Online Power Optimization in Feedback-Limited, Dynamic and Unpredictable IoT Networks

Alexandre Marcastel, *Student Member, IEEE*, E. Veronica Belmega, *Member, IEEE*, Panayotis Mertikopoulos, *Member, IEEE*, and Inbar Fijalkow, *Senior Member, IEEE*

***Abstract*—One of the key challenges in Internet of Things (IoT) networks is to connect many different types of autonomous devices while reducing their individual power consumption. This problem is exacerbated by two main factors: *a*) the fact that these devices operate in and give rise to a highly dynamic and unpredictable environment where existing solutions (e.g., water-filling algorithms) are no longer relevant; and *b*) the lack of sufficient information at the device end. To address these issues, we propose a regret-based formulation that accounts for *arbitrary network dynamics*: this allows us to derive an online power control scheme which is provably capable of adapting to such changes, while relying solely on *strictly causal feedback*. In so doing, we identify an important tradeoff between the amount of feedback available at the transmitter side and the resulting system performance: if the device has access to unbiased gradient observations, the algorithm's regret after T stages is $\mathcal{O}(T^{-1/2})$ (up to logarithmic factors); on the other hand, if the device only has access to scalar, utility-based information, this decay rate drops to $\mathcal{O}(T^{-1/4})$. The above is validated by an extensive suite of numerical simulations in realistic channel conditions, which clearly exhibit the gains of the proposed online approach over traditional water-filling methods.**

***Index Terms*—IoT networks, online exponential learning, imperfect and scarce feedback**

I. Introduction

THE emerging Internet of things (IoT) paradigm is projected to connect billions of wireless "things" (wireless sensors, wearables, biochip transponders, etc.) in a vast network with drastically different requirements between components (e.g. in terms of throughput and power characteristics) [3]. Following Moore's prediction on silicon integration, the wireless surroundings of IoT networks are expected to exhibit massive device densities with high interference levels. An orthogonal spectrum allocation is therefore energetically inefficient, as an unrealistic number of bands or subcarriers would be required to accommodate all devices. The usage of new access protocols such as non-orthogonal multiple access (NOMA) [4] is considered instead, in which interference mitigation becomes critical. For this reason, and also given that the autonomous wireless devices have stringent battery limitations, optimizing the power consumption emerges as one of the key ingredients for achieving a "speed of thought" user experience at the application level [5].

A major challenge that arises here is that IoT networks are characterized by an unprecedented degree of temporal variability – due itself to the unique mobility attributes of modern wearable devices, intermittent user activity, application diversity etc. As such, IoT networks cannot be treated as static (or stationary) systems, implying in turn that conventional optimization techniques that *target a fixed state,* for instance via water-filling type of algorithms, are no longer relevant. The main limitation of classical approaches is their lack of robustness to strictly causal – *no look-ahead* – channel state information, which is inevitable in dynamic, unpredictable environments. Therefore, power optimization in dynamic IoT networks calls for a different toolbox that is provably capable of adapting to unpredictable changes in the network.

Motivated by its prolific success in the fields of machine learning and artificial intelligence [6, 7], we propose in this paper a *regret-based* formulation of power optimization which allows us to consider arbitrary variations in the network. The core component of this approach is that, instead of targeting a specific network state, it aims to derive an online power allocation policy whose performance over time is as close as possible to that of the best fixed policy in hindsight (even though computing the latter requires non-causal knowledge of the system parameters and their evolution ahead of time). Owing to this straightforward and flexible definition, regret minimization has become the leading paradigm for online decision making in uncertain, dynamic environments, ranging from online ad auctions [8, 9] and recommender systems [10] to throughput and energy efficiency optimization problems in wireless communications [11, 12].

A critical performance limitation in the above is the fact that wireless devices in IoT networks typically receive limited and/or corrupted feedback from their environment [13]. To name but an example, channel state information (CSI) is usually acquired by the access point (AP) using

A. Marcastel, E. V. Belmega and I. Fijalkow are with ETIS, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, Cergy-Pontoise, France. P. Mertikopoulos is with Univ. Grenoble Alpes, CNRS, Inria, LIG, Grenoble, France.

This research was supported in part by the Orange Lab Research Chair on IoT within the University of Cergy-Pontoise, by the French National Research Agency (ANR) project ORACLESS (ANR–16–CE33–0004–01), by the ELIOT ANR-18-CE40-0030 and FAPESP 2018/12579-7 project, and by ENSEA, Cergy-Pontoise, France. Part of this work was presented in VTC2016-Fall [1] and GLOBECOM 2016 [2].

pilot transmissions that are subsequently fed back to each device. Since IoT networks bring together massive numbers of devices, the signaling overhead increases to the point where it cannot be distributed over multiple frequency bands in an efficient manner (due to spectrum scarcity) [14, 15]. Therefore, to reduce the impact of this overhead, the amount of information fed back to wireless devices must be reduced as much as possible, and the resulting estimation errors must be likewise taken into account. The same kind of problem has been underlined in cooperative multi-user networks [16], in which the global network optimum objective leads to massive signaling; and in massive multiple-input and multiple-output (MIMO) systems [17–20], in which the increase in the number of antennas leads to a prohibitive amount of required CSI. Instead, in IoT networks, it is not the number of antennas but the large number of connected devices that create this bottleneck.

A. Summary of contributions and paper outline

In the field of online learning, the challenges that result from incomplete and/or imperfect feedback have been studied extensively in the context of the so-called multi-armed bandit problems [6]. These problems are inherently discrete in nature, so the lessons learned from this literature do not apply to the power allocation framework studied here (a continuous, multi-dimensional problem in itself). Nevertheless, by leveraging ideas originating in the well-known exponential weights algorithm for multi-armed bandits [6], we derive an online power allocation policty based on exponentiated gradient descent (EGD), and which comprises two basic steps: *a*) tracking the gradient of the users' power minimization objective in a dual, unconstrained space; and *b*) using a judiciously designed exponential function to map the output of this step to a feasible power allocation profile and keep going. Thanks to this two-step, primal-dual approach, we are then able to derive concrete regret minimization guarantees for the online power minimization problem, irrespective of the network's dynamics.

To establish a benchmark, we begin with the *full information* or the first-order feedback case, where each wireless device is assumed to have perfect feedback on the gradient of its individual power minimization objective. In this case, the proposed power allocation policy is shown in Section III to enjoy a $\mathcal{O}(T^{-1/2})$ regret guarantee, meaning that the algorithm's performance over a horizon of T transmission cycles is no more than $\mathcal{O}(T^{-1/2})$ away from the best fixed policy in hindsight. Importantly, unless rigid statistical hypotheses are made for the underlying IoT network (such as assuming that it evolves following a stationary ergodic process), this guarantee cannot be improved; however, we show in Section IV that it can still be attained even if the feedback received by each device is imperfect and/or otherwise corrupted by non-systematic measurement errors and observational noise.

In addition to providing a comparison baseline, the full information case also allows us to compare the performance of the proposed algorithm to that of classical water-filling algorithms [21–23] and highlight the difficulties encountered by the latter when the network evolves dynamically over time and only a strictly causal (with no look-ahead) feedback information is available at the transmitter.

On the other hand, if the only information received by each device is the observed value of their power minimization objective (the so-called *zeroth-order feedback* setting), these bounds change significantly. Lacking any sort of vector-valued, gradient-based feedback, we rely on simultaneous stochastic approximation techniques [6, 7], to build an estimator for the gradient: importantly, this estimator is potentially biased, but its bias can be controlled by tuning a certain sampling parameter. By jointly optimizing the value of this parameter and that of the original algorithm's step-size, we then show that the proposed policy still leads to no regret, but now at a slower rate of $\mathcal{O}(T^{-1/4})$.

In Section VI, we validate our theoretical analysis via numerical experiments and highlight highly dynamic networks with realistic, unpredictable channel conditions. Classical water-filling algorithms are very sensitive to unpredictable changes in the network and are outperformed by our proposed online algorithms in terms of power consumption and achieved rate. Concerning the impact of available feedback, our numerical results also illustrate a compromise between the amount and/or quality of the feedback information and the algorithms' performance (measured here in terms of the time needed to attain a no-regret state). The zeroth-order feedback case requires only the knowledge of a scalar at each iteration (the value of the objective function) as opposed to a vector (the gradient), but the average time required to reach a no-regret state is higher.

B. Related works

Regarding resource allocation in static IoT environments, several problems have been studied [24–26]. In [24] the authors study the resource allocation for machine to machine (M2M) communications using cooperative game theoretic tools in which the machines want to maximize their own rate. In [25], the authors study the problem of clustering and power allocation for both uplink and downlink in NOMA systems. Similar to [24], each device aims at maximizing its own rate. To solve this problem, the authors used classical optimization tools. In [26], the problem of power control for mutual interference avoidance is studied by using also classical optimization tools. In all these works, the network is assumed to remain static over time and the devices are required to have perfect feedback information. Here, we relax both assumptions by taking into account the inherent dynamics of an IoT network and the impact of feedback imperfections and scarcity.

In (non-IoT) wireless networks, there exists a wide resource allocation literature essentially concerned with either static [21–23, 27, 28] or stochastic [16, 20, 29–35] optimization problems, to cite but a few. In these works, the

underlying network is assumed to remain static or to evolve following a stationary random process. Their main aim is to derive efficient algorithms, based on classical optimization, stochastic optimization, or on machine learning tools, that converge to an optimal fixed or steady state. These works are inherently different from the present paper, in which we squarely focus on arbitrarily dynamic networks (the network can even evolve in a non-stationary way). In such unpredictable networks, there is no fixed *solution state* to converge to, so the very notion of *convergence* as a performance metric needs to be rethought from the ground up.

Adaptive allocation policies based on online optimization tools have been recently proposed but in quite different settings and problems [11, 12, 36–38]. In [36], the authors proposed a multi-armed bandit formulation of the channel selection problem and derived an online channel selection algorithm using upper confidence bound techniques; a similar approach has also been used in the context of beam-alignment for millimeterWave communications [37]. In [38], an adversarial multi-armed bandit formulation is proposed to tackle an access point association problem in hybrid indoor LiFi-WiFi communication systems exploiting the exponential weights algorithm. For IoT networks, the recent work [39] acknowledges the high potential of the online learning framework and then focuses on multi-armed bandits for mobile computation offloading problems at the edge layer. However, in our setting, the agents' decisions are not taken within a stochastic environment (so upper confidence bounds are not applicable) and all variables are continuous as opposed to discrete (so multi-armed bandits are not suitable).

Regarding physical-layer resource allocation problems, the authors of [11, 12] studied dynamic MIMO systems from the point of view of online throughput and energy efficiency maximization. By contrast, our focus here is the power minimization problem in IoT networks, which is inherently different. Specifically, in the online throughput maximization problem in [11], the opportunistic devices have to always transmit at full available power, which is not power-efficient and the proposed learning algorithm does not apply to the problem at hand. The energy efficiency (defined as the ratio between the achieved rate and the overall power consumption) maximization problem in [12] is non-convex and is cast into a convex problem by performing a suitable variable change, which results into a specific exponential learning algorithm that also does not apply here. However, the learning algorithms in these works rely on the availability of gradient information which amounts to a (typically large) matrix worth of feedback; by contrast, the algorithm provided in this paper only requires a single readily available scalar as feedback at the device end.

To the best of our knowledge, our paper is the first in the IoT literature to take into account the network's inherent dynamics and its unpredictable temporal variability when designing power-efficient allocation policies. Furthermore, it is among the first works in the resource allocation literature in multi-user wireless networks, proposing an adaptive algorithm relying only on a single scalar feedback information (the sole past experienced objective value).

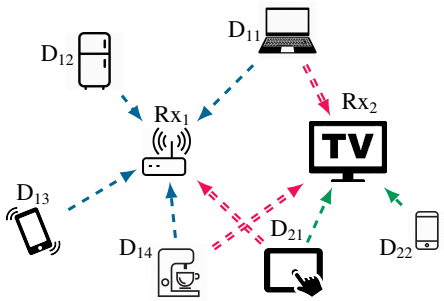


Figure 1: System composed of six transmit devices (D_{11}, D_{12}, etc.) and two receivers (Rx_1, Rx_2). The blue and green arrows represent the direct links while the red (double-lined) ones are interfering links.

II. Model and Problem Formulation

We consider a system composed of M transmitters and N receivers communicating over S orthogonal subcarriers or sub-bands as illustrated in Fig. 1: each device transmits to only one intended receiver, but a given receiver may decode several incoming signals.

Since we aim at devising a distributed policy that needs no central controller, we can focus on one particular transmitting-receiving pair. The received signal for the (arbitrarily chosen) focal device becomes:

$$y^s(t) = h^s(t)x^s(t) + \sum\nolimits_j h_j^s(t)x_j^s(t) + z^s(t), \tag{1}$$

where $s \in \{1, \ldots, S\}$ is the subcarrier index; $x^s(t)$ is the transmitted signal; $h^s(t)$ is the channel gain between the focal transmitter and its intended receiver; x_j^s is the transmitted signal of device j; $h_j^s(t)$ is the interfering channel gain between transmitter j and the focal receiver; and $z^s(t)$ is the received noise of the focal device.

We also define the effective channel gain vector $\mathbf{w}(t) = (w^s(t))$, where $w^s(t)$ represents the effective gain in subcarrier s and is given by

$$w^s(t) = \frac{g^s(t)}{\sigma^2 + \sum_j g_j^s(t)p_j^s(t)}, \quad \forall s, \tag{2}$$

where σ^2 is the variance of the noise $z^s(t)$, $p_j^s(t)$ is the transmitted power by the user j in subcarrier s, $g_j^s(t) = |h_j^s(t)|^2$ and $g^s(t) = |h^s(t)|^2$.

The above expression implies that the receiver employs single-user decoding (SUD), meaning that, when decoding a transmitted signal, the other incoming signals are treated as noise. This consideration is relevant in distributed and energy-limited networks, such as IoT networks, in which the receivers may not be able to decode the interfering signals (e.g., may not know the codebooks of the interferers). Also, the receivers may not afford to sequentially process and decode their incoming signals (via successive interference

cancellation) and the transmitting devices may not be coordinated (and may not know their decoding order).

The aim of IoT networks is to interconnect many different types of devices in a distributed or self-optimizing way. Since most of them are likely to be small devices like sensors, phones, or isolated devices that operate solely on limited batteries, reducing the power consumption is a key challenge in IoT networks [40, 41]. In this work, the main objective is to minimize the power consumption of the focal device taking into account its quality of service (QoS) requirements. These requirements are performance targets, which depend on the specific application, for example: a) minimum rate per device (i.e., the rate of the focal device has to be higher than a given threshold R_{min}); b) minimum SINR per device; c) minimum network sum-rate. Such QoS requirements differ from physical hard constraints (e.g., the transmit power positivity constraints) in that they cannot be ideally guaranteed - always at 100% - in practical communication systems and some outage has to be generally tolerated.

In view of the above, the trade-off between power minimization and QoS requirements will be modeled via the loss function:

$$L_t(\mathbf{p}) = \sum_{s=1}^{S} p^s \quad + \lambda \ \left[R_{min} - R_t(\mathbf{p})\right]^+ \tag{3}$$

where $\mathbf{p} = (p^1, \ldots, p^S)$ represents the power allocation vector of the focal device with components $p^s, \forall s$ representing the power allocated to the s-th subcarrier. The first term in the objective is the overall power consumption and the second term is a soft-constraint (or penalty) term, which is activated whenever the minimum target rate R_{min} is not achieved. Finally, $R_t(\mathbf{p})$ denotes the well-known Shannon rate:

$$R_t(\mathbf{p}) = \sum_{s=1}^{S} \log(1 + w^s(t)p^s) \tag{4}$$

and $[x]^+ \triangleq \max\{x, 0\}$, meaning that no penalty is applied when the achieved rate is greater than the threshold $R_t(\mathbf{p}) \geq R_{min}$. Although we choose a linear penalty function for its relevance to communications [28, 42, 43] (and to simplify the presentation), our results carry over the more general class of concave functions, e.g., logarithmic penalties [44]. The parameter λ can also be interpreted as the unit-cost for each bps/Hz under the QoS target R_{min} and, as we will see in Section VI, it also represents a sensitivity parameter that has to be carefully tuned to adjust the flexibility regarding the minimum rate constraint violations or outages. Indeed, higher values of λ lead to less QoS outages, but at the cost of incurring a higher power consumption.

To sum up, the online optimization problem under study can be stated as:

$$\begin{array}{ll} \text{minimize} & L_t(\mathbf{p}(t)) \\ \text{over} & \mathbf{p}(t) = (p^1(t), \ldots, p^S(t)) \\ \text{subject to} & p^j(t) \geq 0, \ \ \forall j \in \{1, \ldots, S\} \\ & \sum_{s=1}^{S} p^s(t) \leq P_{\max} \ . \end{array} \tag{5}$$

The minimization variable is the power allocation vector of the focal device across the available frequency subcarriers, $\mathbf{p}(t)$, and both constraints are physical ones. The first constraint guarantees that the transmit power of the focal device, in each subcarrier j, is always positive. The second constraint comes from the power supply limitation and implies that the total power of the focal device that is spread over the subcarriers is bounded from above by the maximum transmit power of the device.

Concerning the above objective function $L_t(\mathbf{p})$, notice that it *may vary in a non-stationary and unpredictable way* such that the focal device cannot determine *a priori* (before the transmission takes place) its instantaneous or dynamic optimal power allocation $\mathbf{p}^*(t)$ that minimizes this objective at each time t. Nevertheless, we assume that the device receives some feedback after each transmission, such as the past experienced objective value or its past gradient. The idea in online optimization is to exploit this strictly causal feedback information to build a dynamic and adaptive power allocation policy $\mathbf{p}(t)$ that *minimizes as much as possible* the time-varying objective function $L_t(\mathbf{p}(t))$[1].

The major novelty in the above formulation relative to more classical power allocation problems lies in its dynamic nature and the fact that we make *no assumptions* on the network dynamics. Notice that, the objective function in (3) depends on the network dynamics via the second penalty term. Indeed, the achieved rate $R_t(\mathbf{p}(t))$ depends on the varying wireless channels and also on the power allocation policies of the other devices via the effective channel gains in (2). Classical approaches leading to water-filling type of algorithms [21–23] rely both on static (or stationary) channel models and on strong assumptions on the information available at the transmitter before the transmission takes place (e.g., perfect channel state information in the form of the SINR in each subcarrier). In highly dynamic and distributed IoT networks, such assumptions are too stringent and no longer hold.

On that account, the aim of this work is twofold: to explicitly take into account the device mobility, their network connectivity patterns and behaviour, which may be completely arbitrary and unpredictable; and to greatly reduce the information required at the transmitter.

In order to evaluate the performance of a given online

[1]Going back to our model of the QoS requirement, another motivation for including it into the objective function in (3), as opposed to imposing a hard constraint, is that the latter would result in an arbitrarily time-varying and unpredictable feasible set at the decision instant. This issue is highly non-trivial and open in online optimization, which would require going well beyond the standard regret minimization framework and, hence, falls out of the scope of this work.

policy $\mathbf{p}(t)$, the most commonly used notion is that of the regret [6, 7, 11, 12], which compares its performance in terms of loss with a benchmark policy. Now, comparing any policy $\mathbf{p}(t)$, built using outdated feedback information, with the instantaneous or dynamic optimal solution $\mathbf{p}^*(t)$ is obviously too ambitious. Instead, the notion of regret compares the policy $\mathbf{p}(t)$ to a less ambitious benchmark: the fixed strategy that minimizes the overall objective over a given transmission horizon T:

$$Reg(T) \triangleq \sum_{t=1}^{T} L_t(\mathbf{p}(t)) - \min_{\mathbf{q}\in\mathcal{P}} \sum_{t=1}^{T} L_t(\mathbf{q}), \tag{6}$$

where $\mathcal{P} \triangleq \left\{\mathbf{p} \in \mathbb{R}^S \;\middle|\; p^s \geq 0, \forall s, \sum_{s=1}^{S} p^s \leq P_{max}\right\}$ denotes the feasible set. Otherwise stated, the regret measures the performance gap between a power allocation policy $\mathbf{p}(t)$ and the best mean optimal solution over a fixed horizon T. If the regret is negative, then the dynamic policy $\mathbf{p}(t)$ outperforms the best mean optimal solution overall. To quantify this, the policy $\mathbf{p}(t)$ is said to lead to no-regret if

$$\limsup_{T\to\infty} \frac{1}{T} Reg(T) \leq 0. \tag{7}$$

A no-regret policy $\mathbf{p}(t)$ is asymptotically optimal and performs at least as good as the best fixed strategy on average (when T grows large).

Notice that although the best mean optimal solution is less ambitious than $\mathbf{p}^*(t)$ (minimizing the objective function at each t) its computation requires the same non-causal knowledge of the system parameters and the evolution of the objective throughout the time horizon T before the transmission takes place, or in hindsight. Therefore, the design of dynamic policies that reach no-regret while relying on strictly causal and local information is a remarkable and desirable goal. Moreover, in the particular case of a static network composed of a single transmit device, the online optimization problem in (5) reduces to a classic convex optimization problem. A no-regret online policy $\mathbf{p}(t)$ in this case implies the convergence of the average policy: $\overline{\mathbf{p}}(t) \triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}(t)$ to the solution set of the relevant optimization problem [45]. In conclusion, given the IoT network dynamics and unpredictability, our focus in the following is precisely to develop no-regret online policies for the online optimization problem defined in (5).

III. First-Order feedback

In the resource allocation problem under study, the focal device has to choose in which of the available subcarriers to transmit, how much of the available power to consume and how to split this amount over the chosen subcarriers, all this based on the strictly causal feedback information. This is reminiscent of the well-known multi-armed bandit problem in sequential online learning [7]: there, at each instant, the player chooses an action (or an arm) out of several possibilities and receives a reward as a result. Outside the so-called "stochastic" case (where each arm's payoff is determined by a fixed probability distribution), the most widely used algorithmic scheme is the exponential (or multiplicative) weights algorithm [7], where payoffs are aggregated over time and the optimizer selects an arm with a probability proportional to the exponential of these scores. In what follows, we derive the necessary machinery to extend this idea to the continuous optimization problem at hand and derive an exponentiated gradient descent algorithm for power minimization in this context.

OXL algoritm: Online Exponential Learning Algorithm

Initialization: $\mathbf{y}(0) \leftarrow 0; t \leftarrow 0.$

Repeat

- ▹ Pre-transmission phase: update transmit powers
- $\mathbf{p}(t) \leftarrow \mathbf{Q}(\mathbf{y}(t))$ defined in (OXL)
- ▹ Transmit at $\mathbf{p}(t)$
- ▹ Post-transmission phase: receive gradient feedback $\mathbf{v}(t)$
- Update scores $\mathbf{y}(t+1) \leftarrow \mathbf{y}(t) - \mu(t)\,\mathbf{v}(t)$
- $t \leftarrow t+1$

until transmission ends

In our setup, we begin by assuming that each device has access to some feedback mechanism that provides the first-order gradient information $\mathbf{v}(t) = \nabla L_t(\mathbf{p}(t))$ at the end of each transmission. Our proposed algorithm can be summarized in two steps. First, the device tracks the past gradient of its objective without taking account the power constraints. Second, the device maps the first step into the feasible set $\mathcal{P}$ using a well chosen exponential map as follows:

$$\begin{aligned} &\mathbf{y}(t) = \mathbf{y}(t-1) - \mu\,\mathbf{v}(t), \\ &p^s(t) = Q^s(\mathbf{y}(t)) \triangleq P_{\max} \frac{\exp(y^s(t))}{1+\sum_{i=1}^{S}\exp(y^i(t))}, \end{aligned} \tag{OXL}$$

where μ is the step-size parameter. We denote by $\mathbf{Q}(\mathbf{y}(t)) = (Q^1(t), \ldots, Q^S(t))$ the exponential vector field that maps the updated score $\mathbf{y}(t)$ into the feasible set.

Essentially, the online exponential learning algorithm detailed above, tracks the cumulative negative gradient of the convex loss function and then maps the result to the feasible set. The exponential mapping step could be replaced by an Euclidean projection and the resulting algorithm would be an online gradient descent [46] algorithm. We chose the exponential mapping because of its *reduced complexity* relative to a projected gradient descent algorithm that would require an additional (possibly costly) projection step. Indeed, from (OXL) it is easy to see that the updates are easy to compute and that they meet the contraints. More precisely, the complexity of each iteration t is linear in the problem dimensionality S, the number of subcarriers over which the focal device transmits. Hence, given that S is not expected to grow large for a specific IoT device (transmitting on a small subset of the total number of subcarriers available to the entire IoT network), the OXL algorithm is particularly appealing for distributed, device-centric IoT networks.

We will now study the evolution of the regret of the dynamic power allocation policy (OXL) to show that it

holds the no-regret property. To that end, let V denote an upper bound on the gradient feedback $\mathbf{v}(t)$ in the sense that $\|\mathbf{v}(t)\|^2 \leq V$. We then have the following result (for a proof, see Appendix A):

Theorem 1. *If the OXL algorithm is run with a constant step-size μ then, it enjoys the regret bound:*

$$Reg(T) \leq \frac{P_{\max}\log(1+S)}{\mu} + \frac{\mu P_{\max} TV}{2}. \tag{8}$$

Tuning the step-size μ: The step-size μ plays an important role in the *exploration vs. exploitation* tradeoff and, hence, in the ability of OXL algorithm to reach the no regret state, as we will see in the following. Intuitively, small values of μ imply that the subcarriers are almost equally explored and the gradient information is not exploited enough. High values of μ imply that only the best performing carriers w.r.t. past gradients are exploited and highly potential carriers, which have not performed well in the past, are rooted out too soon.

Notice that the above upper bound grows linearly with T, which may lead to a non-zero average regret. Nevertheless, this bound is a convex function of the step-size μ and, hence, can be minimized w.r.t. μ by setting the first-order derivative to zero. The resulting optimal step-size is

$$\mu^* = \sqrt{2\log(1+S)/(TV)}, \tag{9}$$

which then yields the sub-linear optimal bound

$$Reg(T) \leq P_{\max}\sqrt{2TV\log(1+S)}. \tag{10}$$

Therefore, by carefully choosing the step-size μ, OXL algorithm leads to no regret:
$\limsup_{T\to\infty} \frac{1}{T} Reg(T) = 0$.

Corollary 1. *If the OXL algorithm is run for a known horizon T using the optimal step-size μ^* in* (9)*, then it leads to no regret and the average regret $Reg(T)/T$ decays as $\mathcal{O}(T^{-1/2})$.*

The resulting regret bound in (10) depends on the system parameters: the total power $P_{\max}$, the number of subcarriers S, an upper bound on the gradient norm V, but also on the transmission horizon T, which the device does not necessarily know in advance. To avoid this limitation, we use the doubling trick [47]: the algorithm is run repeatedly starting with a unit-size window (number of iterations) and then doubling the window size at each new run until transmission ends. Hence, each widow size is known and the device can compute the corresponding optimal step μ^* (by replacing T with the window size in (9)). The bound in Corollary 1 applies in each window and the following result is proven in Appendix B.

Proposition 1. *If the OXL algorithm is run when the transmission time T is unknown by using the doubling trick with an optimal step-size for each window until the transmission ends (as in Figure 2), the regret enjoys the following bound:*

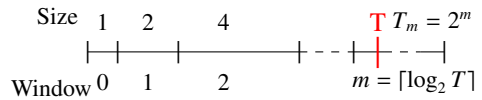


Figure 2: Illustration of the windows in the doubling trick.

$$Reg(T) \leq \frac{2}{\sqrt{2}-1} P_{\max}\sqrt{2TV\log(1+S)}. \tag{11}$$

Hence, OXL algorithm leads to no-regret and the average regret $Reg(T)/T$ decays as $\mathcal{O}(T^{-1/2})$.

We observe that not knowing the horizon T in advance results only in a small loss in the regret bound (in the multiplying constant).

IV. Imperfect gradient feedback

In this section, we relax the assumption of perfect gradient feedback and we consider that the focal device has access only to an imperfect gradient estimate, denoted by $\tilde{\mathbf{v}}(t)$, which meets the following conditions

$$\begin{aligned} \mathbb{E}[\tilde{\mathbf{v}}(t)] &= \nabla L_t(\mathbf{p}(t)), \\ \mathbb{E}[\|\tilde{\mathbf{v}}(t)\|^2] &\leq \tilde{V}, \end{aligned} \tag{12}$$

where the expectation is taken over the randomness of the estimator. These conditions are not very restrictive as they require the absence of systematic errors and a bounded variance, as such, they are satisfied by all common error distributions (Gaussian, log-normal, etc) [12]. For example, the common error model: $\tilde{\mathbf{v}}(t) = \nabla L_t(\mathbf{p}(t)) + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma_z^2\mathbf{I})$ [48] satisfies the above conditions.

Under these assumptions, the transmit powers in OXL algorithm are updated in function of $\tilde{\mathbf{v}}$ instead of the actual gradient $\mathbf{v}$ (via the internal score $\mathbf{y}(t)$). Thus, the online policy $\mathbf{p}(t)$ depends on the randomness of the estimator, which implies that the regret in (6) will also depend on this randomness. To take this into account, we study the *average regret* $\mathbb{E}[Reg(T)]$, where the expectation is taken over the randomness of the estimator. The no-regret property can be easily extended to the average regret as follows: a power allocation policy $\mathbf{p}(t)$ leads to no regret (on average) if

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[Reg(T)] \leq 0. \tag{13}$$

A different possibility would be to study the probability that the regret in (6) falls bellow zero, but we leave this as a non-trivial open issue for future investigation.

Our first result, proved in Appendix C, concerns the case in which the transmission horizon T is known.

Theorem 2. *If the OXL algorithm is run for T iterations with a constant step-size μ and an imperfect gradient estimation defined in* (12)*, the average regret is bounded by:*

$$\mathbb{E}[Reg(T)] \leq \frac{P_{\max}\log(1+S)}{\mu} + \frac{\mu}{2} P_{\max} T\tilde{V}. \tag{14}$$

We observe that the upper bound above is somewhat similar to the one in Theorem 1 and can also be minimized with respect to the step-size μ for the same reasons. The optimal step-size is

$$\mu^* = \sqrt{2\log(1+S)/(T\tilde{V})}, \tag{15}$$

which provides the optimal bound

$$\mathbb{E}[Reg(T)] \leq P_{\max}\sqrt{2T\tilde{V}\log(1+S)}. \tag{16}$$

Therefore, using this optimal step-size leads to no regret even if the device only has access to imperfect gradient observations.

Corollary 2. *If the OXL algorithm is run for a known transmission time T with an imperfect gradient feedback and the optimal step-size μ^* in* (15)*, then the no-regret property holds and the average regret $\mathbb{E}[Reg(T)]/T$ decreases in $\mathcal{O}(T^{-1/2})$.*

The above result implies that *unbiased errors in the gradient estimation do not impact* greatly the evolution of the regret in expectation. This result should be contrasted to the corresponding one in the perfect gradient case. First, in the perfect gradient case, there is no randomness and the regret results are deterministic. Here, because of the random errors in the estimated gradient the above results hold only in expectation. Second, the obtained upper bounds depend on $\tilde{V}$, which is an upper bound on the second order statistics of the estimation over the entire time horizon (as opposed to the Lipschitz constant V). This means that the variance of the errors negatively impact the expected regret; higher variance errors result in higher expected regret.

Finally, if the device does not know in advance the transmission horizon T, the doubling trick described in Sec. III requires the knowledge of $\tilde{V}$ (to compute the optimal step-size). This may not be realistic in an unpredictable, time-varying and possibly non-stationary environment. To avoid this additional requirement, we take a variable step-size approach as in [6, 7, 49] and focus on the schedule

$$\mu(t) = \alpha/\sqrt{t}, \tag{17}$$

with parameter $\alpha > 0$. Using this variable step-size, we obtain the following result (for a proof, see Appendix D):

Theorem 3. *If the OXL algorithm is run with imperfect gradient feedback for an unknown horizon T and using the variable step-size $\mu(t) = \alpha t^{-1/2}$, then the average regret is bounded by*

$$\frac{\mathbb{E}[Reg(T)]}{T} \leq \frac{P_{\max}\log(1+S)}{\alpha\sqrt{T}} + \frac{P_{\max}\tilde{V}\alpha(1+\log T)}{2\sqrt{T}}. \tag{18}$$

Consequently, the device's average regret $\mathbb{E}[Reg(T)]/T$ vanishes as $\mathcal{O}(\log(T)T^{-1/2})$, i.e. OXL algorithm leads to no regret.

We remark the loss in the decay rate of the regret resulting from the lack of knowledge of $\tilde{V}$. This means that, with scarcer available knowledge, the device will reach the no regret state at a slower rate. Nevertheless, this loss is only logarithmic and even without the knowledge of T and relying on an imperfect and unbiased gradient feedback, the OXL algorithm is able to reach no regret.

V. Zeroth-order Feedback

In this section, our objective is to reduce even further the amount of required information to be fed back to the transmitting device. Instead of receiving a vector as feedback - the gradient or its unbiased estimation - the devices are now assumed to know only the value of the experienced objective function. This means that only a single scalar worth of information is needed at the transmitting device – a major advantage in feedback-limited and dynamic networks, where the acquisition of non-causal and complete channel state information (not to mention other network parameters) is a tall order. To the best of our knowledge, the proposed algorithm is the first adaptive power allocation algorithm for multiple-carrier, multiple-user networks requiring scalar feedback. Classic resource allocation algorithms such as water-filling policies require at least one quality indicator per subcarrier (e.g., the SINR value in each subcarrier), and, hence a (possibly large) vector worth of feedback.

To develop an online policy $\mathbf{p}(t)$ that leads to no regret, we modify the exponential mapping step in Sec. III and propose a novel learning algorithm that only requires zeroth-order feedback. To this aim, the first obstacle is to estimate the gradient of the objective based only on its value – in other words to do "gradient descent without a gradient" [50]. The main idea that we exploit here is the simultaneous stochastic approximation technique, which randomly samples the objective function in a neighbourhood of the power policy $\mathbf{p}(t)$ to obtain a *(potentially biased) estimate* of the gradient at this point [6, 7, 50].

For simplicity, we illustrate this technique on a particular directional derivative of $L_t(\mathbf{p})$ along the unit vector $\mathbf{x}$ (recall that the gradient is a collection of directional derivatives), denoted by $\nabla_{\mathbf{x}}L_t(\mathbf{p})$:

$$\nabla_{\mathbf{x}}L_t(\mathbf{p}) = \lim_{\delta\to 0}\frac{L_t(\mathbf{p}+\delta\mathbf{x}) - L_t(\mathbf{p}-\delta\mathbf{x})}{2\delta}, \tag{19}$$

which we want to estimate based on the single function value $L_t(\mathbf{p})$. To do so, we randomly sample the objective function around the point $\mathbf{p}$ in the direction $\mathbf{x}$ by drawing a Bernoulli distributed random variable $u \in \{-1, +1\}$ with equal probability. We can compute the expectation of these samples w.r.t. the randomness of u:

$$\mathbb{E}\left[L_t(\mathbf{p}+\delta u\mathbf{x})u\right] = \frac{L_t(\mathbf{p}+\delta\mathbf{x}) - L_t(\mathbf{p}-\delta\mathbf{x})}{2}. \tag{20}$$

From (19) and (20), we observe that

$$\mathbb{E}\left[\frac{L_t(\mathbf{p}+\delta u\mathbf{x})u}{\delta}\right] \approx \nabla_{\mathbf{x}}L_t(\mathbf{p}). \tag{21}$$

Since the above is satisfied with equality only in the limit when $\delta \to 0$, the quantity $L_t(\mathbf{p} + \delta u\mathbf{x})u/\delta$ represents an approximation (possibly biased) of the directional derivative of $L_t(\mathbf{p})$ with respect to $\mathbf{x}$.

Now, in order to build a gradient estimate, the idea is to uniformly sample the objective function along a vector $\mathbf{u}(t)$ drawn from the S-dimensional Euclidean sphere of of radius δ. Extending the above to the space of dimension S, the estimator becomes:

$$\tilde{\mathbf{v}}(t) = \frac{S}{\delta} L_t(\tilde{\mathbf{p}}(t))\mathbf{u}(t), \tag{22}$$

where $\tilde{\mathbf{p}}(t) = \mathbf{p}(t) + \delta\mathbf{u}(t)$ and $\mathbf{u}(t)$ is uniformly taken over the unit Euclidean sphere: $\{\mathbf{u} \in \mathbb{R}^S |\ \|\mathbf{u}(t)\|^2 = 1\}$ [6]. More details are provided in Appendix E.

In [6, 7, 50], this estimator is proposed without accounting for the fact that the random sample point $\tilde{\mathbf{p}}(t) = \mathbf{p}(t) + \delta\mathbf{u}(t)$ can fall outside of the feasible set. In our power allocation problem, using the same procedure would imply that the transmit power vector $\tilde{\mathbf{p}}(t)$ is allowed to go outside $\mathcal{P}$. However, our power constraints are physical ones: transmit power positivity, maximum available power budget, which means that any violations are prohibited.

One of the major contributions of this work is to introduce a novel learning algorithm that exploits the gradient estimation above, while guaranteeing that the transmit powers always lie in the feasible set. For this, we define a modified and shrunk feasible set $\mathcal{P}_\delta$ such that, for any $\mathbf{p}_\delta(t) \in \mathcal{P}_\delta$, we have $\mathbf{p}_\delta(t) + \delta\mathbf{u}(t) \in \mathcal{P}$:

$$\mathcal{P}_\delta = \left\{ \mathbf{p}_\delta \in \mathbb{R}^S \;\middle|\; p_\delta^s \geq \delta, \sum_{s=1}^{S} p_\delta^s \leq P_{\max} - \sqrt{S}\delta \right\}. \tag{23}$$

Having defined this new feasible set, the suitable exponential map that guarantees that $\mathbf{p}_\delta(t) + \delta\mathbf{u}(t)$ always lies in this set is

$$p_\delta^s(t) \triangleq \delta + P_{\max}(1 - C_\delta)\frac{\exp(y^s(t))}{1 + \sum_{i=1}^{S}\exp(y^i(t))}, \tag{EXPδ}$$

where $C_\delta = \frac{\delta}{P_{\max}}(S + \sqrt{S})$. Using (EXP$\delta$), we introduce a novel exponential mapping: $\mathbf{Q}_\delta(\mathbf{y}(t)) \triangleq (p_\delta^1(t), \ldots, p_\delta^S(t))$. From the definition of $\mathcal{P}_\delta$ and (EXPδ), we can deduce the following conditions restricting the choice of the δ parameter

$$0 < \delta \leq \frac{P_{\max}}{S + \sqrt{S}} \leq \frac{P_{\max}}{\sqrt{S}}. \tag{24}$$

Summing up all ingredients, our novel algorithm can be summarized by the following three steps:

$$\begin{aligned} \tilde{\mathbf{v}}(t) &= \frac{S}{\delta} L_t(\tilde{\mathbf{p}}(t))\mathbf{u}(t), \\ \mathbf{y}(t+1) &= \mathbf{y}(t) - \mu\,\tilde{\mathbf{v}}(t) \\ \mathbf{p}_\delta(t) &= \mathbf{Q}_\delta(\mathbf{y}(t)), \end{aligned} \tag{OXL$_0$}$$

where $\tilde{\mathbf{v}}(t)$ represents the biased estimate of the gradient. For implementation details, see OXL$_0$ algorithm below. Although OXL$_0$ requires an additional step (i.e., the computation of the gradient estimation $\tilde{\mathbf{v}}(t)$) compared with OXL, the complexity of each iteration remains linear in the problem dimensionality S.

OXL$_0$ algorithm: Online Exponential Learning Algorithm with zeroth-order Feedback

Parameters: $\mu > 0; 0 < \delta \leq P_{max}/(S + \sqrt{S})$.
Initialization: $\mathbf{y}(0) \leftarrow 0$; $t \leftarrow 0$.
Repeat
- ▹ Pre-transmission phase:
- Update $\mathbf{p}_\delta(t) \leftarrow \mathbf{Q}_\delta(\mathbf{y}(t))$ defined in (EXPδ)
- Draw a random $\mathbf{u}(t)$ uniformly from the unit-sphere
- ▹ Transmit at $\tilde{\mathbf{p}}(t) \leftarrow \mathbf{p}_\delta(t) + \delta\mathbf{u}(t)$
- ▹ Post-transmission phase: receive scalar feedback $L_t(\tilde{\mathbf{p}}(t))$
- Compute the gradient estimation $\tilde{\mathbf{v}}(t) = \frac{S}{\delta} L_t(\tilde{\mathbf{p}}(t))\,\mathbf{u}(t)$
- Update scores $\mathbf{y}(t+1) \leftarrow \mathbf{y}(t) - \mu(t)\,\tilde{\mathbf{v}}(t)$
- $t \leftarrow t + 1$

until transmission ends

In Appendix E, we prove that the regret can be bounded as follows:

Theorem 4. *If the OXL$_0$ algorithm is run with constant parameters δ and μ then the average regret is bounded by:*

$$\begin{aligned} \mathbb{E}[Reg(T)] \leq \frac{P_{\max}\log(1+S)}{2\mu} + \mu T S^2\left(\frac{B}{\delta} + K\right)^2 \\ + KT\delta\left(3 + P_{max}\left(S + 2\sqrt{S}\right)\right). \end{aligned} \tag{25}$$

where K is the Lipschitz constant and B the maximum value of the objective function $L_t(\cdot)$.

Tuning the parameters μ and δ: The step-size μ impacts the sensitivity of the algorithm to variations in the power policy. When μ is large, a small variation in the score $\mathbf{y}(t)$ results in a large variation in the power allocation. These large variations, can create oscillations in the power allocation policy $\mathbf{p}_\delta(t)$ and the time required to reach no regret increases as a result. At the opposite, a small μ leads to smaller variations in the power allocation, which also imply a long time for the regret to reach zero. Hence, there is a compromise and μ has to be carefully tuned to minimize the time to reach the no regret state.

The parameter δ represents the sampling radius of the device around the power policy $\mathbf{p}_\delta(t)$. When tuning δ, there is also a trade-off to be made between the precision of the gradient estimate and its variance. By reducing δ, the device reduces the distance to $\mathbf{p}_\delta(t)$ and the estimator gains in precision. But since the device only has access to one value of this estimate, reducing δ also increases the variability of the estimator (21).

The bound (25) can be further optimized, but because of the additional constraint $\delta \leq P_{max}/S + \sqrt{S}$, the resulting optimal bound will not be in closed-form. Having a slightly sub-optimal but closed-form expression will prove to be very useful in the sequel (when the time horizon T is unknown). For this, we choose $\delta^* = \frac{P_{max}}{(S+\sqrt{S})T^{1/4}}$ that always

meets the constraint and that decays optimally with respect to T. Then, we optimize the resulting bound in (25) only w.r.t to μ. The optimal μ^* is obtained by setting to zero the first-order derivative of the bound with respect to μ:

$$\mu^* = \sqrt{\frac{P_{max}\log(1+S)}{2T}}\left[S\left(\frac{B}{\delta^*}+K\right)\right]^{-1}. \quad (26)$$

Then, introducing δ^* and μ^* in (25) yields the bound

$$\mathbb{E}[Reg(T)] \leq U_1\, T^{3/4} + U_2\, T^{1/2}, \quad (27)$$

where

$$U_1 = S\,B\left(S+\sqrt{S}\right)\sqrt{\frac{2\log(1+S)}{P_{max}}} + K\left(3+P_{max}\left(S+2\sqrt{S}\right)\right)\frac{P_{max}}{S+\sqrt{S}}, \quad (28)$$

$$U_2 = \sqrt{2P_{max}\log(1+S)}\,S\,K. \quad (29)$$

Notice that the optimal bound w.r.t. δ and μ is also a function $\mathcal{O}(T^{3/4})$ and, hence, our particular choice of $\delta^*(T)$ above does not incur a large loss in terms of regret minimization rate and has the advantage of providing a closed-form expression of the upper bound.

Corollary 3. *If the OXL$_0$ algorithm is run for a known transmission horizon T and with the parameters δ^* and μ^* in* (26)*, then it leads to no regret and the average regret $\mathbb{E}[Reg(T)]/T$ vanishes as $\mathcal{O}(T^{-1/4})$.*

As in the previous sections, this result relies on the fact that the devices know their transmission horizon T in advance. To remove this requirement, the device can use the doubling trick or a time varying step-size. Since the time varying step-size generally involves a loss in the decay rate of the regret (see Sec. IV), we next investigate whether the information required by the doubling trick is readily available or not.

To do so, we have to determine specific values for the constants B and K in (26). A short calculation shows that they depend only on readily available system parameters:

$$\begin{aligned} B &= S\,P_{\max} + \lambda R_{min}, \\ K &= 1 + 2\lambda R_{min}. \end{aligned} \quad (30)$$

From (30) and (26) we conclude that the device is able to compute the parameters μ^* and δ^*. This implies that, if the time horizon T is not known in advance, the device can use the doubling trick described in Sec. III.

Proposition 2. *Assuming that the OXL$_0$ algorithm is run when T is unknown by using the doubling trick with the parameters μ_m^* and δ_m^* chosen as above in each window of size T_m, then the expected regret is bounded by*

$$\mathbb{E}[Reg(T)] \leq \frac{2\sqrt{2}}{2^{3/4}-1}\,U_1\,T^{3/4} + \frac{2}{\sqrt{2}-1}U_2\,T^{1/2}, \quad (31)$$

with U_1 and U_2 defined in (27)*. This means that the OXL$_0$ algorithm leads to no regret and the average regret $\mathbb{E}[Reg(T)]/T$ decreases at $\mathcal{O}(T^{-1/4})$.*

The proof follows similarly to the proof of Proposition 1 and is omitted. Importantly, reducing the available feedback results in a slower decay rate of the regret; the average regret vanishes as $\mathcal{O}(T^{-1/4})$ with zeroth-order feedback, whereas it vanishes as $\mathcal{O}(\log(T)T^{-1/2})$ with imperfect gradient feedback and as $\mathcal{O}(T^{-1/2})$ with perfect gradient feedback. Nevertheless, even under extremely limited feedback information - requiring a single sample of the objective function instead of its gradient - our proposed learning procedure (OXL$_0$ algorithm) achieves no-regret, irrespective of the evolution the network over time and despite the fact that its governing dynamics are unknown at the device end.

VI. Numerical experiments

Our goal in this section is to illustrate the performance guarantees of our learning algorithms in highly dynamic networks with realistic fading channel conditions, and with various degrees of (strictly causal, no look-ahead) feedback available at the device end, ranging from perfect gradient information to the bare-bones observation of the achieved loss. We start by comparing the OXL algorithm (full information) to classical approaches based on water-filling [21–23], suitably adapted to the setting at hand.

At each device, the benchmark water-filling is implemented so that the overall power consumption is minimized under the minimum rate constraint R_{min}. If the obtained solution does not meet the maximum power constraint, two possibilities are considered: a) the device remains silent - the *energy-driven* solution labeled WF0; b) the device transmits anyway by splitting the overall power budget uniformly over the S subcarriers - the *rate-driven* solution labeled WFP$_{\max}$.

We consider at first a simple setting composed of a pair of transmit-receive devices $N = M = 1$ communicating over four subcarriers ($S = 4$). The system parameters are: $\sigma^2 = 0.1$, $P_{max} = 1.5$ W, $R_{min} = 3$ bps/Hz and $\lambda = 1$; the channel gains are generated randomly as follows: $h^s(t+1) = \alpha h^s(t) + (1-\alpha)\varepsilon^s(t)$ with i.i.d. variables $\varepsilon^s(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 = 10$. This particular model allows us to control the temporal correlation of the channels via the parameter $\alpha \in [0,1]$ in between the extremes: the static channel case for $\alpha = 1$ (completely predictable); and the i.i.d. Rayleigh-fading case for $\alpha = 0$ (unpredictable).

For a fair comparison, we assume that the transmitting device only has access to a strictly causal feedback at each time instant. Fig. 3 illustrates the performance in terms of the relative outage defined as:

$$\text{Out} = [1 - R(\mathbf{p})/R_{\min}]^+, \quad (32)$$

of WF0 (Fig. 3(a)) and OXL algorithm (Fig. 3(b)). The performance is averaged over 100 realizations of the channel gains and for three different values of the time-correlation factor $\alpha \in \{0.2, 0.5, 0.8\}$. We remark that WF0 exhibits a high sensitivity to the temporal correlation of the channels:

Number of users	$M = 10$
Number of subcarriers	$S = 4$
Central frequency	$f_c = 2$ GHz
Bandwidth	10 MHz
Maximum power	[0.5, 2] W
Minimum rate	[0.5, 3] bps/Hz
λ	[0.5, 10]

Table I: Network parameters.

lower α (less predictable channel conditions), the worse the performance of WF-based algorithms. This is explained by the fact that water-filling algorithms perform well assuming that the SINR in each carrier is perfectly known ahead of the transmission (in the static channel case). Hence, the absence of look-ahead (non-causal) information negatively impacts the performance of classical water-filling algorithms. By contrast, the OXL algorithm consistently outperforms WF0 in terms of relative outage and is significantly more robust w.r.t. the channel dynamics. We find this feature of OXL to be particularly promising and appealing for applications to IoT networks, where the system changes constantly (and unpredictably), rendering conventional WF-based techniques obsolete.

The simple channel model above allowed us to highlight the impact of the channel dynamics and of having strictly causal feedback information on the system parameters. To validate the performance of OXL in more realistic environments, we consider in what follows a network composed of multiple interfering devices, in which the different channels are generated according to the commonly used COST-HATA model [51] that includes pathloss, fast fading and shadowing effects [47]. The speed of the devices is chosen arbitrarily between 0 km/h and 130 km/h so as to account for a wide spectrum of wireless mobile devices (smartphones, wearable, pedestrian, vehicle etc.). The minimum rate requirement R_{min}, the available power budget $P_{\max}$, and the rate vs. power tradeoff parameter λ also differ from one device to another.

Fig. 4 illustrates the comparison in terms of the rate vs. power consumption between OXL algorithm and both water-filling algorithms (Fig. 4(b) in the multiple device setting composed of $M = 10$ interfering devices over $S = 4$ subcarriers and communicating to the same receiver $N = 1$. The plotted curves are averaged over 100 realizations of the COST-HATA channel gains. We assume that all devices employ the same algorithms but with different parameters (for the OXL algorithm case).

We first note that classical water-filling is more rigid in terms of the rate vs. power tradeoff: either the device remains silent (WF0) or transmits with full power whenever its minimum rate constraint is incompatible with its power budget ($\text{WFP}_{\max}$). The parameter λ allows the device using OXL algorithm to smoothly tune its rate vs. power operating point depending on the target application. By increasing λ, the power consumption increases but the relative outage decreases. When all devices employ a rate-driven water-filling $\text{WFP}_{\max}$ a cascading effect emerges due to the fact that all devices are forced to transmit at full power ($p^s = P_{\max}/S$), which generates high network interference and, hence, has a deleterious effect on the algorithm's performance. We then see that both water-filling algorithms perform equally poorly in terms of relative outage when compared with the OXL algorithm (caused by their lack of robustness to strictly causal feedback information).

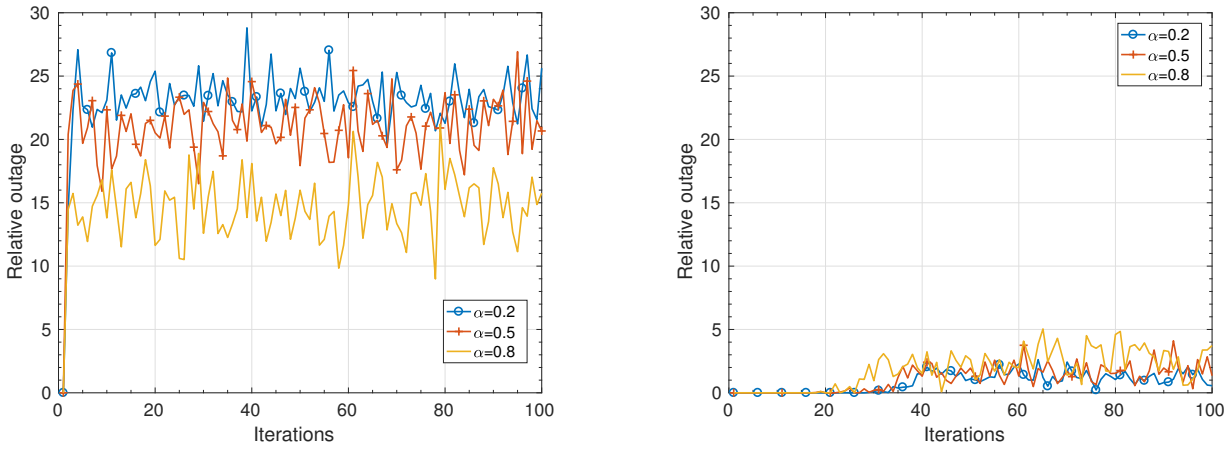
The next two goals of this section are: a) to validate our theoretical results in terms of regret, which evaluates both how close and how fast the proposed online algorithms reach the optimal fixed target state; and b) to investigate the effects of reducing the feedback information on the regret decay rate of the proposed methods.

Fig. 5 illustrates the vanishing regret of both our proposed algorithms, OXL (with perfect and imperfect gradient feedback) and OXL_0 (with a scalar feedback). Moreover, it also illustrates the impact of having a scarce or imperfect feedback and the impact of the problem dimensionality S. Fig. 5(a) confirms that having an imperfect gradient feedback does not influence significantly the regret decay rate, as anticipated by our theoretical results. However, this is no longer true when the only information available at the device end is a single scalar. The average regret of the OXL_0 algorithm decays slower compared with OXL algorithm (though the latter cannot be applied with zeroth-order feedback). Finally, Fig. 5(b) illustrates the average regret of OXL_0 algorithm for different values of the problem's dimensionality $S \in \{1, 2, 4\}$. In all cases, the average regret decays to zero; however if the number of available subcarriers increases, the variance of the estimator $\tilde{\mathbf{v}}(t)$ increases commensurately. Therefore the quality of the estimator decreases, which results in a reduced decay rate of the average regret.

VII. Conclusions

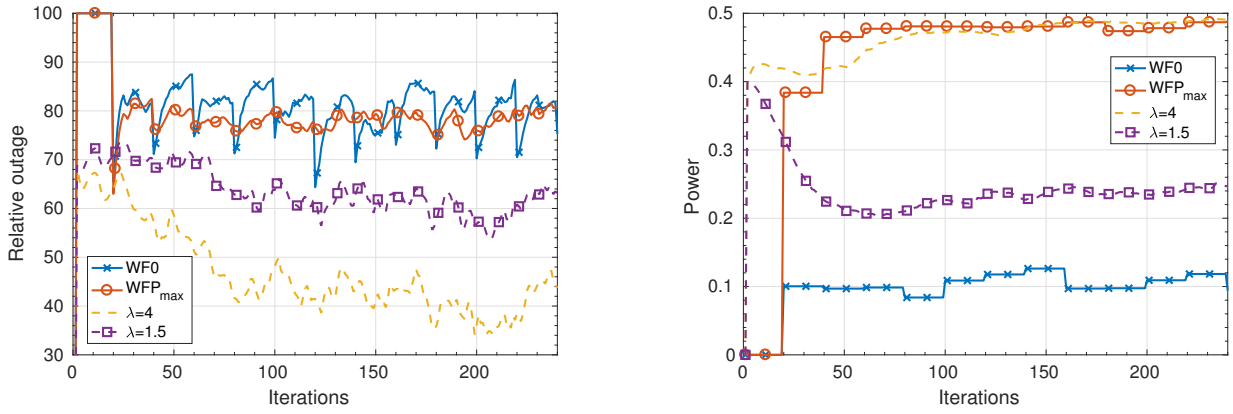
In this paper, we derived two adaptive algorithms (namely OXL and OXL_0) for solving power allocation problems in highly dynamic and unpredictable IoT networks based on online optimization tools and exponential learning. A key contribution lies in the fact that the proposed OXL_0 algorithm only requires the observation of a loss value at the device end. This algorithm is the first power allocation policy over multiple subcarriers which relies on a single scalar, as opposed to a vector worth of information containing the SINR values in all subcarrier required by classic water-filling algorithms.

Our simulations validate our theoretical expectations by showing that water-filling algorithms are highly sensitive to outdated feedback information and, hence, are not robust to rapid and unpredictable changes in the network. The proposed OXL algorithm outperforms classic water-filling algorithms in all investigated settings in which the network dynamics is not known at the device end. The impact of feedback scarcity is then assessed: the zeroth-order feedback algorithm is the slowest to reach no regret, followed



(a) Relative outage of WF0 (b) Relative outage of OXL algorithm

Figure 3: Performance comparison between WF0 and OXL algorithm in a time-varying setting, in which the transmitting device has access to strictly causal information. OXL algorithm outperforms WF0 in terms of relative outage irrespective from the channel dynamics. WF0 is negatively impacted by the outdated feedback information: the more unpredictable the channel gains, the higher the outage.



(a) Relative outage (b) Overall power

Figure 4: Performance comparison between OXL algorithm and water-filling for an arbitrary device. Water-filling algorithms are rigid in terms of the rate vs. power tradeoff, while OXL algorithm allows for a more smooth tuning via the parameter λ and for better performance both in terms of relative outage and power consumption. The rate-driven $\text{WFP}_{\max}$ exhibits a cascading effect in the network and forces all devices to transmit at full power resulting in poor performance. The energy-driven WF0 results in poor performance in terms of relative outage.

by the first algorithm in the imperfect gradient feedback case and then by the same algorithm in the perfect gradient case.

Appendix

A. *First order feedback: known horizon T*

For simplicity of presentation in the remaining appendices, we focus on the regret w.r.t. an arbitrary fixed policy $\mathbf{q} \in \mathcal{P}$ defined as $Reg_{\mathbf{q}}(T) \triangleq \sum_{t=1}^{T} L_t(\mathbf{p}(t)) - \sum_{t=1}^{T} L_t(\mathbf{q})$, and derive upper-bounds that are independent from $\mathbf{q}$ and, hence, also hold for the regret in (6) (or for its expectation).

The first step to prove Theorem 1 is to bound the regret based on the convexity of $L_t(\mathbf{q})$ as follows

$$Reg_{\mathbf{q}}(T) \leq \sum_{t=1}^{T} \langle \mathbf{v}(t) | \mathbf{p}(t) - \mathbf{q} \rangle , \tag{33}$$

where $\mathbf{q} \in \mathcal{P}$ is an arbitrarily chosen power allocation.

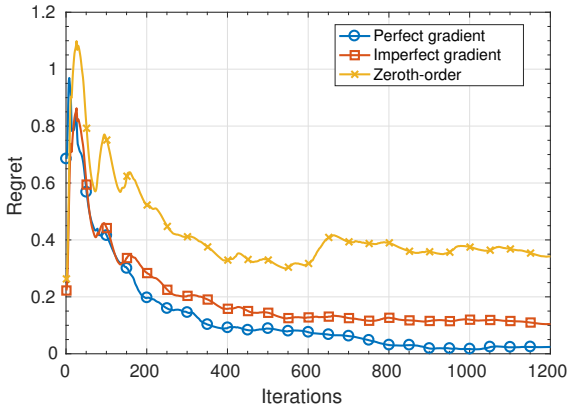
Using the fact that $\mathbf{y}(t+1) = \mathbf{y}(t) - \mu \mathbf{v}(t)$ and $\mathbf{y}(1) = 0$, we obtain

$$Reg_{\mathbf{q}}(T) \leq \sum_{t=1}^{T} \langle \mathbf{v}(t) | \mathbf{p}(t) \rangle + \frac{1}{\mu} \langle \mathbf{y}(T+1) | \mathbf{q} \rangle . \tag{34}$$

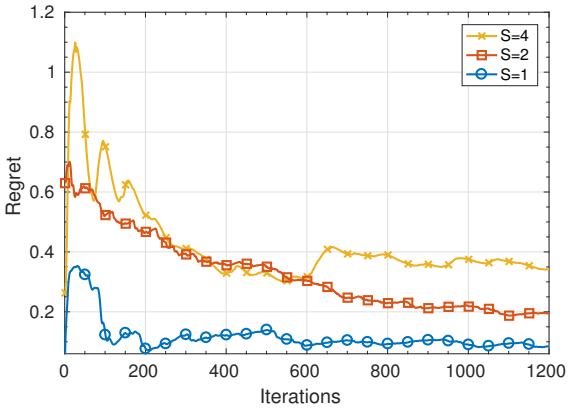
Then, we define a potential function $f^*(\mathbf{y}(t)) = P_{\max} \log(1 + \sum_{s=1}^{S} \exp(y_s(t)))$, which is used to show that the exponentiation step in (OXL) is equivalent to $\mathbf{p}(t) = \nabla f^*(\mathbf{y}(t))$. Also, the second order Taylor approximation of $f^*(\mathbf{y}(t))$ yields

$$f^*(\mathbf{y}(t+1)) \leq f^*(\mathbf{y}(t)) - \mu \langle \mathbf{v}(t) | \nabla f^*(\mathbf{y}(t)) \rangle + \frac{\mu^2}{2} P_{\max} \|\mathbf{v}(t)\|_2^2 . \tag{35}$$

Combining the above inequality with equation (34) and



(a) Regret of OXL and OXL_0 algorithms



(b) Regret of OXL_0 algorithm

Figure 5: Impact of feedback amount and problem dimensionality. Both proposed algorithms, OXL and OXL_0, exhibit a vanishing regret, as anticipated by our theoretical results. The average regret of OXL_0 algorithm, relying only on the scalar value of the objective function, decays slower than the average regret of OXL algorithm with perfect or imperfect gradient feedback. Having to estimate the gradient of dimension S using the scalar value of the objective impacts the decay rate of the average regret of OXL_0 algorithm: the higher the problem dimensionality S, the slower the average regret.

given that $\mathbf{p}(t) = \nabla f^*(\mathbf{y}(t))$, we obtain:

$$Reg_{\mathbf{q}}(T) \leq \frac{1}{\mu}\left[f^*(0) - f^*(\mathbf{y}(T+1))\right] + \frac{\mu}{2}P_{\max}\sum_{t=1}^{T}\|\mathbf{v}(t)\|_2^2 + \frac{1}{\mu}\langle \mathbf{y}(T+1)|\mathbf{q}\rangle\,. \tag{36}$$

By using Fenchel's inequality [52] we get

$$f^*(\mathbf{y}) + f(\mathbf{q}) \geq \langle \mathbf{y}|\mathbf{q}\rangle\,,\ \forall\,\mathbf{y},\mathbf{q} \tag{37}$$

where $f(\mathbf{q})$ is the convex conjugate of $f^*(\mathbf{y})$ defined as $f(\mathbf{q}) = \sup_{\mathbf{y}\in\mathbb{R}}\langle \mathbf{y}|\mathbf{q}\rangle - f^*(\mathbf{y})$. We can then substitute $\langle \mathbf{y}(T+1)|\mathbf{q}\rangle - f^*(\mathbf{y}(T+1))$ by $f(\mathbf{q})$ in (36) and obtain

$$Reg_{\mathbf{q}}(T) \leq \frac{1}{\mu}\left[f(\mathbf{q}) + P_{\max}\log(1+S)\right] + \frac{\mu}{2}P_{\max}VT. \tag{38}$$

We can show that $f(\mathbf{q}) \leq 0$ for all $\mathbf{q} \in \mathcal{P}$ by using a variable change ($\mathbf{x} = \mathbf{q}/P_{\max}$) combined with Jensen's inequality for convex functions and the regret bound reduces to

$$Reg(T) \leq \frac{P_{\max}\log(1+S)}{\mu} + \frac{\mu}{2}P_{\max}TV \tag{39}$$

The optimal step-size is then obtained by minimizing the above bound.

B. First order feedback: unknown horizon T

OXL algorithm is run with the optimal step $\mu^*(T_m)$ defined in (9) in each window. Then, the regret in window m of size $T_m = 2^m$, denoted by $\widetilde{Reg}(T_m)$, can be bounded as in (10):

$$\widetilde{Reg}(T_m) \leq P_{\max}\sqrt{2T_mV\log(1+S)}. \tag{40}$$

For a time horizon T, the number of widows equals $\lceil\log_2(T)\rceil$, where $\lceil x\rceil$ is the ceiling function. The overall regret can be bounded by the sum of all windows' regrets:

$$Reg(T) \leq \sum_{m=0}^{\lceil\log_2(T)\rceil} P_{\max}\sqrt{2T_mV\log(1+S)}. \tag{41}$$

The result then follows by a geometric series argument.

C. Imperfect gradient feedback: known horizon T

From the convexity of the objective function, we can write

$$\mathbb{E}[Reg_{\mathbf{q}}(T)] \leq \mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla L_t(\mathbf{p}(t))|\mathbf{p}(t)-\mathbf{q}\rangle\right]. \tag{42}$$

The idea is to link the above bound to the estimate $\tilde{\mathbf{v}}(t)$. By definition, we have that $\nabla L_t(\mathbf{p}(t)) = \mathbb{E}[\tilde{\mathbf{v}}(t)|\tilde{\mathbf{v}}(t-1),...,\tilde{\mathbf{v}}(1)]$. By the law of total expectation, the following equality holds

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla L_t(\mathbf{p}(t))|\mathbf{p}(t)-\mathbf{q}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle \tilde{\mathbf{v}}(t)|\mathbf{p}(t)-\mathbf{q}\rangle\right]. \tag{43}$$

The term inside the expectation on the RHS can be bounded as in (34) and, similarly to the proof of Therorem 1, we obtain

$$\mathbb{E}[Reg(T)] \leq \mathbb{E}\left[\frac{P_{\max}\log(1+S)}{\mu} + \frac{\mu}{2}P_{\max}\sum_{t=1}^{T}\|\tilde{\mathbf{v}}(t)\|_2^2\right]. \tag{44}$$

Given that $\mathbb{E}\left[\|\tilde{\mathbf{v}}(t)\|_2^2\right] \leq \tilde{V}$, the result follows.

D. Imperfect gradient feedback: unknown horizon T

To bound the regret assuming a variable step-size $\mu(t)$, we will consider the following weighted regret

$$WReg_{\mathbf{q}}(T) \triangleq \mathbb{E}\left[\sum_{t=1}^{T}\mu(t)\left(L_t(\mathbf{p}(t)) - L_t(\mathbf{q})\right)\right], \tag{45}$$

where $\mu(t)$ is the variable step-size in (17). Using a similar approach as in the proof of Theorem 3, we obtain

$$WReg_{\mathbf{q}}(T) \leq P_{\max}\log(1+S) + \frac{P_{\max}}{2}\tilde{V}\sum_{t=1}^{T}\mu^2(t). \quad (46)$$

To bound the regret, we use the summability criterion of Hardy [53], which allows us to compare weighted sums – here, $\mathbb{E}[Reg_{\mathbf{q}}(T)]$ and $WReg_{\mathbf{q}}(T)$. In particular, note that the step-size sequence $\mu(t) = \alpha t^{-1/2}$ satisfies the conditions $\mu(t) \geq \mu(t+1)$; and $\sum_{t=1}^{T}\mu(t)/\mu(T) = \mathcal{O}(T)$. Therefore, by Theorem 14 in [53], we obtain

$$\frac{\mathbb{E}[Reg_{\mathbf{q}}(T)]}{T} \sim \frac{WReg_{\mathbf{q}}(T)}{\sum_{t=1}^{T}\mu(t)} \leq \frac{P_{\max}}{\sqrt{T}}\left[\frac{\log(1+S)}{\alpha} + \frac{\alpha\tilde{V}(1+\log T)}{2}\right]. \quad (47)$$

E. Zeroth-order feedback: known horizon T

To prove Theorem 4, we introduce first some properties. Consider the following expectation of the objective function [7]

$$\tilde{L}_t(\mathbf{p}) \triangleq \mathbb{E}_{\mathbf{u}\in\mathcal{B}}\left[L_t(\mathbf{p}+\delta\mathbf{u})\right], \quad (48)$$

where $\mathbf{u}$ is a random vector drawn uniformly on the unit Euclidean ball $\mathcal{B} = \left\{\mathbf{u}\in\mathbb{R}^s \,|\, \|\mathbf{u}\|^2 \leq 1\right\}$ and the expectation is taken over the randomness of $\mathbf{u}$. We can show that $\tilde{L}_t(\cdot)$ is a biased estimator of $L_t(\cdot)$ and

$$|L_t(\mathbf{p}) - \tilde{L}_t(\mathbf{p})| \leq K\delta, \quad \forall\mathbf{p}, \quad (49)$$

where K is the Lipschitz constant of the objective function. An important property of $\tilde{L}_t(\mathbf{p})$ is that its gradient relies on the values of the objective function as follows

$$\nabla\tilde{L}_t(\mathbf{p}) = \mathbb{E}_{\mathbf{u}\in\mathcal{S}}\left[\frac{S}{\delta}L_t(\mathbf{p}+\mathbf{u}\delta)\mathbf{u}\right], \quad (50)$$

where $\mathbf{u}$ is drawn for the unit Euclidean sphere $\mathcal{S} = \left\{\mathbf{u}\in\mathbb{R}^S \,|\, \|\mathbf{u}\|^2 = 1\right\}$.

Another useful property is that the new exponential mapping step in (EXPδ), which is adapted to the modified set $\mathcal{P}_\delta$, can be written equivalently as:

$$\mathbf{p}_\delta(t) = \arg\max_{\mathbf{q}\in\mathcal{P}_\delta}\left\{\langle\mathbf{y}(t)|\mathbf{q}\rangle - h(\mathbf{q})\right\},$$
$$h(\mathbf{q}) \triangleq \sum_{s=1}^{S}(q^s-\delta)\log(q^s-\delta) + \left(C - \sum_{s=1}^{S}q^s\right)\log\left(C - \sum_{s=1}^{S}q^s\right), \quad (51)$$

with $C = P_{\max} - \delta\sqrt{S}$.

The first step to prove Theorem 4 is to compare $L_t(\mathbf{p}_\delta(t)+\delta\mathbf{u})$, the incurred loss at time t, to $L_t(\mathbf{p}_\delta(t))$ by using that $L_t(\cdot)$ is a K-Lipschitz function:

$$\mathbb{E}[Reg_{\mathbf{q}}(T)] \leq \mathbb{E}\left[\sum_{t=1}^{T}L_t(\mathbf{p}_\delta(t)) - L_t(\mathbf{q})\right] + KT\delta\left(1 + P_{max}\tilde{S}\right),$$

with $\tilde{S} = S + 2\sqrt{S}$. The second step is to compare $L_t(\mathbf{p}_\delta(t))$ and $L_t(\mathbf{q})$ to $\tilde{L}_t(\mathbf{p}_\delta(t))$ and $\tilde{L}_t(\mathbf{q})$ respectively.

$$\mathbb{E}[Reg_{\mathbf{q}}(T)] \leq \mathbb{E}\left[\sum_{t=1}^{T}\tilde{L}_t(\mathbf{p}_\delta(t)) - \tilde{L}_t(\mathbf{q})\right] + KT\delta\left(3 + P_{max}\tilde{S}\right).$$

Since $\tilde{L}_t(\mathbf{p})$ is convex w.r.t. $\mathbf{p}$ we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\tilde{L}_t(\mathbf{p}_\delta(t)) - \tilde{L}_t(\mathbf{q})\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla\tilde{L}_t(\mathbf{p}_\delta(t))|\mathbf{p}_\delta(t)-\mathbf{q}\rangle\right].$$

We can write $\nabla\tilde{L}_t(\mathbf{p}_\delta(t)) = \mathbb{E}[\tilde{\mathbf{v}}(t)|\mathbf{u}(1),...,\mathbf{u}(t-1)]$, where $\tilde{\mathbf{v}}(t)$ is the estimation defined in (22) and where the expectation is taken over the randomness of $\mathbf{u}$. Using this property and the law of total expectation, the bound on the expected regret becomes:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla\tilde{L}_t(\mathbf{p}_\delta(t))|\mathbf{p}_\delta(t)-\mathbf{q}\rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\langle\tilde{\mathbf{v}}(t)|\mathbf{p}_\delta(t)-\mathbf{q}\rangle\right].$$

By using (51), we can bound the sum $\sum_{t=1}^{T}\langle\tilde{\mathbf{v}}(t)|\mathbf{q}\rangle$ and obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle\tilde{\mathbf{v}}(t)|\mathbf{p}_\delta(t)-\mathbf{q}\rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\langle\mathbf{v}(t)|\mathbf{p}_\delta(t+1)-\mathbf{p}_\delta(t)\rangle\right] + \frac{H}{2\mu},$$

where $H = \min_{\mathbf{p}\in\mathcal{P}_\delta}h(\mathbf{p})$. We use again (51) and the Cauchy-Schwartz inequality to bound the sum of $\langle\mathbf{v}(t)|\mathbf{p}_\delta(t+1)-\mathbf{p}_\delta(t)\rangle$ and, by combining all the above, we find

$$\mathbb{E}[Reg(T)] \leq \frac{H}{2\mu} + \mu TS^2\left(\frac{B}{\delta}+K\right)^2 + KT\delta\left(3 + P_{max}\left(S+2\sqrt{S}\right)\right).$$

where $B = \max_{t,\mathbf{p}}L_t(\mathbf{p})$. Finally, Theorem 4 follows by finding that $H = P_{max}\log(1+S)$.

References

[1] A. Marcastel, E. V. Belmega, P. Mertikopoulos, and I. Fijalkow, "Online power allocation for opportunistic radio access in dynamic OFDM networks," in *Vehicular Technology Conference (VTC-Fall), 2016 IEEE 84th*, Sep. 2016, pp. 1–5.

[2] ——, "Online interference mitigation via learning in dynamic IoT environments," in *Globecom Workshop, Internet of Everything, 2016 IEEE*, Dec. 2016, pp. 1–5.

[3] A. I. Sulyman, S. M. Oteafy, and H. S. Hassanein, "Expanding the cellular-IoT umbrella: An architectural approach," *IEEE Trans. Wireless Commun.*, vol. 24, no. 3, pp. 66–71, Jun. 2017.

[4] M. Basharat, W. Ejaz, M. Naeem, A. M. Khattak, and A. Anpalagan, "A survey and taxonomy on nonorthogonal multiple-access schemes for 5G networks," *Transactions on Emerging Telecommunications Technologies*, vol. 29, no. 1, pp. 1–17, Jun. 2018.

[5] Huawei Technologies, *5G: A technology vision*. White paper, 2013. [Online]. Available: www.huawei.com/ilink/en/download/HW_314849

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

[7] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[8] I. Caragiannis, C. Kaklamanis, P. Kanellopoulos, M. Kyropoulou, B. Lucier, R. P. Leme, and E. Tardos, "Bounding the inefficiency of outcomes in generalized second price auctions," *Journal of Economic Theory*, vol. 156, pp. 343–388, Mar. 2015.

[9] N. Cesa-Bianchi, C. Gentile, and Y. Mansour, "Regret minimization for reserve prices in second-price auctions," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 549–564, Jan. 2015.

[10] P. Viappiani and C. Boutilier, "Regret-based optimal recommendation sets in conversational recommender systems," in *Proc. 3rd ACM conference on Recommender Systems*, Oct. 2009, pp. 101–108.

[11] P. Mertikopoulos and E. V. Belmega, "Transmit without regrets: Online optimization in MIMO-OFDM cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 11, pp. 1987–1999, Dec. 2014.

[12] ——, "Learning to be green: Robust energy efficiency maximization in dynamic MIMO-OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 743–757, Apr. 2016.

[13] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1497–1516, Sep. 2012.

[14] C. Goursaud and J.-M. Gorce, "Dedicated networks for IoT: PHY/-MAC state of the art and challenges," *EAI Endorsed Transactions on Internet of Things*, vol. 1, no. 1, pp. 1–11, Oct. 2015.

[15] Y. Chen, F. Han, Y.-H. Yang, H. Ma, Y. Han, C. Jiang, H.-Q. Lai, D. Claffey, Z. Safar, and K. R. Liu, "Time-reversal wireless paradigm for green internet of things: An overview," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 81–98, 2014.

[16] W. Li, M. Assaad, and P. Duhamel, "Distributed stochastic optimization in networks with low informational exchange," in *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*. IEEE, 2017, pp. 1160–1167.

[17] D. J. Love, R. W. Heath, V. K. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, Oct. 2008.

[18] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 5, pp. 742–758, Apr. 2014.

[19] W. Li, M. Assaad, G. Ayache, and M. Larranaga, "Matrix exponential learning for resource allocation with low informational exchange," in *IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 266–270.

[20] W. Li and M. Assaad, "Matrix exponential learning schemes with low informational exchange," *arXiv preprint arXiv:1802.06652*, 2018. [Online]. Available: https://arxiv.org/pdf/1802.06652v2.pdf

[21] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[22] J.-S. Pang, G. Scutari, F. Facchinei, and C. Wang, "Distributed power allocation with rate constraints in Gaussian parallel interference channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3471–3489, Jul. 2008.

[23] G. Scutari, D. P. Palomar, and S. Barbarossa, "The MIMO iterative waterfilling algorithm," *IEEE Trans. Signal Process.*, vol. 57, no. 5, pp. 1917–1935, Jan. 2009.

[24] H. Safdar, N. Fisal, R. Ullah, W. Maqbool, F. Asraf, Z. Khalid, and A. Khan, "Resource allocation for uplink M2M communication: A game theory approach," in *Wireless Technology and Applications (ISWTA), 2013 IEEE Symp.*, Sep. 2013, pp. 48–52.

[25] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.

[26] T. Zheng, Y. Qin, H. Zhang, and S. Kuo, "Adaptive power control for mutual interference avoidance in industrial Internet-of-Things," *China Communications*, vol. 13, no. Supplement 1, pp. 124–131, Sep. 2016.

[27] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 641–646, 1993.

[28] R. Masmoudi, E. V. Belmega, I. Fijalkow, and N. Sellami, "A unifying view on energy-efficiency metrics in cognitive radio channels," in *Signal Processing Conference (EUSIPCO), 2014 Proc. 22nd European*, Sep. 2014, pp. 171–175.

[29] T. Holliday, N. Bambos, P. Glynn, and A. Goldsmith, "Distributed power control for time varying wireless networks: Optimality and convergence," in *Proc. of the annual ALLERTON Conference on Communication Control and Computing*, vol. 41, no. 2, 2003, pp. 1024–1033.

[30] R. Tajan, C. Poulliat, and I. Fijalkow, "Interference management for cognitive radio systems exploiting primary IR-HARQ: A constrained Markov decision process approach," in *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, 2012, pp. 1818–1822.

[31] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, 2012.

[32] J. Wang, C. Jiang, Z. Han, Y. Ren, and L. Hanzo, "Network association strategies for an energy harvesting aided super-WiFi network relying on measured solar activity." *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3785–3797, 2016.

[33] T. Chen, A. Mokhtari, X. Wang, A. Ribeiro, and G. B. Giannakis, "Stochastic averaging for constrained optimization with application to online resource allocation," *IEEE Trans. Signal Process.*, vol. 65, no. 12, pp. 3078–3093, 2017.

[34] P. Mertikopoulos, E. V. Belmega, R. Negrel, and L. Sanguinetti, "Distributed stochastic optimization via matrix exponential learning," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2277–2290, 2017.

[35] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2017.

[36] 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, Mar. 2011.

[37] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," *arXiv preprint arXiv:1712.00702*, 2017.

[38] J. Wang, C. Jiang, H. Zhang, X. Zhang, V. C. Leung, and L. Hanzo, "Learning-aided network association for hybrid indoor LiFi-WiFi systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3561–3574, 2018.

[39] T. Chen, S. Barbarossa, X. Wang, G. B. Giannakis, and Z.-L. Zhang, "Learning and management for Internet-of-Things: Accounting for adaptivity and scalability," *arXiv preprint arXiv:1810.11613*, 2018.

[40] L. Mainetti, L. Patrono, and A. Vilei, "Evolution of wireless sensor networks towards the internet of things: A survey," in *Software, Telecommunications and Computer Networks (SoftCOM), IEEE 19th Intl. Conf. on*, 2011, pp. 1–6.

[41] G. M. Lee and N. Crespi, "The Internet of Things: Challenge for a new architecture from problems," in *IAB Interconnecting Smart Objects with the Internet Workshop*, Mar. 2011, pp. 1–2.

[42] T. Alpcan, T. Başar, R. Srikant, and E. Altman, "CDMA uplink power control as a noncooperative game," *Wireless Networks*, vol. 8, no. 6, pp. 659–670, Nov. 2002.

[43] E. Altman and L. Wynter, "Equilibrium, games, and pricing in transportation and telecommunication networks," *Networks and Spatial Economics*, vol. 4, no. 1, pp. 7–21, Mar. 2004.

[44] M. Chiang, P. Hande, T. Lan, C. W. Tan *et al.*, "Power control in wireless cellular networks," *Foundations and Trends in Networking*, vol. 2, no. 4, pp. 381–533, 2008.

[45] E. V. Belmega, P. Mertikopoulos, R. Negrel, and L. Sanguinetti, "Online convex optimization and no-regret learning: Algorithms, guarantees and applications," *arXiv preprint arXiv:1804.04529, submitted to IEEE Signal Processing Magazine*, 2018.

[46] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. of the 20th International Conference on Machine Learning (ICML-03)*, Aug. 2003, pp. 928–936.

[47] 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, pp. 389–403, Mar. 2007.

[48] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[49] L. B. Klebanov, S. T. Rachev, and F. J. Fabozzi, *Robust and Non-Robust models in Statistics*. Nova Science Publishers, 2009.

[50] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *SODA'05: Proceedings of the 16th annual ACM-SIAM symposium on discrete algorithms*, Jan. 2005, pp. 385–394.

[51] G. F. Pedersen, *COST 231-Digital mobile radio towards future generation systems*. EU, 1999.

[52] R. T. Rockafellar, *Convex analysis*. Princeton University Press, 2015.

[53] G. H. Hardy, *Divergent Series*. Oxford University Press, 1949.

Alexandre Marcastel (S'16) received his engineer degree from ENSEA, Cergy-Pontoise, France in 2014, the M.Sc. and Ph.D. degrees both from the University of Cergy-Pontoise, France in 2015 and 2019, respectively. During his Ph.D. in ETIS laboratory (ETIS, UMR 8051, University Paris Seine, University Cergy-Pontoise, ENSEA, CNRS), he worked on applications of online optimization to dynamic and unpredictable wireless (IoT) networks. He is currently a teaching assistant in ENSEA, Cergy-Pontoise, France and his main research interests lie in resource allocation problems for dynamic wireless networks.

E. Veronica Belmega (S'08–M'10) received the M.Sc. (engineer diploma) degree from the University Politehnica of Bucharest, Bucharest, Romania, in 2007, and the M.Sc. and Ph.D. degrees both from the University Paris-Sud 11, Orsay, France, in 2007 and 2010, respectively. From 2010 to 2011, she was a Postdoctoral Researcher in a joint project between Princeton University, N.J., USA and the Alcatel-Lucent Chair on Flexible Radio in Supélec, France. She is currently an Associate Professor with ETIS/ENSEA–Université de Cergy-Pontoise–CNRS, Cergy-Pontoise, France. She was one of the ten recipients of the L'Oréal–UNESCO–French Academy of Science Fellowship: "For young women doctoral candidates in science" in 2009. Since 2018, she is the recipient of the Doctoral Supervision and Research Bonus (PEDR) by the French National Council of Universities (CNU 61). She serves on the editorial board of the Transactions on Emerging Telecommunications Technologies (ETT) and has been distinguished among the Top 11 Editors, for outstanding contributions to ETT during 2016-2017.

Panayotis Mertikopoulos (M' 11) received the Ptychion degree in physics (summa cum laude) from the University of Athens in 2003, his M.Sc. and M.Phil. degrees in mathematics from Brown University in 2005 and 2006 (both summa cum laude), and his Ph.D. degree from the University of Athens in 2010. During 2010–2011, he was a post-doctoral researcher at the École Polytechnique, Paris, France. Since 2011, he has been a CNRS Researcher at the Laboratoire d'Informatique de Grenoble, Grenoble, France. P. Mertikopoulos was an Embeirikeion Foundation Fellow between 2003 and 2006, and received the best paper award in NetGCoop '12. He is serving on the editorial board and program committees of several journals and conferences on learning and optimization (such as NIPS and ICML). His main research interests lie in learning, optimization, game theory, and their applications to networks and machine learning systems.

Inbar Fijalkow (M'96–SM'10) received her engineering and Ph.D. degrees from TelecomParis, Paris, France, in 1990 and 1993, respectively. In 1993–1994, she was a postdoctoral research fellow at Cornell University, NY, USA, supported by a French Lavoisier Fellowship. Since 1994, she is a member of ETIS research unit, UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, while teaching at ENSEA in Cergy, France. From 2000 to 2004, she was in charge of the master research program in intelligent and communicating systems (SIC). She was the vice-head of the French research group in image and signal processing, GdR ISIS, from 2002 to 2004. She was the dean of the ETIS laboratory from 2004 to 2013. In 2015-2016, she was a visiting researcher at UC Irvine (CA, USA) funded by the French CNRS. She is currently a full Professor at ENSEA at the "classe exceptionnelle" level. Her research interests are in signal processing for digital communications, including iterative processing, optimization, estimation theory, signal processing for dirty-RF. She is (co-)author of over 180 publications. Dr. Fijalkow is a past IEEE Transactions on Signal Processing associate editor. She is a Senior IEEE member and the member of several technical committees. She is a mentor in several associations for women in science and engineering.