quite arbitrary as well; the only assumption that we will make is that the matrices $\widetilde{\mathbf{H}}_k(t)$ remain bounded for all time.

In light of the above, and 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 [1, 2, 10, 26, 27], we will consider the following constraints for the SUs' transmit profiles:

1) Constrained total transmit power:

$$\mathrm{tr}(\mathbf{P}) = \sum\nolimits_k p_k \leq P. \quad (7a)$$

2) Constrained transmit power per subcarrier:

$$\mathrm{tr}(\mathbf{P}_k) = p_k \leq P_k. \quad (7b)$$

3) Null-shaping constraints:

$$\mathbf{U}_k^\dagger\mathbf{P}_k = 0, \quad (7c)$$

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 e.g. [25] for a more detailed discussion).

Of course, to maximize (5) in the absence of energy awareness considerations, the user will saturate his total power constraint (7a) by transmitting at the highest possible (total) power.[3] Thus, after a suitable change of basis, the set of admissible transmit profiles for the rate function (5) may be expressed as:

$$\begin{aligned}\mathcal{X} = \{\,\mathrm{diag}(\mathbf{P}_1, \dots, \mathbf{P}_K) :\ & \mathbf{P}_k \in \mathbb{C}^{m_k\times m_k}, \\ & \mathbf{P}_k \succcurlyeq 0, 0 \leq \mathrm{tr}(\mathbf{P}_k) \leq P_k \text{ and } \sum\nolimits_k \mathrm{tr}(\mathbf{P}_k) = P\}, \end{aligned} \quad (8)$$

where $m_k \equiv \mathrm{nullity}(\mathbf{U}_k)$ 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 in (4), we obtain the decomposition $\mathcal{X} = \mathcal{X}_0 \times \prod_k \mathcal{D}_k$ where $\mathcal{X}_0 = \left\{\mathbf{p} \in \mathbb{R}^K :\ 0 \leq p_k \leq P_k, \sum_k p_k = P\right\}$ denotes the set of admissible *power allocation vectors* and $\mathcal{D}_k = \{\mathbf{Q}_k \in \mathbb{C}^{m_k\times m_k} :\ \mathbf{Q}_k \succcurlyeq 0, \mathrm{tr}(\mathbf{Q}_k) = 1\}$ is the set of admissible *normalized covariance matrices* for subcarrier k. The individual objective of the focal SU at time t will thus be given by the *online rate maximization problem*:

$$\begin{aligned} &\text{maximize} && \Phi(\mathbf{P};t) \\ &\text{subject to} && \begin{cases}\mathbf{P} = \mathrm{diag}(p_1\mathbf{Q}_1, \dots, p_K\mathbf{Q}_K), \\ (p_1, \dots, p_K) \in \mathcal{X}_0,\ \mathbf{Q}_k \in \mathcal{D}_k.\end{cases} \end{aligned} \quad \text{(ORM)}$$

Remark. In the following sections, we will need the derivatives of the rate function Φ; to that end, some matrix calculus yields

$$\frac{\partial\Phi}{\partial\mathbf{P}_k^*} \equiv \mathbf{M}_k = \widetilde{\mathbf{H}}_k^\dagger[\mathbf{I} + \widetilde{\mathbf{H}}_k\mathbf{P}_k\widetilde{\mathbf{H}}_k^\dagger]^{-1}\widetilde{\mathbf{H}}_k, \quad (9)$$

[1] In all cases, 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 [8].

[2] Throughout this paper, $\mathrm{diag}(\mathbf{A}_1, \dots, \mathbf{A}_K)$ will denote the block-diagonal (direct) sum of the matrices $\mathbf{A}_k$.

[3] The online optimization techniques that we will present can be extended to more general energy-aware objectives where (7a) is not saturated, but we will not do so due to space limitations.

where $\mathbf{P}_k^*$ denotes the complex conjugate of $\mathbf{P}_k$. Since the effective channel matrices $\widetilde{\mathbf{H}}_k(t)$ are assumed bounded for all t, the derivatives of $\Phi(\cdot;t)$ with respect to p_k and $\mathbf{Q}_k$ will also remain bounded. Quantitatively, we will thus assume that there exists some $M > 0$ such that

$$\|\mathbf{M}_k\| \leq M \quad \text{for all } k \in \mathcal{K} \text{ and for all } \mathbf{P} \in \mathcal{X}, \tag{10}$$

where $\|\cdot\| \equiv \|\cdot\|_\infty$ denotes the uniform norm on $\mathbb{C}^{m_k \times m_k}$.

B. Online Optimization and Regret Minimization

In our setting, there is no direct causal link between the PUs' behavior and the choices of the SUs, so the effective channel matrices $\widetilde{\mathbf{H}}_k$ (and, hence, the objective function Φ) may change arbitrarily over time. This leads to a "game against nature" which evolves as follows:

1) At each time slot $t = 1, 2 \ldots$, the *agent* (i.e. the focal SU) selects an *action* (transmit profile) $\mathbf{P}(t) \in \mathcal{X}$.
2) The agent's *payoff* (transmission rate) $\Phi(\mathbf{P}(t); t)$ is determined by nature and/or the behavior of other users (via the effective channel matrices $\widetilde{\mathbf{H}}_k$).
3) The agent employs some *decision rule* (dynamic transmit policy) to pick a new transmit profile $\mathbf{P}(t+1) \in \mathcal{X}$ at stage $t+1$, and the process is repeated ad infinitum (or until the user's transmission ends).

In this setting, the worst-case scenario for the user – and one which has attracted considerable interest in the literature – is when the environment cannot be assumed to follow some fixed probability law. In particular, Cover's impossibility result [28] shows that the cumulative payoff difference between an *oracle* (a decision rule which prescribes an action based on knowledge of the future) and any *adaptive policy* (a decision rule which only relies on past observations) can become arbitrarily large, even in relatively simple problems (such as trying to predict a binary sequence). As a result, in the absence of absolute performance guarantees, and given that static solution concepts (such as Nash equilibria) are no longer applicable, the most widely used online optimization criterion is that of *regret minimization*, a notion which was first introduced by Hannan [16] and which has since given rise to an extremely active field of research at the interface of optimization, statistics and theoretical computer science – see e.g. [17, 18] for a survey.

Roughly speaking, the regret of a dynamic policy compares the average payoff obtained by an agent that follows it to the average payoff that he would have obtained by constantly choosing the same action over the entire transmission horizon. Specifically, the *cumulative regret* of the dynamic policy $\mathbf{P}(t) \in \mathcal{X}$ with respect to $\mathbf{P}_0 \in \mathcal{X}$ will thus be:

$$\mathrm{Reg}_T(\mathbf{P}_0) = \sum\nolimits_{t=1}^{T} \left[\Phi(\mathbf{P}_0; t) - \Phi(\mathbf{P}(t); t)\right], \tag{11}$$

i.e. $\mathrm{Reg}_T(\mathbf{P}_0)$ measures the cumulative transmission rate difference up to stage T between the benchmark transmit profile $\mathbf{P}_0 \in \mathcal{X}$ and the dynamic policy $\mathbf{P}(t)$. The corresponding *average regret* will then simply be $T^{-1}\,\mathrm{Reg}_T(\mathbf{P}_0)$ and the goal of regret minimization is to devise a dynamic transmit policy $\mathbf{P}(t)$ that leads to *no regret*, viz.

$$\limsup_{T\to\infty} \frac{1}{T}\,\mathrm{Reg}_T(\mathbf{P}_0) \leq 0, \tag{12}$$

for all $\mathbf{P}_0 \in \mathcal{X}$ and irrespective of the evolution of the objective $\Phi(\cdot;t)$ as a function of the effective channel matrices $\widetilde{\mathbf{H}}_k(t)$. In other words, if we interpret $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\Phi(\mathbf{P}_0;t)$ as the long-term average transmission rate associated to $\mathbf{P}_0$, then (12) means that the average data rate of the dynamic transmit policy $\mathbf{P}(t)$ must be at least as good as that of *any* benchmark profile $\mathbf{P}_0 \in \mathcal{X}$.

Remark 1. Obviously, if the optimum transmit policy which maximizes (ORM) could be predicted at every stage $t = 1, 2, \ldots$ in an oracle-like fashion, we would have $\mathrm{Reg}_T(\mathbf{P}_0) \leq 0$ in (11) for all $\mathbf{P}_0 \in \mathcal{X}$. The no-regret requirement (12) is thus fundamental for performance evaluation in the context of online optimization because negative regret is a key indicator of tracking the maximum of (ORM) as it evolves over time.

Remark 2. If the channel matrices are drawn at each realization from an *isotropic* distribution, spreading power uniformly across carriers and antennas is the optimal choice when nature (including the network's PUs) is actively choosing the worst possible channel realization for the transmitter [29]. A no-regret policy extends this "min-max" concept by ensuring that *no matter* how the channels evolve over time (isotropically, adversarially, or otherwise), the policy's achieved transmission rate will be asymptotically as good as that of any fixed transmit profile, including the uniform one (as a special case where nature is actively playing against the transmitter – e.g. jamming).

III. Online Power Allocation and Signal Covariance Optimization

To build intuition step-by-step, we will break up the online rate maximization problem (ORM) in simpler components and we will derive adaptive transmit policies based on an exponential learning principle that leads to no regret in each one. This analysis will then be merged into an adaptive transmit policy for the full MIMO–OFDM problem in the following section.

A. The OFDM Component: Online Power Allocation

1) A gentle start – the case $P_k \geq P$: For illustration purposes, we first examine the case where the power-per-channel constraints (7b) can be absorbed in the total power constraint (7a), i.e. $P_k \geq P$ for all $k \in \mathcal{K}$; also, for scaling purposes, it will be more convenient to consider the normalized power variables

$$q_k = p_k / P. \tag{13}$$

With this in mind, if the normalized signal covariance profile $\mathbf{Q} = \mathrm{diag}(\mathbf{Q}_1, \ldots, \mathbf{Q}_K)$ of the focal SU is kept fixed, we obtain the online power allocation problem:

$$\begin{aligned} &\text{maximize} && \Phi(\mathbf{q}; t), \\ &\text{subject to} && \mathbf{q} \in \Delta \end{aligned} \tag{OPA}$$

where $\Delta = \{\mathbf{q} \in \mathbb{R}_+^K : \sum_{k=1}^K q_k = 1\}$ denotes the set of feasible (normalized) power allocation profiles, and in a slight abuse of notation, we write $\Phi(\mathbf{q}; t)$ to highlight the dependence of the rate function (5) on the normalized power allocation profile $\mathbf{q} \in \Delta$ instead of $\mathbf{P} \in \mathcal{X}$.

A special case of this problem is when the user cannot split power across subcarriers and can only choose one channel on which to transmit. Essentially, this channel selection framework boils down to the well known "multi-armed bandit" problem [30] which has given rise to a vast corpus of literature on learning algorithms – see e.g. [17] and references therein. As a result, much recent work on CR networks [13–15] has been focused on no-regret channel selection algorithms that lead to no regret by utilizing Q-learning [14] or upper confidence bound (UCB) techniques [13].

Unfortunately, these techniques are inherently tied to discrete problems, so it is not clear how to extend them to the continuous context of (OPA). Instead, motivated by the exponential weight algorithm introduced in [19–21] for sequence prediction, our approach will consist of scoring each channel over time and then allocating power proportionally to the exponential of these scores. In particular, inspired by the analysis of [31], each channel will be scored by means of the *gradient payoffs*:

$$v_k = \frac{\partial \Phi}{\partial q_k} = P \frac{\partial \Phi}{\partial p_k} = P \cdot \operatorname{tr}[\mathbf{M}_k \mathbf{Q}_k], \tag{14}$$

where $\mathbf{Q}_k \in \mathcal{D}_k$ is the user's (fixed) covariance matrix and $\mathbf{M}_k$ is given by (9). Our exponential learning power allocation policy will then consist of the recursion:

$$\begin{aligned} y_k(t) &= y_k(t-1) + v_k(t), \\ q_k(t) &= \frac{\exp(\eta t^{-1/2} y_k(t))}{\sum_\ell \exp(\eta t^{-1/2} y_\ell(t))}, \end{aligned} \tag{XL-PA}$$

where $\eta > 0$ is a learning rate parameter and the $\sqrt{t}$ factor has been included to moderate very sharp score differences.

Our first result is that (XL-PA) performs asymptotically as well as *any* fixed power allocation profile $\mathbf{q}_0 \in \Delta$:

Proposition 1. *If $P_k \geq P$ for all $k \in \mathcal{K}$, (XL-PA) leads to no regret in the online power allocation problem (OPA). Specifically, for every $\mathbf{q}_0 \in \Delta$, and independently of the system's evolution over time, the user's regret is bounded by:*

$$\frac{1}{T} \operatorname{Reg}_T(\mathbf{q}_0) \leq \frac{1}{\sqrt{T}} \left(\frac{\log K}{\eta} + 4P^2 M^2 \eta \right), \tag{15}$$

with M given by (10).

Proof: See Appendices A and E. ∎

Remark 1. Since the focal SU transmits with positive power on every available subcarrier, we will assume that he can obtain the relevant CSI needed to calculate the payoffs (14). That said, the user's CSI might well be imperfect, in which case Propositon 1 does not apply; we will study the case of imperfect CSI for the full MIMO–OFDM problem in Section IV.

Remark 2. The use of the gradient-based payoffs (14) in the exponential learning policy (XL-PA) can be compared to the online gradient descent algorithm introduced in [32] where the learner tracks the gradient of his evolving objective and projects back to the problem's feasible set when needed. We did not take such an approach because projections are unstable numerical operations [33] and they can also become quite costly from a computational standpoint (the problem's constraints would have to be checked individually at every iteration).

2) The general case: The dynamic power allocation policy (XL-PA) and Proposition 1 concern the case where the power-per-channel constraints (7b) can be absorbed in the total power constraint (7a). Otherwise, if $P_k < P$ for some channel $k \in \mathcal{K}$ (e.g. if certain PUs have very low interference tolerance on their licensed channels), (XL-PA) cannot be employed "as is" because it does not respect the constraint $p_k \leq P_k$. When this is the case, the analysis of Appendix B yields the *modified* exponential learning policy:

$$\begin{aligned} y_k(t) &= y_k(t-1) + v_k(t), \\ p_k(t) &= P_k \left(1 + \exp(\lambda - \eta t^{-1/2} y_k)\right)^{-1} \end{aligned} \tag{XL-PA$'$}$$

where $\lambda > 0$ is defined implicitly so that (7a) is satisfied:

$$P = \textstyle\sum_{k \in \mathcal{K}} P_k \left(1 + \exp(\lambda - \eta t^{-1/2} y_k)\right)^{-1}. \tag{16}$$

Just like (XL-PA), the (XL-PA′) policy exhibits exponential sensitivity to the scores y_k modulo a normalization factor corresponding to the constraints (7a) and (7b). Since the RHS of (16) is bounded below and strictly decreasing in λ, it is straightforward to calculate the value of λ itself, e.g. by performing a low-complexity line search for e^λ [33].[4] We then get:

Proposition 2. *The policy (XL-PA′) leads to no regret. In particular, for every $\mathbf{p}_0 \in \mathcal{X}_0$, the user's regret is bounded by*

$$T^{-1} \operatorname{Reg}_T(\mathbf{p}_0) \leq \mathcal{O}(T^{-1/2}), \tag{17}$$

irrespective of the system's evolution over time.

Proof: See Appendix B. ∎

Remark. We should note here that (XL-PA′) is not equivalent to (XL-PA) if $P_k \geq P$; instead, (XL-PA) should be viewed as a simpler alternative to (XL-PA′) that can be employed whenever the maximum power-per-channel constraints (7b) can be subsumed in the total power constraint (7a). For convenience, we will present our results in the simpler case $P_k \geq P$ and we will rely on a series of remarks to translate these remarks to the regime $P_k < P$ (cf. Appendices A and B).

B. The MIMO Component: Online Covariance Optimization

Dually to the analysis of the previous section, if the user's power allocation profile $\mathbf{p} = (p_1, \dots, p_K)$ remains fixed throughout the duration of the transmission, (ORM) boils down to the online signal covariance optimization problem:

$$\begin{aligned} &\text{maximize} && \Phi(\mathbf{Q}; t), \\ &\text{subject to} && \mathbf{Q}_k \succcurlyeq 0, \ \operatorname{tr}(\mathbf{Q}_k) = 1, \end{aligned} \tag{OCOV}$$

where we now use the notation $\Phi(\mathbf{Q}; t)$ to highlight the dependence of the user's transmission rate (5) on the normalized covariance matrix $\mathbf{Q} = \operatorname{diag}(\mathbf{Q}_1, \dots, \mathbf{Q}_K) \in \mathcal{X}_+ \equiv \prod_k \mathcal{D}_k$.

A key challenge in (OCOV) is that any learning algorithm must respect the problem's (implicit) semidefiniteness constraints $\mathbf{Q}_k \succcurlyeq 0$. To that end, motivated by the analysis of [22], we will consider the *matrix exponential learning* policy

$$\begin{aligned} \mathbf{Y}_k(t) &= \mathbf{Y}_k(t-1) + \mathbf{V}_k(t), \\ \mathbf{Q}_k(t) &= \frac{\exp(\eta t^{-1/2} \mathbf{Y}_k(t))}{\operatorname{tr}[\exp(\eta t^{-1/2} \mathbf{Y}_k(t))]}, \end{aligned} \tag{XL-COV}$$

[4] For a closed-form expression of (XL-PA′) based on a modified version of the replicator equation of evolutionary game theory, see [34].

where the matrix-valued gradient payoff $\mathbf{V}_k$ is defined as:

$$\mathbf{V}_k = \frac{\partial \Phi}{\partial \mathbf{Q}_k^*} = p_k \mathbf{M}_k, \tag{18}$$

with $\mathbf{M}_k$ given by (9).Intuitively, (XL-COV) reinforces the spatial directions that peform well by increasing the corresponding eigenvalues (the $t^{-1/2}$ factor simply keeps the spectrum of $\mathbf{Y}_k$ from growing too fast). Along these lines, our analysis in Appendix C yields:

Proposition 3. *The dynamic transmit policy* (XL-COV) *leads to no regret in the online signal covariance optimization problem* (OCOV)*. In particular, for every $\mathbf{Q}_0 \in \mathcal{X}_+ \equiv \prod_k \mathcal{D}_k$, and irrespective of the system's evolution over time, we will have:*

$$\frac{1}{T}\operatorname{Reg}_T(\mathbf{Q}_0) \leq \frac{1}{\sqrt{T}}\left(\frac{\sum_{k=1}^K \log m_k}{\eta} + 4P^2M^2\eta\right), \tag{19}$$

where m_k is the number of spatial degrees of freedom left open to SUs on subcarrier k by the constraint (7c).

Remark. Assuming a *static* environment, [22] showed that matrix exponential learning allows users to optimize their sum rate (under successive interference cancellation) in uplink MIMO multiple access channels. In this sense, Proposition 3 can be seen as a dual result: it shows that if a user follows the dynamic signal covariance policy (XL-COV), he will be *unilaterally* satisfied even in the much more general channel model (2), and independently of how the system evolves over time.

IV. Augmented Exponential Learning for MIMO–OFDM Systems

In this section, our goal will be to merge the componentwise analysis of the previous section into an adaptive transmit policy that leads to no regret in the full MIMO–OFDM problem (ORM), even under imperfect CSI.

A. Augmented Exponential Learning

Working for simplicity with the special case $P_k \geq P$, combining (XL-PA) and (XL-COV) gives the dynamic transmit policy:

Algorithm 1 Augmented Exponential Learning (AXL)

Parameter: $\eta > 0$.
Initialize: $t \leftarrow 0$; channel scores $y_k \leftarrow 0$, $\mathbf{Y}_k \leftarrow 0$.
Repeat
  $t \leftarrow t + 1$;
  foreach channel $k \in \mathcal{K}$ **do**
    set $\begin{cases} p_k \leftarrow P\exp(\eta t^{-1/2} y_k)/\sum_\ell \exp(\eta t^{-1/2} y_\ell); \\ \mathbf{Q}_k \leftarrow \exp(\eta t^{-1/2}\mathbf{Y}_k)/\operatorname{tr}[\exp(\eta t^{-1/2}\mathbf{Y}_k)]; \end{cases}$
  foreach channel $k \in \mathcal{K}$ **do**
    measure $\mathbf{M}_k \leftarrow \widetilde{\mathbf{H}}_k^\dagger[\mathbf{I} + p_k\widetilde{\mathbf{H}}_k\mathbf{Q}_k\widetilde{\mathbf{H}}_k^\dagger]^{-1}\widetilde{\mathbf{H}}_k$;
    update scores: $\begin{cases} y_k \leftarrow y_k + P\operatorname{tr}[\mathbf{M}_k\mathbf{Q}_k]; \\ \mathbf{Y}_k \leftarrow \mathbf{Y}_k + p_k\mathbf{M}_k; \end{cases}$
until transmission ends.

The augmented exponential learning (AXL) algorithm 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*: each SU only needs to update his individual transmit policy using local CSI (the matrices $\widetilde{\mathbf{H}}_k$).
(P2) It is *asynchronous*: there is no need for a global update timer to synchronize the system's SUs.
(P3) It is *stateless*: the SUs do not need to know the state of the system (e.g. the network's topology), and/or be aware of each other's actions.
(P4) It is *reinforcing*: the SUs tend to increase their unilateral transmission rates.

Remark 2. If the maximum power-per-channel constraints imposed on the network's SUs do not satisfy the condition $P_k \geq P$ for all $k \in \mathcal{K}$, AXL must be modified with respect to the power allocation update step: specifically, the exponential allocation rule $p_k \leftarrow P\exp(\eta t^{-1/2}y_k)/\sum_\ell \exp(\eta t^{-1/2}y_\ell)$ must be replaced by the update rule of (XL-PA′), i.e. by setting $p_k \leftarrow P_k[1 + \exp(\lambda - \eta t^{-1/2}y_k)]^{-1}$. To simplify our presentation, we will keep the assumption $P_k \geq P$ in what follows, but with the implicit understanding that if $P_k < P$ for some $k \in \mathcal{K}$, then it is the modified version of AXL that should be used instead.

With all this in mind, our main result in this section is that the AXL algorithm leads to no regret if $P_k \geq P$ for all channels:

Theorem 1. *The adaptive transmit policy generated by AXL leads to no regret in the online rate maximization problem* (ORM)*. In particular, for every fixed transmit profile $\mathbf{P}_0 \in \mathcal{X}$, and independently of how the system's rate function* (5) *evolves over time, the user's regret will be bounded by:*

$$\frac{1}{T}\operatorname{Reg}_T(\mathbf{P}_0) \leq \frac{1}{\sqrt{T}}\left(\frac{\log K + \sum_{k=1}^K \log m_k}{\eta} + 4P^2M^2\eta\right), \tag{20}$$

where M is given by (10) *and m_k is the number of spatial dimensions that are left open to SUs by the constraint* (7c).

Proof: See Appendices D and E. ■

Remark 1. As we already explained, if $P_k < P$ for some channel $k \in \mathcal{K}$, the power update step in the AXL algorithm should be replaced by the power allocation rule (XL-PA′). In this case, AXL still leads to no regret with an $\mathcal{O}(T^{-1/2})$ bound on the regret, but the exact expression is more complicated, so we will not present it here (see Appendix B for the details).

Remark 2. The proof of Theorem 1 relies on a deep connection between the exponential maps of (XL-PA) and (XL-COV) with the Gibbs–Shannon and von Neumann entropy functions respectively. In fact, as we shall see in Appendices A–B, our approach is intimately related to the Hessian–Riemannian optimization method of [35] and the online mirror descent techniques presented in [18, 23]. Unfortunately, these methods both require the introduction of significant technical apparatus, so we will not discuss them at length here; for a detailed discussion from a game-theoretic viewpoint, the reader is instead referred to [36].

Remark 3. It should also be noted that the bound (20) is not the sum of the bounds (15) and (19). As we show in Appendices D and E, the reason for this is that Theorem 1 is not a corollary of Propositions 1 and 3 but, rather, a combination of these two independent results with a descent from continuous to discrete time – a technique that was recently developed by J. Kwon and one of the authors in [37].

Remark 4. In practice, the learning parameter η of the AXL algorithm can be tuned freely by the user. As such, if the user can estimate ahead of time the quantity M (which can be seen as an effective bound on the gradient matrices $\mathbf{M}_k$ over time), η can be chosen so as to optimize the regret guarantee (20) – thus leading to lower regret levels faster.

To that end, some calculus shows that the optimal choice of η which minimizes the RHS of (20) is:

$$\eta = \tfrac{1}{2} PM \left(\log K + \textstyle\sum_k \log m_k\right)^{1/2}, \tag{21}$$

which then leads to the optimized regret guarantee:

$$\mathrm{Reg}_T(\mathbf{P}_0) \leq 4PM\big(\log K + \textstyle\sum_k \log m_k\big)^{1/2} T^{1/2}. \tag{22}$$

This bound resembles the bound derived in [23] for learning processes that stop after a predetermined number of steps; on the other hand (and in contrast to our results), unless some sort of "doubling correction" is used [17], the method proposed in [23] might lead to positive regret in an infinite horizon setting such as the one we are considering here.

B. Learning with Imperfect Channel State Information

In practice, a major challenge occurs if the user does not have perfect CSI with which to calculate the matrix gradients (9) that are needed to run the AXL algorithm. To wit, since these gradients are determined by the effective channel matrices $\widetilde{\mathbf{H}}_k = \mathbf{W}_k^{-1/2}\mathbf{H}_k$, imperfect measurements of the actual channel matrices $\mathbf{H}_k$ or of the multi-user interference-plus-noise covariance matrices $\mathbf{W}_k$ would invariably interfere with each update cycle. Accordingly, our aim in this section will be to study the robustness of AXL in the presence of observation noise.

To account for as wide a range of measurement errors as possible, we will assume here that at each update period $t = 1, 2, \ldots$, the user can only observe a noisy estimate

$$\hat{\mathbf{M}}_k(t) = \mathbf{M}_k(t) + \mathbf{\Xi}_k(t) \tag{23}$$

of $\mathbf{M}_k(t)$, where the noise process $\mathbf{\Xi}_k(t)$ represents a random observational error (not necessarily i.i.d.). Specifically, with regards to $\mathbf{\Xi}_k$, we will only assume that $\|\mathbf{\Xi}_k\| \leq \Sigma$ (a.s.) for some (arbitrarily large) $\Sigma > 0$ and that

$$\mathbb{E}\left[\mathbf{\Xi}_k(t) | \mathcal{F}_{t-1}\right] = 0, \tag{24}$$

where $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 1}$ denotes the history of the user's choices. Remarkably, as long as there is no systematic bias in the user's measurements, the AXL algorithm still leads to no regret, even in the presence of *arbitrarily large* observation errors:

Theorem 2. *The AXL algorithm with noisy observations $\hat{\mathbf{M}}_k$ of the form (23) leads to no regret (a.s.). Specifically, if $\|\mathbf{\Xi}_k\| \leq \Sigma$, then, for all $\mathbf{P}_0 \in \mathcal{X}$ and for all $z > 0$:*

(*i*) *The user's expected regret is bounded by:*

$$\mathbb{E}\left[T^{-1}\,\mathrm{Reg}_T(\mathbf{P}_0)\right] \leq RT^{-1/2}. \tag{25}$$

(*ii*) *The user's realized regret is bounded by the perfect CSI guarantee of AXL with exponentially high probability:*

$$\mathbb{P}\left(\frac{1}{T}\,\mathrm{Reg}_T(\mathbf{P}_0) \leq \frac{R}{\sqrt{T}} + z\right) \geq 1 - \exp\left(-\frac{z^2 T}{2D^2\Sigma^2}\right), \tag{26}$$

where $D > 0$ is a constant and R is given by the deterministic regret guarantee (20) of AXL with perfect CSI, viz.:

$$R = \eta^{-1} \cdot \big(\log K + \textstyle\sum_k \log m_k\big) + 4P^2 M^2 \eta. \tag{27}$$

In short, Theorem 2 (proven in Appendix F) shows that, with high probability, AXL guarantees an $\mathcal{O}(T^{-1/2})$ bound on the user's regret, even under measurement errors of arbitrarily high magnitude. Accordingly, a few remarks are in order:

Remark 1. Part (ii) of Theorem 2 should be interpreted as a large deviations result: essentially, it states that the regret generated by AXL with imperfect CSI exceeds that of the perfect CSI variant with exponentially small probability (i.e. the tails of the large deviations distribution are effectively Gaussian). Intuitively, this means that the algorithm's regret falls below the deterministic, perfect CSI guarantee (20) with high probability. Moreover, it is also important to note that Theorem 1 is recovered by (26) in the deterministic limit $\Sigma \to 0^+$: the probability that the user's regret exceeds the determinstic guarantee $R/\sqrt{T}$ converges uniformly to 0 as $\Sigma \to 0^+$.

Remark 2. Interestingly, the first- and second-order statistics of the measured gradients $\hat{\mathbf{M}}_k$ play different roles in the presence of imperfect CSI: the expected value $\mathbb{E}[\hat{\mathbf{M}}_k] = \mathbf{M}_k$ of $\hat{\mathbf{M}}_k$ controls the expected regret guarantee of AXL via (25), whereas the variance $\mathrm{Var}(\hat{\mathbf{M}}_k) = \mathbb{E}[\|\mathbf{\Xi}_k\|^2]$ of $\hat{\mathbf{M}}_k$ controls the deviations of the regret from its "bulk" behavior – but has no impact on the expected regret of AXL.

V. Numerical Results

To validate the predictions of Section IV for the AXL algorithm, we conducted extensive numerical simulations from which we illustrate here a selection of the most representative scenarios – though the observations made below remain valid in typical mobile wireless environments.

In Fig. 1, we simulated 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 $\nu = 2\,\mathrm{GHz}$. 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 (2) were modeled after the well-known Jakes model for Rayleigh fading [38]. For simplicity, we assumed that the PUs were going online and offline following a Poisson process, while the simulated SUs employed the AXL algorithm with $\eta = 1$ and an update epoch of $\delta = 5\,\mathrm{ms}$.[5]

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

[5] We did not optimize the choice of η because we wanted to focus on the case where the network's SUs have minimal information.

closed (the legend of Fig. 1 indicates the antennas that were not employed in each benchmark policy).[6] The results of these simulations were plotted in Fig. 1(a): as predicted by Theorem 1, AXL leads to no regret and falls below the no-regret threshold within a few epochs, indicating that its average performance is strictly better than any of the benchmark transmit profiles.

For comparison purposes, we also simulated the same scenario, but with the SUs employing a randomized transmit policy. In particular, motivated by [29], we simulated the randomized scheme:

$$\begin{aligned} \mathbf{Q}_k(t+1) &= (1-r)\mathbf{Q}_k(t) + r\mathbf{R}_k(t), \\ \mathbf{Q}_k(0) &= m_k^{-1}\mathbf{I}, \end{aligned} \tag{28}$$

where the matrix $\mathbf{R}_k(t)$ 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 $r = 1$ (in our simulations, we took $r = 0.9$). Even though this dynamic transmit policy is sampling the state space essentially uniformly for large values of r, Fig. 1(b) shows that it leads to positive regret against 6 out of the 7 benchmark policies.[7] In other words, the no-regret property of AXL is not a spurious artifact of exploring the problem's state space in a uniform way, but it is inextricably tied to the underlying learning mechanism.

The no-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 along with the user's maximum achievable data rate and the rates achieved by the uniform policy and the randomized policy (28); to test different fading conditions, we simulated average user velocities of $v = 5$ m/s and $v = 15$ m/s (Figs. 2(a) and 2(b) respectively). We see there that AXL adapts to the changing channel conditions and tracks the user's maximum achievable rate remarkably well, in stark contrast to the uniform and randomized transmit policies.[8]

Finally, to assess the performance of the AXL algorithm with respect to the users' sum rate under SIC and its robustness in the presence of imperfect CSI, we simulated in Fig. 3 a static multi-user MIMO multiple access channel consisting of a wireless base receiver with 5 antennas, 10 PUs and 25 SUs (each with a random number of transmit antennas picked uniformly between 2 and 6). Each user's channel matrix $\mathbf{H}_k^{qr} \equiv \mathbf{H}_k^q$ was drawn from a complex Gaussian distribution at the outset of the transmission (but remained static once picked), and we then ran the AXL algorithm with $\eta = 1$. The algorithm's performance over time was assessed by plotting the *efficiency ratio*

$$\mathrm{eff}(t) = \frac{\Psi(t) - \Psi_{\min}}{\Psi_{\max} - \Psi_{\min}}, \tag{29}$$

where $\Psi(t)$ denotes the users' sum rate at the t-th iteration of the algorithm, and $\Psi_{\max}$ (resp. $\Psi_{\min}$) is the maximum (resp. minimum) value of Ψ over the set of feasible transmit profiles.[9] For comparison purposes, we also plotted the efficiency ratio achieved by water-filling methods – namely iterative water-filling (IWF) and simultaneous water-filling (SWF) [39]. Remarkably, when the users have perfect CSI, the AXL policy achieves the system's maximum sum rate within a few iterations; by contrast, SWF fails to converge altogether while the convergence time of IWF scales linearly with the number of SUs (Fig. 3(a)). On the other hand, in the presence of imperfect CSI (modeled as zero-mean i.i.d. Gaussian pertrubations to the gradient matrices $\mathbf{M}_k$ with relative magnitude of 50%), AXL still achieves the system's sum capacity (albeit at a slower rate) whereas water-filling methods offer no significant advantage over the user's initial transmit profile (cf. Fig. 3(b)).

VI. Conclusions

In this paper, we introduced an adaptive transmit policy for secondary users in MIMO-OFDM cognitive radio systems that evolve dynamically over time as a function of changing user and environmental conditions. By decomposing the users' online rate maximization into a signal covariance and a power allocation component and drawing on the method of matrix exponential learning, we derived an augmented exponential learning (AXL) scheme which leads to no regret: for every SU, the proposed transmit policy performs asymptotically as well as the best fixed transmit profile over the entire transmission horizon, and 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 transmit profile even under rapidly changing conditions.

Importantly, the implementation of the proposed algorithm requires only local CSI; moreover, the algorithm retaints its no-regret properties even in the case of *imperfect* CSI (with arbitrarily large measurement errors) and significantly outperforms classical water-filling algorithms (where the use of perfect CSI is critical).

To a large extent, our dynamic transmit policy owes its no-regret properties to an associated entropy function (for instance, the von Neumann quantum entropy for the problem's signal covariance component). As a result, by choosing a proper entropy-like kernel (e.g. as in [35]), we can examine significantly more general situations, including for example pricing and/or energy-awareness constraints.

Appendix
Technical Proofs

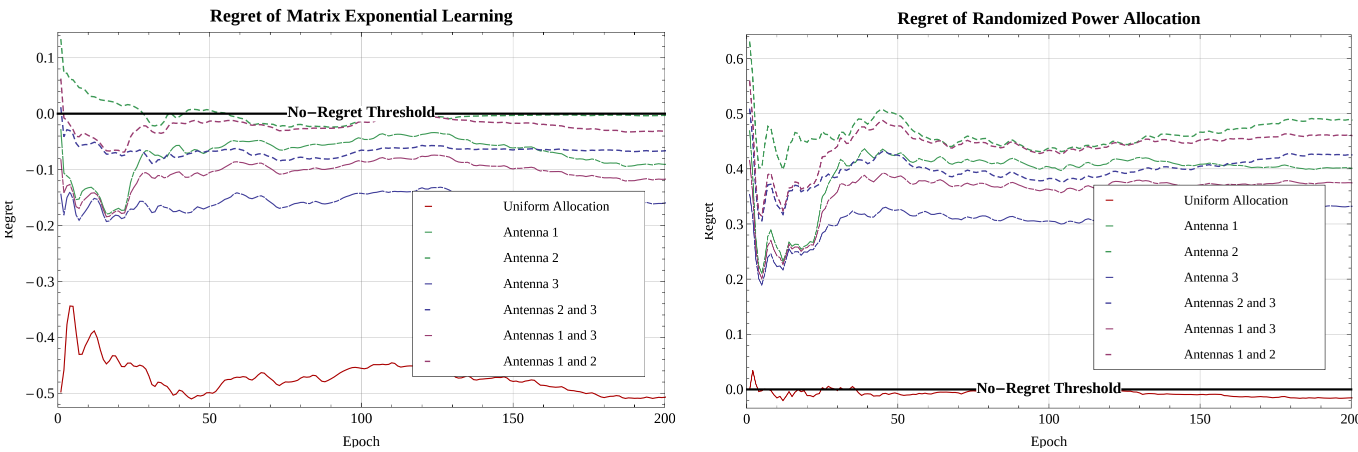
Our proof approach will rely on a technique that was introduced by Sorin [40] and was recently extended by J. Kwon and one of the authors [37]. In a nutshell, we will first establish the no-regret property in continuous time, and we will then derive the corresponding discrete-time result by estimating the difference between the continuous- and discrete-time processes.

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

[7] The policy (28) leads to no regret against the uniform power allocation policy because the average of (28) is the uniform transmit policy itself.

[8] Of course, if the user's velocity becomes exceedingly high, the quality of this tracking may deteriorate as a result of the channel's extreme variability; even in this case however, AXL is guaranteed to perform at least as well as the best fixed transmit profile.

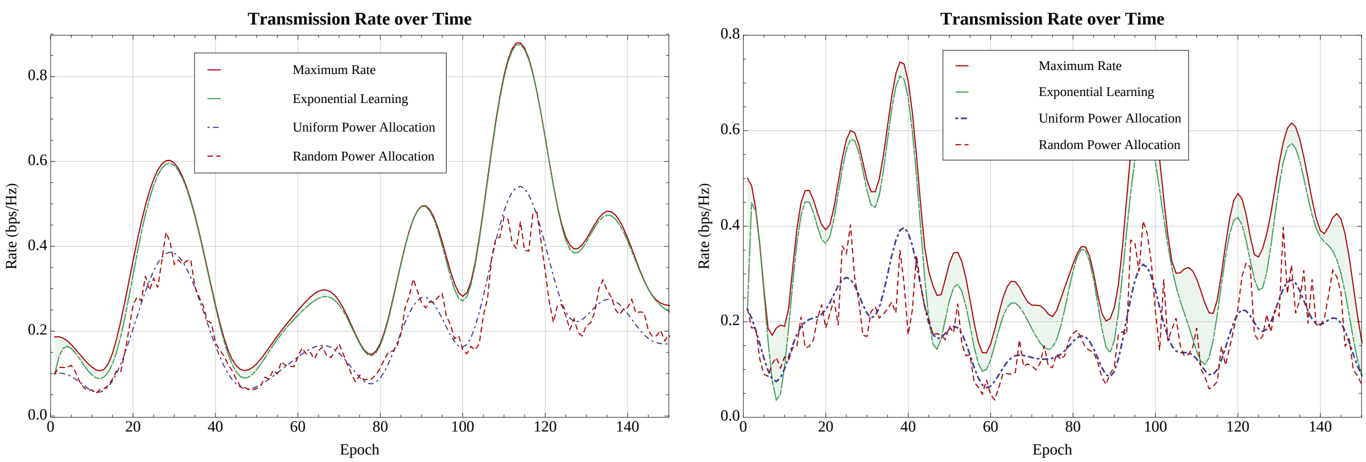
[9] The reason for using this ratio was to eliminate scaling artifacts arising e.g. from the sum rate taking values in a narrow band close to its maximum value.



(a) No regret under augmented exponential learning.

(b) Positive regret under randomized power allocation.

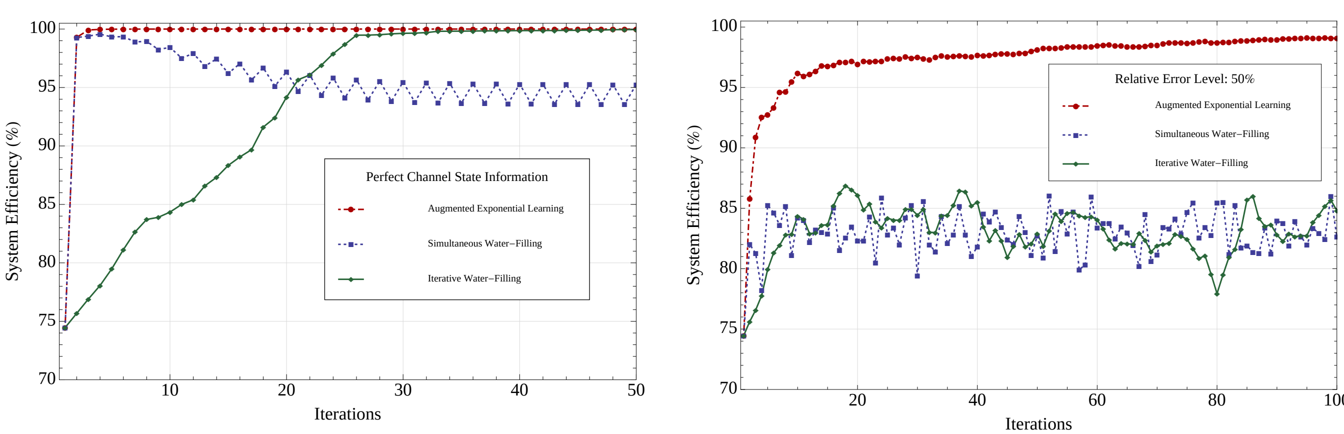
Fig. 1. The long-term regret induced by augmented exponential learning and a random sampling transmit policy (Figs 1(a) and 1(b) respectively) against different benchmark transmit profiles (see text for details). In tune with Theorem 1, AXL quickly achieves the no-regret threshold whereas the randomized policy (28) leads to positive regret in 6 out of the 7 benchmark signal profiles.



(a) Performance of AXL with average user velocity $v = 5$ km/s.

(b) Performance of AXL with average user velocity $v = 15$ km/s.

Fig. 2. Data rates achieved by AXL in a changing environment with different fading velocities: the dynamic transmit policy induced by the AXL algorithm allows users to track their maximum achievable transmission rate remarkably well even under rapidly changing channel conditions.



(a) Learning with perfect CSI.

(b) Learning with an average relative error level of 50%.

Fig. 3. Convergence and robustness of AXL with imperfect CSI in a MIMO MAC system with 10 PUs and 25 SUs: in contrast to water-filling methods, AXL attains the channel's sum capacity even in the presence of very high measurement errors.

A. Online Power Allocation: the Case $P_k \geq P$.

To begin with, note that the exponential mapping of (XL-PA) may be characterized as the solution of the convex program:

$$\begin{aligned} &\text{maximize} && \langle \mathbf{y} | \mathbf{q} \rangle - h(\mathbf{q}) \\ &\text{subject to} && q_k \geq 0,\ \textstyle\sum_k q_k = 1, \end{aligned} \tag{30}$$

where $\langle \mathbf{y}|\mathbf{q}\rangle$ denotes the bilinear pairing $\langle \mathbf{y}|\mathbf{q}\rangle = \sum_k q_k y_k$ and $h(\mathbf{q}) = \sum_k q_k \log q_k$ denotes the Gibbs–Shannon entropy on the simplex $\Delta \equiv \Delta(\mathcal{K})$ spanned by $\mathcal{K}$. More precisely, we have the following classical result [41, Chapter 25]:

Lemma 1. *For every $\mathbf{y} \in \mathbb{R}^K$, the problem (30) admits the unique solution $G(\mathbf{y})$ with $G_k(\mathbf{y}) = e^{y_k} / \sum_\ell e^{y_\ell}$.*

Consider now the following continuous-time variant of (XL-PA) for $t \geq 0$:

$$\begin{aligned} \dot{y}_k &= \frac{\partial \Phi}{\partial q_k}, \\ \mathbf{q}(t) &= G\left(\gamma(t)\mathbf{y}(t)\right), \end{aligned} \tag{31}$$

where $\gamma(t) = \min\{\eta, \eta t^{-1/2}\}$; moreover, define the cumulative continuous-time regret with respect to some fixed $\mathbf{q}_0 \in \Delta$ as

$$\mathrm{Reg}_T^c(\mathbf{q}_0) = \textstyle\int_0^T \left[\Phi(\mathbf{q}_0; t) - \Phi(\mathbf{q}(t); t)\right] dt, \tag{32}$$

where $\Phi(\cdot; t)$, is a piecewise continuous stream of rate functions and the index c in Reg_T^c indicates that we are working in continuous time. We then have:

Proposition 4. *The cumulative regret generated by the learning scheme (31) satisfies $\mathrm{Reg}_T^c(\mathbf{q}_0) \leq \eta^{-1} \log K \cdot \sqrt{T}$ for all $\mathbf{q}_0 \in \Delta$.*

Proof: Let $h^*(\mathbf{y})$ denote the convex conjugate of h, i.e. $h^*(\mathbf{y}) = \max_{\mathbf{q}\in\Delta}\{\langle \mathbf{y}|\mathbf{q}\rangle - h(\mathbf{q})\} = \langle \mathbf{y}|G(\mathbf{y})\rangle - h(G(\mathbf{y}))$. Moreover, set $\gamma(t) = \min\{\eta t^{-1/2}, \eta\}$ and let $\mathbf{q}(t)$ be defined by (31) with $\mathbf{v}(t) = \dot{\mathbf{y}}(t) = \nabla_{\mathbf{q}(t)}\Phi(\mathbf{q}(t); t)$. By Lemma 1, we will have $h^*(\gamma\mathbf{y}) = \log \sum_\ell e^{\gamma y_\ell}$ and hence:

$$\frac{d}{dt} h^*(\gamma \mathbf{y}) = \sum_{k\in\mathcal{K}} \left.\frac{\partial h^*}{\partial y_k}\right|_{\gamma\mathbf{y}} (\dot{\gamma} y_k + \gamma \dot{y}_k) = \dot{\gamma}\langle \mathbf{y}|\mathbf{q}\rangle + \gamma\langle \mathbf{v}|\mathbf{q}\rangle, \tag{33}$$

where we used (31) and the fact that $\nabla_{\mathbf{y}} h^*(\mathbf{y}) = G(\mathbf{y})$. By isolating $\langle \mathbf{v}|\mathbf{q}\rangle$ and integrating by parts, we then get:

$$\begin{aligned} \int_0^T \langle \mathbf{v}|\mathbf{q}\rangle\, dt &= \frac{h^*(\gamma(T)\mathbf{y}(T))}{\gamma(T)} - \frac{h^*(\gamma(0)\mathbf{y}(0))}{\gamma(0)} \\ &\quad + \int_0^T \frac{\dot{\gamma}}{\gamma^2} h^*(\gamma\mathbf{y})\, dt - \int_0^T \frac{\dot{\gamma}}{\gamma}\langle \mathbf{y}|\mathbf{q}\rangle\, dt \\ &= \frac{h^*(\gamma(T)\mathbf{y}(T))}{\gamma(T)} - \frac{h^*(0)}{\gamma(0)} - \int_0^T \frac{\dot{\gamma}}{\gamma^2} h(G(\gamma\mathbf{y}))\, dt, \end{aligned} \tag{34}$$

where the last step follows from the fact that $\mathbf{q} = G(\gamma\mathbf{y})$ and the defining relation $h^*(\gamma\mathbf{y}) = \langle \gamma\mathbf{y}|G(\gamma\mathbf{y})\rangle - h(G(\gamma\mathbf{y}))$. Then, given that the minimum of h over Δ is $-\log K$, we will also have $h^*(0) = -h_{\min} = \log K$; thus, with $\dot{\gamma} \leq 0$, (34) becomes:

$$\begin{aligned} \int_0^T \langle \mathbf{v}|\mathbf{q}\rangle\, dt &\geq \frac{h^*(\gamma(T)\mathbf{y}(T))}{\gamma(T)} - \frac{h^*(0)}{\gamma(0)} + h^*(0)\int_0^T \frac{\dot{\gamma}}{\gamma^2}\, dt \\ &\geq \frac{\langle \gamma(T)\mathbf{y}(T)|\mathbf{q}_0\rangle - h(\mathbf{q}_0)}{\gamma(T)} - \frac{\log K}{\gamma(T)} \\ &\geq \langle \mathbf{y}(T)|\mathbf{q}_0\rangle - \frac{\log K}{\eta}\sqrt{T}, \end{aligned} \tag{35}$$

where we used the fact that $h^*(\gamma\mathbf{y}) \geq \langle \gamma\mathbf{y}|\mathbf{q}_0\rangle - h(\mathbf{q}_0)$ for all $\mathbf{q}_0 \in \Delta$ in the second line and that $h \leq 0$ in the last step. With Φ concave over Δ, we will also have $\Phi(\mathbf{q}_0; t) - \Phi(\mathbf{q}(t); t) \leq \left\langle \nabla_{\mathbf{q}(t)}\Phi \middle| \mathbf{q}_0 - \mathbf{q}(t)\right\rangle = \langle \mathbf{v}(t)|\mathbf{q}_0 - \mathbf{q}(t)\rangle$; hence, by (35), we get:

$$\mathrm{Reg}_T^c(\mathbf{q}_0) \leq \textstyle\int_0^T \langle \mathbf{v}|\mathbf{q}_0 - \mathbf{q}\rangle\, dt \leq \frac{\log K}{\eta}\sqrt{T}, \tag{36}$$

and our proof is complete. ■

B. Online Power Allocation: The General Case.

If $P_k < P$ for some k, we still obtain a no-regret power allocation policy if we use the modified entropy function $h(p) = \sum_k \left(p_k \log p_k + (P_k - p_k)\log(P_k - p_k)\right)$, and define the modified Gibbs map:

$$G_0(\mathbf{y}) = \arg\max\nolimits_{\mathbf{p}\in\mathcal{X}_0} \left\{\langle \mathbf{y}|\mathbf{p}\rangle - h_0(\mathbf{p})\right\}. \tag{37}$$

Specifically, consider the following modified version of (31):

$$\begin{aligned} \dot{y}_k &= \frac{\partial \Phi}{\partial p_k}, \\ \mathbf{p}(t) &= G_0\left(\gamma(t)\mathbf{y}(t)\right), \end{aligned} \tag{38}$$

where $\Phi(\cdot; t)$ is a continuous stream of rate functions of the form (5) and $\gamma = \min\{\eta, \eta t^{-1/2}\}$. We then have:

Proposition 5. *The learning scheme (38) leads to no regret in continuous time:* $\mathrm{Reg}_T^c(\mathbf{p}_0) \leq \mathcal{O}(\sqrt{T})$ *for all* $\mathbf{p}_0 \in \mathcal{X}_0$.

Proof: Shadowing the proof of Proposition 4, let $h_0^*(\mathbf{y}) = \max_{\mathbf{p}\in\mathcal{X}_0}\{\langle \mathbf{y}|\mathbf{p}\rangle - h_0(\mathbf{p})\} = \langle \mathbf{y}|G_0(\mathbf{y})\rangle - h_0(G_0(\mathbf{y}))$ be the convex conjugate of $h_0(\mathbf{p})$. Since the derivative of h_0 blows up to infinity at the boundary of $\mathcal{X}_0$, the unique solution to the maximization problem defining G_0 will lie at the interior of $\mathcal{X}_0$. The Karush–Kuhn–Tucker (KKT) conditions thus give $y_k - \frac{\partial h_0}{\partial p_k} = \lambda$, where λ is the Lagrange multiplier for the equality constraint $\sum_\ell p_\ell = P$. We will then also have $\frac{\partial h_0^*}{\partial y_k} = G_{0,k}(\mathbf{y}) + \sum_{\ell=1}^K y_\ell \frac{\partial}{\partial y_k} G_{0,\ell}(\mathbf{y}) - \sum_{\ell=1}^K \frac{\partial h_0}{\partial p_\ell}\frac{\partial}{\partial y_k} G_{0,\ell}(\mathbf{y}) = G_{0,k}(\mathbf{y})$, where, in the last step, we used the fact that $\sum_{\ell=1}^K G_{0,\ell}(\mathbf{y}) = P$ (so $\sum_{\ell=1}^K \partial_{y_k} G_{0,\ell} = 0$ for all k). Thus, letting $\mathbf{v}(t) = \nabla_{\mathbf{p}}\Phi(\mathbf{p}; t)$ so that $\mathbf{y}(t) = \int_0^t \mathbf{v}(s)\, ds$ and $\mathbf{p}(t) = G_0(\gamma(t)\mathbf{y}(t))$, we obtain the basic identity:

$$\frac{d}{dt} h_0^*(\gamma \mathbf{y}) = \sum_{k\in\mathcal{K}} \left.\frac{\partial h_0^*}{\partial y_k}\right|_{\gamma\mathbf{y}} (\dot{\gamma} y_k + \gamma \dot{y}_k) = \dot{\gamma}\langle \mathbf{y}|\mathbf{p}\rangle + \gamma\langle \mathbf{v}|\mathbf{p}\rangle, \tag{39}$$

and the rest of the proof follows as in the case of Prop. 4. ■

C. Online Signal Covariance Optimization

For the MIMO component (OCOV) of (ORM) we will consider the continuous-time scheme:

$$\begin{aligned} \dot{\mathbf{Y}}_k &= \frac{\partial \Phi}{\partial \mathbf{Q}_k^*}, \\ \mathbf{Q}_k &= \frac{\exp(\gamma \mathbf{Y}_k)}{\mathrm{tr}\left[\exp\left(\gamma \mathbf{Y}_k\right)\right]}. \end{aligned} \tag{40}$$

where, as before, $\gamma = \min\{\eta, \eta t^{-1/2}\}$. Then, with the user's regret defined as in (32), we get:

Proposition 6. *The cumulative regret generated by the continuous-time learning scheme* (40) *satisfies* $\mathrm{Reg}_T^c(\mathbf{Q}_0) \leq \eta^{-1}\sqrt{T}\sum_{k=1}^K \log m_k$ *for all* $\mathbf{Q}_0 \in \mathcal{X}_+ \equiv \prod_{k=1}^K \mathcal{D}_k$.

To prove Proposition 6, we will first show that the matrix exponential of (19) solves the semidefinite problem:

$$\begin{aligned} &\text{maximize} && \operatorname{tr}[\mathbf{YQ}] - h_+(\mathbf{Q}) \\ &\text{subject to} && \mathbf{Q} \succcurlyeq 0,\ \operatorname{tr}(\mathbf{Q}) = 1, \end{aligned} \tag{41}$$

where $\mathbf{Y}$ is a Hermitian matrix and $h_+(\mathbf{Q}) = \operatorname{tr}[\mathbf{Q}\log\mathbf{Q}]$ is the *von Neumann entropy*. We thus obtain:

Lemma 2. *For every Hermitian matrix* $\mathbf{Y} \in \mathbb{C}^{m\times m}$*, the problem* (30) *admits the unique solution* $\mathbf{Q}_\mathbf{Y} = \exp(\mathbf{Y})/\operatorname{tr}[\exp(\mathbf{Y})]$*. Accordingly, the convex conjugate of* h_+ *will be:*

$$h_+^*(\mathbf{Y}) = \max\nolimits_{\mathbf{Q}\in\mathcal{D}}\{\operatorname{tr}[\mathbf{YQ}] - h_+(\mathbf{Q})\} = \log\operatorname{tr}[\exp(\mathbf{Y})]. \tag{42}$$

Proof: To begin with, let $A(\mathbf{Y},\mathbf{Q}) = \operatorname{tr}[\mathbf{YQ}] - h_+(\mathbf{Q})$ denote the objective of the problem (41), and let $Z = \{\mathbf{A} \in \mathbb{C}^{m\times m} : \mathbf{A}^\dagger = \mathbf{A}, \operatorname{tr}(\mathbf{A}) = 0\}$ be the space of tangent directions to $\mathcal{D}$. Then, if $\{q_j, \mathbf{u}_j\}_{j=1}^m$ is an eigen-decomposition of $\mathbf{Q} + t\mathbf{Z}$ for $\mathbf{Q} \in \mathcal{D}^\circ$ and $\mathbf{Z} \in Z$, we will have $A(\mathbf{Y}, \mathbf{Q} + t\mathbf{Z}) = \operatorname{tr}[\mathbf{YQ}] + \operatorname{tr}[\mathbf{YZ}]\,t - \sum_j q_j \log q_j$. Hence, the directional derivative of $A(\mathbf{Y},\mathbf{Q})$ along $\mathbf{Z}$ at $\mathbf{Q}$ will be: $\nabla_\mathbf{Z} A(\mathbf{Y},\mathbf{Q}) = \frac{d}{dt}\big|_{t=0} A(\mathbf{Y}, \mathbf{Q} + t\mathbf{Z}) = \operatorname{tr}[\mathbf{YZ}] - \sum_{k=1}^K \dot{q}_k \log q_k$ where we have used the fact that $\sum_j \dot{q}_j = 0$ (recall that $\sum_j q_j = \operatorname{tr}(\mathbf{Q} + t\mathbf{Z}) = 1$ for all t such that $\mathbf{Q} + t\mathbf{Z} \in \mathcal{D}^\circ$). However, differentiating the defining relation $(\mathbf{Q} + t\mathbf{Z})\mathbf{u}_j = q_j \mathbf{u}_j$ with respect to t gives $\mathbf{Z}\mathbf{u}_j + (\mathbf{Q} + t\mathbf{Z})\dot{\mathbf{u}}_j = \dot{q}_j \mathbf{u}_j + q_j \dot{\mathbf{u}}_j$, so, after multiplying from the left by $\mathbf{u}_j^\dagger$, we get $\dot{q}_j = \mathbf{u}_j^\dagger \mathbf{Z} \mathbf{u}_j + \mathbf{u}_j^\dagger(\mathbf{Q} + t\mathbf{Z})\dot{\mathbf{u}}_j - q_j \mathbf{u}_j^\dagger \dot{\mathbf{u}}_j = \mathbf{u}_j^\dagger \mathbf{Z}\mathbf{u}_j$. Summing over j gives $\sum_j \dot{q}_j \log q_j = \sum_j \mathbf{u}_j^\dagger \mathbf{Z}\mathbf{u}_j \log q_j = \operatorname{tr}[\mathbf{Z}\log\mathbf{Q}]$; then, by substituting in the previous expression for $\nabla_\mathbf{Z} A(\mathbf{Y},\mathbf{Q})$, we finally obtain $\nabla_\mathbf{Z} A(\mathbf{Y},\mathbf{Q}) = \operatorname{tr}[\mathbf{Z}(\mathbf{Y} - \log\mathbf{Q})]$.

With this expression at hand, it is easy to see that $A(\mathbf{Y},\cdot)$ becomes infinitely steep at the boundary $\operatorname{bd}(\mathcal{D})$ of $\mathcal{D}$, i.e. $|\nabla_\mathbf{Z} A(\mathbf{Y},\mathbf{Q}_n)| \to \infty$ whenever $\mathbf{Q}_n \to \operatorname{bd}(\mathcal{D})$ (simply note that the eigenvalues of $\log\mathbf{Q}$ blow up when $\mathbf{Q}$ becomes singular). Since h_+ is strictly convex, it follows that A will be of Legendre type [41], so (41) will admit a unique solution $\mathbf{Q}_\mathbf{Y}$ at the interior $\mathcal{D}^\circ$ of $\mathcal{D}$ [41, Chapter 26]. Accordingly, by the KKT conditions for (41), we will have $\nabla_\mathbf{Z} A(\mathbf{Y},\mathbf{Q}_\mathbf{Y}) = 0$ for all tangent directions $\mathbf{Z}$ to $\mathcal{D}^\circ$ at $\mathbf{Q}_\mathbf{Y}$, i.e. $\operatorname{tr}[\mathbf{Z}(\mathbf{Y} - \log\mathbf{Q}_\mathbf{Y})] = 0$ for all Hermitian $\mathbf{Z} \in \mathbb{C}^{m\times m}$ such that $\operatorname{tr}(\mathbf{Z}) = 0$. From this last condition, we immediately get $\mathbf{Y} - \log\mathbf{Q}_\mathbf{Y} \propto \mathbf{I}$, and with $\operatorname{tr}(\mathbf{Q}_\mathbf{Y}) = 1$, we obtain $\mathbf{Q}_\mathbf{Y} = \exp(\mathbf{Y})/\operatorname{tr}[\exp(\mathbf{Y})]$; the expression for $h_+^*(\mathbf{Y})$ then follows by substituting $\mathbf{Q}_\mathbf{Y}$ in the definition of $A(\mathbf{Y},\mathbf{Q})$. ■

Armed with this characterization, we now get:

Proof of Proposition 6: Let $h_k(\mathbf{Q}_k) = \operatorname{tr}(\mathbf{Q}_k \log\mathbf{Q}_k)$, $\mathbf{Q}_k \in \mathcal{D}_k$, so $h_k^*(\mathbf{Y}_k) = \log\operatorname{tr}[\exp(\mathbf{Y}_k)]$ by Lemma 2; moreover, let $\mathbf{Q} = \operatorname{diag}(\mathbf{Q}_1,\dots,\mathbf{Q}_K)$ and set $h_+(\mathbf{Q}) = \sum_k h_k(\mathbf{Q}_k) = \operatorname{tr}[\mathbf{Q}\log\mathbf{Q}]$ for $\mathbf{Q} \in \mathcal{X}_+ \equiv \prod_k \mathcal{D}_k$. Then, if $\mathbf{Y} = \operatorname{diag}(\mathbf{Y}_1,\dots,\mathbf{Y}_K)$ with $\mathbf{Y}_k$ Hermitian, we will have $h_+^*(\mathbf{Y}) = \max_{\mathbf{Q}\in\mathcal{X}_+}\{\operatorname{tr}[\mathbf{YQ}] - h(\mathbf{Q})\} = \sum_k h_k^*(\mathbf{Y}_k) = \sum_k \log\operatorname{tr}[\exp(\mathbf{Y}_k)]$. Accordingly, if we let $\mathbf{V}_k(t) = \partial_{\mathbf{Q}_k^*}\Phi(\mathbf{Q};t)$, we will have:

$$\begin{aligned} \frac{d}{dt}h_+^*(\gamma\mathbf{Y}) &= \sum\nolimits_{k=1}^K \operatorname{tr}[\exp(\gamma\mathbf{Y}_k)]^{-1}\frac{d}{dt}\operatorname{tr}[\exp(\gamma\mathbf{Y}_k)] \\ &= \sum\nolimits_{k=1}^K \operatorname{tr}[\exp(\gamma\mathbf{Y}_k)]^{-1}\operatorname{tr}[(\dot{\gamma}\mathbf{Y}_k + \gamma\dot{\mathbf{Y}}_k)\exp(\mathbf{Y}_k)] \\ &= \dot{\gamma}\operatorname{tr}[\mathbf{YQ}] + \gamma\operatorname{tr}[\mathbf{VQ}] \end{aligned} \tag{43}$$

where we set $\mathbf{V} = \operatorname{diag}(\mathbf{V}_1,\dots,\mathbf{V}_K)$. Following the same steps as in the proof of Proposition 4, we then obtain:

$$\int_0^T \operatorname{tr}[\mathbf{VQ}]\,dt = \frac{h_+^*(\gamma(T)\mathbf{Y}(T))}{\gamma(T)} - \frac{h_+^*(0)}{\gamma(0)} - \int_0^T \frac{\dot{\gamma}}{\gamma^2}h_+(\mathbf{Q})\,dt, \tag{44}$$

The minimum of h_+ over $\mathcal{X}_+ = \prod_k \mathcal{D}_k$ is just $-\sum_k \log m_k$, so we will also have $h^*(0) = -\min_{\mathbf{Q}\in\mathcal{X}_+} h_+(\mathbf{Q}) = \sum_k \log m_k$; then, with $\dot{\gamma} \leq 0$, (44) becomes:

$$\begin{aligned} \int_0^T \operatorname{tr}[\mathbf{VQ}]\,dt &\geq \frac{h_+^*(\gamma(T)\mathbf{Y}(T))}{\gamma(T)} - \frac{h^*(0)}{\gamma(0)} + h_+^*(0)\int_0^T \frac{\dot{\gamma}}{\gamma^2}\,dt \\ &\geq \frac{\operatorname{tr}[\gamma(T)\mathbf{Y}(T)\mathbf{Q}_0] - h_+(\mathbf{Q}_0)}{\gamma(T)} - \frac{\sum_{k=1}^K \log m_k}{\gamma(T)} \\ &\geq \operatorname{tr}[\mathbf{Y}(T)\mathbf{Q}_0] - \frac{\sum_{k=1}^K \log m_k}{\eta}\sqrt{T}, \end{aligned} \tag{45}$$

where we used the fact that $h_+^*(\gamma\mathbf{Y}) \geq \operatorname{tr}[\gamma\mathbf{Y}\mathbf{Q}_0] - h_+(\mathbf{Q}_0)$ for all $\mathbf{Q}_0 \in \mathcal{X}_+$ in the second line and the fact that $h_+ \leq 0$ in the last step. Since Φ is concave in $\mathbf{Q}$ and $\mathbf{V} = \nabla_{\mathbf{Q}^*}\Phi$, the rest of the proof follows in the same way as that of Proposition 4. ■

D. The Full MIMO–OFDM Problem

Our final step in this continuous-time setting will be to establish the no-regret properties of the following continuous-time variant of the AXL algorithm for $P_k \geq P$:

$$\begin{aligned} \dot{y}_k &= \frac{\partial\Phi}{\partial q_k}, & \dot{\mathbf{Y}}_k &= \frac{\partial\Phi}{\partial\mathbf{Q}_k^*}, \\ q_k &= \frac{\exp(\gamma y_k)}{\sum_{\ell=1}^K \exp(\gamma y_\ell)}, & \mathbf{Q}_k &= \frac{\exp(\gamma\mathbf{Y}_k)}{\operatorname{tr}[\exp(\gamma\mathbf{Y}_k)]}, \end{aligned} \tag{46}$$

with $\gamma = \min\{\eta, \eta t^{-1/2}\}$ as usual. Without further ado, we have:

Proposition 7. *If* $P_k \geq P$ *for all* $k \in \mathcal{K}$*, then, for all* $\mathbf{P}_0 \in \mathcal{X}$*, the cumulative regret generated by* (46) *will satisfy* $\mathrm{Reg}_T^c(\mathbf{P}_0) \leq \eta^{-1}\sqrt{T}\left(\log K + \sum_{k=1}^K \log m_k\right)$.

Proof: Recall that any $\mathbf{P} \in \mathcal{X}$ may be decomposed as $\mathbf{P} = \operatorname{diag}(p_1\mathbf{Q}_1,\dots,p_K\mathbf{Q}_K)$ with $\mathbf{p} = (p_1,\dots,p_K) \in \mathcal{X}_0$ and $\mathbf{Q} = \operatorname{diag}(\mathbf{Q}_1,\dots,\mathbf{Q}_K) \in \mathcal{X}_+ \equiv \prod_k \mathcal{D}_k$. Then, using the normalized power allocation vector $\mathbf{q} = \mathbf{p}/P \in \Delta$ for convenience, let $H(\mathbf{q},\mathbf{Q}) = h(\mathbf{q}) + h_+(\mathbf{Q}) = \sum_{k=1}^K [q_k \log q_k + \operatorname{tr}(\mathbf{Q}_k \log\mathbf{Q}_k)]$ denote the aggregate entropy over the space $\Delta \times \prod_k \mathcal{D}_k$ and consider the associated Legendre–Fenchel problem:

$$\begin{aligned} &\text{maximize} && \langle\mathbf{y}|\mathbf{q}\rangle + \operatorname{tr}[\mathbf{YQ}] - H(\mathbf{q},\mathbf{Q}), \\ &\text{subject to} && \mathbf{q} \in \Delta,\ \mathbf{Q} \in \textstyle\prod_k \mathcal{D}_k. \end{aligned} \tag{47}$$

Clearly, the problem (47) may be decomposed as a sum of (30) and (41), so each component of the solution of (47) will be given by Lemmas 1 and 2 respectively; likewise, the convex conjugate of H will be $H^*(\mathbf{y},\mathbf{Y}) = h^*(\mathbf{y}) + h_+^*(\mathbf{Y})$, with h^* and h_+^* defined as before. Our claim is then obtained by following the same steps as in the proofs of Propositions 4 and 6. ■

E. The Descent to Discrete Time

In this appendix, our aim will be to derive the no-regret properties of the discrete-time policies (XL-PA), (XL-COV) and of the AXL algorithm (Propositions 1, 3 and Theorem 1 respectively) by means of a comparison technique introduced by Sorin [40] and developed further by J. Kwon and one of the authors [37]. Specifically, we have:

Lemma 3. *Let $\mathcal{C}$ be a compact convex set in $\mathbb{R}^N$, let $\mathbf{v}(t)$ be a sequence of payoff vectors in $\mathbb{R}^N$ with $\|\mathbf{v}(t)\| \leq V$ in the uniform norm of $\mathbb{R}^N$ ($t = 1, 2 \dots$), and consider the sequence of play $\mathbf{x}(t+1) = Q\left(\eta t^{-1/2} \sum_{s=1}^{t} \mathbf{v}(s)\right)$ where $Q\colon \mathbb{R}^N \to \mathcal{C}$ is C-Lipschitz with respect to the L^1 norm on $\mathcal{C}$. Moreover, letting $\mathbf{v}^c(t) = \mathbf{v}(\lceil t \rceil)$ be a piecewise constant interpolation of $\mathbf{v}(t)$ for $t \in [1, +\infty)$, consider the continuous-time process $\mathbf{x}^c(t) = Q\left(\gamma(t) \int_0^t \mathbf{v}^c(s)\, ds\right)$ with $\gamma(t) = \min\{\eta t^{-1/2}, \eta\}$, and assume that it guarantees the regret bound:*

$$\int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}_0 - \mathbf{x}^c(t) \rangle \, dt \leq R(T)\sqrt{T} \quad \text{for all } \mathbf{x}_0 \in \mathcal{X}_+. \tag{48}$$

Then, for all $\mathbf{x}_0 \in \mathcal{A}$, the discrete-time sequence $\mathbf{x}(t)$ guarantees

$$\sum_{t=1}^{T} \langle \mathbf{v}(t) | \mathbf{x}_0 - \mathbf{x}(t) \rangle \leq \sqrt{T}\left(R(T) + 4CV^2\eta\right). \tag{49}$$

Proof: By assumption, if we set $\mathbf{y}(t) = \int_0^t \mathbf{v}^c(s)\, ds$, we will also have $\mathbf{x}^c(t) = Q(\gamma(t)\mathbf{y}(t)) = \mathbf{x}(t+1)$ whenever t is a positive integer. Hence, for every integer $T \geq 1$, we will have $\int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}^c(t) \rangle \, dt - \sum_{t=1}^{T} \langle \mathbf{v}(t) | \mathbf{x}(t) \rangle = \int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}^c(t) \rangle \, dt - \int_0^T \langle \mathbf{v}(\lceil t \rceil) | \mathbf{x}(\lceil t \rceil) \rangle \, dt = \int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}^c(t) - \mathbf{x}^c(\lfloor t \rfloor) \rangle \, dt$, where we used the fact that $\mathbf{x}^c(\lfloor t \rfloor) = \mathbf{x}(\lceil t \rceil)$ in the second step. On the other hand, Hölder's inequality gives $|\langle \mathbf{v}^c(t) | \mathbf{x}^c(t) - \mathbf{x}^c(\lfloor t \rfloor) \rangle| \leq \|\mathbf{v}^c(t)\|_\infty \cdot \|\mathbf{x}^c(t) - \mathbf{x}^c(\lfloor t \rfloor)\|_1 \leq V \|\mathbf{x}^c(t) - \mathbf{x}^c(\lfloor t \rfloor)\|_1 \leq V \|Q(\gamma(t)\mathbf{y}(t)) - Q(\gamma(\lfloor t \rfloor)\mathbf{y}(\lfloor t \rfloor))\|_1 \leq CV \|\gamma(t)\mathbf{y}(t) - \gamma(\lfloor t \rfloor)\mathbf{y}(\lfloor t \rfloor)\|_\infty$. The last term may then be rewritten as:

$$\begin{aligned}
\|\gamma(t)\mathbf{y}(t) - \gamma(\lfloor t \rfloor)\mathbf{y}(\lfloor t \rfloor)\|_\infty &= \left\| \int_{\lfloor t \rfloor}^{t} \frac{d}{ds}\left(\gamma(s)\mathbf{y}(s)\right) \, ds \right\|_1 \\
&\leq \int_{\lfloor t \rfloor}^{t} \left\| \gamma(s)\mathbf{v}^c(s) + \dot{\gamma}(s) \int_0^s \mathbf{v}^c(w)\, dw \right\|_\infty ds \\
&\leq V \int_{\lfloor t \rfloor}^{t} \left(\gamma(s) - s\dot{\gamma}(s)\right) ds.
\end{aligned} \tag{50}$$

Recalling that $\gamma(t) = \min\{\eta, \eta t^{-1/2}\}$, this last integral will be equal to ηt if $t \in [0, 1]$ and $3\eta(t^{1/2} - \lfloor t \rfloor^{1/2})$ otherwise. Thus, combining the above inequalities, we obtain:

$$\begin{aligned}
\int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}^c(t) - \mathbf{x}^c(\lfloor t \rfloor) \rangle \, dt &\leq CV^2 \int_0^T \int_{\lfloor t \rfloor}^{t} \left(\gamma(s) - s\dot{\gamma}(s)\right) ds \, dt \\
&\leq CV^2\eta \left(\frac{1}{2} + 3 \sum_{k=1}^{T-1} \int_k^{k+1} \frac{t-k}{\sqrt{t} + \sqrt{k}} \, dt \right) \leq 4CV^2\eta\sqrt{T}.
\end{aligned} \tag{51}$$

Hence, by the definition of $\mathbf{v}^c(t)$ and the assumptions of the lemma, we finally obtain

$$\begin{aligned}
\sum_{t=1}^{T} \langle \mathbf{v}(t) | \mathbf{x}_0 - \mathbf{x}(t) \rangle &= \int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}_0 - \mathbf{x}^c(t) \rangle \, dt \\
&+ \int_0^T \langle \mathbf{v}^c(t) | \mathbf{x}^c(t) - \mathbf{x}^c(\lfloor t \rfloor) \rangle \, dt \leq R(T)\sqrt{T} + 4CV^2\eta\sqrt{T},
\end{aligned}$$

which completes our proof. ■

With this comparison at hand, the analysis of the previous sections yields:

Proof of Proposition 1: Note first that $v_k = \frac{\partial \Phi}{\partial q_k} = P \operatorname{tr}[\mathbf{M}_k \mathbf{Q}_k]$, so the payoff vectors $\mathbf{v}$ of (14) are bounded in the uniform norm of $\mathbb{R}^K$ by PM – cf. (10). Given that the Lipschitz constant of the exponential mapping $G(y)$ of (1) is easily seen to be $C = 1$ [18], the proposition follows by combining the continuous-time bound of Proposition 4 with Lemma 3. ■

Proof of Proposition 2: Note first that the modified Gibbs map of (37) simply represents the power allocation policy of (XL-PA′): indeed, by the KKT conditions for the maximization problem defining G_0, we will have:

$$\frac{p_k}{P_k - p_k} = e^{\lambda - y_k} \implies p_k = P_k \frac{e^{y_k}}{e^{\lambda} + e^{y_k}}, \tag{52}$$

so, given that the power vector $\mathbf{p}$ must satisfy the total power constraint (7a), the Lagrange multiplier λ must satisfy the condition $P = \sum_k p_k = \sum_k P_k(1 + e^{\lambda - y_k})^{-1}$. Comparing this last equation with (16), we conclude that p_k will be given by the power update step of (XL-PA′) with $\mathbf{y}$ replaced by $\gamma\mathbf{y}$, so our claim follows by combining Proposition 5 with Lemma 3. ■

Proof of Proposition 3: The matrix payoffs $\mathbf{V}_k = \frac{\partial \Phi}{\partial \mathbf{Q}_k^*} = p_k \mathbf{M}_k$ satify $\|\mathbf{V}_k\| \leq PM$ by (10). Moreover, the von Neumann entropy h_+ is 1-strongly convex with respect to the L^1 norm, so the matrix exponential mapping $\mathbf{Y} \mapsto \mathbf{Q}_{\mathbf{Y}} = \exp(\mathbf{Y}) / \operatorname{tr}[\exp(\mathbf{Y})]$ is 1-Lipschitz – see e.g. [23]. Our claim then follows by combining the continuous-time bound of Proposition 6 with Lemma 3. ■

Proof of Theorem 1: As in the proofs of Propositions 1 and 3, the map $(\mathbf{y}, \mathbf{Y}) \mapsto (\mathbf{q}, \mathbf{Q}) \in \Delta \times \prod_k \mathcal{D}_k$ of (46) is 1-Lipschitz and the payoffs $(\mathbf{v}, \mathbf{V}_k)$ are bounded by PM in the uniform norm of $\mathbb{R}^K \times \prod_k \mathbb{C}^{m_k \times m_k}$. The theorem then follows by combining the continuous-time bound of Proposition 7 with Lemma 3. ■

F. Learning with Imperfect CSI

Our goal in this appendix will be to prove the no-regret properties of AXL under imperfect CSI.

Proof of Theorem 2: Let $\mathbf{P}(t) = \operatorname{diag}(\mathbf{P}_1(t), \dots, \mathbf{P}_k(t)) \in \mathcal{X}$ be the sequence of transmit profiles generated by the AXL algorithm with perturbed observations $\hat{\mathbf{M}} = \mathbf{M} + \mathbf{\Xi}$. Then, for every $\mathbf{P}_0 \in \mathcal{X}$, we will have:

$$\begin{aligned}
&\operatorname{Reg}_T(\mathbf{P}_0) \leq \sum_{t=1}^{T} \operatorname{tr}[\nabla\Phi(\mathbf{P}(t)) \cdot (\mathbf{P}_0 - \mathbf{P}(t))] \\
&= \sum_{t=1}^{T} \operatorname{tr}\left[\hat{\mathbf{M}}(t) \cdot (\mathbf{P}_0 - \mathbf{P}(t))\right] - \sum_{t=1}^{T} \operatorname{tr}[\mathbf{\Xi}(t) \cdot (\mathbf{P}_0 - \mathbf{P}(t))],
\end{aligned} \tag{53}$$

where the inequality follows from the concavity of Φ. Since $\mathbf{P}(t)$ is generated by the sequence of matrix payoffs $\hat{\mathbf{M}}(t)$, the first term of this expression will simply be the regret generated by $\mathbf{P}(t)$ against $\hat{\mathbf{M}}(t)$, so we will have $\sum_{t=1}^{T} \operatorname{tr}\left[\hat{\mathbf{M}}(t) \cdot (\mathbf{P}_0 - \mathbf{P}(t))\right] \leq R\sqrt{T}$ by Theorem 1 (or, more accurately, by combining (35) and (45) with Lemma 3).

As for the second term, it is easy to see that the process $V(t) = \operatorname{tr}[\mathbf{\Xi}(t) \cdot (\mathbf{P}(t) - \mathbf{P}_0)]$ is a martingale difference: indeed, by (24) and the fact that $\mathbf{P}(t)$ is fully determined by $\hat{\mathbf{M}}(1), \dots, \hat{\mathbf{M}}(t-1)$, we get $\mathbb{E}[V(t)|\mathcal{F}_{t-1}] = \mathbb{E}\left[\operatorname{tr}[\mathbf{\Xi}(t) \cdot (\mathbf{P}(t) - \mathbf{P}_0)] | \mathcal{F}_{t-1}\right] = \operatorname{tr}[\mathbb{E}[\mathbf{\Xi}(t)|\mathcal{F}_{t-1}] \cdot (\mathbf{P}(t) - \mathbf{P}_0)] = 0$. Moreover, with $\|\mathbf{\Xi}\| \leq \Sigma$, we will also have $|V(t)| \leq \|\boldsymbol{\Xi}(t)\| \cdot \|\mathbf{P}_0 - \mathbf{P}(t)\|_1 \leq \Sigma \cdot D$, where $D = \max\{\|\mathbf{P}_0 - \mathbf{P}\|_1 : \mathbf{P}_0, \mathbf{P} \in \mathcal{X}\}$ denotes the L^1-diameter of $\mathcal{X}$.

The bound (25) is thus obtained by taking the expectation of $\mathrm{Reg}_T(\mathbf{P}_0)$ and using the zero-mean property of V. Similarly, the fact that $\mathbf{P}(t)$ generates no regret almost surely (and not only in expectation) follows by noting that $T^{-1}\sum_{t=1}^{T} V(t) \to 0$ as a consequence of the strong law of large numbers for martingale differences – see e.g. [42, Theorem 2.18]. Finally, for the large deviations bounds (26), note first that (53) yields:

$$\mathbb{P}\left(\frac{1}{T}\,\mathrm{Reg}_T(\mathbf{P}_0) \geq \frac{R}{\sqrt{T}} + z\right) \leq \mathbb{P}\left(\sum\nolimits_{t=1}^{T} |V(t)| \geq Tz\right). \quad (54)$$

However, with $\|\boldsymbol{\Xi}\| \leq \Sigma$, Azuma's inequality [43] yields $\mathbb{P}\left(\sum_{t=1}^{T} V(t) \geq Tz\right) \leq \exp\left(-\frac{T^2 z^2}{2\sum_{t=1}^{T} \operatorname{ess\,sup}|V(t)|^2}\right) \leq \exp\left(-\frac{Tz^2}{2\Sigma^2 D^2}\right)$, and our claim follows. ∎

References

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

[2] FCC Spectrum Policy Task Force, "Report of the spectrum efficiency working group," Federal Communications Comission, Tech. Rep., November 2002.

[3] J. Mitola III and G. Q. Maguire Jr., "Cognitive radio: making software radios more personal," *IEEE Personal Commun. Mag.*, vol. 6, no. 4, pp. 13–18, August 1999.

[4] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[5] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, February 2005.

[6] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894–914, 2009.

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

[8] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications and Related Technologies*, vol. 10, no. 6, pp. 585–596, 1999.

[9] 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.*, vol. 29, no. 2, pp. 362–373, 2011.

[10] G. Scutari and D. P. Palomar, "MIMO cognitive radio: A game theoretical approach," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 761–780, February 2010.

[11] J. Wang, G. Scutari, and D. P. Palomar, "Robust MIMO cognitive radio via game theory," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1183–1201, March 2011.

[12] 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*, 2005, pp. 269–278.

[13] 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.*, vol. 29, no. 4, pp. 731–745, April 2011.

[14] 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*, 2009, pp. 1893–1898.

[15] 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.

[16] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games, Volume III*, ser. Annals of Mathematics Studies, M. Dresher, A. W. Tucker, and P. Wolfe, Eds. Princeton, NJ: Princeton University Press, 1957, vol. 39, pp. 97–139.

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

[18] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[19] V. G. Vovk, "Aggregating strategies," in *COLT '90: Proceedings of the Third Workshop on Computational Learning Theory*, 1990, pp. 371–383.

[20] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, no. 2, pp. 212–261, 1994.

[21] P. Auer, N. Cesa-Bianchi, and C. Gentile, "Adaptive and self-confident on-line learning algorithms," *Journal of Computer and System Sciences*, vol. 64, no. 1, pp. 48–75, 2002.

[22] 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*, 2012, pp. 3028–3032.

[23] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, "Regularization techniques for learning with matrices," *The Journal of Machine Learning Research*, vol. 13, pp. 1865–1890, 2012.

[24] H. Bölcskei, D. Gesbert, and A. J. Paulraj, "On the capacity of OFDM-based spatial multiplexing systems," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 225–234, February 2002.

[25] K. B. Letaief and Y. J. A. Zhang, "Dynamic multiuser resource allocation and adaptation for wireless systems," *Wireless Communications, IEEE*, vol. 13, no. 4, pp. 38–47, August 2006.

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

[27] 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.*, vol. 47, no. 1, pp. 130–138, jan 2009.

[28] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*, 1965, pp. 263–272.

[29] D. P. Palomar, J. M. Cioffi, and M. Lagunas, "Uniform power allocation in MIMO channels: a game-theoretic approach," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, p. 1707, July 2003.

[30] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[31] P. Mertikopoulos, E. V. Belmega, A. L. Moustakas, and S. Lasaulce, "Distributed learning policies for power allocation in multiple access channels," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 96–106, January 2012.

[32] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML '03: Proceedings of the 20th International Conference on Machine Learning*, 2003.

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

[34] P. Mertikopoulos and E. V. Belmega, "Adaptive spectrum management in MIMO-OFDM cognitive radio: An exponential learning approach," in *ValueTools '13: Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools*, 2013.

[35] F. Alvarez, J. Bolte, and O. Brahic, "Hessian Riemannian gradient flows in convex programming," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 477–501, 2004.

[36] P. Coucheney, B. Gaujal, and P. Mertikopoulos, "A class of robust entropy-driven learning procedures in games," Inria, Tech. Rep., 2013. [Online]. Available: http://arxiv.org/abs/1303.2270

[37] J. Kwon and P. Mertikopoulos, "A continuous-time approach to online optimization," forthcoming.

[38] 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.*, vol. 56, no. 2, p. 389, March 2007.

[39] G. Scutari, "Simultaneous iterative water-filling for Gaussian frequency-selective interference channels," in *ISIT '06: Proceedings of the 2006 International Symposium on Information Theory*, 2006.

[40] S. Sorin, "Exponential weight algorithm in continuous time," *Mathematical Programming*, vol. 116, no. 1, pp. 513–528, 2009.

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

[42] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*, ser. Probability and Mathematical Statistics. New York: Academic Press, 1980.

[43] K. Azuma, "Weighted sums of certain dependent random variables," *Tôhoku Mathematical Journal*, vol. 19, no. 3, pp. 357–367, 1967.