IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

One-bit Feedback Exponential Learning for Beam Alignment in Mobile mmWave

IRCHED CHAFAA[1,2], E. VERONICA BELMEGA[1] (Senior Member, IEEE), and MÉROUANE DEBBAH[1,3] (Fellow, IEEE)

[1]ETIS, CY Cergy Paris Université, ENSEA, CNRS, Cergy, France
[2]L2S, UMR 8506, Université Paris-Saclay, CentraleSupélec, CNRS, Gif-sur-Yvette, France
[3]Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France
irched.chafaa@ensea.fr, belmega@ensea.fr, merouane.debbah@centralesupelec.fr

Corresponding author: Irched Chafaa (e-mail: irched.chafaa@ensea.fr).

Parts of this work have been presented in [1], [2]

ABSTRACT Efficient beam alignment in wireless networks capable of supporting device mobility is currently one of the major challenges in mmWave communications. In this context, we formulate the beam-alignment problem via the adversarial multi-armed bandit (MAB) framework, which copes with arbitrary network dynamics including non-stationary or adversarial components. Building on the well known exponential weights algorithm (EXP3) and by exploiting the structure and sparsity of the mmWave channel, we propose a modified (MEXP3) policy that requires solely one-bit of feedback information (reducing the amount of exchanged data during the beam-alignment process). Our MEXP3 comes with optimal theoretical guarantees in terms of asymptotic regret. Moreover, for finite horizons, our regret upper-bound is tighter than that of the original EXP3 suggesting better performance in practice. We then introduce an additional modification that accounts for the temporal correlation between successive beams and propose another beam-alignment policy. Our numerical results demonstrate that our beam-alignment policies outperform existing ones with respect to the regret but also to the outage, throughput and delay in typical mobile mmWave settings.

INDEX TERMS beam alignment, exponential weights, mobile mmWave, multi-armed bandits

I. INTRODUCTION

TO cope with the ever increasing mobile data traffic, an envisioned solution for future networks is to exploit the large available spectrum in the millimeter wave (mmWave) band [3], [4]. However, communicating at these frequencies is very challenging as the transmitted signal suffers from strong attenuation because of the high free-space path loss and additional losses when the signal penetrates objects or is absorbed by particles in the atmosphere [5]. All this leads to a limited propagation range and to a few multipath components (or a sparse mmWave channel). Hence, highly-directional beams have to be employed by both the transmitter and the receiver to concentrate the signal's energy and compensate all these losses. Such beams can be formed using high-gain antenna arrays, which contain a large number of elements and yet occupy little space thanks to the small wavelengths at such high frequencies [6].

This represents the so called *beam-alignment problem* where the beams of the transmitter and the receiver need to be constantly aligned to ensure a reliable communication link and overcome the difficulties of the mmWave channel. Moreover, the beam-alignment policies need to support user mobility and to cope with unpredictable and possibly non-stochastic variations of the wireless network (e.g., caused by the users' behavior and intermittent connectivity). Under such time-varying conditions, adjusting the beam-directions and identifying optimal beamforming vectors implies significant signaling and training overhead, which affects the overall performance. Hence, online beam-alignment policies capable of adapting on-the-fly to such changes become necessary to enable future mobile mmWave applications such as virtual reality headsets, autonomous vehicles, etc.

A. EXISTING WORK

The beam-alignment problem has been addressed in the literature from two main perspectives: beam training and compressed sensing (CS)[7]. The first approach consists of training from a set of candidate beamforming vectors through exhaustive search [8] or adaptive hierarchical search [9], [10] to identify the best beam-direction in terms of a given metric (e.g., signal-to-noise ratio (SNR)). For instance, in the IEEE 802.11ad standard [10] the beam search is done with wide beams whose widths are reduced progressively following a multi-level hierarchical scheme. The main limitation of such approaches resides in the large training feedback and coordination overhead making it unsuitable for mobile mmWave applications.

The CS-based methods [11], [12] use the sparsity of the mmWave channel to formulate the beam-alignment as a sparse recovery problem and reduce the training delay. The channel parameters, such as the propagation path gains and angles of arrival/departure, are estimated and used to construct the beamforming vectors for data transmission. This approach scales poorly with the number of antennas and requires a precise prior knowledge of the channel structure and sparsity. Moreover, these methods rely on strong assumptions regarding the temporal variation of the channel (either static or stochastic) during the estimation phase, which poses several issues when the channel is highly dynamic and possibly non-stochastic.

More recently, online optimization and machine learning (ML) tools have been investigated to design beam-alignment algorithms in dynamic wireless networks that rely on less stringent assumptions and are more data-oriented. Two main ML frameworks are exploited for beam-alignment: deep learning and multi-armed bandits (MAB). The policies based on the former make use of properly trained artificial neural networks (ANN) as universal approximators learning the relation between the mmWave environment and the optimal beam-direction. Although promising, this approach relies on a large amount of relevant training data and an optimal design of the ANN architecture. Acquiring such relevant training data is currently not trivial being both expensive and time-consuming; not to mention the privacy and security issues it may raise. More details can be found in [7], [13] and references therein.

In this work, we adopt the MAB formulation leaving as future investigation a more in-depth comparison between the two ML-based approaches. In this framework, the beam alignment is cast into a sequential decision-making problem, in which the devices (e.g., a central node or the transmitter/receiver) choose at each step a beam-direction, out of a finite set of choices, and then observe a reward (e.g., the resulting SNR at the receiver). The devices then learn the best beam direction by jointly *exploiting* the observed past rewards and *exploring* new beam directions. The authors in [14] proposed the unimodal beam alignment (UBA) algorithm that restricts the search set of the best directions by using the correlation between consecutive beams and the unimodality of the power of the received signal. An online beam-alignment algorithm for mmWave vehicular communications was introduced in [15], which uses the vehicle's direction of arrival as a contextual information. In [16], another beam-alignment policy was investigated, which requires the perfect knowledge of the data rates for all chosen beam directions. The policy in [17] incorporates the receiver's location as an out-of-band additional information to improve the beam alignment. In [18], the proposed policy aims at reducing the beam-alignment delay by exploiting previously acquired knowledge about the channel.

All the existing MAB-based policies above [14], [15], [16], [17], [18] depend on a central authority, which first chooses jointly the best pair of beam directions of the transmitter and the receiver, and then feedbacks the result to both devices resulting in a high signaling overhead. Moreover, these approaches are deterministic and exploit the so-called upper confidence bound (UCB). This directly implies that they are relevant only in stochastic and stationary wireless environments and cannot account for other possibly non-stationary components such as the behaviour and connectivity patterns of other devices.

B. OUR CONTRIBUTIONS

In this work, we focus squarely on *distributed* beam-alignment policies that do not require the existence of a central node nor rely on any assumptions regarding the network dynamics, as opposed to [14], [15], [16], [17], [18]. For this, we build on the exponential weights algorithm for exploration and exploitation (EXP3) in [19] to define novel beam-alignment policies capable to adapt to such *arbitrary and unpredictable environments*.

The exponential weights algorithm, also known as the multiplicative weights, has been repeatedly discovered in many fields ranging from optimization, game theory and machine learning [20], and has since become ubiquitous. Indeed, its applications range from data classification and prediction [21], privacy-preserving data analysis [22], learning graphical models [23], pooling problems for blending industries [24], learning the Nash equilibrium in various non-cooperative games [25], etc.

To the best of our knowledge, our work is the first to exploit exponential weights for beam alignment in mmWave networks [1], [2]. Compared with traditional schemes, our policies aim at learning the best beam-directions in an adaptive, online manner without relying on pre-deployed training every time the channel changes. Indeed, it is possible to simultaneously transmit data while tracking good beams from the beginning of the transmission. Of course, this comes with a cost in terms of high outage levels in the early stages of the learning process. The main advantage of our adaptive policies is that this cost happens only once, in the beginning of the transmission, and that they do not require dedicated training every channel coherence time (nor to optimize the training phase duration, which has a crucial impact on the data transmission efficiency).

Finally, our online beam-alignment policies do not require the perfect knowledge of the channel and relies solely on one-bit of feedback that basically captures whether the target SNR has been reached at the receiver.

Specifically, our main contributions can be summarized as follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3033419, IEEE Access

- We model the beam alignment in arbitrarily dynamic mmWave networks as an adversarial MAB problem [26], in which the transmitter and the receiver select their own beam directions individually while relying only on a 1-bit of feedback. Thanks to the adversarial formulation, which by definition is capable of coping with environments that vary in a completely arbitrary way, possibly non-stationary (as in our case) and even adversarial, decoupling the learning at the transmitter and the receiver becomes possible and results in splitting the complexity of the beam-alignment process between the two nodes.
- Building on the well-known EXP3 algorithm, we propose a novel modified exponential weights (MEXP3) algorithm that exploits the sparse structure of the mmWave channel. Based on this, we design a modified reward that reinforces the exploitation of past good beam-directions and penalizes the poor ones. We then prove that the new MEXP3 has the no-regret property and that the average regret decays to zero optimally as $\mathcal{O}(1/\sqrt{\mathcal{T}})$, where $\mathcal{T}$ denotes the time horizon, similarly to the original EXP3. Moreover, for fixed and finite horizons, our regret upper-bound for MEXP3 is tighter (smaller multiplicative constant factor) than the original EXP3 bound, suggesting better performance in practical settings.
- We introduce a further reward modification and propose the nearest neighbor-aided beam tracking modified exponential-weight algorithm (NBT-MEXP3), which exploits the temporal correlation between consecutive aligned beams to restrict the beam search to the neighborhood of a previously found good beam. The property of no-regret is conjectured for NBT-MEXP3 and validated via extensive numerical simulations.
- Although the asymptotic regret performance of the proposed algorithms, $\mathcal{O}(1/\sqrt{\mathcal{T}})$, is optimal and cannot be improved under arbitrary network dynamics [27], [28], our two novel policies MEXP3 and NBT-MEXP3 offer significant performance improvements in practical mmWave settings. Our numerical simulations, show that the proposed policies offer better practical performance especially in terms of outage and throughput for both single and multipath channels. Our modified rewards lead to online learning algorithms that adapt better and faster to the varying mmWave channel, which results in lower outage and higher data rates compared to other existing policies.

II. SYSTEM MODEL

We consider a point-to-point mmWave multiple-input multiple-output (MIMO) system, as depicted in Fig. 1, consisting of a fixed transmitter (Tx), equipped with M_T antennas and $N_T \leq M_T$ radio frequency (RF) chains, and a mobile receiver (Rx) equipped with M_R antennas and $N_R \leq M_R$ RF chains. Both nodes communicate via directional beams which point towards certain spatial directions determined by the hybrid (analog and digital) beamforming vectors $\mathbf{f}_i \in \mathbb{C}^{M_T}, i \in \{1, \ldots, A\}$ and $\mathbf{w}_j \in \mathbb{C}^{M_R}, j \in \{1, \ldots, A\}$ used at the transmitter and the receiver respectively.

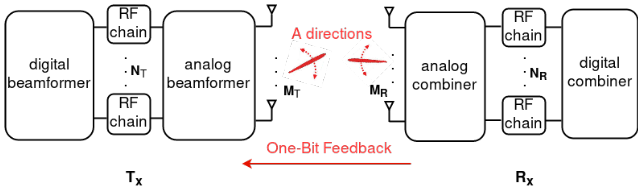

FIGURE 1. Beam-alignment in a point-to-point mmWave MIMO system.

Hybrid beamforming codebook

The codebook consists of a set of A hybrid beamforming vectors designed offline using the procedure in [29, Algorithm 1], which offers high beamforming gains compared to other existing codebooks in the literature [29]. The codebook size, $A = 2^n$, $n \in \mathbb{N}^\star$, represents the total number of all possible beam directions.

The analog beamformers designed to steer the transmitted signal into a particular spatial direction are implemented using phase shifters, which cover uniformly the angular domain between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Since each beamforming vector corresponds to a unique direction, our proposed online policies will select a suitable beamforming vector (or equivalently the beam direction) to meet the SNR requirements.

The main role of the digital weights is to optimize the beamforming gain of the different beams (with respect to an ideal beam pattern). They are fixed and tuned as in [29]. Also, the digital part of the codebook enables transmission via multiple data streams, which could be exploited in multi-user systems. Such a transmission mode is not possible when using only an analog beamformer with a single radio frequency (RF) chain.

mmWave channel model

The transmitted signal in the mmWave band experiences limited scattering. Therefore, we use the well-known narrow-band geometric model [9], [6], [30], [16] with L propagation paths

$$\mathbf{H}(t) = \sqrt{\frac{M_T M_R}{\rho}} \sum_{l=1}^{L} \alpha_\ell \, \mathbf{a}_R(\theta_\ell) \mathbf{a}_T(\phi_\ell)^\dagger \, \mathrm{e}^{j2\pi\nu_\ell t}, \qquad (1)$$

where ρ represents the average path loss [31]; $\alpha_\ell \sim \mathcal{N}(0, \sigma^2_{\alpha_\ell})$, $\ell \in \{1, 2, .., L\}$ is the complex path gain assumed to follow a Gaussian distribution; σ_{α_ℓ} is the average power gain; ϕ_ℓ and θ_ℓ are the angles of departure (AoD) and arrival (AoA) respectively; ν_ℓ is the Doppler shift of the ℓ^{th} path; $\mathbf{a}_T(\theta_\ell)$ and $\mathbf{a}_R(\phi_\ell)$ are the array steering vectors for the transmitter and the receiver. Assuming a uniform linear array (ULA), $\mathbf{a}_T(\theta_\ell)$ and $\mathbf{a}_R(\phi_\ell)$ can be expressed as:

$$\mathbf{a}_R(\theta_\ell) = \frac{1}{\sqrt{M_R}}[1, \mathrm{e}^{j\frac{2\pi}{\lambda}d\sin(\theta_\ell)}, ..., \mathrm{e}^{j(M_R-1)\frac{2\pi}{\lambda}d\sin(\theta_\ell)}]^T, \tag{2}$$

$$\mathbf{a}_T(\phi_\ell) = \frac{1}{\sqrt{M_T}}[1, \mathrm{e}^{j\frac{2\pi}{\lambda}d\sin(\phi_\ell)}, ..., \mathrm{e}^{j(M_T-1)\frac{2\pi}{\lambda}d\sin(\phi_\ell)}]^T, \tag{3}$$

where d is the distance between the antenna elements within the array and λ represents the wavelength of the transmitted signal. While most of the results in this paper hold irrespective from the channel model and network dynamics, we use the narrowband model in (1) to illustrate the performance of the proposed algorithms in Sec. V.

User mobility

To model the mobility of the receiver, we exploit the boundless mobility model adopted in [32] assuming a bounded two-dimensional movement area. This model is memory-based and incorporates temporal correlations in the update of the user's speed and direction, which leads to realistic settings and time-varying channel matrices $\mathbf{H}(t)$. It also allows to impose limitations on the linear speed, acceleration and rotation speed, and thus offers a good tradeoff between accuracy and flexibility [32]. In this model, the speed $v(t)$ and direction of movement $\Theta(t)$ are updated every channel coherence time T_c as follows:

$$\begin{cases} v(t+T_c) = \min\{\max\{v(t)+\Delta v,\ 0\},\ v_{\max}\}, \\ \Theta(t+T_c) = \Theta(t) + \Delta\,\Theta, \end{cases} \tag{4}$$

where $v_{\max}$ is the maximum speed; the speed variation is $\Delta v \sim \mathcal{U}[-a_{\max}\, T_c,\ a_{\max}\, T_c]$ with $a_{\max}$ being the maximum linear acceleration and $\mathcal{U}$ denoting the uniform distribution; $\Delta\Theta \sim \mathcal{U}[-\omega_{\max}\, T_c,\ \omega_{\max}\, T_c]$ is the direction variation with $\omega_{\max}$ denoting the maximum rotation speed.

In this work, we exploit this mobility model to generate different trajectories to simulate the receiver's mobility. Every update time T_c, the model parameters $v(t)$ and $\Theta(t)$ are used to determine the receiver's position with respect to the transmitter. Then using this new position, the channel parameters AoA $\theta_\ell(t)$ and AoD $\phi_\ell(t)$ are updated as shown explicitly in [33]. The other channel parameters $\rho(t)$ and $\nu_\ell(t)$ are also updated every T_c depending on the new transmitter-receiver distance and the speed respectively.

Received signal

The received signal $y_{i,j}$ at time t can be written as:

$$y_{i,j}(t) = \mathbf{w}_j^\dagger(t)\ \mathbf{H}(t)\ \mathbf{f}_i(t)\ s(t) + \mathbf{w}_j^\dagger(t)\ \mathbf{n}(t), \tag{5}$$

where i and j denote the indices of the transmit and receive beams respectively. To simplify the presentation, we drop the explicit temporal variability of the channel model parameters hereafter. During the beam-alignment process, the transmitter uses a beamforming vector $\mathbf{f}_i$ to transmit its symbols $s \in \mathbb{C}$ such that $\mathbb{E}[\mid s \mid^2] = P_{\text{tr}}$, where P_{tr} is the transmit power. The receiver uses its own beamforming vector $\mathbf{w}_j$ to recover the transmitted signal. The channel noise vector, denoted by $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2)$, is a Gaussian distributed random variable. The resulting SNR at the receiver depends on the beams $\mathbf{f}_i$ and $\mathbf{w}_j$ and is expressed as

$$\text{SNR}_{i,j} = \frac{\mid \mathbf{w}_j^\dagger \mathbf{H}\mathbf{f}_i \mid^2 P_{\text{tr}}}{\sigma_n^2}. \tag{6}$$

Assuming a stochastic channel model, we define an outage as the event in which the SNR falls below a certain threshold ξ, whose value represents the target SNR. The outage probability can be then defined as

$$\text{P}_{\text{out}}(i,j) \triangleq \text{Pr}[\text{SNR}_{i,j} < \xi]. \tag{7}$$

The choice of the threshold ξ will depend on the nature of the application. For instance, if the mmWave link is used for an application that requires high values of SNR, the value of ξ should be high as well.

The outage probability is an important performance metric in communications systems in which an average performance is less relevant than guaranteeing a minimum instantaneous quality of service [34], [35]. In 5G for instance, ultra-reliable low latency applications depend crucially on instantaneous reliability, which can be measured by the outage probability [36]. However, minimizing the outage probability above is quite a challenging problem as its explicit expression becomes intractable in practice (e.g., in our channel model with mobility or when the statistics of the channel is unknown). Indeed, even in the most simplified MIMO Rayleigh channel with perfect knowledge of the channel statistics at the transmitter, optimizing the outage probability remains an open issue [37]. Thus, we propose here to exploit the multi-armed bandit framework and sequential decision processes in an effort to approach the minimum outage as detailed below.

III. PROBLEM FORMULATION

Multi-armed bandit learning approaches have been considered recently to jointly tune the beam-alignment vectors at the transmitter and receiver $\mathbf{f}_i$ and $\mathbf{w}_j$ in such stochastic environments [14], [15], [16], [17], [18]. However, these works do not aim at minimizing the outage probability and they rely on a central authority or node that is able to compute the best pair of beams and to feedback the result to both the transmitter and receiver.

Our main goal is to propose distributed and decoupled beam-alignment policies at the transmitter and the receiver, which choose their own beamforming vectors $\mathbf{f}_{i_t}$ (beam-direction i_t) and $\mathbf{w}_{j_t}$ (beam-direction j_t) independently. Furthermore, our policies do not require any knowledge on the channel state or statistics and are only based on a single bit of feedback.

Ideally, to minimize the outage probability in a time-varying environment in a decoupled way, the transmitter would like to select the best beam-direction i_t at time t solving the following problem:

$$\forall t, \quad \underset{I\in\{1,2,\ldots,A\}}{\text{minimize}} \quad P_{\text{out}}\,(I, j_t) \tag{8}$$

and the receiver would do the same:

$$\forall t, \quad \underset{J \in \{1,2,\ldots,A\}}{\text{minimize}} \quad P_{\text{out}}\left(i_t, J\right). \tag{9}$$

Several issues arise in this ideal formulation[1]. First, the objective functions at each time instant are typically unknown at the transmitter and receiver. Indeed, the two objectives are inter-dependent, which raises a causality issue, and the channel statistics may be unknown at the transmitter. Second, the definition of the outage probability becomes problematic as the environment seen by one of the nodes (either the transmitter or the receiver) depends on the decision process of the other node, which effectively results in a non-stationary environment.

All the above motivates the use of online optimization and, more specifically, the use of the adversarial multi-armed bandit (MAB) framework [26] to propose adaptive beam-alignment schemes that approach these goals and which do not require any assumptions on the network dynamics. Indeed, since our beam-alignment problem is decoupled between the transmitter and receiver, the existing centralized approaches based on stochastic MABs [14], [15], [16], [17], [18] (which assume stochastic network dynamics) are no longer relevant. As mentioned above, even if the wireless channel is stationary the decoupled decision processes of each of the nodes may not be so.

A. ADVERSARIAL MAB FORMULATION

The advantage of the adversarial MAB formulation described below is that, by definition [26], it does not rely on any assumptions on the network dynamics, which can vary in a completely arbitrary way including adversary or non-stationary components. This feature is precisely what allows us to decouple the learning process between the transmitter and receiver.

In this formulation, the transmitter and the receiver are decision nodes that exploit separately an iterative online decision process as follows. At each time instant $t \in \{1, \ldots, \mathcal{T}\}$, where $\mathcal{T}$ is the time horizon or the transmission duration, a decision node chooses an action, in this case a beam direction: $i_t \in \{1, \ldots, A\}$ at the transmitter and $j_t \in \{1, \ldots, A\}$ at the receiver. As a result of the transmission, we assume that the receiver is able to compute a binary ACK-type of reward:

$$r_{i_t,j_t}(t) \triangleq \begin{cases} 1, & \text{if} \quad \text{SNR}_{i_t,j_t}(t) \geqslant \xi, \\ 0, & \quad\quad\ \text{otherwise}, \end{cases} \tag{10}$$

which is then fed back to the transmitter. Based on this observed reward, the decision nodes will update their action choices and so on.

The intuition behind our chosen reward in (10) is that the overall averaged reward over the transmission horizon $\mathcal{T}$, i.e., $\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} r_{i_t,j_t}(t)$, offers an approximation or an empirical measure of the outage probability. Moreover, assuming a stochastic channel model, the expected reward of a fixed beam pair (i, j) is directly linked to the outage probability defined in (7):

$$\begin{aligned} \mathbb{E}[r_{i,j}] &= P[\text{SNR}_{i,j} \geqslant \xi] \\ &= 1 - P_{\text{out}}(i,j), \end{aligned} \tag{11}$$

where the expectation $\mathbb{E}[.]$ is taken over the randomness of the channel. As shown above, maximizing the expected reward is equivalent to minimizing the outage probability in the stochastic case or the centralized beam-alignment problem. In our decoupled beam-alignment problem, this average reward represents an empirical measure of the outage probability at each of the decision nodes.

B. REGRET PERFORMANCE METRIC

In the MAB framework, the notion of *regret* has been considered as the relevant performance metric that evaluates the performance of an online policy [38], [14], [16]. The regret measures the gap in the average reward between the online policy and the *best fixed oracle policy in hindsight* over the time horizon $\mathcal{T}$. The latter is an ideal policy that maximizes the overall reward and relies on the non-causal knowledge of all the rewards during the entire horizon [19]. To be precise, the average regret at the transmitter side in our case writes as:

$$Reg_T = \frac{1}{\mathcal{T}}\left(\max_I \sum_{t=1}^{\mathcal{T}} r_{I,j_t}(t) - \sum_{t=1}^{\mathcal{T}} r_{i_t,j_t}(t)\right). \tag{12}$$

Similarly, the average regret at the receiver equals

$$Reg_R = \frac{1}{\mathcal{T}}\left(\max_J \sum_{t=1}^{\mathcal{T}} r_{i_t,J}(t) - \sum_{t=1}^{\mathcal{T}} r_{i_t,j_t}(t)\right). \tag{13}$$

An online policy has the property of no-regret if its average regret decays to (or less than) zero asymptotically: $\limsup_{\mathcal{T}\to\infty} Reg_Q \leq 0$, *with* $Q \in \{T, R\}$ *being the decision nodes.*

The no-regret property is an asymptotic performance guarantee ensuring that the online policy performs at least as good as the best fixed (or oracle) policy in hindsight (i.e., having perfect and non-causal knowledge of the networks dynamics throughout the horizon $\mathcal{T}$). Quite remarkably, this is achieved while relying only on strictly causal feedback amounting to a single bit of information. To sum up, in the rest of the paper, we propose several online policies in an effort to minimize the regret at both the transmitter and the receiver.

IV. PROPOSED BEAM-ALIGNMENT POLICIES

As mentioned before, the beam-alignment represents a crucial step in establishing a reliable link for data transmission in mmWave systems. In this section, we present three beam-alignment policies exploiting machine learning techniques and adversarial MABs. The first policy is based on the original *exponential weights for exploration and exploitation* (EXP3) algorithm in [19], which will be detailed below. We then propose two novel policies by modifying the chosen actions' (or beams) rewards. Our policies draw inspiration from

[1] This decoupled beam-alignment problem can also be interpreted as a team game or a common goal non-cooperative game with unknown payoff functions.

the sparse nature of the mmWave channel and the correlation between successive beams, leading to better performance results.

A. EXP3-BASED BEAM-ALIGNMENT POLICY

The main idea of EXP3 is to assign a probability to each possible action, and then choose an action according to this probability distribution, at each iteration t. Once an action is chosen, the decision node receives a reward and, as a result, it updates the probability distribution following an exponential map that depends on the cumulative scores or rewards up to that instant. In our case, the beam directions that often provide an SNR above the threshold ξ are the ones that are reinforced and have higher probabilities. In other words, EXP3 increases the probability of actions with good performance history, while not discarding completely the exploration of other actions that may perform better in the future; this effectively balances *the data exploitation versus data exploration*.

More precisely, at iteration t, the transmitter chooses a beam-direction i_t for data transmission according to the probability distribution $\hat{\mathbf{p}}_T(t)$ whose entries are defined for all $i \in \{1, 2, \ldots, A\}$ as:

$$\hat{p}_{T,i}(t) = (1-\gamma)\, p_{T,i}(t) + \frac{\gamma}{A}, \tag{14}$$

$$p_{T,i}(t) = \frac{\exp(\eta\, G_{T,i}(t-1))}{\sum_{k=1}^{A} \exp(\eta\, G_{T,k}(t-1))}, \tag{15}$$

where $G_{T,i}(t-1) = \sum_{\tau=1}^{t-1} \hat{r}_{T,i}(\tau)$ represents the cumulative score of action i. Since only the reward $r_{i_t,j_t}(t)$ of the chosen beam i_t can be observed at time t, we need to estimate other beams' rewards. For this, we define

$$\hat{r}_{T,i}(t) = \begin{cases} \dfrac{r_{i,j_t}(t)}{\hat{p}_{T,i}(t)}, & \text{if } i = i_t, \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

which represents an unbiased reward estimator for all beams i at time t [19].

The parameters $\eta > 0$ and $\gamma \in (0, 1]$ are tuning or learning parameters that tradeoff between data exploration and exploitation. Increasing the value of γ draws the probability distribution away from the exponential Gibbs distribution in (15) towards the uniform distribution, and hence moves away from data exploitation towards more exploration. An opposite behaviour is observed for the parameter η. When increasing η the exponential Gibbs distribution moves away from the uniform distribution (i.e., when $\eta = 0$) towards a Dirac or a deterministic pure exploitation policy. Notice that both parameters have to be very carefully tuned to optimize the tradeoff exploration vs. exploitation.

After the transmission, the transmitter receives 1-bit of feedback or the value of the reward $r_{i_t,j_t}(t)$ from the receiver and updates the cumulative scores as follows:

$$G_{T,i}(t) = G_{T,i}(t-1) + \hat{r}_{T,i}(t), \quad \forall i \in \{1, 2, \ldots, A\}. \tag{17}$$

The new cumulative rewards will be then exploited to update the transmitter's probability distribution $\hat{\mathbf{p}}_T(t+1)$ for the next round and so on. These different steps are summarized in the algorithm BA-EXP3.

Remark that the BA-EXP3 online policy consists of two equally important ingredients: i) the exponential mapping in (15) that reinforces the probabilities to choose beams that have performed well in the past, while still exploring new beams; and ii) the estimated rewards $\hat{r}_{T,i}(t)$ based on which the cumulative score in (17) is computed and which effectively evaluates the performance of past explored beams.

The receiver runs a similar algorithm independently from the transmitter. The two nodes' learning processes are linked via the feedback signaling. More precisely, the receiver uses its own probability distribution $\hat{\mathbf{p}}_R(t)$, defined similarly as in (14), to choose a beam-direction j_t at round t. Then, the receiver evaluates the binary reward for the chosen beam-directions i_t and j_t by comparing the received SNR with the threshold ξ as in (10), and then updates its probability distribution for the next round. We further assume that the obtained reward is sent back to the transmitter via a reliable control channel as a 1-bit feedback information [2].

```
BA-EXP3: Exponential Weight for Beam Alignment at Tx
Parameters: $\eta > 0$ and $\gamma \in (0, 1]$
Initialization: $G_i(0) = 0$ and $p_{T,i}(1) = 1/A, \forall i$
Repeat for $t = 1, 2, ..., \mathcal{T}$
    Select action $i_t$ with probability distribution $\hat{\mathbf{p}}_T(t)$
    Receive feedback $r_{i_t,j_t}(t) \in \{0, 1\}$ from Rx
    Update the cumulative rewards as in (17)
    Update the distribution $\hat{\mathbf{p}}_T(t+1)$ via (14)
```

The following theoretical result from [19] indicates that the expected average regret of algorithm BA-EXP3 decays to zero as $\mathcal{O}(1/\sqrt{\mathcal{T}})$. This decay rate is optimal and cannot be improved in the absence of strong stationarity assumptions regarding the underlying network dynamics [27], [28]. As argued in Sec. III, in our *distributed* beam-alignment problem, the wireless environment depends on the other node's decisions and, hence, does not evolve following a stochastic stationary process.

Corollary 1 (Theorem 1 in [19]) *If the BA-EXP3 beam-alignment policy is run at both the transmitter and receiver with the parameters* $\eta = \dfrac{\gamma}{A}$ *and* $\gamma = \min\left\{1, \sqrt{\dfrac{A \log A}{(e-1)\,\mathcal{T}}}\right\}$ *for a horizon* $\mathcal{T}$, *then the expected average regret is upper*

[2] The control channel can be either a microwave channel as proposed in the ECMA 387 standard [39], or a mmWave channel as in the IEEE 802.15.3c [40] and IEEE 802.11ad [41] standards. On the one hand, the directional mmWave channel is low-cost, but may suffer from poor reliability due to difficult propagation characteristics. On the other hand, the omni-directional microwave channel is more reliable at the expense of additional hardware and energy consumption [42].

bounded as:

$$\mathbb{E}[Reg_Q] \leq 2\sqrt{e-1}\sqrt{\frac{A \log A}{\mathcal{T}}}, \tag{18}$$

with $e = \exp(1)$, *and the expectation is taken over the randomness of the BA-EXP3 policy.*

The upper bound of the expected average regret in (18) shows that the BA-EXP3 policy is asymptotically optimal when $\mathcal{T}$ grows large. Also, when the transmission horizon $\mathcal{T}$ is finite or small, this bound also provides a worst-case guarantee in terms of the gap between the empirical outage of BA-EXP3 compared with the ideal oracle policy, which depends only on $\mathcal{T}$ and the number of available beams A.

B. MODIFIED EXPONENTIAL WEIGHTS ALGORITHM (MEXP3)

Measurement campaigns [43], [44], [45] have demonstrated the existence of only a few multipath components in the mmWave propagation environment, which leads to a limited number of available propagation paths with high enough SNR for data transmission. We exploit this channel's sparsity to adapt the BA-EXP3 algorithm and identify faster the good beam-directions.

To do so, let us denote the global reward matrix $\mathbf{R}(t) \in \{0,1\}^{A\times A}$ such that

$$\mathbf{R}(t) = [r_{i,j}(t)]_{\substack{1\leq i\leq A\\ 1\leq j\leq A}} \tag{19}$$

where the rewards are defined in (10). The matrix $\mathbf{R}(t)$ contains the rewards of all possible pairs (i,j) at time t and is not fully available at any of the two nodes. A typical example of a reward matrix $\mathbf{R}(t)$ is illustrated in Fig. 2 for a particular mmWave channel setting and $A = 16$ (complete details will be provided in the next section). Due to the characteristics of the mmWave channel, the matrix $\mathbf{R}(t)$ has a particular sparse structure. The few non-zero entries are all grouped in one or a few clusters and correspond to the set of good beam-directions.

Hence, the goal of our online policies is to identify the indices i (beam-directions at Tx) and j (beam-directions at Rx) that correspond to a unit value in this matrix (to avoid an outage event and guarantee a minimum SNR at the receiver). Based on this observation, we leverage the structure of the reward matrix to define a modified reward:

$$\tilde{r}_{T,i}(t) = \begin{cases} \dfrac{-1}{1-\hat{p}_{T,i}(t)}, & \text{if } i = i_t \text{ and } r_{i_t,j_t}(t) = 0,\\ \dfrac{\beta}{\hat{p}_{T,i}(t)}, & \text{if } i = i_t \text{ and } r_{i_t,j_t}(t) = 1,\\ 0, & \text{otherwise,} \end{cases} \tag{20}$$

where $\beta \geq 1$ is a weighting parameter which affects the beam selection probabilities and, hence, represents another parameter that tradeoffs between data exploration and exploitation and which needs to be carefully tuned.

The intuition behind the above modified reward is to reinforce good beam directions (i.e., the ones that provide $r_{i_t,j_t}(t) = 1$) by associating them a reward β-times higher than the original BA-EXP3. Moreover, the poor beams ($r_{i_t,j_t}(t) = 0$) are penalized by associating them a strictly negative reward as opposed to zero. Dividing by the quantity $1 - \hat{p}_{T,i_t}(t)$ leads to an important and fast penalization of a past good beam that has accumulated a high probability to be chosen, but which has become obsolete because of changes in the mmWave environment. Also, dividing by $1 - \hat{p}_{T,i_t}(t)$ insures a soft penalization of a beam with low probability to avoid discarding it completely as it may become a future good beam. To sum up, this denominator penalizes the poor beams according to their past performance and not randomly by just assigning a negative reward. Therefore, the modified reward encourages the algorithm to adapt faster to the changes in the channel and to keep track of the good beam directions.

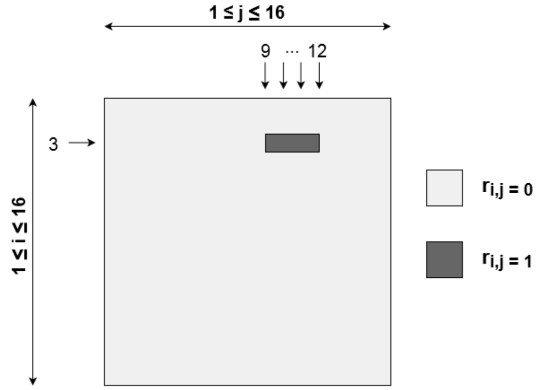

FIGURE 2. Illustration of a typical reward matrix in mmWave channels, for the setting: $M_T = 32$, $M_R = 4$, $L = 1$, $A = 16$, $\xi = 6$ dB and a carrier frequency $f_c = 28$ GHz.

```
MEXP3: Modified Exponential Weight for Beam Alignment at Tx
Parameters $\eta > 0$, $\beta \geq 1$ and $\gamma \in (0,1]$
Initialization: $G_i(0) = 0$ and $p_{T,i}(1) = 1/A, \forall i$
Repeat for $t = 1, 2, ..., \mathcal{T}$
    Select action $i_t$ with probability distribution $\hat{\mathbf{p}}_T(t)$
    Receive feedback $r_{i_t,j_t}(t)$ from Rx
    Construct the modified rewards $\tilde{r}_{T,i}(t)$ as in (20)
    Update the cumulative rewards:
        $G_{T,i}(t) = G_{T,i}(t-1) + \tilde{r}_{T,i}(t), \forall i$
    Update the distribution $\hat{\mathbf{p}}_T(t+1)$ via (14)
```

Although the new algorithm MEXP3 may seem quite similar to the original algorithm BA-EXP3 at first, our new modified reward $\tilde{r}_{T,i_t}(t)$ in (20) results in a very different behavior with respect to the regret and other performance metrics. This modified reward changes one of two key ingredients of the original EXP3 algorithm: the cumulative scores $G_{T,i}(t) = \sum_{\tau=1}^{t} \tilde{r}_{T,i}(\tau)$ that evaluate the performance of the past explored beams, which are then mapped on the probability simplex (via the exponential map). In particular, the no-regret proof and showing that the expected cumulative regret of MEXP3 grows sub-linearly with respect to the time horizon is very different than the proof in [19]. One of the main challenges we have overcome is that $\tilde{r}_{T,i}(t)$ no longer represents an unbiased reward estimator for beam i. All the details behind our proof are presented in the Appendix.

Theorem 1 *If the MEXP3 policy is run at both the transmitter and receiver with the parameters* $A \geq 3$, $\eta = \frac{\gamma}{\beta A}$, $\gamma = \min\left\{1 - \frac{2}{A}, \sqrt{\frac{2A\ln A}{e\,\mathcal{T}}}\right\}$ *and* $\beta \geq \max\left\{1, \sqrt{\frac{2}{A}}\left(\gamma - \frac{\gamma}{A}\right)^{-1}\right\}$*, then the expected average regret is bounded by:*

$$\mathbb{E}[Reg_Q] \leq \sqrt{2\,e}\,\sqrt{\frac{A\ln A}{\mathcal{T}}}. \tag{21}$$

Notice that the expectation of the regret in Theorem 1 is taken only with respect to the random sequence of the chosen beam directions. This means that the no-regret property holds irrespective from the underlying system dynamics, which can be arbitrary and even non stationary.

The above result shows that MEXP3 provides an optimal asymptotic regret performance similarly to the original BA-EXP3; the regret decays as $\mathcal{O}(1/\sqrt{\mathcal{T}})$. Even though we cannot improve this decay rate under arbitrary and non-stationary network dynamics, the upper bound we obtain for MEXP3 is tighter than the one for BA-EXP3 in the multiplicative constant, which is important in the finite horizon regime. This indicates that MEXP3 outperforms BA-EXP3 in terms of regret and that the gap between the two algorithms is larger for relatively short transmissions (finite $\mathcal{T}$).

C. NEAREST NEIGHBOUR-AIDED BEAM TRACKING (NBT-MEXP3)

Here, we propose an additional modification by using a contextual information to accelerate the beam search and tracking. Our empirical observations of the temporal evolution of the reward matrix $\mathbf{R}(t)$ indicates a temporal correlation between the locations of its unit-valued clusters that depends on the mobility of the receiver (speed, orientation, etc.) and on the wireless characteristics (blockage, line-of-sight (LOS), non-LOS (NLOS), etc.). The location of the clusters does not change abruptly or randomly but rather smoothly following the mobility of the receiver.

Concretely, this means that future good beam directions are more likely to be among the neighboring directions that have performed well in the past. Therefore, we can exploit this intuition to keep track of good beams with the aid of their nearest neighbours. This new feature increases the tracking speed of the good beams by adapting to the user's mobility and other changes in the channel. For this, we modify the rewards of the non-chosen beams as follows:

$$\tilde{r}_{T,k}\,(t) = \frac{\beta'\, r_{i_t,j_t}(t)}{\hat{p}_{T,i_t}(t)}, \quad \forall k \in V_{i_t}, \tag{22}$$

where $V_{i_t} = \{i_t{-}1, i_t{+}1\}$ is the set of the nearest neighbors[3] of the chosen beam direction i_t at time t and parameter $\beta' \in [1, \beta]$, which plays a similar role as β for the neighbouring beams.

Combining the modified reward in (20) for the chosen action i_t with the reward in (22) for its neighbors, we construct a new reward vector $\tilde{\mathbf{r}}_T(t) = [\tilde{r}_{T,k}(t)]_{k\in\{1,\ldots,A\}}$, which is used to update the cumulative rewards for each beam-direction, as follows

$$\tilde{r}_{T,k}(t) = \begin{cases} \dfrac{(-1)^{1+r_{i_t,j_t}(t)}\ \beta^{\ r_{i_t,j_t}(t)}}{1 - r_{i_t,j_t}(t) + (-1)^{1+r_{i_t,j_t}(t)}\hat{p}_{T,k}(t)}, & \text{if } k = i_t, \\ \dfrac{\beta'\ r_{i_t,j_t}(t)}{\hat{p}_{T,k-1}(t)}, \quad \text{if } k = i_t + 1, & \\ \dfrac{\beta'\ r_{i_t,j_t}(t)}{\hat{p}_{T,k+1}(t)}, \quad \text{if } k = i_t - 1, & \\ 0, \quad \text{otherwise}. & \end{cases} \tag{23}$$

The resulting NBT-MEXP3 algorithm is detailed below.

NBT-MEXP3: Nearest Neighbour-aided Beam Tracking with MEXP3 at Tx

```
Parameters $\eta > 0$, $1 \leq \beta' \leq \beta$ and $\gamma \in (0, 1]$
Initialization: $G_i(0) = 0$ and $p_{T,i}(1) = 1/A, \forall i$
Repeat for $t = 1, 2, ..., \mathcal{T}$
    Select action $i_t$ with probability distribution $\hat{\mathbf{p}}_T(t)$
    Receive feedback $r_{i_t,j_t}(t)$ from Rx
    Construct the reward vector $\tilde{\mathbf{r}}_T(t)$ as in (23)
    Update the cumulative rewards:
        $G_i(t) = G_i(t-1) + \tilde{r}_{T,i}\,(t)\,,\ \forall i$
    Update the distribution $\hat{\mathbf{p}}_T(t+1)$ via (14)
```

Although finding a sub-linear upper bound for the regret of NBT-MEXP3 is not trivial, our extensive numerical simulations indicate that the NBT-MEXP3 policy holds the no-regret property asymptotically.

Conjecture 1 *The proposed NBT-MEXP3 beam-alignment policy has the no-regret property and the average regret decays to zero as* $\mathcal{O}(1/\sqrt{\mathcal{T}})$*, similarly to BA-EXP3 and MEXP3.*

The proof of the above conjecture is left open for future work. By following a similar approach as in the proof of Corollary 1 and Theorem 1, an encountered difficulty comes from the ratio $\frac{p_{T,k}(t)}{\hat{p}_{T,i_t}(t)}$, $k \in V_{i_t}$, which appears in the expectation of the regret and which cannot be bounded appropriately. This term is due to our modified reward $\tilde{r}_{T,i}(t)$ in (23) assigning a non-zero reward to the neighbouring beam of a good direction.

[3] We focus only on the two nearest neighbors for simplicity reasons and also based on our empirical observations. Choosing a larger (or optimized) size for the neighbors' set could be of interest for future research.

V. NUMERICAL RESULTS

In this section, we evaluate the performance of the proposed beam-alignment algorithms in terms of regret, outage, throughput and delay in a typical mmWave setting described in Sec. II and specified here. Notice that our online beam-alignment policies and their theoretical guarantees do not rely on any assumptions on the underlying network dynamics. This implies that the conclusions drawn below carry over many other mmWave settings incorporating various practical aspects and specifications.

The plotted curves are averaged over $10,000$ scenarios or time-varying channel realizations over the horizon $\mathcal{T}$. Our online policies do not rely on any initial knowledge of the wireless environment (i.e. the beam search starts with a random choice following the uniform distribution). The channel is assumed to remain constant during a transmission frame T_c, which represents the channel coherence time. The duration T_c consists of several sub-frames such that each sub-frame represents one iteration of the online beam-alignment policies at both Tx and Rx or, more precisely, the time interval between two successive feedback signals. In our simulations, we consider a channel coherence time $T_c = 1.3$ ms [46] and a sub-frame duration of 250 μs [47]. This results in 5 sub-frames per coherence interval, meaning that the channel conditions change every 5 iterations of our online policies, because of device mobility, time-varying wireless characteristics, etc. We consider the mmWave MIMO point-to-point link of Fig. 1 with $M_T = 42$, $N_T = 4$, $M_R = 32$ and $N_R = 2$. Both nodes are equipped with ULAs with $\lambda/2$ spacing between their elements. The transmission power is $P_{\text{tr}} = 37$ dBm. The size of the beamforming codebook is $A = 32$ for a single-path channel, which ensures a good tradeoff between beam-alignment accuracy and exploration cost as detailed in [1]. The threshold ξ for the SNR at the receiver is fixed at 6 dB.

Regarding the wireless channel matrix, the commonly used geometric model in (1) is adopted with $\alpha_\ell \sim \mathcal{N}(0,1)$, the carrier frequency $f_c = 28$ GHz and a bandwidth of 1 MHz, which meets the narrowband channel assumption according to the maximum delay spread measurements in [48]. The noise power density equals -174 dBm/Hz. The pathloss ρ is calculated following the close-in free space model in [31], [48] as follows

$$\rho = 20 \log \frac{4\pi f_c}{c} + 10\, n_p \log D \ \ [\text{dB}], \tag{24}$$

where $c = 3 \times 10^8$ is the speed of light, D is the distance between the transmitter and receiver and n_p is the pathloss exponent, which equals 2.1 for LOS and to 3 for NLOS [48, Table 3]. The Doppler shift is updated every T_c such that $\nu_\ell = \dfrac{v\, f_c}{c}$ as in [6], [30]. Unless stated otherwise, we consider a single-path channel ($L = 1$). For the multipath channel $L = 3$, we consider a LOS path combined with two NLOS paths determined by two reflectors positioned randomly between the transmitter and the receiver for each new channel realization.

The location of the transmitter is fixed (e.g., a base station). The receiver (a mobile user) is assumed to move in the area covered by the transmitter's ULA within a distance less than 200 m. The mobility model in (4) is used with the typical parameters: speed $v_{max} = 30$ km/h, acceleration $a_{max} = 2$ m/s^2 and rotation speed $\omega_{max} = \pi/4$ rad/s. The position of the receiver is updated every transmission frame of duration 1.3 ms, which corresponds to the channel coherence time under our dynamic conditions [46]. For each receiver position, a new channel matrix $\mathbf{H}$ is computed by updating its parameters as detailed in Section II. In other words, the channel conditions change (implying that the good beam directions that meet the SNR requirement also change) every 5 iterations of our algorithms in the figures below.

The learning parameters of our online policies are chosen empirically based on extensive numerical simulations. Here, we set $\eta = 0.02$, $\gamma = 0.001$ for BA-EXP3; $\eta = 0.023$, $\gamma = 0.03$ and $\beta = 10$ for MEXP3; $\eta = 0.01$, $\gamma = 0.001$, $\beta = 10$ and $\beta' = 5$ for NBT-MEXP3. Naturally, we exploit the values and the ranges obtained in Corollary 1 and Theorem 1 as starting point. Notice that these values are optimal only with respect to the upper-bounds of the regret and are not necessarily optimal in terms of the actual regret. A more efficient way to fine tune these parameters is a non-trivial issue to be investigated in future work.

In what follows, we compare our policies with existing ones in the literature but also with several relevant benchmarks, which we briefly described below.

- **Centralized-UCB**: the centralized beam-alignment policy proposed in [16] based on stochastic MABs and the upper-confidence bound (UCB) algorithm.
- **Exhaustive search**: the brute-force policy that tries all A^2 beam pairs (one at each iteration) in a round-robin fashion and selects the best one; this optimal beam is then exploited until the channel changes and the process is reinitialized.
- **Rand**: the random beam-direction is drawn following the uniform distribution.

Average regret

We start by comparing our policies to the original BA-EXP3 and the Centralized-UCB at the transmitter side. The average regret is plotted in Fig. 3. We also plot the upper bounds of the expected average regret of BA-EXP3 and MEXP3 given in Corollary 1 [19] and Theorem 1.

We first notice that all policies based on exponential learning: BA-EXP3, MEXP3 and NBT-MEXP3, yield an average regret lower and decaying faster compared with Centralized-UCB. This can be explained by the fact that the Centralized-UCB policy has a larger set of choices, the A^2 beam-direction pairs, whereas the distributed policies have a set of only A beam-directions. The larger search set of Centralized-UCB requires more data exploration, which leads to more regret.

Also, both our modified policies outperform the original BA-EXP3. They are more adapted to the varying mmWave channel since they are inspired from its particular structure.

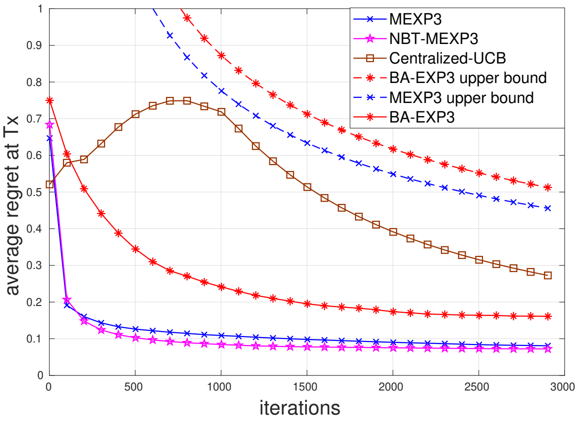

FIGURE 3. The average regret at the transmitter decays faster for our proposed policies MEXP3 and NBT-MEXP3, with a slight advantage for the latter. The three distributed policies based on exponential learning clearly outperform the Centralized-UCB policy.

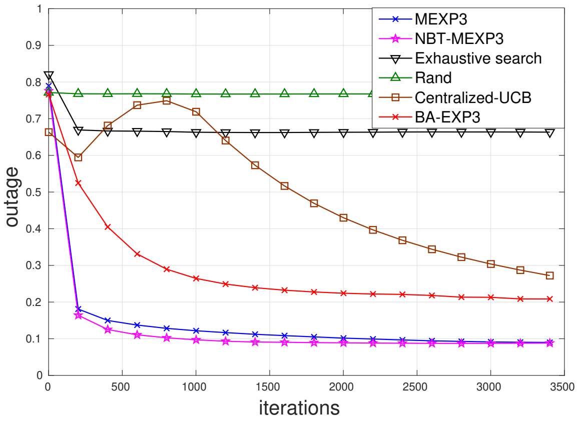

FIGURE 4. Our novel policies, MEXP3 and NBT-MEXP3, outperform the original BA-EXP3, Centralized-UCB as well as the other benchmarks. This shows the importance of exploiting the structure of the mmWave channel and of our modified rewards to reach lower outage.

The upper-bounds validate our theoretical results: the obtained bound for MEXP3 in Theorem 1 is tighter than the bound for the BA-EXP3.

Outage

Fig. 4 illustrates the empirical outage of the different beam-alignment policies. Our new algorithms, MEXP3 and NBT-MEXP3, outperform clearly the original BA-EXP3 and the other policies. This highlights the interest of exploiting the special structure of the mmWave channel and modifying the rewards as in MEXP3. The nearest neighbours additional reward modification in the NBT-MEXP3 policy provides a slight performance improvement compared to MEXP3. Also, the exhaustive search policy results in high outage similarly to the random policy. This is mainly caused by the fact that only 5 beam pairs can be explored from the total of $A \times A = 1024$ possibilities before the change in the channel conditions occurs. In turn, this effectively renders the gathered information about those 5 trials outdated and irrelevant.

Regarding the number of iterations needed to reach an outage below 10% (around 2000 iterations for MEXP3), it is equivalent to a duration of 500 ms. Although 500 ms may seem long at first, it is a low price to pay for the entire transmission duration. Indeed, once these early learning stages have passed, our method is capable to adapt to the network changes and track good beams while reliably transmitting data. At the opposite, traditional methods have to perform dedicated training and identify good beam directions *every time the channel has changed* (every T_c) before transmitting any data at all, having a crucial impact on the effective performance.

Effective throughput

Fig. 5 illustrates the evolution of the average achievable rate as a function of the iterations. In our proposed policies, we do not separate the communication in two distinct phases: beam-alignment training and data transmission. Instead, the transmitter and receiver communicate effectively during the whole frame while adjusting the beam-directions (at the cost of higher outage levels in the early learning stages). In Fig. 5, we make the same assumption for Centralized-UCB and exhaustive search policies for comparison purposes. Our novel policies MEXP3 and NBT-MEXP3 outperform the original BA-EXP3 and the other benchmarks in terms of the speed in reaching higher data rates.

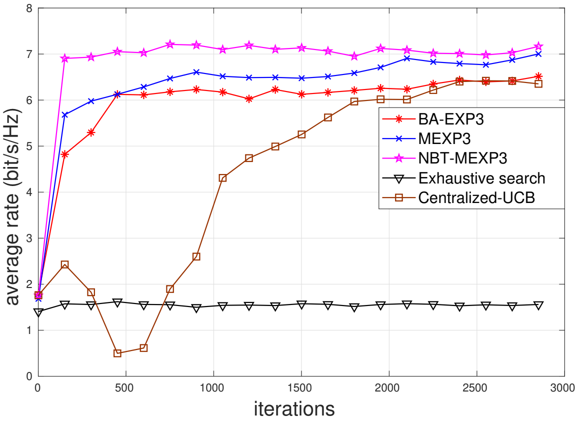

FIGURE 5. Exploiting neighbouring beams (NBT-MEXP3) is beneficial in achieving higher average rate.

Average delay

Here, we compare the average beam-alignment delay of the three exponential learning policies, which represents the average time interval required to identify good beam-directions that provide an SNR above the threshold for a given channel. Fig. 6 depicts the average delay as a function of the SNR threshold for two different codebook sizes $A = \{8, 32\}$. We notice that reaching higher SNR thresholds require more exploration time to find good beams. This highlights the *latency vs. reliability tradeoff* between the delay and the SNR at the receiver.

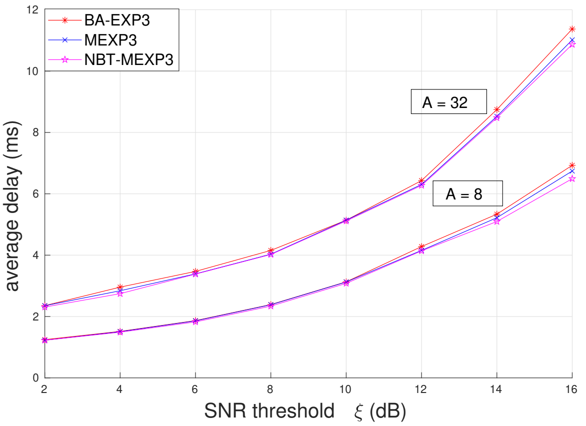

FIGURE 6. The average delay as function of the SNR threshold ξ and for codebook sizes $A = \{8, 32\}$. All exponential learning policies lead to similar performance with a slight advantage for the NBT-MEXP3 policy.

Regarding the impact of the codebook size A, Fig. 6 shows that the average delay increases with the codebook size. Indeed, the beams of a larger codebook are narrower and induce more delay given that the search set of beams is larger. However, using a large codebook allows to focus the signal's energy in a more compact angular domain to reach higher beamforming gains, which illustrates again the latency vs. reliability tradeoff.

Impact of user mobility

We now investigate the ability of our NBT-MEXP3 and MEXP3 algorithms to support high-mobility conditions and their impact on the empirical outage. We compare the outage performance obtained with the following mobility parameters: $v_{max} = 30$ km/h, $a_{max} = 2$ m/s^2 and $\omega_{max} = \pi/4$ rad/s (low-mobility); and with more dynamic parameters: $v_{max} = 110$ km/h, $a_{max} = 5$ m/s^2 and $\omega_{max} = \pi/2$ rad/s (high-mobility). In Fig. 7, we can see that increasing the mobility of the receiver leads to higher outage levels as expected. Higher mobility implies more frequent changes in the mmWave channel which affects the quality of the beam alignment and results in lower SNR. Moreover, the proposed algorithms need more iterations to reach low outage levels compared to the low-mobility setting. Fig. 7 shows that our proposed policies may be suitable for high-mobility mmWave applications with an increased delay cost.

Impact of multipath channel

We compare the outage performance of the proposed beam-alignment policies, MEXP3 and NBT-MEXP3, in a multipath channel (when $L = 3$) composed of one LOS path and two NLOS components and the single LOS channel (when $L = 1$). Fig. 8 shows the ability of the proposed policies to adjust the beams even in a multipath channel with an additional exploration cost, as it takes longer to reach lower outage levels. This can be explained by the less favorable propagation conditions (involving higher path loss for NLOS paths combined with possible destructive combination of multipath components).

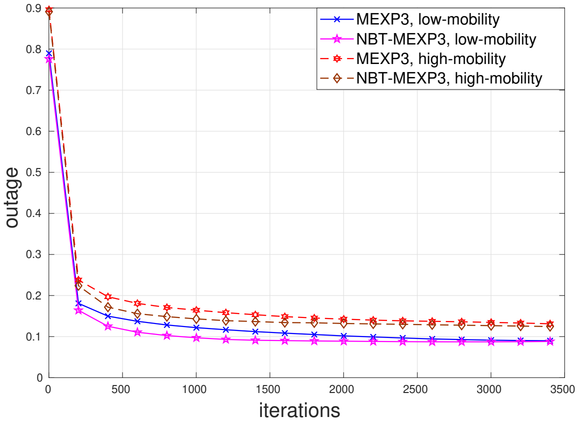

FIGURE 7. Impact of the user's speed: higher mobility leads to higher outage levels.

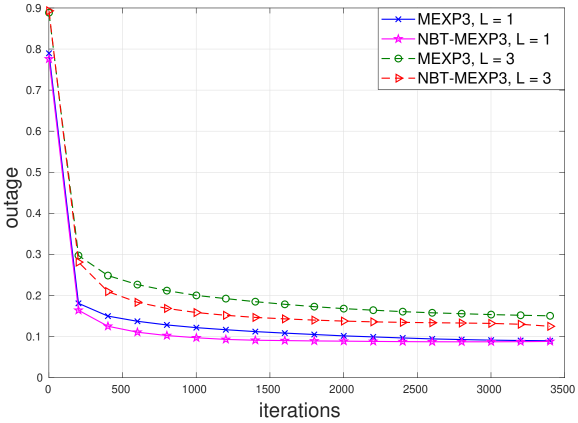

FIGURE 8. Multipath components and NLOS paths lead to higher outage and increase the exploration cost:

COMPLEXITY VS. PERFORMANCE

Regarding the complexity of the various policies, we discuss here the scalability of each iteration in function of the codebook size A. Exhaustive search and the random policies have a constant cost, $\mathcal{O}(1)$, as only one pair of beams can be tested at each sub-frame or iteration. For Centralized-UCB policy, the complexity of each iteration scales linearly with the number of available choices, in this case the number of joint beamforming pairs, $\mathcal{O}(A^2)$. The complexity of all distributed online policies based on the adversarial MAB framework: BA-EXP3, MEXP3 and NBT-MEXP3, also scales linearly with the number of choices, which in this case represents the number of individual beams at each decision node, i.e., $\mathcal{O}(A)$ (the size of the probability distributions updated at each iteration).

The above highlights the tradeoff between complexity and performance. The least complex policies are the ones which perform quite poorly in terms of performance (random and exhaustive search). Remarkably, our online policies allows one to distribute the complexity between the transmitter and receiver, resulting in relatively low complexity policies that are also capable of adapting to the dynamic and unpredictable changes in the network. Centralized-UCB suffers from the larger number of joint beam pairs, A^2, and, because it inherently relies on stochastic and stationary channel assumptions, it is not suitable to multi-user scenarios, in which the network dynamics will depend on other decision nodes and will hence be non-stationary.

On the contrary, our distributed online policies rely on no assumptions on the underlying network dynamics and can be extended to multi-user scenarios as discussed below.

VI. EXTENSIONS TO WIDEBAND, MULTI-USER MMWAVE NETWORKS

For the sake of simplicity and clarity of presentation, we have focused in this paper on a narrowband single point-to-point mmWave link. The extension to wideband mmWave networks (multi-carrier or single-carrier) involves adapting the codebook design (specifically the digital part of the beams) as in [49], [50]. Once this is done, our online policies can be easily exploited. The amount of feedback bits over the control channel would equal the number of carriers (one bit per carrier) in the multi-carrier case.

The extension of the proposed policies to multi-user interference channels is straightforward [2]. The adversarial MAB framework and our online policies based on the exponential weights algorithm rely on no assumptions on the underlying network dynamics, which can easily incorporate the interference from other transmitter-receiver pairs. The one-bit feedback mechanism would operate in a similar manner for each individual pair, under the mild assumption that each one has access to an interference-free control channel.

In the uplink (multiple access channels) or the downlink (broadcast channels), the design of the multi-user beamforming codebooks is much more involving and requires non trivial interference management, allocation of the multiple antennas over the served users, etc. Nevertheless, once the codebooks have been properly designed and each transmitter and receiver has access to its own finite set of actions (beams), the multi-armed bandits framework and our online policies based on the exponential weights algorithm can be easily adapted. The feedback mechanism would require a number of control channels equal to the number of users (as in the multi-user interference channels) knowing that only a single bit required to/from each of the users. At last, in the downlink, the transmitter would have to wait for all one-bit feedback signals to arrive from the receivers before transmitting new data.

VII. CONCLUSIONS

In this paper, we address the beam-alignment problem in dynamic mmWave networks. We exploit the adversarial multi-armed bandit framework to design distributed policies, in which the transmitter and the receiver choose their beams separately while relying only on a one-bit feedback. Building on the well known exponential weights algorithm (EXP3), we propose two novel beam-alignment policies (MEXP3 and NBT-MEXP3) that exploit the mmWave characteristics and lead to tracking optimal beam directions more efficiently. We prove rigorously that our MEXP3 online policy has the no-regret property, while a conjecture is provided for NBT-MEXP3 (validated via extensive simulations).

The performance of the proposed algorithms is demonstrated via numeric results in terms of regret, outage, throughput and average delay in a practical mmWave setting. We show that our policies outperform the original BA-EXP3 and other existing centralized policies by being capable to adapt to the rapid and unpredictable changes of the mmWave channel. Regarding the performance gap between our two novel algorithms, NBT-MEXP3 only slightly outperforms MEXP3. One possible improvement lead is to exploit the mobility model to predict the user's position and orientation.

VIII. APPENDIX

PROOF OF THEOREM 1

We start by proving the following lemma which will be exploited in the main proof [4].

Lemma 1 *For the parameters* $A \geq 3$, $\eta = \dfrac{\gamma}{\beta A}$, *and* $\beta \geq \max\left\{1, \sqrt{\dfrac{2}{A}}\,\left(\gamma - \dfrac{\gamma}{A}\right)^{-1}\right\}$ *and* $0 < \gamma \leq 1 - \dfrac{2}{A}$, *we have*

$$p_{T,i_t}(t)\;\tilde{r}^2_{T,i_t}(t) \leq \frac{\beta^2\;A}{2\;(1-\gamma)},\;\forall t \geq 1.$$

Since $0 \leq p_{T,i_t}(t) \leq 1$, to prove this result, it suffices to show that for all $t \geq 1$

$$0 \leq \frac{2\;(1-\gamma)\;\beta^{\;2 r_{i_t,j_t}(t)}}{\beta^2\;A\;\left(1 - r_{i_t,j_t}(t) + (-1)^{1+r_{i_t,j_t}(t)}\hat{p}_{T,i_t(t)}\right)^2} \leq 1, \tag{25}$$

We distinguish the two cases depending on the value of the reward of the chosen actions. a) If $r_{i_t,j_t}(t) = 1$, the inequalities in (25) are met for $\gamma \leq 1 - \dfrac{2}{A}$. b) If $r_{i_t,j_t}(t) = 0$, the inequalities in (25) are true for $\beta \geq \max\left\{1, \sqrt{\dfrac{2}{A}}\,\left(\gamma - \dfrac{\gamma}{A}\right)^{-1}\right\}$.

For the main proof of Theorem 1, we define the sum $S_t \triangleq \sum_{i=1}^{A} \exp\left(\eta\,G_{T,i}\,(t-1)\right),\ \forall t \geq 1$. Then, for any $A > 1$, $\eta > 0$, $\beta \geq 1$ and $0 < \gamma < 1$, we have the following ratio

$$\frac{S_{t+1}}{S_t} \quad = \quad \textstyle\sum_{i=1}^{A} p_{T,i}(t) \exp\left(\eta\;\tilde{r}_{T,i}(t)\right). \tag{26}$$

From Lemma 3.3 in [19] and the inequality $\tilde{r}_{T,i}(t) \leq \dfrac{\beta A}{\gamma}$, the following holds:

$$\exp\left(\eta\;\tilde{r}_{T,i}(t)\right) \leq 1 + \eta\;\tilde{r}_{T,i}(t) + \Phi_M(\eta)\;\tilde{r}^2_{T,i}(t)$$

[4] We provide a proof to Theorem 1 at the transmitter side as similar steps hold for the regret bound at the receiver.

with $\Phi_M(\eta) = \dfrac{\exp(M\eta) - 1 - M\,\eta}{M^2}$ and $M = \dfrac{\beta A}{\gamma}$.
Using this inequality, the ratio in (26) can be expressed as

$$\frac{S_{t+1}}{S_t} \leq 1 + \eta\, p_{T,i_t}(t)\, \tilde{r}_{T,i_t}(t) + \Phi_M(\eta)\, p_{T,i_t}(t)\, \tilde{r}^2_{T,i_t}(t), \tag{27}$$

since the unchosen beams at time t have a zero reward.

Next, the idea is to upper bound the last two terms of the inequality (27) as follows:

$$p_{T,i_t}(t)\, \tilde{r}_{T,i_t}(t) \leq \frac{\beta\, r_{i_t,j_t}(t)}{1-\gamma}, \forall t \geq 1, \tag{28}$$

$$p_{T,i_t}(t)\, \tilde{r}^2_{T,i_t}(t) \leq \frac{\beta^2\, A}{2\,(1-\gamma)}, \forall t \geq 1, \tag{29}$$

the latter follows from Lemma 1.
Since $\eta > 0$ and $\Phi_M(\eta) > 0$, combining (27), (28), (29) and the fact that $1 + x \leq \exp(x)$, $\forall x \in \mathbb{R}$ leads to

$$\frac{S_{t+1}}{S_t} \quad \leq \quad \exp\left(\frac{\eta\,\beta\, r_{i_t,j_t}(t)}{1-\gamma} + \frac{\Phi_M(\eta)\,\beta^2\, A}{2\,(1-\gamma)}\right). \tag{30}$$

Now, by first taking the logarithm and then summing over $t = 1, ..., \mathcal{T}$ in the above, we further obtain

$$\ln \frac{S_{\mathcal{T}+1}}{S_1} \leq \frac{\eta\,\beta}{1-\gamma}\left(\sum_{t=1}^{\mathcal{T}} r_{i_t,j_t}(t)\right) + \frac{\Phi_M(\eta)\,\beta^2\, A}{2\,(1-\gamma)}\mathcal{T}. \tag{31}$$

Since $S_1 = A$ and $S_{\mathcal{T}+1} \geq \exp\left(\eta\, G_{T,k}(\mathcal{T})\right)$, for an arbitrary and fixed $k \in \{1, \ldots, A\}$, we get

$$\ln \frac{S_{\mathcal{T}+1}}{S_1} \geq \eta\, G_{T,k}(\mathcal{T}) - \ln A,\ \forall k \tag{32}$$

with $G_{T,k}(\mathcal{T}) = \sum_{t=1}^{\mathcal{T}} \tilde{r}_{T,k}(t)$.

Combining (31) and (32), we can bound the overall rewards of the algorithm, denoted by G_{Alg}, as follows

$$\begin{aligned} G_{\text{Alg}} &\triangleq \sum_{t=1}^{\mathcal{T}} r_{i_t,j_t}(t) \\ &\geq \frac{1-\gamma}{\beta}\sum_{t=1}^{\mathcal{T}} \tilde{r}_{T,k}(t) - \frac{1-\gamma}{\eta\,\beta}\ln A - \frac{\Phi_M(\eta)\,\beta\, A\,\mathcal{T}}{2\eta} \end{aligned}$$

Taking the expectation with respect to the distribution of the chosen beams $\langle i_1, ..., i_{\mathcal{T}}\rangle$ in the random online policy, we obtain

$$\mathbb{E}[G_{\text{Alg}}] \geq \frac{1-\gamma}{\beta}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\tilde{r}_{T,k}(t)] - \frac{1-\gamma}{\eta\,\beta}\ln A - \frac{\Phi_M(\eta)\,\beta\, A\,\mathcal{T}}{2\eta},$$

with

$$\mathbb{E}[\tilde{r}_{T,k}(t)] \quad = \quad \begin{cases} \beta, & \text{if } r_{k,j_t}(t) = 1, \\ \dfrac{-\hat{p}_{T,k}(t)}{1-\hat{p}_{T,k}(t)}, & \text{if } r_{k,j_t}(t) = 0. \end{cases}$$

We can now show that

$$\mathbb{E}[G_{\text{Alg}}] \geq (1-\gamma)\mathcal{T} - \frac{1-\gamma}{\eta\,\beta}\ln A - \frac{\Phi_M(\eta)\,\beta\, A\,\mathcal{T}}{2\eta},\ \forall t \geq 1.$$

Next, let $EG_{\max} \triangleq \max\limits_{I,J} \sum_{t=1}^{\mathcal{T}} \mathbb{E}[r_{i_t,j_t}(t)]$, denote the expected cumulative rewards of the oracle best solution in hindsight. Knowing that it can't be higher than $\mathcal{T}$, as all rewards are either 0 or 1, we show that

$$EG_{\max} - \mathbb{E}[G_{\text{Alg}}] \leq \gamma\,\mathcal{T} + \frac{1-\gamma}{\eta\,\beta}\ln A + \frac{\Phi_M(\eta)\,\beta\, A\,\mathcal{T}}{2\eta}. \tag{33}$$

Substituting $\Phi_M(\eta) = \dfrac{\gamma^2}{\beta^2\, A^2}\left(\exp(\dfrac{\beta\eta A}{\gamma}) - 1 - \dfrac{\beta\eta A}{\gamma}\right)$ and $\eta = \dfrac{\gamma}{\beta\, A}$ in (33) yields to

$$EG_{\max} - \mathbb{E}[G_{\text{Alg}}] \leq \frac{(1-\gamma)\, A \ln A}{\gamma} + \frac{\exp(1)\,\gamma\,\mathcal{T}}{2}.$$

Hence, the expected average regret is upper bounded as

$$\frac{EG_{\max} - \mathbb{E}[G_{\text{Alg}}]}{\mathcal{T}} \quad \leq \quad \frac{A \ln A}{\gamma\,\mathcal{T}} + \frac{\exp(1)\,\gamma}{2}.$$

The upper bound above is a convex function with respect to γ. We can thus minimize it and obtain the optimal step-size $\gamma = \sqrt{\dfrac{2\,A \ln A}{\exp(1)\,\mathcal{T}}}$ and the following optimal bound of the expected average regret

$$\frac{EG_{\max} - \mathbb{E}[G_{\text{Alg}}]}{\mathcal{T}} \leq \sqrt{2\,\exp(1)}\sqrt{\frac{A \ln A}{\mathcal{T}}},$$

which completes our proof.

REFERENCES

[1] I. Chafaa, E. V. Belmega, and M. Debbah, "Adversarial multi-armed bandit for mmWave beam alignment with one-bit feedback," *ACM Proceeding of the 12th EAI VALUETOOLS conference*, pp. 23–30, 2019.

[2] ——, "Exploiting channel sparsity for beam alignment in mmWave systems via exponential learning," in *IEEE ICC 2020 Workshop: Machine Learning in Communications, accepted paper*, 2020.

[3] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *arXiv preprint arXiv:1902.10265*, 2019.

[5] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave cellular wireless networks: Potentials and challenges," *arXiv preprint arXiv:1401.2560*, 2014.

[6] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process*, vol. 10, no. 3, pp. 436–453, 2016.

[7] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37 328–37 348, 2018.

[8] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun*, vol. 32, no. 6, pp. 1164–1179, 2014.

[9] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process*, vol. 8, no. 5, pp. 831–846, 2014.

[10] "Wireless lan medium access control (MAC) and physical layer (PHY) specifications amendment 3: Enhancements for very high throughput in the 60 GHz band," *IEEE 802.11 working group and others, IEEE Computer Society*, 2012.

[11] D. E. Berraki, S. M. Armour, and A. R. Nix, "Application of compressive sensing in sparse spatial channel recovery for beamforming in mmWave outdoor systems," in *IEEE WCNC*, 2014, pp. 887–892.

[12] J. Choi, "Beam selection in mmWave multiuser MIMO systems using compressive sensing," *IEEE Trans. Commun*, vol. 63, no. 8, pp. 2936–2947, 2015.

[13] A. Klautau, N. Gonzalez-Prelcic, and R. W. Heath, "Lidar data for deep learning-based mmwave beam-selection," *IEEE Commun. Lett*, vol. 1, p. 1–1, 2019.

[14] 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.

[15] A. Asadi, S. Müller, G. H. A. Sim, A. Klein, and M. Hollick, "FML: Fast Machine Learning for 5G mmWave Vehicular Communications," in *IEEE INFOCOM*, 2018, pp. 1961–1969.

[16] J.-B. Wang, M. Cheng, J.-Y. Wang, M. Lin, Y. Wu, H. Zhu, and J. Wang, "Bandit inspired beam searching scheme for mmWave high-speed train communications," *arXiv preprint arXiv:1810.06150*, 2018.

[17] V. Va, T. Shimizu, G. Bansal, and R. W. Heath Jr, "Online learning for position-aided millimeter wave beam training," *arXiv preprint arXiv:1809.03014*, 2018.

[18] W. Wu, N. Cheng, N. Zhang, P. Yang, W. Zhuang *et al.*, "Fast mmWave beam alignment via correlated bandit learning," *arXiv preprint arXiv:1909.03313*, 2019.

[19] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *IEEE FOCS*, 1995, pp. 322–331.

[20] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: a meta-algorithm and applications," *Theory of Computing*, vol. 8, no. 6, pp. 121–164, 2012.

[21] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, pp. 349–360, 2009.

[22] M. Hardt and G. N. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *IEEE FOCS*, 2010, pp. 61–70.

[23] A. Klivans and R. Meka, "Learning graphical models using multiplicative weights," in *IEEE FOCS*, 2017, pp. 343–354.

[24] L. Mencarelli, Y. Sahraoui, and L. Liberti, "A multiplicative weights update algorithm for minlp," *EURO J. Comput. Optim.*, vol. 5, no. 1-2, pp. 31–86, 2017.

[25] P. Mertikopoulos and Z. Zhou, "Learning in games with continuous action sets and unknown payoff functions," *Mathematical Programming*, vol. 173, no. 1-2, pp. 465–507, 2019.

[26] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[27] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *Journal of the ACM*, vol. 44, no. 3, pp. 427–485, 1997.

[28] 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*, 2018.

[29] J. Song, J. Choi, and D. J. Love, "Codebook design for hybrid beamforming in millimeter wave systems," in *IEEE ICC*, 2015, pp. 1298–1303.

[30] X. Song, S. Haghighatshoar, and G. Caire, "Efficient beam alignment for mmWave single-carrier systems with hybrid MIMO transceivers," *arXiv preprint arXiv:1806.06425*, 2018.

[31] S. Sun, T. A. Thomas, T. S. Rappaport, H. Nguyen, I. Z. Kovacs, and I. Rodriguez, "Path loss, shadow fading, and line-of-sight probability models for 5G urban macro-cellular scenarios," in *IEEE Globecom Workshops*, 2015, pp. 1–7.

[32] L. De Nardis and M.-G. Di Benedetto, "Momo: a group mobility model for future generation mobile wireless networks," *arXiv preprint arXiv:1704.03065*, 2017.

[33] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "Position and orientation estimation through millimeter-wave MIMO in 5G systems," *IEEE Trans. Wireless Commun*, vol. 17, no. 3, pp. 1822–1835, 2018.

[34] C.-H. Yao, Y.-Y. Chen, B. P. Sahoo, and H.-Y. Wei, "Outage reduction with joint scheduling and power allocation in 5G mmWave cellular networks," in *IEEE PIMRC*, 2017, pp. 1–6.

[35] J. N. Murdock, E. Ben-Dor, Y. Qiao, J. I. Tamir, and T. S. Rappaport, "A 38 GHz cellular outage study for an urban outdoor campus environment," in *IEEE WCNC*, 2012, pp. 3085–3090.

[36] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[37] E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecommun*, vol. 10, no. 6, pp. 585–595, 1999.

[38] A. Marcastel, E. V. Belmega, P. Mertikopoulos, and I. Fijalkow, "Online power optimization in feedback-limited, dynamic and unpredictable iot networks," *IEEE Trans. Signal Process*, vol. 67, no. 11, pp. 2987–3000, 2019.

[39] E. TC48, "High rate 60 ghz phy, mac and hdmi pal," *ECMA standard*, vol. 387, 2008.

[40] J. P. GILB, "Part 15.3: Wireless medium access control and physical layer specifications for high rate wireless personal area networks: Amendment 2: Millimeter-wave based alternative physical layer extension," *P802-15-3c-DF3_Draft_Amendment*.

[41] I. S. Association, "IEEE std 802.11 ad-2012, part 11: Wireless Lan Medium Access Control and physical layer specifications, amendment 3: Enhancements for very high throughput in the 60 ghz band," *IEEE Computer Society*, 2012.

[42] H. S. Ghadikolaei, "Mac aspects of millimeter-wave cellular networks," in *Wireless Mesh Networks-Security, Architectures and Protocols*. IntechOpen, 2019.

[43] H. Zhao, R. Mayzus, S. Sun, M. Samimi, J. K. Schulz, Y. Azar, K. Wang, G. N. Wong, F. Gutierrez Jr, and T. S. Rappaport, "28 GHz millimeter wave cellular communication measurements for reflection and penetration loss in and around buildings in New York city." in *IEE ICC*, 2013, pp. 5163–5167.

[44] Y. Azar, G. N. Wong, K. Wang, R. Mayzus, J. K. Schulz, H. Zhao, F. Gutierrez Jr, D. Hwang, and T. S. Rappaport, "28 GHz propagation measurements for outdoor cellular communications using steerable beam antennas in New York city." in *IEEE ICC*, 2013, pp. 5143–5147.

[45] G. R. MacCartney and T. S. Rappaport, "73 GHz millimeter wave propagation measurements for outdoor urban mobile and backhaul communications in New York city." in *IEEE ICC*, 2014, pp. 4862–4867.

[46] F. Khan, Z. Pi, and S. Rajagopal, "Millimeter-wave mobile broadband with large scale spatial processing for 5G mobile communication," in *IEEE Allerton*, 2012, pp. 1517–1523.

[47] C. Herranz, M. Zhang, M. Mezzavilla, D. Martin-Sacristán, S. Rangan, and J. F. Monserrat, "A 3GPP NR compliant beam management framework to simulate end-to-end mmWave networks," in *ACM MSWiM*, 2018, pp. 119–125.

[48] J. Lee, J. Liang, M.-D. Kim, J.-J. Park, B. Park, and H. K. Chung, "Measurement-based propagation channel characteristics for millimeter-wave 5G Giga communication systems," *ETRI Journal*, vol. 38, no. 6, pp. 1031–1041, 2016.

[49] A. Alkhateeb and R. W. Heath, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, 2016.

[50] W. Huang, Y. Huang, R. Zhao, S. He, and L. Yang, "Wideband millimeter wave communication: Single carrier based hybrid precoding with sparse optimization," *IEEE Trans. Veh Technol*, vol. 67, no. 10, pp. 9696–9710, 2018.

IRCHED CHAFAA has been a PhD student at ETIS laboratory with ENSEA graduate school and L2S laboratory with CentraleSupelec graduate school in France since November 2017 working on resource allocation policies for mmWave networks using tools from the machine learning framework. His research interests lie in mmWave networks and online learning applied to wireless communications with a focus on distributed dynamic networks.

E. VERONICA BELMEGA (S'08-M'10-SM'20) has been an Associate Professor (MCF HDR) with ENSEA graduate school since 2011 and Deputy Director of ETIS laboratory since 2020, Cergy, France. She received the M.Sc. (engineer diploma) degree from the University Politehnica of 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. She received the HDR habilitation degree from the University of Cergy-Pontoise in 2019. From 2010 to 2011, she was a Post-doctoral researcher in a joint project between Princeton University, N.J., USA and Supélec, France. In 2015-2017, she was a visiting researcher at Inria, Grenoble, France. Her research interests lie in convex optimization, game theory and online learning applied to distributed networks. She served as Executive Editor of Trans. on Emerging Telecommun. Technologies (ETT) in 2016-2020; among the Top Editors 2016-2017. Dr. E. Veronica Belmega received the L'Oréal - UNESCO - French Academy of Science national fellowship in 2009. From 2018 until 2022, she receives the Doctoral Supervision and Research Bonus by the French National Council of Universities. She is a Senior Member IEEE since 2020.

MÉROUANE DEBBAH (S'01-M'04-SM'08-F'15) received the M.Sc. and Ph.D. degrees from the Ecole Normale Supérieure Paris-Saclay, France. He was with Motorola Labs, Saclay, France, from 1999 to 2002, and also with the Vienna Research Center for Telecommunications, Vienna, Austria, until 2003. From 2003 to 2007, he was an Assistant Professor with the Mobile Communications Department, Institut Eurecom, Sophia Antipolis, France. From 2007 to 2014, he was the Director of the Alcatel-Lucent Chair on Flexible Radio. Since 2007, he has been a Full Professor with CentraleSupelec, Gif-sur-Yvette, France. Since 2014, he has been a Vice-President of the Huawei France Research Center and the Director of the Mathematical and Algorithmic Sciences Lab. He has managed 8 EU projects and more than 24 national and international projects. His research interests lie in fundamental mathematics, algorithms, statistics, information, and communication sciences research. He is an IEEE Fellow, a WWRF Fellow, and a Membre émérite SEE. He was a recipient of the ERC Grant MORE (Advanced Mathematical Tools for Complex Network Engineering) from 2012 to 2017. He was a recipient of the Mario Boella Award in 2005, the IEEE Glavieux Prize Award in 2011, and the Qualcomm Innovation Prize Award in 2012. He received 20 best paper awards, among which the 2007 IEEE GLOBECOM Best Paper Award, the Wi-Opt 2009 Best Paper Award, the 2010 Newcom++ Best Paper Award, the WUN CogCom Best Paper 2012 and 2013 Award, the 2014 WCNC Best Paper Award, the 2015 ICC Best Paper Award, the 2015 IEEE Communications Society Leonard G. Abraham Prize, the 2015 IEEE Communications Society Fred W. Ellersick Prize, the 2016 IEEE Communications Society Best Tutorial Paper Award, the 2016 European Wireless Best Paper Award, the 2017 Eurasip Best Paper Award, the 2018 IEEE Marconi Prize Paper Award, the 2019 IEEE Communications Society Young Author Best Paper Award and the Valuetools 2007, Valuetools 2008, CrownCom 2009, Valuetools 2012, SAM 2014, and 2017 IEEE Sweden VT-COM-IT Joint Chapter best student paper awards. He is an Associate Editor in-Chief of the journal Random Matrix: Theory and Applications. He was an Associate Area Editor and Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2011 to 2013 and from 2013 to 2014, respectively.

• • •