Improve your aim: a Deep Reinforcement Learning approach for 5G NR mmWave beam refinement

Mauro Belgiovine, Kaushik Chowdhury
Institute for the Wireless Internet of Things at Northeastern University, Boston
belgiovine.m@northeastern.edu, krc@ece.neu.edu

***Abstract*—Massive MIMO (mMIMO) technology is considered as a key enabler for 5G and beyond cellular networks, which allows formation of highly directional radiation beams in the millimeter-wave (mmWave) band. Specifically, considering the 5G new radio (NR) standard, a codebook-based approach is used that allows setting the antenna weights, so that both transmission and reception can be achieved in the desired angle. However, when a fixed codebook is used, these angular directions may not be exactly aligned along the optimal path that maximizes the SINR between the transmitter-receiver pair, depending on the granularity of the beam and the codebook size. To address these issues, we propose selection of the analog parameters of the transceiver chain through Deep Reinforcement Learning (DRL). Simulation results show that our approach allows fine-grained beam refinement to the coarse initial estimates of Angle-of-Arrival and Angle-of-Departures in mmWave Frequency Range 2 (FR2) for the 5G NR standard obtained during the a reduced initial beam establishment procedure (P-1). We observe our approach consistently improves the Reference Signal Received Power (RSRP) perceived at the UE side up to 15% while allowing a reduction in the number of Synchronization Signal Blocks (SSBs) up to a factor of $\times 64$ compared to the equivalent number used in P-1 to obtain comparable steering accuracy. Finally, once the trained DRL agent is implemented, it eliminates 100% of control signals needed for the beam refinement procedures, namely P-2 for transmitter beam refinement and P-3 for receiver.**

***Index Terms*—Deep Reinforcement Learning, Next Generation Wireless.**

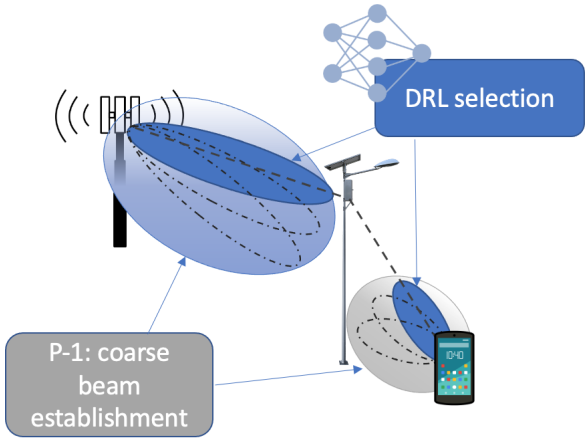

Fig. 1: Overview of proposed architecture. An initial coarse beam selection is performed through the 5G NR standards-defined P-1 procedure. Then, a beam refinement procedure is carried by a DRL agent that adjusts the beam angle with a correction term based on CSI and prior experience.

I. Introduction

FUTURE cellular networks will transform our everyday life, providing large data bandwidths to allow for virtual and augmented reality, Internet of Everything, tele-healthcare and many other applications. 5G New Radio (NR) and beyond envisions the use of massive Multiple-Input Multiple-Output (mMIMO) antenna systems and higher frequency bands, such as millimiter-wave (mmWave) and TeraHertz (THz) carrier frequencies, to meet ever-growing demands for data traffic. These components will allow to (I) improve spatial diversity, (II) transmit data over larger bandwidths and (III) reduce interference among users using beamforming, pushing further the achievable spectral efficiency and data rates for cellular networks.

Although 5G NR [1] explicitly defines the use of mmWave (24.25 GHz to 52.6 GHz) in the dedicated Frequency Range 2 (FR2) and its physical layer radio access procedures, practical implementation and deployment of such transceivers is still subject of study. Phased antenna arrays to enable beamforming, which focuses the radiated energy in the direction of the optimal path to the receiver based on channel characteristics, have become a critical component at higher frequencies that experience increased path loss and attenuation. However, due to the complexity of hardware design at higher frequencies, hybrid analog-and-digital hardware architectures [2] have been widely accepted as a candidate solution for deploying large antenna arrays. Such hybrid architectures typically have a large number of antenna elements connected to a much smaller set of RF-chains, due to narrower antenna spacing at shorter wavelengths, and they are generally more cost-effective. Such hardware necessitates a type of beamforming technique called Hybrid BeamForming (HBF), that requires the joint design of analog and digital beam weights for precoders and combiners in order to form optimal radiation patterns. Specifically, digital signal processing is employed in baseband to reduce interference, while discrete phase shifters are used in the analog (or RF) domain to steer the radiation beams. In general, HBF aims to find a set of hybrid weights that is as "close" as possible to the optimal fully-digital solution (i.e. considering equal number of antennas and RF-chains), which is solved using non-convex optimization techniques subject to several energy and solution space constraints. This is known to be a NP-hard problem [3]. On the other hand, fully-analog techniques rely on *codebook* designs that aim to use a discrete set of *codewords*, each composed of pre-defined analog beam weights that map to a beam pointing towards a specific angle. Such techniques can be adopted if the main goal

978-1-5386-7462-8/23/$31.00 ©2023 IEEE

is spatial diversity rather than spatial multiplexing, although their simplicity requires a trade off between overhead of communication and codebook angular resolution, depending on the network requirements, that limits the achievable network performance. Reinforcement Learning (RL) [4] and Deep Reinforcement Learning (DRL) have shown remarkable results in learning optimal strategies for many challenging problems and environments with complex observations by employing Deep Neural Networks (DNNs) trained as policy functions, allowing to extract complex features from observations gathered from the environment and to map extremely large sets of states to possible actions.

In this work, we focus on the enhancement of the analog component of HBF devices through beam refinement in order to improve spatial diversity in the system. Our contributions are as follows:

- We propose to train a DRL agent capable of applying a correction term on discrete beam angles obtained with an initial *coarse* codebook-based beam establishment phase, based on 3GPP 5G NR standard procedures.
- We tailor our models to a specific location, learning an optimal beam refinement policy that adapts to the spatial features of the environment under study.
- By deploying independent agents on each network device, we aim to improve the alignment of radiation beams at both communication ends, while avoiding large beam sweeping overhead and removing the need of explicit beam refinement procedures.

Our approach tackles the objective of allowing faster and more accurate directional transmissions in FR2, enabling better network bandwidth utilization and higher throughputs for V2X applications involving highly mobile scenarios and large data rates, such as remote mobile surgeries and other types of remote emergency response.

II. Related work

A. Signal processing approaches

Prior work on HBF for computing the analog and digital precoders and combiners can be divided into two main categories:

- *codebook based* (or *constrained weights*): these methods stem from fully-analog beamforming techniques, where analog weights are selected from a set of pre-defined *codewords* that form a so-called *codebook*. Such methods require *beam sweeping* and rely only on received signal power measurements if spatial multiplexing is not required.
- *non-codebook based* (or *unconstrained weights*): these methods do not assume a pre-defined set of codes to set the analog components of hybrid precoders and combiners. They require Channel State Information (CSI) acquisition to allow spatial multiplexing and fine grained beam alignment.

For *codebook based* methods, beam configurations need to be tested by *sweeping* through all possible combinations of precoders and combiners in the analog codebook. After this procedure, a pair of codes is selected in order to maximize a certain performance metric, such as Reference Signal Received Power (RSRP) or spectral efficiency. Specifically, 5G NR in FR2 adopts a two stage codebook-based approach to first acquire an initial coarse beam configuration in Procedure 1 (P-1) and then refine it with Procedure 2 (P-2) and Procedure 3 (P-3), for transmit and receive beams respectively. Note that P-2 and P-3 can be considered a special case of P-1: instead of sweeping through all the combinations, the UE (in P-2) or gNB (in P-3) measures an additional set of reference signal resources, configured on transmitter beams with finer angular resolution, using a fixed receiver configuration obtained from P-1 (or P-2 for gNB). Although computationally simple, such techniques require the transmission of multiple control blocks, i.e. Synchronization Signal Blocks (SSB) for P-1, Non-Zero-Power CSI-RS (NZP-CSI-RS) for P-2 and Sounding Reference Signals (SRS) for P-3, in order to test all the beam configurations across the three procedures, adding a substantial overhead. Moreover, these methods are well suited only for purely analog beamforming applications: if spatial multiplexing is required, accurate CSI needs to be retrieved in order to (I) compute the channel rank and (II) derive optimal precoders and combiners through Singular Value Decomposition (SVD), which will serve as starting point to compute digital weights for hybrid beamforming. Finally, the codebook size must be chosen carefully, in order to obtain a reasonable angle resolution to achieve accurate beam steering while minimizing the control messages exchanged during the beam establishment phase.

On the other hand, for *non-codebook based* approaches, optimization-based methods exist and usually lead to near-optimal solutions, such as AltMin based [3] ones (e.g. Manifold Optimization, Phase Extraction). Although they require optimizing spatial filters whenever new precoders and combiners need to be designed. The cost of optimizing hybrid weights on-line is computationally expensive in practice when considering highly mobile scenarios. Another commonly adopted schema to design unconstrained hybrid beamforming weights is based on Orthogonal Matching Pursuit (OMP) [5]. The OMP partitioning algorithm exploits the sparsity assumption that characterizes mmWave wireless channels; but requires perfect CSI acquisition and perfect knowledge of the Angle-of-Arrival (AoA) and Angle-of-Departure (AoD) of every ray. Thus, a practical OMP deployment requires a wide set of possible codewords from which it would choose those that present the highest projection value in the direction of the optimal weights.

B. Learning based approaches

To overcome limitations of traditional approaches, Deep Learning (DL) based approaches for HBF have been proposed to lower the overhead introduced by the beam sweeping phase by predicting the correct code to be applied on the analog component of the system. Authors in [6] aim to reduce the number of beam training control messages by including compressive channel sensing techniques for the design of their neural network model, but their hybrid beamforming solution is still based on constrained weights, which ultimately limit the achievable spectral efficiency. Other works focus on mapping the channel matrix to unconstrained beam weights [7], but such methods are limited in practical implementation due to

the high number of parameters, as the solution search space grows extremely large for devices with large antenna arrays and multiple RF-chains. Specifically in the context of DRL applied to hybrid beam selection, authors in [8] consider a codebook-based selection scheme for high mobility scenarios that rely on a partially observed set of beam combinations and online re-training of the DRL agent, whose convergence cannot be guaranteed for newly observed instances of the problem. In [9], two deep Q-learning based DRL agents are deployed to select first the number of active antennas and then a codeword from a predefined set of codebooks, based only on SNR level observations. This method also performs online training of the agents, which is not guaranteed to converge within the channel coherence time, and it is still limited by the codebook performance which, in this case, depends on both angular resolution of codewords and number of active antennas selected by the algorithm.

III. System model

For our experiments we consider a 5G NR network operating in the mmWave band FM2 with center frequency $f_c = 28$ GHz. The gNB transmitter is equipped with N_t antennas that are used to communicate through N_s data streams with one or more UE receivers equipped with N_r receiver antennas each. Both devices are assumed to mount Uniform Linear Arrays (ULA). To enable spatial multiplexing, the transmitter is equipped with N_t^{RF} RF-chains $N_s \leq N_t^{RF} \ll N_t$ and similarly the receiver has N_r^{RF} RF-chains such that $N_s \leq N_r^{RF} \ll N_r$. Such architecture permits to transmit N_s independent streams by applying a $\mathbf{F}_{BB}$ baseband precoder with size $N_t^{RF} \times N_s$ followed by a $\mathbf{F}_{RF}$ analog precoder with size $N_t \times N_t^{RF}$. Similarly, the receiver uses a hybrid combiner composed by $\mathbf{W}_{BB}$ baseband combiner with size $N_r^{RF} \times N_s$ and a $\mathbf{W}_{RF}$ analog combiner of size $N_r \times N_r^{RF}$. The analog precoders and combiners are implemented in the analog domain using phase shifters, thus all elements of $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$ have equal norm. Furthermore, the total power constraint is enforced by normalizing the baseband precoder $\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2 = N_s$. For a generic $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ channel, the received hybrid beamformed signal is modeled as follows:

$$\mathbf{y} = \sqrt{\rho}\mathbf{W}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{W}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{n} \quad (1)$$

where ρ is the average received power, $\mathbf{s}$ is the $N_s \times 1$ discrete-time transmitted symbol vector such that $\mathbb{E}[\mathbf{s}\mathbf{s}^{\mathrm{H}}] = \frac{1}{N_s}\mathbf{I}_{N_s}$ and $\mathbf{n}$ is a noise vector with i.i.d. samples from a normal distribution $\mathcal{CN}\left(0, \sigma_n^2\right)$. In this work, we focus on enhancing the analog component of the hybrid beamforming schema, as this component plays the most important role when evaluating beam steering accuracy. Therefore, for the sake of simplicity and without loss of generality, we assume $N_t^{RF} = N_r^{RF} = N_s = 1$ for the rest of this work, focusing on increasing spatial diversity and leaving the subject of spatial multiplexing to future research.

We consider a simulation based on multipath scattering MIMO propagation channels [2] in which radiated signals from a transmitting array are reflected from multiple scatterers back toward a receiving array. To generate a scenario, we randomly sample N_{scat} scatterers coordinates and complex coefficients, that will remain fixed during training and testing phases, while the user is randomly moving around a finite area defined in the 2D space. Note that, while we chose a single possible configuration of the environment, this model can be trained with virtually any scatterers configuration. Fig. 2 represents the scenario considered for the results presented in this work.

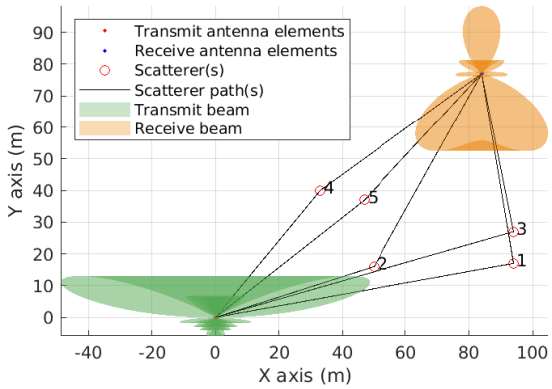

Fig. 2: Training scenario with $N_{scat} = 5$, random UE position and beam computation based on 5G NR P-1.

IV. Proposed DRL approach for 5G NR beam refinement

Our proposed approach starts from an initial *coarse codebook* beam sweeping procedure, employing a burst of Synchronization Signal Blocks (SSBs) similar to that of the standard P-1 procedure, but using only a fraction of the full beam codebook. Once the UE has computed the RSRPs for each code combination, it reports this information back to the gNB. We refer to the matrix $\mathbf{P}$ composed by all the RSRPs for each codeword combination as *power pattern*. Both gNB and UE then use the power pattern $\mathbf{P}$ as input to their respective DRL agents, in order to perform a *single corrective action* aiming to adjust the analog beamforming weights (i.e. $\mathbf{F}_{RF}$ for transmitter and $\mathbf{W}_{RF}$ for receiver) so that they point to the most promising direction within the range of the correction term, based on the experience gained during the training of the agents. The motivation behind this approach is that beam widths at mmWave, due to shorter wavelengths and higher numbers of antennas, result in narrower beams. Even a slight angle adjustment pointing in the right direction can potentially improve the achievable network capacity. To achieve this goal, we train a DRL agent on each device that takes an observation as input consisting of:

- The power pattern matrix $\mathbf{P}$, consisting of the normalized Reference Signal Received Power (RSRP) values measured at the receiver side during the coarse P-1 procedure;
- The *current beam angle* corresponding to the best transmitter (or receiver) codeword, defined between -90 and +90 degrees;

Based on the observation input from the environment, the DRL agent outputs a corrective steering action on the beam angle identified during the coarse beam establishment phase. Since we want this correction term to be applied as quickly

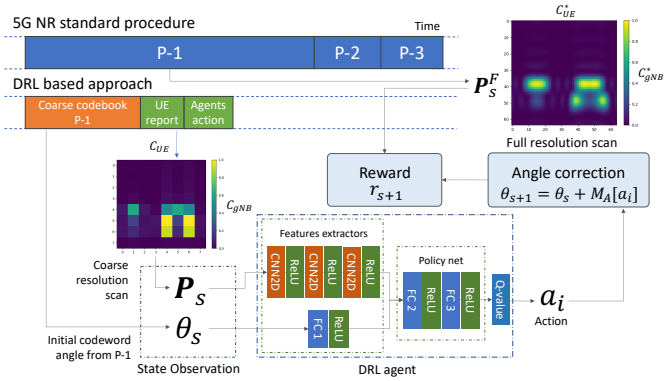

Fig. 3: Overview of proposed DRL agent for 5G NR beam refinenemt in mmWave. Diagram shows processing steps and principle of operation, compared to standard 5G NR hierarchical beam sweeping approach.

as possible after the coarse scanning procedure, we force our model to use only a single action per episode, making the steering correction decision within a single query to the DRL policy. The objective of this approach is to increase the perceived signal power at the receiver without additional beam training procedures or costly algorithms relying on CSI estimation. Fig. 3 shows an overview of the proposed solution.

V. Simulation and DRL Agent Design

A. Simulation parameters

We consider a geometric scattering MIMO wireless channel used by gNB and UE to communicate. The channel consists of $N_{scat} = 5$ scatterers positioned in some fixed location, representing a possible scatterer configuration in a real world setting. gNB is positioned on a 2D surface at xy-coordinates $[0, 0]$ (expressed in meters) and at a height of 50, meanwhile the UE is randomly moving in an area of 75×75 m in a coordinates range of $[25, 100]$ m for both xy-axis. We simulate a downlink (DL) transmission for a gNB equipped with $N_t = 8$ and an UE $N_r = 4$, equipped with $N_t^{RF} = N_r^{RF} = 1$ RF-chains, both configured as ULA.

According to 3GPP specifications [1], the full resolution 5G NR P-1 codebook in FR2 is set to a size of $C^*_{gNB} = C^*_{UE} = 64$[1] at both transmitter and receiver sides. The codebooks are configured using the *steering vector* function, defined as follows for each beam angle θ:

$$a(\theta) = \left[\ 1, e^{-j\pi \sin\theta}, e^{-j2\pi \sin\theta}, \ldots, e^{-j(M-1)\pi \sin\theta}\ \right] \quad (2)$$

At gNB side, beam angles span over a range of $[-60, 60]$ degrees, while at the UE side the range is set between $[-180, 180]$ degrees. The angle codewords are chosen by picking C^*_{gNB} equi-distant angles in the steering range defined for the gNB and, similarly, C^*_{UE} equi-distant angles chosen for the UE. After the standard beam sweeping procedure P-1, the full resolution scan $\mathbf{P}^F \in \mathbb{R}^{C^*_{gNB} \times C^*_{UE}}$ will present the RSRP value measured at the receiver side for each of the possible code combinations.

[1]In FR1 the codebook size is either $C = 4$ or $C = 8$.

Once full resolution scan $\mathbf{P}^F$ is obtained, the coarse resolution scan $\mathbf{P}$ matrix can be produced during simulation by simply selecting a fraction of the RSRP values obtained in $\mathbf{P}^F$ from both codebooks. For example, if we consider a fraction of $1/8$ codes from the full resolution codebook, we will obtain a coarse codebook of size $C_{gNB} = C_{UE} = 8$ by picking 1 code in every 8 consecutive codes, leading to a $\times 64$ reduction in codebook size/resolution. While the coarsely defined codebook will benefit the total SSB signaling overhead time, since $C_{gNB} \times C_{UE} \ll C^*_{gNB} \times C^*_{UE}$, this will inevitably induce a loss of performance from the network capcity perspective, since the narrow beams at mmWave are more sensitive to steering errors from the underlying full resolution scan and will limit the beam alignment accuracy on the most promising channel directions.

B. Agent training strategy

In order to reduce the overhead while retaining greater beam steering accuracy, we propose to train a DRL agent that, given an initial beam angle estimate from a coarsely defined beam sweeping procedure, is able to apply an angle correction term to adjust the beam so that it points to what the agent might consider a more promising direction, based on its prior experience. More specifically, traces of prior experiences can be gathered through a set of full resolution scans performed in the current channel configuration (defined by the scatterers' positions and UE location). Training an agent with this data will produce an agent policy closely tailored to the specific area covered in the experience traces, policy that could be fine-tuned over time with additional experience gained after deployment.

The proposed DRL algorithm chosen for this problem is Deep Q-Learning (DQN) with experience replay memory and dueling policy networks [10]. In this approach, an agent is trained in order pick an action from a discrete set of possible actions $\mathcal{A}$. We define each action as a corrective term to be applied on the beam angles/codewords selected during procedure P-1. The agent output is the Q-value relative to each possible correction term, which is a value expressing the utility of each action in order to maximize the return, given the current observation.

We configure the DRL algorithm as follows:

1) Agent observations: In the DRL environment, we define an observation at a given environment state s as a tuple containing:

- $\mathbf{P}_s \in \mathbb{R}^{C_{gNB} \times C_{UE}}$: normalized power pattern corresponding to RSRP values computing during the coarse 5G NR P-1 procedure;
- θ_s: current beam angle in degrees associated to the transmitter/receiver codeword selected after coarse 5G NR P-1 procedure.

The idea is that the agent will observe the coarse resolution scan, along with the angle in degrees where the coarse beam is pointing to, and uses this information to perform a corrective action that will advance the current state s of the environment to $s + 1$. See Fig. 3 for an example of observation from the coarse beam sweeping procedure.

2) Agent actions: Each corrective action $a_i \in \mathcal{A}$, where $\mathcal{A}$ is the set of possible actions, is expressed in degrees in the

form of a signed integer k, indicating a positive or negative steering action by $\pm kd$ degrees, where d corresponds to the degree of steering and its value depends the number of codes and the angular range defined while generating the codebook. For example, for a gNB full resolution codebook of size $C_{gNB} = 64$ with equally spaced beams defined in $[-60, 60]$ degree range, $d = 1.875$ degrees. Based on the desired range of corrective actions in degree, denoted as A_r, we formulate an action dictionary $\mathcal{M}_A$ that maps each action to the angle correction value in degrees performed by the agent starting at angle θ_s. In our experiments we choose $A_r = 3$, hence the action dictionary $\mathcal{M}_A$ will look as follows:

a_1	a_2	a_3	a_4	a_5	a_6	a_7
$-3d$	$-2d$	$-d$	0	$+d$	$+2d$	$+3d$

Based on a given observation, the agent policy will output an action $a_i \in \mathcal{A}$. We then use this action to derive a new angle θ_{s+1} for the beam configuration in the next environment state $s+1$. We compute θ_{s+1} as follows:

$$\theta_{s+1} = \theta_s + \mathcal{M}_A[a_i] \tag{3}$$

Independent agents deployed at gNB and UE sides follow this process to perform their corrective action on either the precoder or combiner analog weights, obtaining the new angles θ^{TX}_{s+1} and θ^{RX}_{s+1} that will be used to compute the adjusted directional beam.

3) Reward function: Since our goal is to maximize the RSRP observed at the UE side by making a prediction based on past experiences in a localized setting, we chose to model our reward function based on the normalized full resolution power pattern $\mathbf{P}^F_s$. Specifically, once the new steering angles θ^{TX}_{s+1} and θ^{RX}_{s+1} are derived, the RSRP value of the relative full resolution codewords combination is retrieved to obtain a power read based on the refined beam configuration in state $s+1$. Thus, we define the received reward as follows:

$$r_{s+1} = \mathbf{P}_s[\mathcal{C}_{TX}(\theta^{TX}_{s+1}), \mathcal{C}_{RX}(\theta^{RX}_{s+1})] \tag{4}$$

where $\mathcal{C}_{TX} : \mathbb{R} \to \mathcal{X}$ and $\mathcal{C}_{RX} : \mathbb{R} \to \mathcal{Y}$ are functions that returns the code indexes in the full resolution codebook for gNB and UE, respectively, based on the relative angle in degrees given as input. See Fig. 3 for a visual example of full resolution beam scan output. Therefore, a large set of historical observations of $\mathbf{P}^F$ needs to be collected in order to provide the necessary prior experience to the model in the form of a training dataset. Ideally this dataset could also evolve over time in order to improve the historical data used to train the policy network, depending on how fast the physical environment evolves. Note that, while the full resolution scans are required to obtain the dataset used during the training procedure, such information will not be needed once the agent policy has been deployed and corrective decisions will be solely based on coarse resolution P-1 output.

4) Training procedure: By fixing the scatterers configuration, we simulate $7,500$ scenarios with random mobile UE coordinates selected in the range defined in Sec. V-A and generate the full resolution P-1 procedure output $\mathbf{P}^F$ for each channel configuration. During training, in order to test the model under multiple SNR conditions and to be robust to multiple SNR variations, for each episode/channel configuration we pre-process each full scan power pattern as follows:

1) Additive White Gaussian Noise (AWGN) is applied on $\mathbf{P}^F$ based on a random SNR level chosen between $\{-30, -20, -10, 0, 10, 20, 30\}$ dB. The noisy full resolution scan will be denoted as $\tilde{\mathbf{P}}^F$.
2) $\tilde{\mathbf{P}}^F$ is normalized in the range $[0, 1] \subset \mathbb{R}$;
3) The noisy coarse scan output $\tilde{\mathbf{P}}$ is generated from $\mathbf{P}^F$ based on the desired codebook size fraction.

Finally, once $\tilde{\mathbf{P}}$ is obtained, the relative beam angle in degrees θ_s is computed and used to construct the noisy observation $\langle \tilde{\mathbf{P}}_s, \theta_s \rangle$. The DQN algorithm is trained over a total of $750,000$ episodes, each of them consisting in one channel configuration randomly picked from the training dataset and augmented with AWGN, with a decaying ϵ-greedy exploration strategy with an ϵ starting from 0.9 and decaying to 0.15 over the first 70% of the total training steps.

5) Neural Network Policy: The chosen Policy network is composed of 3 different modules:

- A 2D Convolutional Neural Network (CNN) with 3 hidden layers composed by $32\times[3\times3], 128\times[2\times2], 128\times[1\times1]$ convolutional kernels respectively used as feature extractor for $\tilde{\mathbf{P}}_s$;
- A single Fully Connected (FC) layer with 2 neurons and 1 output used as feature extractor for θ_s
- A FC policy network consisting of 2 hidden layers of 512 and 256 neurons each.

Each layers' output is equipped with a ReLU non-linear activation and the final output is sent to the last layer of the policy, which outputs the Q-values relative to each actions given an observation input. See Fig. 3 for an illustration of the model architecture.

VI. Performance evaluation

Once both gNB and UE DRL agents have been trained and deployed, we evaluate the performance of proposed beam refinement approach and compare it to the equivalent coarse and full resolution P-1 procedures on a set of $2,500$ channel configurations. To measure the impact of agents' corrective actions, for every test scenario we compute the RSRP of the chosen beam combination normalized with respect to the maximum RSRP value that we could obtain with a full resolution scan, evaluated in the same channel settings. In Fig. 4 we see how the beam refinement approach on both transmitter and receiver side individually contributes in increasing the transmission power and reaches almost 90% of the total achievable power when transmitter and receiver corrections are combined, while requiring a significantly lower fraction of SSBs overhead and without employing P-2 and P-3 phases. In order to quantify the overhead reduction compared to standard P-1, P-2 and P-3 procedure, we compute the number of control blocks needed to be exchanged in a standard 5G NR beam sweeping and beam refinement procedure using different codebook sizes, assuming $C^*_{gNB} = C^*_{UE}$, and comparing it with the proposed approach. Specifically, we compute the total number of SSB blocks to be transmitted for P-1 as $N^{\text{P1}}_{\text{SSB}} = C^*_{gNB} \times C^*_{UE}$; since P-2 and P-3 are considered special cases of P-1, we consider for simplicity $N^{\text{P2}}_{\text{CSI-RS}} = N^{\text{P3}}_{\text{SRS}} = C^*_{gNB} = C^*_{UE}$, for a total

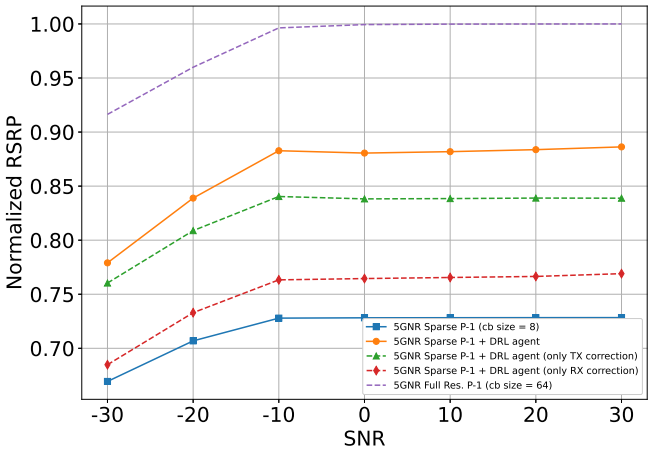

Fig. 4: Normalized RSRP for different SNR levels. Comparison between sparse 5G NR procedure, optimal 5G NR procedure and proposed approach. Codebook size is $C_{gNB} = C_{UE} = 8$ (i.e. $1/8$ of the full resolution codebooks) and range of corrective action is $\pm 3d$ ($d = 1.875$ degrees).

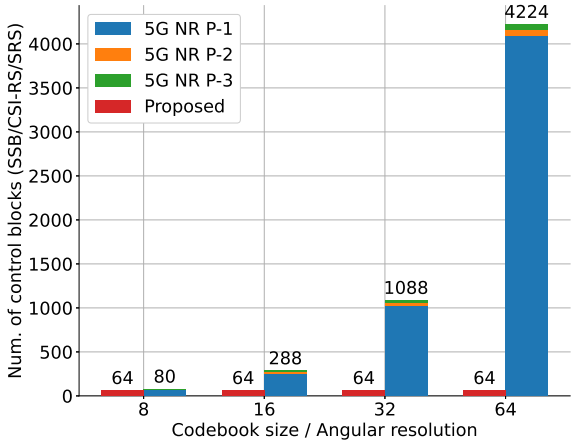

Fig. 5: Number of control blocks for beamsweeping needed for beam establishment and beam refinement: comparison between standard procedures with different codebook/angular resolutions and proposed method.

number of control blocks $N_{\text{CB}} = N_{\text{SSB}}^{\text{P1}} + N_{\text{CSI-RS}}^{\text{P2}} + N_{\text{SRS}}^{\text{P3}}$. From Fig. 5 we can observe how our approach is able to deliver very high steering accuracy while reducing the control signalling overhead by more than $\times 64$ when compared to the highest scan resolution setting, effectively saving bandwidth that would allow for higher network throughput while retaining comparable steering precision. Finally, we measured the inference time when running our DQN policy on NVIDIA RTX 3080Ti GPU to be on average 76μs, which consists of the duration of ~ 9 OFDM symbols[2] and confirms the suitability of the proposed approach for real-time implementations.

VII. Conclusion

In this work we presented a DRL method for beam refinement in mmWave, which is validated through a comprehensive simulation study. This method builds on top of the initial coarse beam estimation procedure P-1 defined in 5G NR and achieves higher spectral efficiency by completely eliminating the additional beam refining procedure needed for P-2 and P-3 at the transmitter and receiver side, respectively. If we consider each possible path in the wireless channel as an independent instance of the beam refinement problem, the proposed approach could be easily scaled up to operate on multiple RF-chains by training multiple agents at each device that perform beam refinement on the 2nd, 3rd (and so on) best paths. For future work we will scale up the approach to multiple beams and to multiple UEs. Moreover, we aim to test our approach on IBM's PAAM 28 GHz phased array in COSMOS [11] PAWR platform.

Acknowledgement

This work is supported by the US NSF under award CNS 1923789.

[2]With $\Delta f = 120$ KHz sub-carrier spacing in FR2, 1 OFDM symbol duration is $T_{OFDM} = 1/\Delta f = 8.33\mu$s

References

[1] 3GPP, "Study on New Radio access technology physical layer aspects.," Technical Report (TR) 38.802, 3rd Generation Partnership Project (3GPP), 03 2017.

[2] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.

[3] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 485–500, 2016.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, second ed., 2018.

[5] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave mimo systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[6] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive mimo systems," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 800–805, 2019.

[7] A. M. Elbir, K. V. Mishra, M. R. B. Shankar, and B. Ottersten, "A family of deep learning architectures for channel estimation and hybrid beamforming in multi-carrier mm-wave massive mimo," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2021.

[8] M. Fozi, A. R. Sharafat, and M. Bennis, "Fast mimo beamforming via deep reinforcement learning for high mobility mmwave connectivity," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 127–142, 2022.

[9] J. Jeong, S. H. Lim, Y. Song, and S.-W. Jeon, "Online learning for joint beam tracking and pattern optimization in massive mimo systems," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pp. 764–773, 2020.

[10] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1995–2003, PMLR, 20–22 Jun 2016.

[11] "Design and implementation of ibm's 28ghz paams and their integration in the cosmos testbed." https://www.cosmos-lab.org/technology/mmwave/. Accessed: 2022-05-08.