Hybrid beamforming algorithm using reinforcement learning for millimeter wave wireless systems

Enrique M. Lizarraga[1], Gabriel N. Maggio[1], Alexis A. Dowhuszko[2], *Senior Member, IEEE*
[1] Digital Communications Research Laboratory, National University of Cordoba, Argentina
[2]Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Castelldefels (Barcelona), Spain
Email: emlizarraga@unc.edu.ar, gabriel.maggio@unc.edu.ar, alexis.dowhuszko@cttc.es

***Abstract*—In this paper, a Reinforcement Learning (RL) algorithm is presented to speed up the selection process of spatial beams to maximize the mean data rate of a multi-antenna wireless system that implements hybrid beamforming in Millimeter Wave (mmWave) frequency bands. In the proposed hybrid beamforming architecture, the analog beamforming layer is codebook-based, and is implemented using a simple array of phase-shifters that delay the RF signal in the different transmit antennas using a fixed number of discrete steps. In contrast, the digital beamforming layer is much more flexible, and implements a fully adaptive (*i.e.*, *non-quantized*) digital precoding scheme that enables the simultaneous transmission of few independent baseband data streams in the spatial domain. Obtained simulation results show that the use of RL-based techniques reduces the iterations that are needed to find the most convenient analog beamformers and digital precoders to be used in transmission, without affecting notably the upper bound data rate that is achieved when brute-force search is utilized.**

***Index Terms*—Hybrid beamforming; codebook-based beamforming; massive MIMO; millimeter wave; artificial intelligence; machine learning; reinforcement learning.**

I. Introduction

The wireless communication systems that have been deployed so far utilize most of the Radio Frequency (RF) spectrum that is available in the low frequency bands (*i.e.*, below 6 GHz), leaving scarce communication resources to be utilized by the future generations of mobile networks. In order to cope with the foreseen demand for wireless connectivity, 3GPP has considered the incorporation of disruptive technologies into the definition the 5G New Radio (NR) air interface. For example, 5G will use the abundant spectral resources that are available in the Millimeter Wave (mmWave) frequency bands, which have not been extensively used so far due to the strong path loss attenuation that they experience [1]. In order to address this impairment, large-scale antenna arrays will be deployed at both extremes of the wireless link, enabling high beamforming gains and allowing the multiplexing of few parallel data streams in the spatial domain. The combination of these two technologies, which is known as mmWave Massive MIMO, requires new *hybrid beamforming* architecture, as the implementation of fully digital precoders is not practical due to the large number baseband processing units and RF transmission chains that would be required [2].

A hybrid beamforming architecture can be divided into two parts, namely the *digital precoder* and the *analog beamformer*. The digital precoder interfaces the parallel streams of input symbols with the RF transmission chains, allowing flexibility when defining the precoding weights for the different frequency portions of the baseband signal. On the other hand, the analog beamformer connects the output of the RF blocks with the transmit antennas. Due to its analog nature, the phase shift that the beamformer applies per antenna is the same for the whole wideband RF signal [3]. Different approaches, such as the ones reported in [4], [5], [6], have been proposed to implement the hybrid beamforming scheme. Most of them assume that the analog beamformer can adjust continuously the phase shift per antenna, and that the weights of the digital precoder can be optimized to obtain a combined effect (*i.e.*, digital precoder plus analog beamformer) that is as close as possible to the one obtained with a fully digital implementation. Typically, these hybrid beamforming algorithms operate iteratively, such that the digital precoder is optimized once the analog beamformer is updated, and *vice versa*.

In this paper, we present a novel hybrid beamforming algorithm that seeks the maximization of the achievable sum data rate of a mmWave Massive MIMO system. For this purpose, the digital precoder and analog beamformer to be utilized in transmission are jointly determined, assuming that the weights of the analog beamformer can only belong to a set of uniformly quantized phase shift values [7]. More precisely, it is assumed that for a given analog beamformer, an equivalent lower-dimension wireless channel can be obtained, whose capacity-achieving transmit digital precoder can be derived using Singular Value Decomposition (SVD). Though the best analog beamformer for the given channel state could be in principle found using a brute force search, this option is not practical unless the number of transmit antennas and phase shift values per antennas is moderate (which is not the case in Massive MIMO systems). In this paper, a novel Machine Learning (ML) algorithm based on Reinforcement Learning (RL) is proposed, in order to speed up the selection of the analog beamformer. This RL algorithm assess the performance of the candidate solution in each instance of the process, taking advantage of the *experience* that the ML algorithm has gain in the past. It is important to note that the sum data rate that is achievable with the proposed RL algorithm is similar to the one using brute force search, though the iterations that are required are notably less.

The rest of the paper is organized as follows: Section II presents the system model and the details of the hybrid

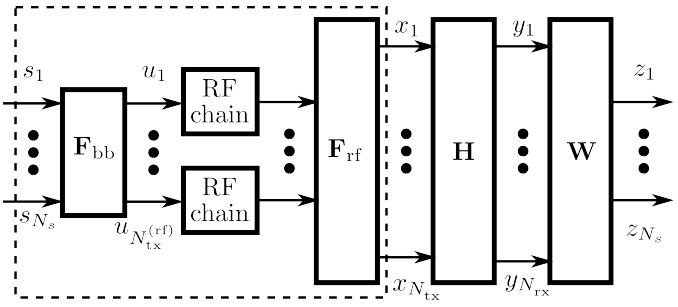

Fig. 1. Hybrid beamforming architecture for a large-scale MIMO system deploying N_{tx} and N_{rx} antennas in the transmitter and receiver, respectively. A codebook-based analog beamformer ($\mathbf{F}_{\mathrm{rf}}$) and a fully-adaptive digital precoder ($\mathbf{F}_{\mathrm{bb}}$) are used in transmission to transport N_{s} symbol streams.

beamforming implementation. Section III introduces the fundamental of RL and derives the algorithm to select the digital precoder and analog beamformer jointly. The simulation setting, as well as the obtained simulation results, are discussed in Section IV. Finally, conclusions are drawn in Section V.

II. System Model

The simplified system model of the proposed large-scale MIMO system with hybrid beamforming is illustrated in Fig. 1, comprising a *digital* precoder and an *analog* beamformer in transmission, represented by matrices $\mathbf{F}_{\mathrm{bb}}$ and $\mathbf{F}_{\mathrm{rf}}$ of size $N_{\mathrm{tx}}^{(\mathrm{rf})} \times N_{\mathrm{s}}$ and $N_{\mathrm{tx}} \times N_{\mathrm{tx}}^{(\mathrm{rf})}$, respectively, and a fully-digital combiner in reception represented by matrix $\mathbf{W}_{\mathrm{bb}}$ of size $N_{\mathrm{s}} \times N_{\mathrm{rx}}$. The MIMO wireless channel between the transmit and receive antennas is described by a complex matrix $\mathbf{H}$ of size $N_{\mathrm{rx}} \times N_{\mathrm{tx}}$, whose coefficients are strongly correlated according the results of the channel measurement campaigns performed in mmWave frequency bands. Moreover, when compared to lower frequency bands, the mean path loss attenuation to be observed is expected to be much stronger.

Though the channel gains that correspond to the different transmit-receive antenna pairs are not completely independent, it is still possible to multiplex $N_{\mathrm{s}} \ll \min\{N_{\mathrm{tx}}, N_{\mathrm{rx}}\}$ parallel data streams provided that the number of singular values of $\mathbf{H}$ that are notably different from zero are at least equal to N_{s}. Though in actual wireless systems the instantaneous values of the channel gains vary continuously in both time and frequency domains, we use a flat block fading channel model to approximate the reality accurately; that is, we assume that the coefficients of $\mathbf{H}$ remain constant during the duration of the transmission time interval, vary independently from time interval to time interval, and show a flat frequency response in the whole communication bandwidth of the mmWave signal.

When *ideal* (non-restricted) transmit precoding and receive combining can be used in both extremes of the link, the coefficients of each of these weighting matrices can be obtained after applying the SVD to the wireless channel matrix, *i.e.*,

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}, \tag{1}$$

where $\mathbf{V}$ and $\mathbf{U}^{\mathrm{H}}$ are unitary matrices whose columns contain the coefficient of the transmit precoding and receive combining vectors, whereas $\mathbf{\Sigma}$ is a diagonal matrix that contains the singular values associated to each data stream. Therefore, when N_{s} data streams are multiplexed in the spatial domain, the optimal transmit precoding matrix is given by

$$\mathbf{F}^{(\mathrm{opt})} = \left[\mathbf{v}_1 \cdots \mathbf{v}_{\mathrm{N_s}}\right], \tag{2}$$

where $\mathbf{v}_n$ is the transmit precoding vectors of size $N_{\mathrm{tx}} \times 1$ that corresponds to the n-th (strongest) singular value σ_n of $\mathbf{H}$.

Unfortunately, the implementation of ideal (non-restricted) precoding schemes becomes impractical as the size of the transmit antenna array grows. This is because a full-digital beamforming architecture requires a separate baseband signal processing blocks per transmit antenna and, at the same time, a dedicated High Power Amplifier (HPA) per RF chain, which impacts negatively on the implementation cost and the energy consumption of the large-scale MIMO system. Therefore, hybrid beamforming architectures should be favored in this situation, with the premise that the hybrid beamforming matrix that results after combining the digital precoding matrix $\mathbf{F}_{\mathrm{bb}}$ with the analog beamforming matrix $\mathbf{F}_{\mathrm{rf}}$ is *similar* to the optimal beamforming matrix presented in (2).

Different approaches have been proposed to define the similarity requirement, which is mathematically stated as

$$\mathbf{F}^{(\mathrm{opt})} \approx \mathbf{F}_{\mathrm{rf}}\,\mathbf{F}_{\mathrm{bb}}. \tag{3}$$

For example, the authors of [8] utilize for this purpose the Frobenius norm of the matrix that results when subtracting the hybrid beamforming matrix from the optimal precoding matrix, which is equivalent to solve the following problem:

$$\{\widehat{\mathbf{F}}_{\mathrm{rf}}, \widehat{\mathbf{F}}_{\mathrm{bb}}\} = \arg \min_{\{\mathbf{F}_{\mathrm{rf}}, \mathbf{F}_{\mathrm{bb}}\}} \|\mathbf{F}^{(\mathrm{opt})} - \mathbf{F}_{\mathrm{rf}}\,\mathbf{F}_{\mathrm{bb}}\|_{\mathrm{F}}, \tag{4}$$

where $\mathbf{F}_{\mathrm{rf}}$ and $\mathbf{F}_{\mathrm{bb}}$ should be selected from the feasibility set of the analog beamforming and digital precoding matrices, respectively. In this paper, the columns of the analog beamforming matrix are codebook-based, such that each of the coefficients that corresponds to the different antennas can only take discrete values from a uniform quantization set [7]. On the other hand, the coefficients of the digital precoding matrix are not restricted to belong to a codebook and, in principle, can take any complex number such that $\|\mathbf{F}_{\mathrm{bb}}\|_{\mathrm{F}} = 1$ is verified. Finally, since $N_{\mathrm{tx}} \gg N_{\mathrm{rx}}$ is verified in most practical large-scale MIMO systems, a full-digital implementation of the receive combining matrix is assumed, and $\mathbf{W} = \mathbf{U}^{\mathrm{H}}$ is used to weight the signals received in each antenna before performing the symbol detection in each independent data stream.

A. Discrete-phase hybrid beamforming

The hybrid beamforming algorithms in proposals as [8], [9] perform pretty well but, in return, require an array of phase shifters that can adjust the phases of the signals in the different transmit antennas continuously. In contrast, the hybrid beamforming algorithm that is presented in this paper assumes that the phase adjustments per transmit antenna can only take two possible values, namely $\theta_{i,j} \in \{-\pi/2, +\pi/2\}$

for $i = 1, \dots N_{\rm tx}$ and $j = 1, \cdots \times N_{\rm tx}^{\rm (rf)}$, keeping the complexity design of the analog transmit beamforming codebook to the simplest. Note that this definition can be extended to other cases without loss of generality, assuming $2^{N_{\rm p}}$ phase levels can be applied in every coefficient of the analog beamformer.

The goal of the proposed hybrid beamforming algorithm is to select the most convenient digital precoder $\mathbf{F}_{\rm bb}$ and analog beamformer $\mathbf{F}_{\rm rf}$, such that the elements of $\mathbf{F}_{\rm rf}$ attain the form

$$f_{i,j} = \frac{\exp\left(\mathrm{j}\,\theta_{i,j}\right)}{\sqrt{N_{\rm tx}}}, \;\; \theta_{i,j} \in \left\{ \frac{2n-1}{2^{N_{\rm p}}} : n = 1, \dots, 2^{N_{\rm p}} \right\}, \quad (5)$$

where $N_{\rm p} = 1$ and $|f_{i,j}| = 1/\sqrt{N_{\rm tx}}$ to prevent changes on the power of the signals at the output of each RF chain. Note that there are $M = 2^{N_{\rm tx} \times N_{\rm tx}^{\rm (rf)} \times N_{\rm p}}$ different elements in the codebook that defines the possible values that the analog beamformer $\mathbf{F}_{\rm rf}$ can take. Since M may grow large even for moderate numbers of transmit antennas and RF chains, we will define a procedure that simplifies the search of the analog beamforming matrix that should be used.

B. Identified equivalent channel

Given a certain $\mathbf{F}_{\rm rf}$ in the system, an equivalent wireless channel $\widetilde{\mathbf{H}}$ of dimension $N_{\rm rx} \times N_{\rm tx}^{\rm (rf)}$ results after combining the actual wireless channel matrix with a given element of the analog beamformer codebook. This statement suppports the development of our algorithm. Then, after applying SVD,

$$\widetilde{\mathbf{H}} = \mathbf{H}\,\mathbf{F}_{\rm rf} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^{\rm H} \quad (6)$$

is obtained. In this situation, since there are no restrictions to define the digital precoding matrix, it is possible to keep the column vectors of $\widetilde{\mathbf{V}}$ that are associated to the $N_{\rm tx}^{\rm (rf)}$ strongest singular values of $\widetilde{\mathbf{H}}$, and make

$$\mathbf{F}_{\rm bb} = \left[\widetilde{\mathbf{v}}_1 \dots \widetilde{\mathbf{v}}_{N_{\rm tx}^{\rm (rf)}} \right] \quad (7)$$

using a similar procedure to the one utilized in (2). The data rate that is achievable with the hybrid beamforming algorithm that is proposed for the given channel state $\mathbf{H}$, when $\mathbf{F}_{\rm bb}$ and $\mathbf{F}_{\rm rf}$ are defined according to (7) and (5), respectively, can be evaluated with the aid of the Shannon's formula, *i.e.*,

$$C = \sum_{n=1}^{N_{\rm s}} \log_2\left(1 + \widetilde{\lambda}_n^2\,\mathrm{SNR}\right) \quad (8)$$

where $\widetilde{\lambda}_n$ is the n-th eigenvalue of the equivalent channel matrix $\widetilde{\mathbf{H}}$, which depend on the singular values derived in (6).

It is important to note that the selection of $\mathbf{F}_{\rm rf}$ affects the achievable data rate of the system notably. Unfortunately, the implementation of a brute-force search to select the analog beamformer that maximizes the achievable data rate of the system is not practical, particularly when the number of codebook elements is large. This is the reason why an alternative RL-based strategy is introduced in the following section.

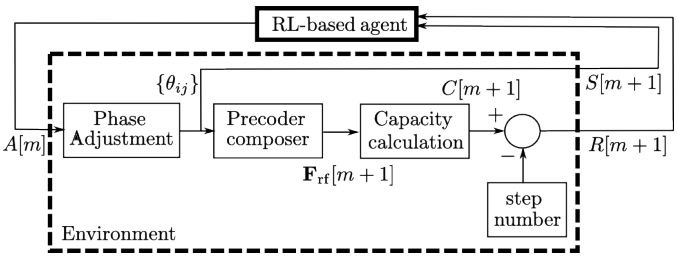

Fig. 2. Illustration of the different environment-agent relationships that were defined for the proposed RL-based hybrid beamforming algorithm.

III. Reinforcement Learning-Aided Analog Precoder Selection

The use of ML tools to re-design the algorithms needed in the different processing blocks of a wireless communication systems is being extensively studied in these days [10]. Keeping in mind this trend, in this paper we focus on the use of RL [11] to implement the hybrid beamforming algorithm, which can identify a suitable candidate matrix for the analog beamformer $\mathbf{F}_{\rm rf}$, verifying the constrains stated in (5) for each coefficient, but taking advantage of the learning from the training that was performed in previous channel states. This way, a performance close to the one achieved with an exhaustive brute-force search can be achieved, avoiding the high computational demand that the latter algorithm requires.

In the following paragraphs, we give an interpretation of the hybrid beamforming algorithm that we aim to design, using for this purpose the terminology and standard elements that are usually utilized in texts regarding RL. The initial working condition of the proposed algorithm depends on the current channel state $\mathbf{H}$, as well as on an initial analog beamformer $\mathbf{F}_{\rm rf}[0]$ that is arbitrarily proposed, which is characterized by the initial set of phases $\left\{\theta_{i,j}[0] : i = 1, \dots N_{\rm tx}; j = 1, \cdots \times N_{\rm tx}^{\rm (rf)}\right\}$. Then, for each iteration m of the algorithm, an entity that in the RL terminology is referred to as *environment*, is affected by an stimulus that is fed into it, providing a response in the following iteration that is characterized by two variables, namely: the *state* $S[m+1]$ and the *reward* $R[m+1]$. The counterpart is a so-called *agent*, which observes the new conditions in the environment and decides the convenient *action* A_m to be taken. In our case, the goal is to update sequentially the coefficients of $\mathbf{F}_{\rm rf}$, trying to keep as low as possible the number of iterations that are needed per new channel state during the operation.

The basic idea behind any RL-based algorithm consists in modifying an input variable and, after that, observe the effect that this action has on the output variable. By mean of this process, new knowledge is gained by applying a simple *trial-error* approach in a systematic way. Figure 2 shows the most important blocks of the proposed algorithm, combining the well-known concepts of RL with the more specific details of a hybrid beamforming scenario. It is important to highlight that this figure illustrates a perspective that has been seldom exploited to solve such kind of optimization problems in the

area of wireless communications. The proposed algorithm can also be interpreted as a systematic way to define a trajectory across iterations with index m, visiting only a subset of states from all the ones that exist. Since the evolution of the environment is directly affected by the current channel state, which in turn can be statistically modelled by its physical properties, it is highlighted the challenging task that would imply to efficiently include this model within an optimization algorithm. Fortunately, the proposed RL-based strategy avoids completely this necessity, and as it is well-known, spontaneously exploits the so-called *policies*, which define the actions to be taken in every condition seen, with an inherent internal representation of the environment that occupies the place of some kind of virtual model. This model is not known *a priori*, but it is rather automatically learned by the algorithm.

A. Actions and Rewards

The design of any RL-based algorithm starts with the definition of its actions and rewards. The actions are elementary stimulus that the agent feeds into the environment under study to observe variations on its state. In our case, the actions are applied to the phases $\theta_{i,j}$ defined in (5), which affect the RF signals that come out from the RF chains $j = 1, \ldots, N_{\text{tx}}^{(\text{rf})}$ and are feed into each transmit antenna $i = 1, \ldots, N_{\text{tx}}$.

Here, an action at iteration m comprises two possibilities, namely:

- *(i)* Select the indexes i and j and, after that, increment the phase in $\theta_{i,j}$ in $\pi/2^{N_p-1}$.
- *(ii)* Select the indexes i and j and, after that, reduce the phase in $\theta_{i,j}$ in $\pi/2^{N_p-1}$.

With the modification of any of the phases stored in $\theta_{i,j}$, a new candidate precoder results, which is denoted as $\mathbf{F}_{\text{rf}}[m+1]$.

Then, the achievable data rate is estimated according to (8), which depends on both the proposed analog beamformer and the current channel state. In addition, it is also necessary to define a penalization metric that monitors the number of iterations that are utilized to converge to a solution. In this paper, to put together the indicated concepts, the reward is proposed to be

$$R[m+1] = C\big(\mathbf{F}_{\text{rf}}[m+1]\big) - m \qquad m = 0, 1, \ldots \tag{9}$$

where the iteration index m is added as a negative term. Based on this, the proposed algorithm will inherently aim to maximize the reward and, while doing so, it will identify the precoders that maximize the achievable data rate. Note that this definitions are also considered in Fig. 2.

B. Handling of Success Conditions

In this paper, we follow an episodic treatment for the problem that we aim to solve [11], which means that the proposed algorithm iterates until the given episode is completed. Moreover, in each episode, the internal state of the algorithm is modified accordingly, such that each episode has a specific trajectory associated to it, as described in the previous section.

From the perspective of a RL-based algorithm, it is important to define the conditions that should be fulfilled in order to declare that the goal of our specific task has been achieved. Note that in our case, the intended task consists in identifying the most convenient analog beamformer for the current channel state. Then, both training and regular phases of the RL-based algorithm should know the specific conditions to claim that the ML processing in the given episode is over. From this definition arises the concept of success. Specifically, an episode can be declared as completed when: *(i)* A predefined maximum number of iterations have been performed, this entails the concept of not having achieved success; *(ii)* The agent verifies that the maximum achievable data rate observed up to a certain iteration m coincides with the one specified as target, which is indicated with C^*, this is the success condition. Meanwhile, the definition of C^* needs also to consider different channel state observations.

Let us assume that C^* is ideally given by the maximum data rate that can be achieved for the current channel state, given the restrictions that are defined to construct the analog beamformer in (5). Then, C^* can be found after a brute-force search, assessing the achievable data rate of the current channel when using each possible analog beamformer candidate in transmission. Since such process is computationally demanding, we formulate an alternative procedure to determine C^* in an efficient way, which can also effectively support the state changes that the channel experiences during the operation of the system. The proposed procedure can be summarized as follows. In the training stage, C^* is initialized with a value arbitrarily close to zero. Then, after each iteration, the observed reward $R[m+1]$ is used to check if the success condition has been achieved or not. This idea is implemented with a simple comparison, the observed achievable data rate is written as $C[m+1] = R[m+1] + m$, then success is dictated by the result $C[m+1] \geq C^*$. The observed data rate is also used to decide if to update the value in C^*. A special threshold value is defined as β_1^{rl}, in those cases in which $C[m+1] > (1+\beta^{\text{rl}})C^*$ is verified, the value of C^* should be updated accordingly. We note that this strategy works very well, requiring only few episodes to move C^* from its initial value to another one that is very close to the average data rate for the given channel state, as concluded after exhaustive simulations. It should also be noted that during the training phase, whenever a different channel state is successively fed into the algorithm, the variable C^* is reset again to the selected arbitrarily low value. The set of episodes devoted to train the algorithm using the same channel state define different *epochs* (due to that, every epoch is accompanied by a reset on the value of C^* in its beginning).

After the training process is over, the assessment stage is started to emulate the regular operation of the RL-based algorithm. Note that during the the assessment stage, each new channel state is associated to only one episode of the proposed algorithm. The value in β_1^{rl} is empirically increased in this case. Morever, C^* is only reset in the beginning of this stage.

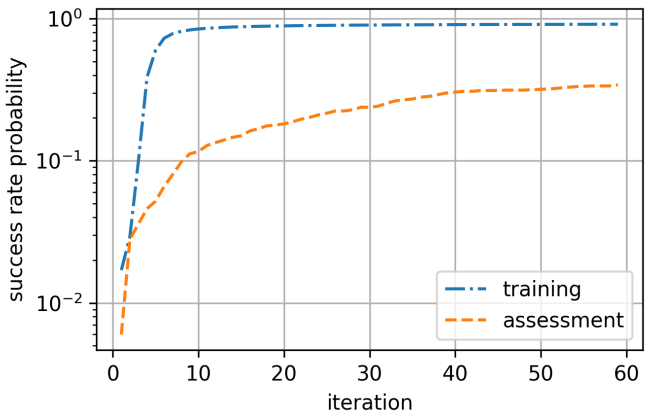

Fig. 3. Probability of visiting the algorithm state that defines the candidate $\mathbf{F}_{\mathrm{rf}}$ which achieves the maximum capacity identified for a certain channel state observation, i.e. C^* (stated as success condition), at iteration m.

C. Implementation of the Q-learning algorithm

From all the algorithmic options that exist to implement RL, in this work we utilize *Q-learning*. This way, we are able to build a proof-of-concept that makes use of a simplified tabular implementation, which avoids the use of more advanced resources, such as neural networks. Thanks to this approach, it will be possible to show the natural effectiveness of RL, which can be strengthen with the introduction of techniques borrowed from the theory of *deep learning*.

The tabular storage of the information that is performed with Q-learning tries to determine the most convenient actions that the algorithm must execute at a given state, to select an adequate analog beamformer for the given channel observation, in turn next state is determined in each iteration according to the selected action. This results in the use of a tabular (or matrix) variable frequently indicated by $Q(s,a)$ [11], where s refers to possible states of the environment and a refers to possible actions that can be taken, this matrix is updated in each iteration m. Thus, the outcome of the whole training process consists of the generation of an information table which, later on, will be utilized in the regular operation phase to perform the decisions. In other works, while in regular operation, the algorithm has to mainly give an interpretation to its status based on that pre-stored information that it has; it implies to update $\mathbf{F}_{\mathrm{rf}}$ during few iterations and finally, when the success (or end) conditions are reached, provide the final output to the system.

IV. Simulation Scenario and Analysis of Results

We first present the specific details of the mmWave channel model that has been used to obtain the simulation results that are reported in this section. For this purpose, we follow the definitions adopted in [4], [5], which are based on the widely-accepted extended Saleh-Valenzuela geometric channel model [12]. In this model, the channel matrix is given by

$$\mathbf{H}(t) = \gamma \sum_{i=1}^{N_{\mathrm{cl}}} \sum_{j=1}^{N_{\mathrm{ray}}} \alpha_{i,j} \Lambda_{\mathrm{r}}(\phi^{\mathrm{r}}_{i,j}, \theta^{\mathrm{r}}_{i,j}) \Lambda_{\mathrm{t}}(\phi^{\mathrm{t}}_{i,j}, \theta^{\mathrm{t}}_{i,j}) \mathbf{a}_{\mathrm{r}}(\phi^{\mathrm{r}}_{i,j}, \theta^{\mathrm{r}}_{i,j}) \mathbf{a}^{\mathrm{H}}_{\mathrm{t}}(\phi^{\mathrm{t}}_{i,j}, \theta^{\mathrm{t}}_{i,j}), \quad (10)$$

where $N_{\mathrm{cl}} = 8$ and $N_{\mathrm{ray}} = 10$ for our specific simulation setting, γ is a normalization factor, and coefficients $\alpha_{i,j} \sim \mathcal{CN}(0,1)$ denote a complex gain. For each propagation path j associated with the i-th cluster, the azimuth angles of arrival and departure are represented $\phi^{\mathrm{r}}_{i,j}$ and $\phi^{\mathrm{t}}_{i,j}$, respectively, whereas the elevation angles of arrival and departure are represented by $\theta^{\mathrm{r}}_{i,j}$ and $\theta^{\mathrm{t}}_{i,j}$. These angles are modeled as Laplacian random variables with an angle deviation of $7.5°$, centered at an uniformly distributed mean cluster angle of $0°$ and $90°$ for azimuth and elevation, respectively. Finally, $\mathbf{a}_{\mathrm{r}}(\phi^{\mathrm{r}}_{i,j}, \theta^{\mathrm{r}}_{i,j})$ ($\mathbf{a}_{\mathrm{t}}(\phi^{\mathrm{t}}_{i,j}, \theta^{\mathrm{t}}_{i,j})$) and $\Lambda_{\mathrm{r}}(\phi^{\mathrm{r}}_{i,j}, \theta^{\mathrm{r}}_{i,j})$ ($\Lambda_{\mathrm{t}}(\phi^{\mathrm{t}}_{i,j}, \theta^{\mathrm{t}}_{i,j})$) represent the normalized planar array response and antenna element gain at the receiver (transmitter) side, respectively, for all rays indexes j and cluster indexes i, assuming an inter-element spacing of half-wavelength. In this paper, the same uniform planar array model described in [4] is also considered.

Based on this channel model, numerical simulations were carried out to analyze the performance of the proposed RL-based algorithm for hybrid beamforming, using the following additional settings. The number of transmit antennas is $N_{\mathrm{tx}} = 9$, while the number of RF chains in transmitter is $N^{(\mathrm{rf})}_{\mathrm{tx}} = 2$, then the number of receive antennas $N_{\mathrm{rx}} = 4$ was used and the number of spatial streams was defined as $N_{\mathrm{s}} = 2$. Phase levels were limited to $2^{N_{\mathrm{p}}} = 2$. Furthermore, the Q-learning algorithm was adjusted with a learn rate $\alpha^{\mathrm{rl}} = 0.98$, and a discount $\gamma^{\mathrm{rl}} = 0.9$. Additionally, an ε-greedy policy was used with parameters $\varepsilon^{\mathrm{rl}}_{\mathrm{max}} = 0.5$ and $\varepsilon^{\mathrm{rl}}_{\mathrm{min}} = 0.1$. The exploration rate in each episode was decreased by means of the factor 0.98. The value in a certain element of $Q(s,a)$ is updated with a *bonus* positive term set in 1300 when success is achieved, as commonly done in practical implementations of the algorithm. This bonus is added to the reward expressed in (9). It is also implemented a partial bonus set in 130 when the agent achieves the capacity $C[m+1] > (1-\beta^{\mathrm{rl}}_2)C^*$. Then used values were $\beta^{\mathrm{rl}}_1 = 0.02$, $\beta^{\mathrm{rl}}_2 = 0.02$. To state the reward as in (9), SNR is supposed to be 15 dB. These settings defined an initial phase devoted to train the algorithm while a convenient content of the tabular representation of $Q(s,a)$ is pursued. First, 10000 episodes with a maximum of 60 iterations were simulated to define an initial adjustment of the algorithm, trying to facilitate the convergence. A single channel observation was used in this stage. Secondly, 40000 new episodes were run upon the results of the first training. In this stage, 40 channel state observations were used. The aim in this stage was to bring the algorithm the possibility of learn common charactertistics between different channel samples. In this way, 1000 episodes where run with each channel state before replacing the channel state embedded in the environment. Figure 3 presents a measurement of the probability of achieving a state of the algorithm that in

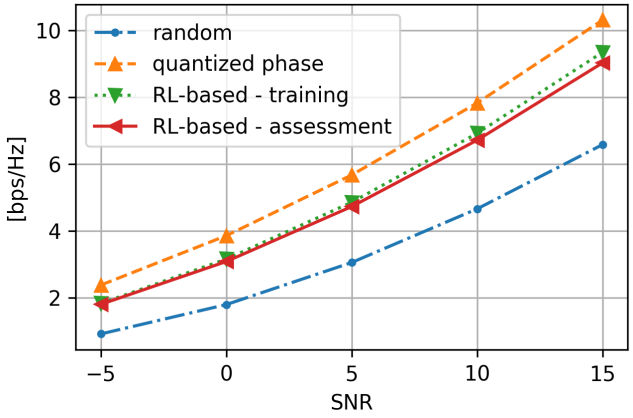

Fig. 4. Achievable data rate as function of the mean SNR for different analog beamformer selection methods. The mmWave channel gains were generated according to the extended Saleh-Valenzuela geometric channel model.

turn defines a certain precoder solution $\mathbf{F}_{\mathrm{rf}}$ which provides the target achievable capacity C^*. Since this condition is considered as success, it is observed a maximum success rate of roughly 92%, and this state is achieved with roughly 10 iterations. It is very interesting to compare this amount of iterations with the cardinality of the candidates set, which would (completely) be evaluated by the brute force search. According to our setting, this value is approximately $250\,.\,10^3$ and an equal number of iterations would be required to follow that principle. The reduced number of iterations that our algorithm employs implies an important efficiency (it is given by the rate $10/(250\,.\,10^3)$).

Later, an assessment stage is performed, where some adjustments are done. The purpose of this stage of simulation is to represent the expected regular operation of the proposed beamforming update algorithm. Values $\beta_1^{\mathrm{rl}} = 2$ and $\beta_2^{\mathrm{rl}} = 0.02$ are set, and $\varepsilon = 0.05$ is constantly defined. Then, 500 channel samples were used while they were extracted from a set not used during former stages of training. To keep complexity as low as possible only one episode with 60 iterations is assigned to every channel state observation. In this case, results presented in Fig. 3 show that the probability of achieving C^* is appoximately 35%.

The results previously indicated suggest an acceptable operation to the algorithm from the perspective of the RL algorithm. However, taking now into consideration the purpose of providing an efficient hybrid precoding adjustment algorithm, we analyze the observed capacity for every channel state and then calculate an average. The obtained curves are presented in Fig. 4. The analog beamformer definition is evaluated for the special case where phases are randomly chosen, this case is used to state lower bound in Fig. 4. Also the brute force search was taken into account to analyze the natural effect of having discrete-phases. This scenario defined an upper bound to the performance. Then the results for the RL-based strategy were also plotted. Note that performance achieved is high, and also note that the assessment stage shows a behavior closer to that achieved in training. This curves show the effectiveness of the proposed strategy. It is highlighted that success rate evaluated in Fig. 3 is an important metric, but according to Fig. 4 it is interpreted that even in episodes where the system does not achieves C^* a convenient candidate $\mathbf{F}_{\mathrm{rf}}$ is equally given by our algorithm.

V. Conclusions

In this paper, a novel approach to design a hybrid beamforming algorithm that is suitable for a large scale MIMO system on mmWave frequency bands has been presented. In order to simplify the implementation of the analog beamformer, the use of discrete phase steps has been introduced. Then, after selecting the analog beamformer candidate, an equivalent wireless channel was determined to apply SVD and identify the digital precoder that should be utilized in transmission. Thanks to the use of a codebook for the analog beamformer, a RL-based algorithm has been derived, which enables to select the most convenient element in transmission using the experience that has been gained in the past. The obtained performance results showed that most of the data rate that brute-force search provides can be reached using our proposed RL-based approach, requiring only a fraction of the iterations that the brute-force needs to reach a solution.

References

[1] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[2] X. Gao, L. Dai, S. Han, C.-L. I, and R. Heath, "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[3] A. Dowhuszko and J. Hämäläinen, "Performance of transmit beamforming codebooks with separate amplitude and phase quantization," *IEEE Signal Process. Letters*, vol. 22, no. 7, pp. 813–817, July 2015.

[4] C. Chen, "An iterative hybrid transceiver design algorithm for millimeter wave MIMO systems," *IEEE Wireless Commun. Letters*, vol. 4, no. 3, pp. 285–288, June 2015.

[5] O. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[6] N. Moghadam, G. Fodor, M. Bengtsson, and D. Love, "On the energy efficiency of MIMO hybrid beamforming for millimeter-wave systems with nonlinear power amplifiers," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7208–7221, Nov. 2018.

[7] A. Dowhuszko, G. Corral-Briones, J. Hämäläinen, and R. Wichman, "Performance of quantized random beamforming in delay-tolerant machine-type communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5664–5680, Aug. 2016.

[8] S. Buzzi, C. D'Andrea, T. Foggi, A. Ugolini, and G. Colavolpe, "Single-carrier modulation versus OFDM for millimeter-wave wireless MIMO," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1335–1348, Mar. 2018.

[9] H. Ghauch, T. Kim, M. Bengtsson, and M. Skoglund, "Subspace estimation and decomposition for large millimeter-wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 528–542, Apr. 2016.

[10] T. Wang, C. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Communications*, vol. 14, no. 11, pp. 92–111, Nov. 2017.

[11] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[12] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Comm.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.