Gradient-free Online Resource Allocation Algorithms for Dynamic Wireless Networks

Alexandre Marcastel, Elena-Veronica Belmega, Panayotis Mertikopoulos, Inbar Fijalkow

Gradient-free Online Resource Allocation Algorithms for Dynamic Wireless Networks

Alexandre Marcastel*, E. Veronica Belmega*, Panayotis Mertikopoulos†, and Inbar Fijalkow*
* ETIS, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, Cergy-Pontoise, France
† Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France

***Abstract*—Future communication networks will be faced with supporting highly mobile, heterogeneous (including aerial) devices, which calls for new and efficient resource allocation policies that are able to adapt *on-the-fly* to the network dynamics while relying on little and possibly outdated information. The aim of this paper is twofold: to explicitly take into account the device mobility, their network connectivity patterns and behavior (which may be completely arbitrary and unpredictable); and to greatly reduce the information required at the transmitter. For this, we exploit the framework of online optimization and exponential learning to derive a provably efficient and gradient-free online power allocation algorithm relying only on a scalar-worth of feedback.**

***Index Terms*—Highly mobile devices, arbitrarily time-varying networks, online optimization, zeroth-order feedback**

I. Introduction

Mobility is expected to become one of the major challenges in future communication networks due to ambitious objectives such as ubiquitous 3D connectivity, unmanned aerial vehicles (UAVs) and flying devices communication, very low latency, and battery-free communications [1, 2]. This creates the need of designing flexible, efficient and adaptive resource allocation algorithms capable of coping with the underlying network dynamics and unpredictability while relying on little and strictly causal feedback information.

The wide literature on resource allocation problems in wireless networks so far relies (for the most part) on either static [3–5] or stochastic [6–8] models and on strong assumptions on the information available at the transmitter (e.g., perfect channel state information in the form of the signal-to-interference-plus-noise ratio (SINR) in each band, gradient feedback). The main aim is to derive efficient algorithms that converge to an optimal fixed or steady state. However, in highly dynamic networks that can evolve in an arbitrary and unpredictable way there is no fixed *solution state* to converge to and disposing of significant amount and non-causal information at the device end is no longer realistic.

Online resource allocation algorithms have been recently proposed in [9–12] exploiting the online optimization and regret minimization framework [13, 14]. These works investigate various system models and problems: rate or energy-efficiency maximization in multi-antenna (MIMO) or multi-carrier interference networks. The derived algorithms have the advantage of relying on strictly causal information, without any assumptions on the dynamics that governs the network evolution, which can even be non-stationary. Moreover, they provably attain no regret, meaning that their performance is at least as good as the best fixed policy in hindsight (i.e., having non-causal knowledge of the network's entire evolution).

Nevertheless, the above online algorithms have in common the (potentially large) amount of feedback information required, i.e., either the perfect or noisy version of the gradient at each iteration (which can be a vector or matrix). In this paper, we investigate the power allocation problem over multiple frequency bands in a distributed and dynamic network composed of multi-user interfering links. Each mobile device wishes to optimize its power consumption while taking into account a minimum quality of service (QoS) constraint. The aim of our 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.

Our main contributions can be summarized as follows. Building on the exponential learning algorithm in [12], we propose a novel *gradient-free* online power allocation algorithm that only requires scalar feedback, i.e., the value of the objective function. Based on this feedback, we derive a one-point stochastic estimator of the gradient [8, 13, 14], which, however, leads to quite challenging feasibility issues in our setting. We tackle these issues by appropriately modifying the exponential mapping step in [12] to fit a shrunk version of the feasible set. Our resulting algorithm is shown to reach no regret at a rate of $O(T^{-1/4})$. This decay rate is slower than the one of its gradient-based counterpart ($O(T^{-1/2})$), highlighting the tradeoff between the amount of available information and the speed of reaching no regret.

II. Online Optimization Problem and Framework

We consider a system composed of M transmitters and N receivers communicating over S orthogonal 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 focus on one particular transmitting-

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), the ELIOT ANR-18-CE40-0030 and FAPESP 2018/12579-7 project, and by ENSEA, Cergy-Pontoise, France.

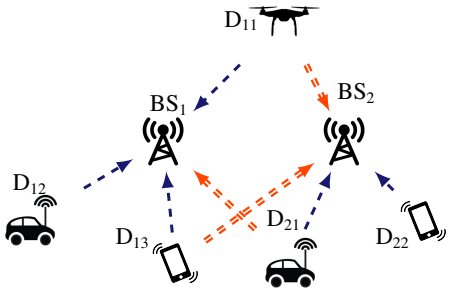


Fig. 1. System composed of five mobile devices (D_{11}, D_{12}, etc.) and two base stations (BS_1, BS_2). The blue arrows represent the direct links while the red (double-lined) ones are interfering links.

receiving pair. The received signal for the (arbitrarily chosen) focal device becomes:

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

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

We also define the effective channel gain vector $\mathbf{w}(t) = (w^s(t))$ with $w^s(t)$ the effective gain in band s:

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

where σ^2 is the variance of the noise $z^s(t)$; $p_j^s(t)$ is the transmitted power by interferer j; $g_j^s(t) = |h_j^s(t)|^2$ and $g^s(t) = |h^s(t)|^2$ in band s. For simplicity, we assume that the receiver employs single-user decoding (SUD), meaning that, when decoding a transmitted signal, the other incoming signals are treated as noise. This is relevant in distributed and energy-limited networks, in which the receivers may not afford to sequentially process and decode their incoming signals and the transmitting devices may not be coordinated.

The problem under investigation is the tradeoff between power minimization and QoS requirement, captured by the following loss function:

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

where $\mathbf{p} = (p^1, \ldots, p^S)$ represents the power allocation vector of the focal device with component p^s representing the power allocated to the s-th band. 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 $R_t(\mathbf{p}) \geq R_{\min}$. We choose a linear penalty function for its relevance to communications [15, 16] and to simplify the presentation.

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

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

Regarding the constraints, they are physical hard constraints as opposed to QoS requirements, and can never be violated: the first is the positivity of the transmit powers; and the second is the maximum available power constraint.

The particularity of the problem above lies in the fact that the objective function $L_t(\mathbf{p})$ may vary in a non-stationary and unpredictable way. The focal device cannot determine *a priori* the best optimal power allocation at each time t. Nevertheless, we assume that the device receives some feedback after each transmission, i.e., the past experienced objective value. The main idea in online optimization is to exploit this strictly causal information to build a dynamic and adaptive power allocation policy $\mathbf{p}(t)$ - henceforth called an *online policy* - that minimizes as much as possible the time-varying objective function $L_t(\mathbf{p}(t))$.

In order to evaluate the performance of a given *online policy* $\mathbf{p}(t)$, the most commonly used notion is that of the regret [13, 14], which compares its performance in terms of loss with a benchmark policy, i.e., the *fixed* strategy that minimizes the overall objective over a given horizon T:

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

where $\mathcal{P} \triangleq \left\{\mathbf{p} \in \mathbb{R}_+^S \,\middle|\, \sum_{s=1}^{S} p^s \leq P_{max}\right\}$ 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 $\limsup_{T\to\infty} Reg(T) \leq 0$.

Remark that the actual computation of the benchmark policy requires the non-causal knowledge of the evolution of the objective throughout the time horizon T in hindsight, before the transmission actually takes place. Therefore, the design of dynamic policies that reach no regret while relying on strictly causal and local information is a non-trivial goal.

III. One-point Gradient Estimation

The existing online resource allocation policies [9–12], require a vector (in multi-carrier OFDM systems) or a matrix (in MIMO systems) feedback representing the perfect or imperfect gradient of the objective function: at decision instant $t+1$, the available feedback is either $\nabla L_t(\mathbf{p}(t))$ or a noisy version of it.

In this work, one of our main objectives is to reduce the amount of required information to be fed back to the transmit devices. More specifically, we assume that the devices

know only the value of the experienced objective function: at decision instant $t+1$, the available feedback is the value $L_t(\mathbf{p}(t))$. This means that *only a single scalar* is needed at the transmitting device – a major advantage in feedback-limited and highly dynamic networks, where the acquisition of network information (e.g., channel state) required for the gradient computation becomes difficult.

To develop an online policy $\mathbf{p}(t)$ that leads to no regret, we start by estimating the gradient of the objective based only on its value, the so-called one-point estimation. For this, we exploit the simultaneous stochastic approximation technique, based on randomly sampling the objective function in a neighbourhood of the power policy $\mathbf{p}(t)$ [13, 14, 17, 18].

We illustrate this idea on a particular directional derivative of $L_t(\mathbf{p})$ along the unit vector $\mathbf{x}$, denoted by $\nabla_{\mathbf{x}} L_t(\mathbf{p})$:

$$\nabla_{\mathbf{x}} L_t(\mathbf{p}) = \lim_{\delta \to 0} \frac{L_t(\mathbf{p}+\delta\mathbf{x}) - L_t(\mathbf{p}-\delta\mathbf{x})}{2\delta}. \tag{7}$$

To estimate this derivative, 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. The expectation of these samples w.r.t. the randomness of u is

$$\mathbb{E}\left[L_t(\mathbf{p}+\delta u\mathbf{x})u\right] = \frac{L_t(\mathbf{p}+\delta\mathbf{x}) - L_t(\mathbf{p}-\delta\mathbf{x})}{2}. \tag{8}$$

From (7) and (8), we observe that

$$\mathbb{E}\left[\frac{L_t(\mathbf{p}+\delta u\mathbf{x})u}{\delta}\right] \approx \nabla_{\mathbf{x}} L_t(\mathbf{p}). \tag{9}$$

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 $\mathbf{x}$.

This principle can be extended to build the following gradient estimate:

$$\tilde{\mathbf{v}}(t) = \frac{S}{\delta} L_t\left(\mathbf{p}(t)+\delta\mathbf{u}(t)\right)\ \mathbf{u}(t), \tag{10}$$

where $\mathbf{u}(t)$ is uniformly taken over the unit Euclidean sphere $\{\mathbf{u}\in\mathbb{R}^S |\ \|\mathbf{u}(t)\|^2 = 1\}$ [13].

Recently, this approach has been exploited in [8] to estimate the gradient of the objective function in the context of a different stochastic (not online) optimization problem. Aside from this, another major difference is that the network users in [8] are assumed to communicate and coordinate their policies to maximize a common overall utility function, as opposed to our distributed approach.

A major rising issue of such approaches is that they do not account for the fact that the random sample point $\mathbf{p}(t)+\delta\mathbf{u}(t)$ may be unfeasible. In our power allocation problem, this would imply that the transmit power vector $\mathbf{p}(t)+\delta\mathbf{u}(t)$ may fall outside $\mathcal{P}$, which cannot be permitted (given the physical power positivity and maximum available power constraints). To tackle this crucial issue, we define a shrunk version of the feasible set:

$$\mathcal{P}_\delta = \left\{ \mathbf{p}_\delta \in \mathbb{R}^S \;\middle|\; p^s_\delta \geq \delta, \sum_{s=1}^{S} p^s_\delta \leq \bar{P} - \sqrt{S}\delta \right\}, \tag{11}$$

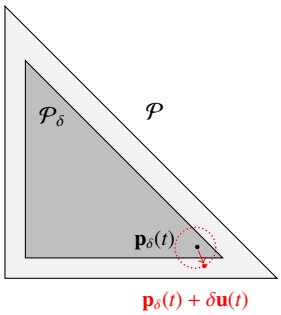


Fig. 2. The shrunk set $\mathcal{P}_\delta$ ensures that the random point $\mathbf{p}_\delta + \delta\mathbf{u}$ sampling the objective function to estimate the gradient remains in the feasible set $\mathcal{P}$.

which guarantees that for any δ-perturbation of its components $\mathbf{p}_\delta(t) \in \mathcal{P}_\delta$, the resulting policy is feasible: $\mathbf{p}_\delta(t)+\delta\mathbf{u}(t) \in \mathcal{P}$, as illustrated in Fig. 2.

IV. Zeroth-order Feedback Online Algorithm

As a building block for our algorithm, we exploit the exponential learning approach inspired from the multi-armed bandit framework and adapted to continuous sets of choices [13]. These algorithms have shown their potential in different online resource allocation problems [9, 11, 12] and their main drawback is their reliance on the gradient feedback, which is precisely what address here.

The exponential learning algorithm adapted to the problem at hand (5) can be summarized in two steps. First, the device transmits using $\mathbf{p}(t)$ by mapping an inner cumulative score $\mathbf{y}(t)$ into the feasible set $\mathcal{P}$ using a judiciously chosen exponential map. Second, as a result of the transmission, the gradient $\mathbf{v}(t)$ is obtained, which is used to update the cumulative score.

$$\begin{aligned} p^s(t) &= \bar{P}\frac{\exp\left(y^s(t)\right)}{1+\sum_{i=1}^{S}\exp\left(y^i(t)\right)},\ \forall s, \\ \mathbf{v}(t) &= \nabla L_t(\mathbf{p}(t)), \\ \mathbf{y}(t+1) &= \mathbf{y}(t) - \mu\ \mathbf{v}(t), \end{aligned} \tag{OXL}$$

where μ is the step-size parameter. This algorithm falls in the class of online gradient descent methods such as the Euclidean projected gradient descent algorithm in [19]. Similarly to [11, 12], we can show that this algorithm leads to no regret in our setting and that the regret vanishes as $O(T^{-1/2})$ (when either the perfect gradient $\mathbf{v}(t)$ or its noisy version is available).

In order to exploit the gradient estimation in Sec. III instead of the actual gradient $\mathbf{v}(t)$, we need to adapt the exponential step above to the shrunk set $\mathcal{P}_\delta$ defined in (11). For this, we introduce a novel exponential mapping: $\mathbf{Q}_\delta(\mathbf{y}(t)) \triangleq (p^1_\delta(t),\ldots,p^S_\delta(t))$ such that

$$p^s_\delta(t) \triangleq \delta + \bar{P}\left(1-C_\delta\right)\frac{\exp(y^s(t))}{1+\sum_{i=1}^{S}\exp(y^i(t))},\ \forall s \tag{EXP$_0$}$$

where $C_\delta = \frac{\delta}{\bar{P}}(S+\sqrt{S})$ and $\delta \leq \frac{\bar{P}}{S+\sqrt{S}}$.

Our novel algorithm exploiting (OXL) jointly with the one-point estimation of the gradient in Sec. III can be summarized as

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

where $\tilde{\mathbf{v}}(t)$ represents the biased estimate of the gradient. For implementation details, see OXL_0 algorithm below. Regarding the complexity, each iteration t is linear in the problem dimensionality S, the number of bands over which the focal device transmits. Hence, given that S is not expected to grow large (a given device transmits on a small subset of the total number of bands available to the entire network), the OXL_0 algorithm is particularly appealing for distributed, device-centric networks.

Since the online policy $\mathbf{p}_\delta(t)+\delta\mathbf{u}(t)$ depends on the random vector $\mathbf{u}(t)$, the regret in (6) will also depend on this randomness. To take this into account, we study the *average regret* $\mathbb{E}[Reg(T)]$ instead, where the expectation is taken over the randomness of the estimator, and prove that algorithm OXL_0 leads to no regret.

Theorem 1. *If the OXL_0 algorithm is run with constant parameters δ and μ then the average regret is bounded by:*

$$\mathbb{E}[Reg(T)] \leq \frac{\bar{P}\log(1+S)}{2\mu} + \mu T S^2\left(\frac{B}{\delta}+K\right)^2 + KT\delta\left(3+\bar{P}\left(S+2\sqrt{S}\right)\right). \quad (12)$$

where K is the Lipschitz constant and B is the maximum value of the objective $L_t(\cdot)$. Moreover, by choosing the parameters δ and μ as follows

$$\delta^* = \frac{\bar{P}}{(S+\sqrt{S})T^{1/4}} \quad (13)$$

$$\mu^* = \sqrt{\frac{\bar{P}\log(1+S)}{2T}}\left[S\left(\frac{B}{\delta^*}+K\right)\right]^{-1}, \quad (14)$$

then OXL_0 leads to no regret and the average regret vanishes as $O(T^{-1/4})$.

Algo. 1 Gradient-free Online Exponential Learning (OXL_0)

Parameters: $\mu > 0; 0 < \delta \leq \bar{P}/(S+\sqrt{S})$.
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 (EXP_0)
- Draw a random $\mathbf{u}(t)$ uniformly from the unit-sphere
- { *Transmit at* $\mathbf{p}_\delta(t)+\delta\mathbf{u}(t)$ }
- { *Post-transmission phase: receive feedback* $L_t(\mathbf{p}_\delta(t)+\delta\mathbf{u}(t))$ }
- Compute the gradient estimation $\tilde{\mathbf{v}}(t) = \frac{S}{\delta} L_t(\mathbf{p}_\delta(t)+\delta\mathbf{u}(t))\,\mathbf{u}(t)$
- Update scores $\mathbf{y}(t+1) \leftarrow \mathbf{y}(t) - \mu(t)\,\tilde{\mathbf{v}}(t)$
- $t \leftarrow t+1$

until transmission ends

The proof is omitted because of the limited space. Regarding the constants B and K, a short calculation shows that they depend only on readily available system parameters: $B = S\bar{P} + \lambda R_{\min}$, $K = 1 + 2\lambda R_{\min}$. Notice that the upper bound in (12) grows linearly in T, unless a careful choice of the two parameters μ and δ is made.

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 large variations and oscillations in the power allocation. At the opposite, a small μ leads to smaller variations in the power allocation. Both extremes imply a long time for the regret to reach zero.

The parameter δ represents the sampling radius around the power policy $\mathbf{p}_\delta(t)$. When tuning it, there is again 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 (9).

By choosing μ^* and δ^* as in (13) and (14) allows us to minimize the upper bound in (12) and to show the no regret property of OXL_0.

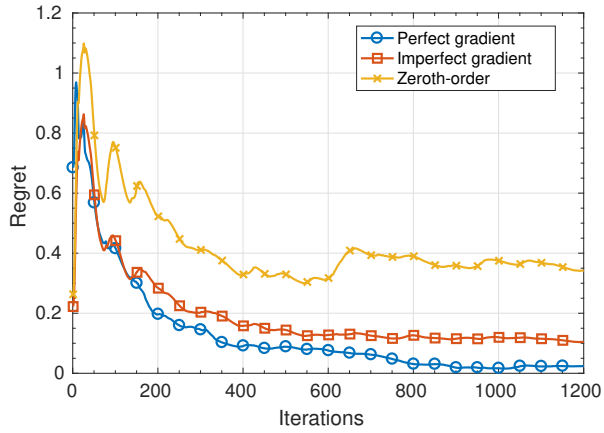
A rising issue is that the devices need to know their transmission horizon T in advance to optimally choose the parameters μ^* and δ^*. This issue can be overcome by exploiting for instance the so-called *doubling trick* [13]. This basically amounts to running the OXL_0 algorithm sequentially over known and doubling windows (enabling the device to optimally tune the parameters in each window) until the transmission ends. By doing so, similar theoretical guarantees (up to a multiplicative constant factor) can be provided in terms of regret.

From Theorem 1, we remark that the regret decays slower to zero, as $O(T^{-1/4})$, when only the scalar value of the objective function is fed back to the device, as opposed to $O(T^{-1/2})$ when the gradient (either perfect or noisy) is known. This highlights the tradeoff between the speed at which the online algorithm reaches the performance of the optimal benchmark and the available information.

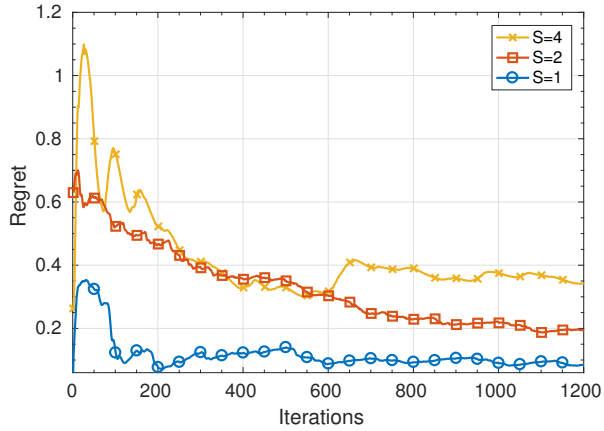
V. Numerical results

We consider a network composed of $M = 10$ interfering devices communicating to a common receiver $N = 1$ over $S = 4$ frequency bands (unless otherwise specified). The channels are generated according to the COST-HATA model [20], including pathloss, fast fading and shadowing effects [21]. The system bandwidth is 10 MHz centered at $f_c = 2$ GHz. The speed of the mobile devices is chosen arbitrarily between 0 km/h and 130 km/h, accounting for a wide variety of wireless mobile devices (smartphones, wearable, pedestrian, vehicle etc.). The minimum rate requirement $R_{\min} \in [0.5, 3]$ bps/Hz, the available power budget $\bar{P} \in [0.5, 2]$ W, and the parameter $\lambda \in [0.5, 10]$, also differ from one device to another. The parameters $\mu = 10^{-3}$ and $\delta = 0.05$ are empirically tuned.

Fig. 3 illustrates the impact of having a scarce or imperfect feedback and the impact of the problem dimensionality S. Fig. 3(a) shows that having an imperfect (a noisy estimated)



(a) Regret of OXL and OXL_0 algorithms



(b) Regret of OXL_0 algorithm

Fig. 3. Impact of feedback amount and problem dimensionality. The average regret of OXL_0 algorithm decays slower than that of OXL algorithm (because of estimating a gradient of higher dimension from a scalar value).

gradient feedback does not influence significantly the regret decay rate compared with the perfect gradient case. 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 the gradient-based OXL algorithm. Finally, Fig. 3(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 bands increases, the variance of the estimator $\tilde{\mathbf{v}}(t)$ also increases. As a result, the quality of the estimator decreases leading to a reduced decay rate of the average regret.

VI. Conclusions

We propose a novel online power allocation algorithm that explicitly takes into account the network dynamics and unpredictability, while relying only on a scalar and strictly causal feedback information, i.e., the value of the objective function (as opposed to gradient-based methods). We use a stochastic approximation technique to derive an estimation of the gradient based on one random sample of the objective function. This random sampling leads to a non-trivial feasibility issue, which we overcome by appropriately shrinking the feasible set. In so doing, we derive a novel exponential learning algorithm provably achieving no regret.

References

[1] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Kténas, N. Cassiau, and C. Dehos, "6G: The next frontier," *arXiv preprint, arXiv:1901.03239*, 2019.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] W. Li, M. Assaad, and P. Duhamel, "Distributed stochastic optimization in networks with low informational exchange," in *55th Annual Allerton Conf. on Commun., Control, and Computing*, 2017, pp. 1160–1167.

[9] 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.

[10] ——, "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.

[11] A. Marcastel, E. V. Belmega, P. Mertikopoulos, and I. Fijalkow, "Online power allocation for opportunistic radio access in dynamic OFDM networks," in *IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016.

[12] ——, "Online interference mitigation via learning in dynamic IoT environments," in *IEEE Globecom IoE Workshop*, Dec. 2016.

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

[14] 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.

[15] 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.

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

[17] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Autom. Control*, vol. 37, no. 3, pp. 332–341, 1992.

[18] 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: 16th ACM-SIAM Symp. on Discrete Algorithms*, Jan. 2005, pp. 385–394.

[19] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Intl. Conf. on Machine Learning (ICML-03)*, Aug. 2003, pp. 928–936.

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

[21] G. Calcev, et al., "A wideband spatial channel model for system-wide simulations," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 389–403, Mar. 2007.