Joint Band Assignment and Beam Management using Hierarchical Reinforcement Learning for Multi-Band Communication

Dohyun Kim, *Member, IEEE*, Miguel R. Castellanos, *Member, IEEE*,
and Robert W. Heath Jr., *Fellow, IEEE*

***Abstract*—Multi-band operation in wireless networks can improve data rates by leveraging the benefits of propagation in different frequency ranges. Distinctive beam management procedures in different bands complicate band assignment because they require considering not only the channel quality but also the associated beam management overhead. Reinforcement learning (RL) is a promising approach for multi-band operation as it enables the system to learn and adjust its behavior through environmental feedback. In this paper, we formulate a sequential decision problem to jointly perform band assignment and beam management. We propose a method based on hierarchical RL (HRL) to handle the complexity of the problem by separating the policies for band selection and beam management. We evaluate the proposed HRL-based algorithm on a realistic channel generated based on ray-tracing simulators. Our results show that the proposed approach outperforms traditional RL approaches in terms of reduced beam training overhead and increased data rates under a realistic vehicular channel.**

***Keywords*—mmWave MIMO, 3GPP NR V2X, band assignment, deep reinforcement learning**

I. Introduction

Multi-band systems can achieve high data rates while maintaining coverage and reliability [1]. The integration of mmWave and sub-6 GHz transceivers means that a single device can leverage the high bandwidths and data rates of mmWave communication and the resilient and wide coverage of sub-6 GHz communication [2]. 3GPP continues to work on multi-band in recent Release 18, aiming to extend functionality such as the sidelink from frequency range 1 (FR1, 0.4 GHz – 7.1 GHz) to frequency range 2 (FR2, 24 GHz – 52 GHz) [3]. Multi-band operations can be realized into two directions: simultaneous transmission and band assignment. Allowing simultaneous usage of bands in a single time slot offers greater data rate potential but higher radio-frequency (RF) complexity.

Beam management establishes and maintains beamformed links and is a critical component of both mmWave and sub-6 GHz communication. In the mmWave band, beam management is a standard procedure to overcome misalignments and outages caused by mobility and blockages [4]. The beam management procedure in sub-6 GHz 5G employs a precoder matrix indicator (PMI) codebook and feedback, categorized as Type-1 and Type-2. Type-1 codebooks, with shorter training overhead and predefined precoders per antenna geometry, are commonly used for spatial multiplexing compared to Type-2 codebooks. The overhead from beam management, which deteriorates the data rate, can be excessive when exhaustive beam alignment methods that use narrow beam codebooks [5]. In highly dynamic scenarios such as 5G vehicular networks, low overhead beam management is paramount for ensuring high data rate with resilient links [6].

Beam management overhead influences the rate performance of band assignment but has been overlooked in existing solutions [7]–[10]. Prior work has studied various objectives such as throughput maximization [7], [10], jammer interference minimization [8], and outage ratio minimization [9], but they typically assume instant evaluation of a selected band. When accounting for beam management, the objective of minimizing the beam management overhead becomes an additional factor that can influence band selection. In the case of throughput maximization, while the mmWave band offers high throughput, the sub-6 GHz band can be more favorable due to its shorter beam management overhead, despite having lower throughput. High mobility and non-line-of-sight (NLOS) operation are other examples in which the sub-6 GHz band could outperform the mmWave band. This highlights the need for a joint formulation of band

Dohyun Kim is with the Institute of Computer Technology, Seoul National University (SNU), Seoul 08826, South Korea (e-mail: dohyun.p.kim@snu.ac.kr). Miguel R. Castellanos is with the Department of Electrical and Computer Engineering, North Carolina State University, 890 Oval Dr., Raleigh, NC 27606 USA (email: mrcastel@ncsu.edu). Robert W. Heath Jr. is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92161, USA (email: rwheathjr@ucsd.edu). This material is based upon work supported by the National Science Foundation under grant nos. NSF-ECCS-2153698, NSF-CCF-2225555, NSF-CNS-2147955 and is supported in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program.

assignment and beam management, where the tradeoff between throughput and overhead is carefully balanced.

Recent work has shown that reinforcement learning (RL) is an effective framework to address the overhead of beam training in mmWave vehicular networks [11]–[13]. The RL framework, including partially observable environments and contextual bandits in the wide sense, is capable of reducing control overhead by using the accumulated deployment history to effectively balance the exploration of new control actions with the exploitation of actions that have yielded the highest expected return in the past. In our prior work [11], we have studied deep RL (DRL) with threshold-based actions to reduce beam training overhead in mmWave MIMO vehicular networks with relay selection. In [12], the incoming vehicle direction was used as input to apply contextual bandits for beam selection in mmWave vehicular networks. In [13], an autoencoder was employed to predict vehicle mobility then to find beam training policies based on RL with partial observability. RL has also been used in band assignment, where WiFi traffic demands are learned from WiFi channel activity observations [14].

DRL has recently seen partial advancements in addressing resource allocation tasks for wireless networks that are characterized by expanding scale, versatility, and heterogeneity. In light of the increasing scale and complexity of upcoming wireless networks, DRL approaches leveraging measurements based on observation are preferred over conventional signal processing methods, which may face challenges in obtaining complete or quasi-complete knowledge of the wireless environment [15]. The work presented in [16] exploits the learning model of DRL to efficiently estimate system states in scenarios involving discrete, continuous, or hybrid states, providing scalability benefits compared to decentralized allocation schemes. Additionally, the work in [17] addresses the complexity arising from heterogeneous communication nodes and link types in a vehicular edge network by employing asynchronous application of DRL algorithms to each agent. Despite these successful applications, the quality of the reward becomes a critical factor influencing the performance of DRL algorithms, as the aforementioned methods rely on deterministic rewards [18]. In this paper, we focus on band assignment that refers to the selection of operating bands over time slots in a sequential manner. Band assignment can be understood as a subproblem of the frequency resource allocation in multi-band systems [19]. Solving band assignment using traditional RL approaches can still be challenging, hence motivating a dedicated structure in the learning approach, because beam training can only be performed in one band at a time and the sample efficiency in each band will be low.

Hierarchical reinforcement learning (HRL) is a promising approach for addressing the joint band assignment and beam management problem, improving sample efficiency compared to DRL by learning policies separately for tasks at different decision hierarchy levels. The idea of hierarchy resembles intelligence observed in nature where learning agents only necessitate a few examples [20]. The minimal example requirement, also known as few-shot learning capabilities, allows generalization across different levels of abstraction. One of the main benefits of exploiting hierarchy is that the shortened episodes, owing to the abstracted tasks, makes both exploration and learning easier [21]. While HRL is still a relatively new approach in wireless communication, it has shown promising results outperforming the traditional DRL methods in resource allocation [22], channel sensing [23], and scheduling [24]. Relevant work on HRL applications in wireless communication, however, rely on discrete action spaces that can limit their generalization to real-world problems. For example, the discrete action space only represent the quantized transmission power constraint in the power allocation problem [22]. We employ HRL using continuous actions, which can improve the scalability of the learning algorithm and enhance its applicability to real-world deployments.

In this paper, we propose an HRL-based algorithm for joint band assignment and beam management that leverages the band characteristics of sub-6 GHz and mmWave. We presume the communication nodes employ codebook-based beamforming, co-located sub-6 GHz and mmWave arrays, and Orthogonal Frequency Division Multiplexing (OFDM). We also assume a fully digital sub-6 GHz array and a hybrid mmWave array with analog and digital beamformers. The system can either perform digital beam training at the sub-6 GHz band, analog beam training at the mmWave band, or digital beam training at the mmWave band. We assume perfect spectral efficiency feedback from the user to the base station, free from quantization or overhead, during both digital beam training at the sub-6 GHz band and analog beam training at the mmWave band. This feedback may be transmitted through a dedicated channel in the unoccupied band or sent on the reverse link with reduced coding and spreading. For the digital beam training at the mmWave band, we assume the quality of the feedback can be modeled using the number of bits in the user codebook. The algorithm employs two policies: an upper-level policy for band selection and a lower-level policy to determine the beam training method. The choice of beam training is guided by comparing the spectral efficiency feedback and two adaptive thresholds determined by the lower-level policy. We use one threshold to separate analog and digital beam training and the other threshold to decide between digital beam training and data transmission. The band selection is made by the

upper-policy, which aggregates state, goal, and reward over an adaptive period. The HRL-based method uses the best known band until the spectral efficiency feedback deteriorates below the learned threshold, in which case the algorithm tries out different band or beam training indicated by the upper-level and lower-level policies.

We summarize our contributions as follows:

1) We formulate a joint band assignment and beam management problem for wireless networks operating on sub-6 GHz and mmWave that accounts for the effect of the beam management overhead on the cumulative data rate. We devise a hierarchical sequential decision-making model of the joint band assignment and beam management problem, avoiding the non-stationary Markov decision process (MDP) by separately learning policies for band selection and beam management.
2) We propose an HRL-based algorithm to solve the joint band assignment and beam management problem. The proposed algorithm uses the spectral efficiency feedback from the receiver to learn thresholds that determines the beam training method.
3) We numerically evaluate the proposed algorithm compared to baseline learning algorithms on a realistic vehicular channel. The HRL-based proposed algorithm outperforms the heuristic owing to the reduced horizon for policy computation from abstracted subtasks.

The rest of the paper is structured as follows. In Section II, we present the system model used to represent the multi-band wireless network. In Section III, we formulate the joint band assignment and beam management problem and discuss the challenges of designing an learning algorithm. In Section IV, we describe a DRL algorithm that can partially address the challenges of the joint band assignment and beam management problem. In Section V, we develop an HRL algorithm to solve the joint band assignment and mode selection problem. In Section VI, we numerically evaluate the proposed algorithm compared to baselines. Finally, we conclude the paper in Section VII.

We use the following notation throughout this paper: $\mathbf{A}$ is a matrix, $\mathbf{a}$ is a vector, a is a scalar, and $\mathcal{A}$ is a set. We denote $\mathbf{a}^{\mathrm{T}}$ the transpose of $\mathbf{a}$, $\mathbf{a}^{*}$ the conjugate transpose, $\|\mathbf{a}\|_2$ the 2-norm, and $\|\mathbf{a}\|_{\mathrm{F}}$ the Frobenius norm. We underline the sub-6 GHz variables as $\underline{\mathbf{a}}$ to distinguish them from mmWave.

II. System model

In this section, we describe the system model for a wireless network operating both on the sub-6 GHz and mmWave bands. As shown in Fig. 1, the system can operate on one band at a time. We assume that the communication nodes are equipped with co-located sub-6 GHz and mmWave arrays. We provide the signal model in the mmWave band in Section II-A. We then outline the codebooks and beam training procedure in the mmWave band in Section II-B. We summarize the sub-6 GHz signal model and beam training process in Section II-C and Section II-D. For the convenience of readers, we summarize the system model parameters in Table I with sub-6 GHz parameters omitted for brevity.

TABLE I
SUMMARY OF SYSTEM MODEL PARAMETERS

Notation	System model parameter
m	Time index of decision horizon
k	Subcarrier index
$b[m]$	Band assignment variable
M_{BT}	Time length of beam training
M_{DT}	Time length of data transmission
K	Total number of subcarriers at mmWave
B	Bandwidth at mmWave
N_{BS}	Number of antennas at the base station
$N_{\mathrm{BS,RF}}$	Number of RF chains at the base station
N_{UE}	Number of antennas at the user
$N_{\mathrm{UE,RF}}$	Number of RF chains at the user
N_{S}	Number of streams
$\mathsf{s}[k,m]$	Symbol vector
$\mathsf{F}_{\mathrm{BB}}[k,m]$	Frequency-selective baseband precoder
$\mathsf{F}_{\mathrm{RF}}[m]$	Frequency-flat RF precoder
$\mathsf{H}[k,m]$	Wideband channel
$\mathsf{W}_{\mathrm{RF}}[m]$	Frequency-flat RF combiner
$\mathsf{W}_{\mathrm{BB}}[k,m]$	Frequency-selective baseband combiner
$P[k,m]$	Transmit power
$G[m]$	Large-scale fading
σ_{n}	Standard deviation of noise
$S[k,m]$	Spectral efficiency per the subcarrier k
ν_{BS}	Analog codebook size at the base station
ν_{UE}	Analog codebook size at the user
N_{SS}	Number of SS blocks per burst
M_{SS}	Periodicity between SS burst exchangement
M_{RF}	Analog beam training overhead
β_{RF}	Ratio of pilots per symbol transmission
ζ_{RF}	Number of OFDM frames
$\mathsf{H}[k,m]$	Digital effective channel
$\hat{\mathsf{H}}[k,m]$	Quantized effective channel
κ_{RVQ}	Number of quantization bits
$\kappa_{\mathrm{channel}}$	Number of bits available via feedback channel
M_{BB}	Digital beam training overhead

Consider a downlink scenario in a multi-band MIMO-OFDM wireless network, where a single base station serves a single mobile user. For each OFDM time frame, we assume the base station selects a transmission mode of either beam training or data transmission. We also assume the base station sends pilots only during beam training for M_{BT} discrete time slots. Whenever the mode is data transmission, the base station sends only data symbols for M_{DT} discrete time slots. The sequence of

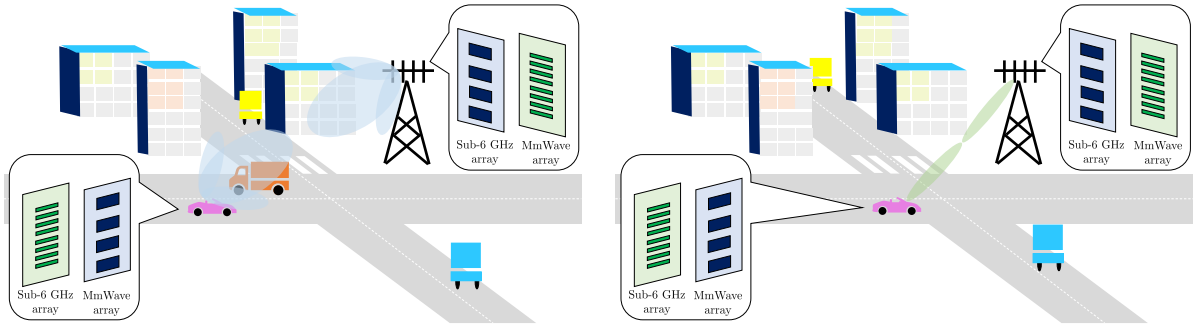

Fig. 1. Illustration of an example system model showing two snapshots: (a) the base station operates on the sub-6 GHz band to serve the user due to a large truck posing as a mobile blockage, and (b) the base station operates on the mmWave band when LOS is available.

modes can be consecutive beam training, consecutive data transmissions, or alternating with an arbitrary number of consecutive modes. The band selection occurs when a new transmission mode is deployed. When the system uses the sub-6 GHz band, denoted by a binary variable $b = 0$, the system operates over a bandwidth $\underline{B}$ with $\underline{K}$ subcarriers. Similarly, we use $b = 1$ to denote that the system operates in the mmWave band. In this case, the system uses a bandwidth B with K subcarriers.

A. Millimeter wave signal model

In the mmWave band, we assume the system employs a fully connected hybrid beamforming architecture. We denote N_{BS} as the number of antennas and $N_{\text{BS,RF}}$ as the number of RF chains at the base station. At the user, we denote N_{UE} as the number of antennas and $N_{\text{UE,RF}}$ as the number of RF chains. The base station and the user communicate via N_{S} data streams, where $N_{\text{S}} \leq N_{\text{BS,RF}} \leq N_{\text{BS}}$ and $N_{\text{S}} \leq N_{\text{UE,RF}} \leq N_{\text{UE}}$. For simplicity, we focus on a fully connected hybrid beamforming architecture in mmWave. Partially connected architectures can reduce power consumption and hardware cost, though beam training can be more complex as the whole channel may need to be reconstructed from subarray measurements. We leave the extension to a partially connected architecture for future work.

At each OFDM time frame m and subcarrier k, the base station sends a symbol vector $\mathsf{s}[k,m]$ of size $N_{\text{S}} \times 1$ to the user. The symbol vector is assumed to be normalized such that $\mathbb{E}[|\mathsf{s}[k,m]|^2] = 1$. The base station precodes the symbol vector with the $N_{\text{BS,RF}} \times N_{\text{S}}$ frequency-selective baseband precoder $\mathbf{F}_{\text{BB}}[k,m]$ followed by the $N_{\text{BS}} \times N_{\text{BS,RF}}$ frequency-flat RF precoder $\mathbf{F}_{\text{RF}}[m]$. We assume the precoded signal propagates through a time-varying wideband channel model $\mathbf{H}[k,m]$ with large-scale fading denoted as $G[m]$ and the noise denoted as $\mathbf{n}[k,m]$. We assume the noise is independently and identically distributed (IID) following the distribution $\mathcal{N}_C(0, \sigma_{\text{n}}^2)$. At the user, the received signal is processed with the $N_{\text{UE}} \times N_{\text{UE,RF}}$ frequency-flat RF combiner $\mathbf{W}_{\text{RF}}[m]$ followed by the $N_{\text{UE,RF}} \times N_{\text{S}}$ frequency-selective baseband combiner $\mathbf{W}_{\text{BB}}[k,m]$. We set power constraint on the base station by denoting $P[k,m]$ as the transmit power and constraining $\mathbf{F}_{\text{BB}}[k,m]$ such that $\|\mathbf{F}_{\text{RF}}[k,m]\mathbf{F}_{\text{BB}}[k,m]\|_F^2 = N_{\text{S}}$.

The end-to-end input-to-output relation in the mmWave band is

$$\begin{aligned}\mathbf{y}[k,m] = &\sqrt{P[k,m]G[m]}\mathbf{W}_{\text{BB}}^*[k,m]\mathbf{W}_{\text{RF}}^*[m]\mathbf{H}[k,m] \\ &\times \mathbf{F}_{\text{RF}}[m]\mathbf{F}_{\text{BB}}[k,m]\mathsf{s}[k,m] \\ &+ \mathbf{W}_{\text{BB}}^*[k,m]\mathbf{W}_{\text{RF}}^*[m]\mathbf{n}[k,m]. \end{aligned} \tag{1}$$

We define the spectral efficiency per the subcarrier k in the mmWave band as

$$\begin{aligned}S[k,m] = \log\det\big(&\mathbf{I}_{N_{\text{S}}} + P[k,m]G[m]\sigma_{\text{n}}^{-2}\mathbf{W}_{\text{BB}}^*[k,m] \\ &\times \mathbf{W}_{\text{RF}}^*[m]\mathbf{H}[k,m]\mathbf{F}_{\text{RF}}[m]\mathbf{F}_{\text{BB}}[k,m]\mathbf{F}_{\text{BB}}^*[m] \\ &\times \mathbf{F}_{\text{RF}}^*[m]\mathbf{H}^*[k,m]\mathbf{W}_{\text{RF}}[m]\mathbf{W}_{\text{BB}}[k,m]\big). \end{aligned} \tag{2}$$

Note that the spectral efficiency in the mmWave band at time m can be written as $\sum_{k=1}^{K} S[k,m]$.

B. Millimeter wave beam management procedure

In this section, we outline the beam management procedure used in the mmWave band. The purpose of the beam management procedure is to determine the beamforming matrices—$\mathbf{F}_{\text{RF}}[m]$, $\mathbf{F}_{\text{BB}}[k,m]$, $\mathbf{W}_{\text{RF}}[m]$, and $\mathbf{F}_{\text{BB}}[k,m]$—adaptive to the dynamic channel conditions using feedback from the user to the base station. We assume that beam training can be split into two stages: analog beam training and digital beam training. The analog beam training is based on beam codebooks, such that the base station and the user select beam pairs. We further assume the analog beam training involves exchanging synchronization signals (SSs) between the

base station and the user [5]. In the digital beam training, the user estimates the digital effective channel and computes the digital combiner then feeds back the quantized effective channel to the base station.

We first describe the analog beam training procedure in the mmWave band. Let us denote the base station analog codebook with size ν_{BS} as $\mathcal{F} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\nu_{\text{BS}}}\}$. We similarly denote the user analog codebook with size ν_{UE} as $\mathcal{W} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{\nu_{\text{UE}}}\}$. To obtain the analog precoder $\mathbf{F}_{\text{RF}}[m]$ and the analog combiner $\mathbf{W}_{\text{RF}}[m]$, the base station and the user exhaustively sweep RF beams over ν_{BS} ν_{UE} time slots simultaneously for all RF chains.

The analog beam management procedure is performed by exchanging SS bursts, where a single SS burst comprises multiple SS blocks [5]. We denote N_{SS} as the number of SS blocks per burst and M_{SS} as the periodicity between two SS burst exchangements. The total number of beams, $\nu_{\text{BS}}\nu_{\text{UE}}$, is divided into bursts of size N_{SS} that are exchanged every M_{SS} time slots such that the overhead of the analog beam training procedure is [11]

$$M_{\text{RF}} = M_{\text{SS}} \left\lceil \frac{\nu_{\text{BS}}\nu_{\text{UE}}}{N_{\text{SS}}} \right\rceil. \tag{3}$$

During the analog beam training, where the transmit and receive beam pair are being swept simultaneously for RF chains, the user feeds back the spectral efficiency for each transmit and receive beam pair to the base station. We use the MMSE estimator for the effective channel, which accounts for the measurement error in its estimation, under a rectangular Doppler spectrum as outlined in [25, Sec. 4.8]. The signal-to-noise-ratio (SNR) prior to beamforming is $P[k,m]G[m]\sigma_{\text{n}}^{-2}$ due to the normalization $\mathbb{E}[|\mathsf{s}[k,m]|^2] = 1$. The MMSE estimator is expressed in terms of the ratio of pilots per symbol transmission, which we denote as β_{RF}, and the total number of OFDM frames during the analog beam training, which we denote as ζ_{RF}. Then, the MMSE can be expressed as

$$\text{MMSE} = \frac{1}{1 + \beta_{\text{RF}}\zeta_{\text{RF}}\text{SNR}}, \tag{4}$$

and the effective SNR as

$$\text{SNR}_{\text{eff}} = \frac{\text{SNR}(1 - \text{MMSE})}{1 + \text{SNR} \cdot \text{MMSE}}. \tag{5}$$

The effective SNR is applied to the spectral efficiency feedback from the user to the base station. This means that the base station uses the effective SNR to compute the spectral efficiency given codebook vectors $\mathbf{g}[m]$ and $\mathbf{v}[m]$. The resulting spectral efficiency can be written as

$$\begin{aligned} &S_{\text{UE}}[m; \mathbf{g}[m], \mathbf{v}[m]] \\ &= \frac{1}{K}\sum_{k=1}^{K} \log_2(1 + \text{SNR}_{\text{eff}}|\mathbf{g}^*[m]\mathbf{H}[k,m]\mathbf{v}[m]|^2). \end{aligned} \tag{6}$$

We presume a greedy approach to configure the analog beamformers $\mathbf{F}_{\text{RF}}[m]$ and $\mathbf{W}_{\text{RF}}[m]$. For RF chain pair $(n_{\text{BS,RF}}, n_{\text{UE,RF}})$ ranging from $n_{\text{BS,RF}} = 1, \ldots, N_{\text{BS,RF}}$ and $n_{\text{UE,RF}} = 1, \ldots, N_{\text{UE,RF}}$, the system obtains $\{\mathbf{g}^\star_{n_{\text{BS,RF}}}[m], \mathbf{v}^\star_{n_{\text{UE,RF}}}[m]\}$ by solving

$$\max_{\forall \mathbf{g} \in \mathcal{W}, \forall \mathbf{v} \in \mathcal{F}} S_{\text{UE}}[m; \mathbf{g}[m], \mathbf{v}[m]] \tag{7a}$$

$$\text{subject to} \quad \mathbf{g} \neq \mathbf{g}_1, \ldots, \mathbf{g} \neq \mathbf{g}_{n_{\text{BS,RF}}-1}, \tag{7b}$$

$$\mathbf{v} \neq \mathbf{v}_1, \ldots, \mathbf{v} \neq \mathbf{v}_{n_{\text{UE,RF}}-1}. \tag{7c}$$

The system then sets the RF beamforming matrices as $\mathbf{F}_{\text{RF}}[m] = [\mathbf{v}^\star_1[m], \mathbf{v}^\star_2[m], \ldots, \mathbf{v}^\star_{N_{\text{BS,RF}}}[m]]$ and $\mathbf{W}_{\text{RF}}[m] = [\mathbf{g}^\star_1[m], \mathbf{g}^\star_2[m], \ldots, \mathbf{g}^\star_{N_{\text{UE,RF}}}[m]]$. Note that the constraints $\mathbf{g} \neq \mathbf{g}_1, \ldots, \mathbf{g} \neq \mathbf{g}_{n_{\text{BS,RF}}-1}$ and $\mathbf{v} \neq \mathbf{v}_1, \ldots, \mathbf{v} \neq \mathbf{v}_{n_{\text{UE,RF}}-1}$ ensure that distinct beams are used for separate RF chains, achieving spatial multiplexing gain [26].

After analog training, the system finds the digital precoder and combiner by estimating the digital effective channel $\bar{\mathbf{H}}[k,m]$, for all subcarriers $k = 1, \ldots, K$. We choose to represent the measurement error from pilot-based estimation using the mean squared error (MSE) [25, Sec. 3.7]

$$\text{MSE} = \frac{1}{\frac{\beta_{\text{BB}}\zeta_{\text{BB}}}{N_{\text{BS}}}\text{SNR}}, \tag{8}$$

where β_{BB} is the ratio of pilots per symbol transmission and ζ_{BB} is the total number of OFDM frames in the digital training. Let us denote $\boldsymbol{\delta}[k,m] \sim \mathcal{CN}(0, \mathbf{I})$ as a complex Gaussian random variable independent from the digital effective channel. We model the estimated effective channel using uncertainty of the form [27]

$$\begin{aligned} \bar{\mathbf{H}}[k,m] = &\mathbf{W}^*_{\text{RF}}[m]\mathbf{H}[k,m]\mathbf{F}_{\text{RF}}[m] \\ &+ \frac{1}{\sqrt{\frac{\beta_{\text{BB}}\zeta_{\text{BB}}}{N_{\text{BS}}}\text{SNR}}}\boldsymbol{\delta}[k,m]. \end{aligned} \tag{9}$$

With the effective channel $\bar{\mathbf{H}}[k,m]$, the user can compute the least squares digital combiner as

$$\mathbf{W}_{\text{BB}}[k,m] = \bar{\mathbf{H}}[k,m](\bar{\mathbf{H}}^*[k,m]\bar{\mathbf{H}}[k,m])^{-1}. \tag{10}$$

We further assume the effective channel is quantized with random vector quantization (RVQ) codebook, denoted as $\mathcal{H}$, constructed with Lloyd's algorithm, following [28]. We note the PMI codebook that is used for sub-6 GHz is not currently implemented in mmWave. RVQ codebooks can be randomly generated independently from the channel realization, and are known to be asymptotically optimal regarding the number of transmit antennas and codebook size [29]. Then, the quantized effective channel fed back from the user to the base station can be written as

$$\hat{\mathbf{H}}[k,m] = \underset{\tilde{\mathbf{H}} \in \mathcal{H}}{\text{argmax}} \|\bar{\mathbf{H}}^*[k,m]\tilde{\mathbf{H}}[k,m]\|_2. \tag{11}$$

Finally, the base station computes the MMSE digital precoder as

$$\mathbf{F}_{\text{BB}}[k,m] = \hat{\mathbf{H}}^{*}[k,m](\hat{\mathbf{H}}[k,m]\hat{\mathbf{H}}^{*}[k,m])^{-1}. \quad (12)$$

The overhead of digital beam training, which we denote as M_{BB}, only involves the feedback (11) and the matrix manipulations throughout (10), (9), and (12). The feedback involves channel access unlike matrix manipulations, hence, the dominant factor in M_{BB} is the number of quantization bits of $\mathcal{H}$. Let us denote κ_{RVQ} as the number of quantization bits of $\mathcal{H}$ and κ_{channel} as the number of bits that can be sent through the feedback channel over a single time slot. Then, the overhead of the digital beam training procedure can be written as

$$M_{\text{BB}} = \left\lceil \frac{\kappa_{\text{RVQ}}}{\kappa_{\text{channel}}} \right\rceil. \quad (13)$$

Compared to the analog beam training overhead, $M_{\text{BB}} \ll M_{\text{RF}}$ because digital beam training requires far fewer feedback procedures than the multiple SS burst exchanges required in analog beam training. The gap between M_{BB} and M_{RF} will increase when the number of antennas in the system increases.

C. Sub-6 GHz system model

In the sub-6 GHz band, we assume the system employs the fully digital beamforming architecture. The base station is equipped with $\underline{N}_{\text{BS}}$ antennas and RF chains to send $\underline{N}_{\text{S}}$ data streams. The user is equipped with $\underline{N}_{\text{UE}}$ antennas and RF chains. The size of symbol vector $\mathsf{s}[k,m]$ is $\underline{N}_{\text{S}} \times 1$. The symbol vector is assumed to be normalized such that $\mathbb{E}[|\mathsf{s}[k,m]|^2] = 1$. The base station precodes the symbol vector with an $\underline{N}_{\text{BS}} \times \underline{N}_{\text{S}}$ frequency-selective precoder $\underline{\mathbf{F}}_{\text{BB}}[k,m]$. The precoded signal propagates through the channel denoted as $\underline{\mathbf{H}}[k,m]$ and the noise denoted as $\underline{\mathbf{n}}[k,m]$. We assume the noise is IID following the distribution $\mathcal{N}_C(0,\underline{\sigma}_{\text{n}}^2)$ At the user, we assume the received signal is decoded with $\underline{N}_{\text{UE}} \times \underline{N}_{\text{S}}$ frequency-selective decoder $\underline{\mathbf{W}}_{\text{BB}}[k,m]$. We set power constraint on the base station by denoting $\underline{P}[k,m]$ as the transmit power and as $\|\underline{\mathbf{F}}_{\text{BB}}[k,m]\|_F^2 = \underline{N}_{\text{S}}$. Then, the end-to-end input-to-output relation in the sub-6 GHz band is

$$\begin{aligned}\underline{\mathbf{y}}[k,m] = &\sqrt{\underline{P}[k,m]\underline{G}[m]}\underline{\mathbf{W}}_{\text{BB}}^{*}[k,m]\underline{\mathbf{H}}[k,m] \\ &\times \underline{\mathbf{F}}_{\text{BB}}[k,m]\mathsf{s}[k,m] + \underline{\mathbf{W}}_{\text{BB}}^{*}[k,m]\underline{\mathbf{n}}[k,m], \quad (14)\end{aligned}$$

and the spectral efficiency per the subcarrier k in the sub-6 GHz band can be written as

$$\begin{aligned}\underline{S}[k,m] = \log\det\big(&\mathbf{I}_{\underline{N}_{\text{S}}} + \underline{P}[k,m]\underline{G}[m]\sigma_{\text{n}}^{-2}\underline{\mathbf{W}}_{\text{BB}}^{*}[k,m] \\ &\times\underline{\mathbf{H}}[k,m]\underline{\mathbf{F}}_{\text{BB}}[k,m]\underline{\mathbf{F}}_{\text{BB}}^{*}[m] \\ &\times\underline{\mathbf{H}}^{*}[k,m]\underline{\mathbf{W}}_{\text{BB}}[k,m]\big). \quad (15)\end{aligned}$$

Due to the normalization of the symbol vector, the SNR prior to beamforming in the sub-6 GHz band is $\underline{P}[k,m]\underline{G}[m]\underline{\sigma}_{\text{n}}^{-2}$.

D. Sub-6 GHz beam management procedure

The 5G NR beam management procedure in the sub-6 GHz band uses channel state information reference signals (CSI-RSs), a technique inherited from 4G, and precoding matrix indicator (PMI) feedback. We presume that the Type-1 PMI codebook is employed and the PMI feedback indicates the PMI table index, which includes both candidate precoders and the channel quantization [30]. Let us denote $\underline{\beta}$ as the ratio of pilots per symbol transmission, $\underline{\zeta}$ as the total number of OFDM frames in the sub-6 GHz band beam training, and $\boldsymbol{\delta}[k,m] \sim \mathcal{CN}(0,\mathbf{I})$ as a complex Gaussian random variable independent from the channel $\underline{\mathbf{H}}[k,m]$. We model the CSI before quantization with the error model [27]

$$\mathbf{P}[k,m] = \underline{\mathbf{H}}[k,m] + \frac{1}{\sqrt{\frac{\underline{\beta}\cdot\underline{\zeta}}{\underline{N}_{\text{BS}}}\text{SNR}}}\boldsymbol{\delta}[k,m]. \quad (16)$$

The precoder selection based on the PMI feedback can be written as

$$\underline{\mathbf{F}}_{\text{BB}}[k,m] = \underset{\forall\mathbf{F}\in\underline{\mathcal{H}}}{\text{argmax}}\,\|\mathbf{P}[k,m]\mathbf{F}[k,m]\| \quad (17)$$

where $\underline{\mathcal{H}}$ denotes the PMI codebook. We further assume the PMI feedback includes the spectral efficiency feedback

$$\begin{aligned}\underline{S}_{\text{UE}}[m] = \frac{1}{\underline{K}}\sum_{k=1}^{\underline{K}}\log_2\Bigg(&1 + \text{SNR} \\ &\times\left|\underline{\mathbf{W}}_{\text{BB}}^{*}[k,m]\mathbf{P}[k,m]\underline{\mathbf{F}}_{\text{BB}}[k,m]\right|^2\Bigg), \quad (18)\end{aligned}$$

where the combiner $\underline{\mathbf{W}}_{\text{BB}}^{*}[k,m]$ is computed based on zero-forcing. Let us denote ν_{PMI} as the size of the PMI codebook and $\underline{\kappa}_{\text{channel}}$ as the number of bits that can be sent through the sub-6 GHz feedback channel over a single time slot. The overhead from the beam training procedure in the sub-6 GHz band, which we denote as $\underline{M}_{\text{BB}}$, can be written as

$$\underline{M}_{\text{BB}} = \left\lceil \frac{\log_2 \nu_{\text{PMI}}}{\underline{\kappa}_{\text{channel}}} \right\rceil. \quad (19)$$

Typically, the beam training overhead in the sub-6 GHz band is between the analog beam training overhead and digital beam training overhead in the mmWave band such that $M_{\text{BB}} < \underline{M}_{\text{BB}} < M_{\text{RF}}$. It is noteworthy that the beam training overhead in the sub-6 GHz can be reduced by grouping multiple antennas in the base station to a single port to reduce the beam training overhead [31].

III. Reinforcement learning formulation of the joint band assignment and beam management problem

In this section, we formulate the joint band assignment and beam management problem for the multi-band wireless network as an RL problem. We first describe the underlying learning model as an MDP. We then discuss the challenges of the RL formulation by describing the inconsistent action space over different bands. We describe a baseline approach of the RL formulation using action masking in Section IV. We further detail the remedies on the challenges by proposing an algorithm based on HRL in Section V.

The base station aims to maximize the system's data rate by selecting the best band of operation and precoder at each time slot. For each time slot m, we denote the actions that the transmitter can take as $\mathcal{A}[m]$. The action dictates a chosen band and also whether to perform beam training or data transmission. We say the action is a set including a chosen band $b[m]$ and a beam management mode $n_{\text{mode}}[m]$. Specifically, we set $b[m] = 1$ to imply the mmWave band being the band of operation and $b[m] = 0$ to imply the sub-6 GHz band being the band of operation. We also set $n_{\text{mode}}[m] = 1$ to indicate analog beam training and $n_{\text{mode}}[m] = 0$ to indicate digital beam training. The system's data rate, which is the performance metric of interest, can be written as

$$R[m] = \Bigg((1 - b[m]) \frac{\underline{B}}{\underline{K}} \sum_{k=1}^{\underline{K}} \underline{S}[k, m] + b[m] \frac{B}{K} \sum_{k=1}^{K} S[k, m] \Bigg). \tag{20}$$

We assume that M is finite to keep the cumulative data rate finite. Denoting the binary variable $c(\mathcal{A}[m]) = 0$ when beam training is in progress and $c(\mathcal{A}[m]) = 1$ when data transmission is performed, the optimization problem can be written as

$$\max_{\{\mathcal{A}[m]\}} \sum_{m=1}^{M} c(\mathcal{A}[m]) R[m] \tag{21a}$$

$$\text{s.t.} \quad \mathcal{A}[m] = \cdots = \mathcal{A}[m + M_{\text{RF}}], \quad \text{if} \quad b[m] = 1 \text{ and } n_{\text{mode}}[m] = 1, \tag{21b}$$

$$\mathcal{A}[m] = \cdots = \mathcal{A}[m + M_{\text{BB}}], \quad \text{if} \quad b[m] = 1 \text{ and } n_{\text{mode}}[m] = 0, \tag{21c}$$

$$\mathcal{A}[m] = \cdots = \mathcal{A}[m + \underline{M}_{\text{BB}}], \quad \text{if} \quad b[m] = 0 \text{ and } n_{\text{mode}}[m] = 0. \tag{21d}$$

The constraints in (21b), (21c), and (21d) represent that each action takes the dedicated time slots to terminate as described in Section II-B and Section II-D. We solve (21) by formulating an MDP, which has been shown to be an effective approach for many resource allocation problems [32]. The crucial elements in an MDP are the state space, action space, and the reward function. The design of the action space poses a significant challenge in the MDP formulation, as the choice between discrete and continuous action spaces plays a crucial role in determining scalability. On the one hand, discrete action spaces may lead to scalability issues due to the need for defining separate actions for various combinations of bands and beam management modes. On the other hand, continuous action spaces partially address scalability concerns by substituting discrete actions with continuous variables that function as decision boundaries [11]. We first describe the MDP of the DRL approach using continuous action space in Section IV. The distinct beam management procedures in the mmWave band and sub-6 GHz band pose limitations on the performance of continuous RL algorithms, as decision boundaries are applied uniformly across all bands. In this regard, we further specify the MDP of the HRL-based algorithm in Section V.

IV. Reinforcement learning approach of the joint band assignment and beam management problem using action masking

In this section, we present a DRL-based method for solving the joint band assignment and beam management problem using threshold-based actions. The DRL-based approach serves as a baseline to the HRL-based algorithm and an example application of continuous action spaces with action masking, later used in the HRL-based algorithm as well.

A naive approach to addressing inconsistent action spaces is through action masking, where the total action space includes all possible actions conservatively, and invalid action probabilities are forced to zero [33], [34]. We describe the MDP formulation of the brute-force approach using action space masking, which we later set as a baseline in the experiments under the name three-threshold policy. The state $\mathcal{T}[m]$, action $\mathcal{A}[m]$, and reward $r[m]$ of the three-threshold policy can be described as the following.

1) States: The state space incorporates the selected beamformers and feedback used throughout the beam management procedures as discussed in Section II-B and Section II-C. In the mmWave band, the analog beam training determined the analog beamformers based on spectral efficiency feedback followed by digital effective channel estimation. In the sub-6 GHz band, the PMI feedback determines the precoder computation. The state can be written as

$$\mathcal{T}[m] = \Big\{ \mathbf{F}_{\text{RF}}[m], \mathbf{W}_{\text{RF}}[m], S_{\text{UE}}[m], \big\{ \hat{\mathbf{H}}[k, m] \big\}_{k=1}^{K}, \big\{ \underline{\mathbf{F}}_{\text{BB}}[k, m] \big\}_{k=1}^{\underline{K}}, \big\{ \mathbf{P}[k, m] \big\}_{k=1}^{\underline{K}} \Big\}. \tag{24}$$

Note that the codebook assumption for constructing the analog beamformers $\mathbf{F}_{\text{RF}}[m], \mathbf{W}_{\text{RF}}[m]$, the quantized feedback channel $\left\{\hat{\mathbf{H}}[k,m]\right\}_{k=1}^{K}$ in the mmWave band, and the precoder $\left\{\underline{\mathbf{F}}_{\text{BB}}[k,m]\right\}_{k=1}^{\underline{K}}$ in sub-6 GHz band can be used to reduce the state space dimension.

2) Actions: The action space consist of three continuous variables

$$\mathcal{A}[m] = \{\underline{\tau}[m], \tau_{\text{RF}}[m], \tau_{\text{BB}}[m]\}. \tag{25}$$

The spectral efficiency feedback of the operating band is compared with the thresholds. At mmWave, the spectral efficiency feedback $S_{\text{UE}}[m]$ is compared with each threshold to perform one of the following. When $S_{\text{UE}}[m] < \underline{\tau}[m]$, the base station switches band to the sub-6 GHz. When $\underline{\tau}[m] < S_{\text{UE}}[m] < \tau_{\text{RF}}[m]$, the base station tries analog beam training. When $\tau_{\text{RF}}[m] < S_{\text{UE}}[m] < \tau_{\text{BB}}[m]$, the base station triggers digital beam training. When $\tau_{\text{BB}}[m] < S_{\text{UE}}[m]$, the base station keeps both the analog and digital precoders and transmits data. In the sub-6 GHz band, the threshold $\tau_{\text{RF}}[m]$ is masked to compare the spectral efficiency to determine band switching and transmission mode. When $\underline{S}_{\text{UE}}[m] < \underline{\tau}[m]$, the base station switches band to the mmWave. When $\underline{\tau}[m] < \underline{S}_{\text{UE}}[m] < \tau_{\text{BB}}[m]$, the base station tries beam training. When $\tau_{\text{BB}}[m] < \underline{S}_{\text{UE}}[m]$, the base station keeps the precoder and transmits data.

3) Reward: The reward can be written as

$$r(\mathcal{T}[m], \mathcal{A}[m]) = c(\mathcal{A}[m])R[m], \tag{26}$$

since the objective of an MDP is maximizing the cumulative reward, as in (21), over time.

The main issue with the MDP specified by (24), (25), and (26) lies in the design of the action space. Though an implicit assumption made in the MDP formulations of [32] and references therein is that the state space and action space are invariant over time. While, in the broad sense, MDPs can have an action space that varies with the state, it is an ongoing research challenge to address the increased complexity arising from estimating action relations and availability [35]. The joint band assignment and beam management problem (21) similarly introduces the dependence of the beam management procedure on the operating band, revoking the challenges from variant action spaces. One approach to alleviate such challenge is to individually learn the band selection and beam training decisions to keep the action space consistent within a single policy. In the following section, we describe how hierarchical learning can be used to solve these issues.

V. Hierarchical reinforcement learning algorithm for joint band assignment and beam management

In this section, we propose an HRL-based algorithm for solving the joint band assignment and beam management problem. We first give a brief introduction of HRL in Section V-A, focusing on the relation of state, action, and reward defined in the upper-level and lower-level policies. We then describe the proposed algorithm, a novel approach for the joint band assignment and beam management problem, incorporating off-policy correction methods and an adaptive upper-level policy period in Section V-B.

A. Hierarchical reinforcement learning

HRL algorithms build upon RL algorithms, which aim to find the policy that maximizes the cumulative reward by training neural networks. The key difference of HRL algorithms to traditional RL algorithms lies in the separation of decision layers, which represents the decomposition of the complex task given to the decision-making agent. The upper decision layer selects subtasks to be performed and the lower decision layer executes the chosen subtask. In the RL framework, the policy of the agent maps a state $\mathcal{T}$ to an action $\mathcal{A}$. HRL algorithms, depicted in Fig. 2, extend the framework to consist the upper-level policy μ^{upper} and the lower-level policy μ^{lower} [21]. The upper-level policy maps a state to a high-level action (or *goal*), where the lower-level policy maps a pair $(\mathcal{T}, g)$ to an action $\mathcal{A}$.

We use DDPG [36] to train the upper-level policy μ^{upper} and the lower-level policy μ^{lower}. Four neural networks are trained in DDPG, where each neural network corresponds to the online actor network $\boldsymbol{\theta}_{\text{A,ON}}$, the target actor network $\boldsymbol{\theta}_{\text{A,TAR}}$, the online critic network $\boldsymbol{\theta}_{\text{C,ON}}$, and the target critic network $\boldsymbol{\theta}_{\text{C,TAR}}$. The actor networks

$$L^{\text{lower}} = \frac{1}{\xi}\sum_{m'}\Bigg((r_{\text{I}}[m'] + \gamma Q^{\text{lower}}(\mathcal{T}[m'+1, g[m'+1]], \mu_{\boldsymbol{\theta}_{\text{A,TAR}}}(\mathcal{T}[m'+1], g[m'+1])|\boldsymbol{\theta}_{\text{C,TAR}})$$
$$-Q^{\text{lower}}(\mathcal{T}[m'], g[m'], \mathcal{A}[m']|\boldsymbol{\theta}_{\text{C,ON}}))^2\Bigg). \tag{22}$$

$$\frac{1}{\xi}\sum_{m'}\nabla_{\mathcal{A}}Q^{\text{lower}}(\mathcal{T}, g, \mathcal{A}|\boldsymbol{\theta}_{\text{C,ON}})|_{\mathcal{T}=\mathcal{T}[m'], g=g[m'], \mathcal{A}=\mu_{\boldsymbol{\theta}_{\text{A,ON}}}(\mathcal{T}[m'], g[m'])}\nabla_{\boldsymbol{\theta}_{\text{A,ON}}}\mu_{\boldsymbol{\theta}_{\text{A,ON}}}(\mathcal{T}, g)|_{\mathcal{T}=\mathcal{T}[m'], g=, g[m']}. \tag{23}$$

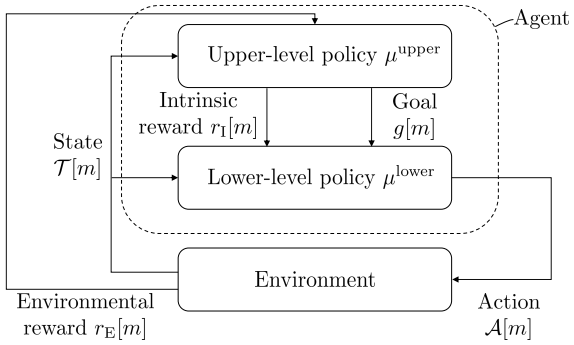

Fig. 2. Hierarchy between the upper-level and lower-level policy in the HRL framework. The upper-level policy generates goals as its action, which is inputted to the lower-level policy to determine the action interacting with the environment.

represent a policy, whereas the critic networks evaluate a policy. The target networks are delayed copies of the online networks with slow updates, which helps to reduce the effects of overfitting and instability.

DDPG uses experience replay that stores a buffer of experiences to update the neural networks. The experience replay consist of trajectories, where a single trajectory is a tuple of the state, action, reward and successor state. The trajectory of the lower-level policy is a tuple of $(\mathcal{T}[m], g[m], \mathcal{A}[m], r_{\text{I}}[m], \mathcal{T}[m+1])$, where r_{I} is the intrinsic reward provided by the upper-level policy. The update of the neural networks for the lower-level policy incorporates the goals in the typical loss minimization and policy gradient methods. Specifically, a ξ-element randomly sampled minibatch is from the experience replay of the lower-level policy, which we denote as $\mathcal{D}_{\text{lower}}$. Using the minibatch, the lower-level $\boldsymbol{\theta}_{\text{C,ON}}$ is updated by minimizing the loss in (22), the lower-level $\boldsymbol{\theta}_{\text{A,ON}}$ is updated with the policy gradient (23), and the target networks are slowly updated from the online networks, where the parameter $\eta << 1$ controls the variance of the target networks:

$$\begin{aligned}\boldsymbol{\theta}_{\text{A,TAR}} &\leftarrow \eta\boldsymbol{\theta}_{\text{A,ON}} + (1-\eta)\boldsymbol{\theta}_{\text{A,TAR}},\\ \boldsymbol{\theta}_{\text{C,TAR}} &\leftarrow \eta\boldsymbol{\theta}_{\text{C,ON}} + (1-\eta)\boldsymbol{\theta}_{\text{C,TAR}}.\end{aligned} \tag{29}$$

The parameter η can help alleviate the overestimation of the Q-values [37].

The upper-level trajectory involves multiple time steps, which we denote as M_{upper}, because the subtask sampling happens coarsely. A single upper-level transition is a tuple of

$$\begin{aligned}\Big(&\big\{\mathcal{T}[m']\big\}_{m'=m}^{m+M_{\text{upper}}-1}, \big\{g[m']\big\}_{m'=m}^{m+M_{\text{upper}}-1},\\ &\big\{\mathcal{A}[m']\big\}_{m'=m}^{m+M_{\text{upper}}-1}, \big\{r_{\text{E}}[m']\big\}_{m'=m}^{m+M_{\text{upper}}},\\ &\mathcal{T}[m+M_{\text{upper}}]\Big),\end{aligned} \tag{30}$$

where r_{E} is the reward given by the environment. To describe the usage practical actor-critic algorithms such as DDPG based on $\mathcal{D}_{\text{upper}}$, we denote the aggregated state as $\mathcal{T}^{\text{agg}}[m] = (\mathcal{T}[m], \ldots, \mathcal{T}[m+M_{\text{upper}}-1])$, aggregated goal as $g^{\text{agg}}[m] = (g[m], \ldots, g[m+M_{\text{upper}}-1])$, and cumulative environmental reward as $r_{\text{E}}^{\text{agg}} = \sum_{m'=m}^{m+M_{\text{upper}}} r_{\text{E}}[m']$.

When updating the upper-level $\boldsymbol{\theta}_{\text{C,ON}}$, minimizing the loss as in (22), an off-policy correction is required to address the varying μ^{lower} in a single upper-level trajectory. Let us denote $\mu_{\text{base}}^{\text{lower}}$ as the lower-level policy that was used when sampling $\mathcal{D}_{\text{upper}}$ and μ^{lower} as the current lower-level policy. We use importance sampling to estimate the loss function regarding the samples generated with μ^{lower} based on the experience replay $\mathcal{D}_{\text{upper}}$ from lower-level policy following $\mu_{\text{base}}^{\text{lower}}$. The direct importance correction can be written as [21]

$$w[m] = \prod_{m'=m}^{m+M_{\text{upper}}-1} \frac{\mu^{\text{lower}}(\mathcal{T}[m'], g[m'], \mathcal{A}[m'])}{\mu_{\text{base}}^{\text{lower}}(\mathcal{T}[m'], g[m'], \mathcal{A}[m'])}, \tag{31}$$

which is applied in the update of upper-level $\boldsymbol{\theta}_{\text{C,ON}}$ in the loss function (27). In case of the goal, a single goal may be selected despite the aggregated goals in the upper-level trajectory. A single goal represents the subtask selection of the upper-level policy. The corrected goal can be computed by [21]

$$\bar{g}[m] = \underset{g[m]}{\operatorname{argmin}} \left(1 - w[m]\right)^2, \tag{32}$$

which is applied in the update of the upper-level $\boldsymbol{\theta}_{\text{A,ON}}$ with the policy gradient (28). Later in the experiments, we use each off-policy correction methods as baselines.

$$\begin{aligned}L^{\text{upper}} = \frac{1}{\xi}\sum_{m'}\Big(&(r_{\text{E}}[m'] + \gamma w[m] Q^{\text{upper}}(\mathcal{T}^{\text{agg}}[m'+1], \mu_{\boldsymbol{\theta}_{\text{A,TAR}}}(\mathcal{T}^{\text{agg}}[m'+1])|\boldsymbol{\theta}_{\text{C,TAR}})\\ &- Q^{\text{upper}}(\mathcal{T}^{\text{agg}}[m'], g^{\text{agg}}[m']|\boldsymbol{\theta}_{\text{C,ON}}))^2\Big).\end{aligned} \tag{27}$$

$$\frac{1}{\xi}\sum_{m'} \nabla_{\mathcal{A}} Q^{\text{upper}}(\mathcal{T}^{\text{agg}}, g^{\text{agg}}|\boldsymbol{\theta}_{\text{C,ON}})|_{\mathcal{T}^{\text{agg}}=\mathcal{T}^{\text{agg}}[m'], g^{\text{agg}}=\mu_{\boldsymbol{\theta}_{\text{A,ON}}}(\mathcal{T}^{\text{agg}}[m'])} \nabla_{\boldsymbol{\theta}_{\text{A,ON}}} \mu_{\boldsymbol{\theta}_{\text{A,ON}}}(\mathcal{T})|_{\mathcal{T}^{\text{agg}}=\mathcal{T}^{\text{agg}}[m']}. \tag{28}$$

B. Joint band assignment and beam management strategy based on hierarchical reinforcement learning

In HRL, the upper-level policy provides its action or the goal to the lower-level policy at fixed intervals M_{upper}. However, previous work has shown that setting the duration too long or too short can result in performance deterioration [22]. On the one hand, if the duration is too long, the lower-level policy may not receive enough goals to be trained effectively. On the other hand, if the duration is too short, the upper-level policy may not capture sufficient abstraction from the environment regarding the beam management overhead to guide the lower-level policy.

To address this tradeoff, we propose the use of round skipping, which is inspired by bandit algorithms [38]. The idea is to set a short default period but periodically evaluate the interaction between the agent and the environment to determine if it is unnecessarily brief. By doing so, we can ensure that the lower-level policy receives sufficient goals for training without compromising the efficiency of task decomposition. This approach offers a more flexible and adaptive solution to the challenge of setting the upper-policy period in HRL.

The round skipping probability is computed based on the mean reward and action availability. Specifically, the non-skipping probability is $\min\{1, \frac{M_{\text{RF}}}{2M_{\text{RF}}-1}\frac{1}{q(\mathcal{A},m)}\}$, where $q(\mathcal{A}, m)$ is the probability that action $\mathcal{A}$ is available at time slot m based on the history up to time slot m. The underlying idea of deriving the terms in the round skipping probability is twofold: lower bounding the ratio between the worst-case expected reward collected by the proposed algorithm to that of the oracle and maintaining a consistent rate of action availability over time. The former involves the competitive ratio of the proposed algorithm for m time steps, which we denote as $\rho(m)$. For any time steps m, the competitive ratio can be lower bounded by at least $\rho(m) \geq \frac{M_{\text{RF}}}{2M_{\text{RF}}-1}(1-\frac{M_{\text{RF}}-1}{M_{\text{RF}}-1+m})$ [39, Theorem 1]. The latter is ensured by tracking the action availability $q(\mathcal{A}, m)$ at time slot m, inductively from the initial time step.

The upper-level actor-critic update is triggered every M_{upper} time slots. If the round-skipping occurs, the band assignment variable b and goal g is kept constant to be used in the lower-level policy computation. Otherwise, the upper-level experience replay is generated by aggregating state, action, and cumulating the environmental reward over time horizon $m, \ldots, m + M_{\text{upper}}$. In the upper-level trajectory, the length of elements are truncated to M_{RF} when $M_{\text{upper}} > M_{\text{RF}}$. The off-policy correction is applied to take account of the varying lower-level policy. The lower-level policy uses the goal g and intrinsic reward r_{I} given by the upper-level actor-critic networks.

The state $\mathcal{T}[m]$, goal $g[m]$, action $\mathcal{A}[m]$, intrinsic reward $r_{\text{I}}[m]$, and extrinsic reward $r_{\text{E}}[m]$ of the HRL-based joint band assignment and beam management algorithm can be described as the following.

1) States: The state space in the proposed HRL algorithm aligns with that of the DRL method, as described in (24). This space encompasses the beamformers and feedback used throughout the beam management procedures, as detailed in Section II-B and Section II-C.

2) Goal: The goal is associated with the operational band. It is configured as $g[m] = 1$ when the mmWave band is used; otherwise, it is set as $g[m] = 0$.

3) Action: The action space consist of two continuous variables, resembling the two thresholds that governs the beam managment procedure in the action space of the DRL method. Specifically,

$$\mathcal{A}[m] = \{\tau_{\text{A}}[m], \tau_{\text{D}}[m]\}. \tag{33}$$

The difference of the action space in the HRL method and that of the DRL method is that the band assignment is determined in the upper-level action. The spectral efficiency feedback $S_{\text{UE}}[m]$ at mmWave is compared with the thresholds to perform one of the following. When $S_{\text{UE}}[m] < \tau_{\text{A}}[m]$, the base station performs analog beam training. When $\tau_{\text{A}}[m] < S_{\text{UE}}[m] < \tau_{\text{D}}[m]$, the base station proceeds digital beam training. When $\tau_{\text{D}}[m] < S_{\text{UE}}[m]$, the base station transmits data using symbols. At sub-6 GHz, $\tau_{\text{A}}[m]$ is masked. When $\underline{S}_{\text{UE}}[m] < \tau_{\text{D}}[m]$, the base station processes beam training. When $\tau_{\text{D}}[m] < \underline{S}_{\text{UE}}[m]$, the base station transmits data using symbols.

4) Intrinsic reward: The intrinsic reward for solving (21) can be written as

$$r_{\text{I}}(\mathcal{T}[m], g[m], \mathcal{A}[m]) = c(\mathcal{A}[m])R[m]. \tag{34}$$

Note that $g[m]$ is analogous to the band assignment variable $b[m]$ discussed in Section III.

5) Extrinsic reward: The reward provided by the environment accounts for the upper-level policy period M_{upper} such that

$$r_{\text{E}}[m] = \frac{1}{M_{\text{upper}}} \sum_{m'}^{m'+M_{\text{upper}}-1} r[m'], \tag{35}$$

The pseudocode of the proposed algorithm is provided in Algorithm V-B. Algorithm V-B highlights the consists of two parts of the upper-level policy update and the lower-level policy update. The upper-level policy is updated only when the boolean random variable *RoundSkip* is false, hence in average updated every M_{upper} iterations. For completeness, Algorithm V-B and Algorithm V-B are provided to further detail the update procedure of the upper-level and lower-level policies. In Algorithm V-B, the band assignment variable $b[m]$ is computed based on the updates on the upper-level policy. In Algorithm V-B,

given the band of operation, the beam management mode is determined based on the lower-level action $\mathcal{A}[m]$ from the updates on the lower-level policy. To ensure the algorithm implementation runs within a single OFDM time slot, we note that graphics processing unit (GPU) with high clock speed and field-programmable gate array (FPGA) may be exploited as discussed in [40].

We provide an estimate of the complexity of the proposed algorithm. Let us denote N_{ℓ}^{sup} as the number of layers of a neural network with a subscript $\ell \in \{\text{A}, \text{C}\}$ describing the type of the network and a superscript $\text{sup} \in \{\text{lower}, \text{upper}\}$ indicating the level of hierarchy of the policy. For example, we denote $N_{\text{A}}^{\text{upper}}$ as the number of layers in the actor network of the upper-level policy. Let us also denote $U_{\ell,u}^{\text{sup}}$ as the number of nodes of the uth layer of a neural network. For example, we denote $U_{\text{A},u}^{\text{upper}}$ as the number of nodes of the uth layer in the actor network of the upper-level policy. Then, the computation complexity of the proposed HRL-based algorithm can be written as [41]

$$O\Bigg(\sum_{\ell\in\{\text{A,C}\}}\sum_{n=2}^{N_{\ell}^{\text{lower}}-1}(U_{\ell,u-1}^{\text{lower}}U_{\ell,u}^{\text{lower}}+U_{\ell,u}^{\text{lower}}U_{\ell,u+1}^{\text{lower}})\Bigg)$$
$$+O\Bigg(\frac{1}{M_{\text{upper}}}\sum_{\ell\in\{\text{A,C}\}}\sum_{n=2}^{N_{\ell}^{\text{upper}}-1}(U_{\ell,u-1}^{\text{upper}}U_{\ell,u}^{\text{upper}}$$
$$+U_{\ell,u}^{\text{upper}}U_{\ell,u+1}^{\text{upper}})\Bigg). \tag{36}$$

Assuming the number of nodes in each layer are similar within the same level of hierarchy of policies, each to U^{lower} for the lower-level policy and to U^{upper} for the upper-level policy, (36) can be further simplified to $O((N_{\text{A}}^{\text{lower}}+N_{\text{C}}^{\text{lower}})(U^{\text{lower}})^2+\frac{1}{M_{\text{upper}}}(N_{\text{A}}^{\text{upper}}+N_{\text{C}}^{\text{upper}})(U^{\text{upper}})^2)$.

VI. Experimental results

In this section, we evaluate the proposed HRL algorithm on a realistic multi-band wireless network. We first describe the scenario in Section VI-A. We outline the performance metric of interest and baselines in Section VI-B. We then provide the numerical results and discussion in Section VI-C.

A. Simulation setup

We simulate an urban vehicular network consisting of a static base station with a fixed transmit power in mmWave and sub-6 GHz bands and mobile vehicle nodes. We implement the Manhattan mobility model, which represents urban roads with a typical grid topology found in metropolitan cities. To generate vehicle trajectories, we employ Simulation of Urban MObility (SUMO) [42]. Among the simulated vehicles, we

Algorithm 1 Joint band assignment and beam management strategy based on HRL

Input: Length M of decision horizon, Boolean constant *UseActionRelabling*, Boolean random variable *RoundSkip*
Randomly initialize online critic network $Q(s,a|\boldsymbol{\theta}_{\text{C,ON}})$ and online actor network $\mu(s|\boldsymbol{\theta}_{\text{A,ON}})$ with $\boldsymbol{\theta}_{\text{C,ON}}$ and $\boldsymbol{\theta}_{\text{A,ON}}$ for upper-level and lower-level

for $m = 1, \ldots, M$ **do**
  if *RoundSkip* **then**
    Continue using upper-level action $g[m]$
  else
    UpperPolicyUpdate
  end if
  LowerPolicyUpdate
end for

Algorithm 1.1 UpperPolicyUpdate

Input: Current decision horizon index m, Boolean constant *UseActionRelabling*, Upper-level trajectory length M_{upper}, Experience replay $\mathcal{D}_{\text{upper}}$ of the upper-level policy, online actor and critic network of the upper-level policy
Set aggregated state as $\mathcal{T}^{\text{agg}}[m] = \mathcal{T}[m' : m' + M_{\text{upper}} - 1]$
if *UseActionRelabling* **then**
  Set goal as (32)
else
  Set aggregated goal as $g^{\text{agg}} = g[m' : m' + M_{\text{upper}} - 1]$
end if
Set reward as $\sum r_{\text{E}}[m']$
Get successor state $\mathcal{T}[m + M_{\text{upper}}]$
Store transition (30) in $\mathcal{D}_{\text{upper}}$
Sample a random minibatch of ξ transitions from $\mathcal{D}_{\text{lower}}$
Update the upper-level online critic network by minimizing the loss (27)
Update the upper-level online actor network by policy gradient (28)
Update the target networks from the online networks according to (29)
Compute upper-level action $g[m]$ by (32)
Update $b[m + M_{\text{upper}}]$

Algorithm 1.2 LowerPolicyUpdate

1: Input: Current decision horizon index m, Experience replay $\mathcal{D}_{\text{lower}}$ of the lower-level policy, online actor and critic network of the lower-level policy
2: Select lower-level action $\mathcal{A}[m]$ according to the current online actor network and exploration noise distribution $\mathcal{N}$
3: Set reward as $r_{\text{I}}[m]$ (20)
4: Update $n_{\text{mode}}[m+1]$
5: Get successor state $\mathcal{T}[m+1]$
6: Store transition $(\mathcal{T}[m], g[m], \mathcal{A}[m], r_{\text{I}}[m], \mathcal{T}[m+1])$ in $\mathcal{D}_{\text{lower}}$
7: Sample a random minibatch of ξ transitions from $\mathcal{D}_{\text{lower}}$
8: Update the lower-level online critic network by minimizing the loss (22)
9: Update the lower-level online actor network by policy gradient (23)
10: Update the target networks from the online networks according to (29)

select a single vehicle to serve as the user. We then apply the SUMO-generated vehicle trajectory to QUAsi Deterministic RadIo channel GenerAtor (QuaDRiGa), where QuaDRiGa generates the channels accounting for the geometric consideration of vehicles acting as reflectors and blockages [43]. While SUMO and QuaDRiGa can represent realistic vehicular networks as ray-tracing channel simulators, scenarios with errornous feedback from the user or malicious attacks may require a novel simulation environment. We use the 3GPP 3D Urban micro (UMi) model provided within QuaDRiGa that determines parameters such as the path, ray, complex path gain, angle of arrival, and angle of departure. At sub-6 GHz, we use the '3gpp-3d' type of antenna array provided by QuaDRiGa in accordance with the 3GPP technical report 36.873 [44].

We summarize the key simulation parameters, which are uniformly applied to simulations unless mentioned otherwise, and assumptions as the following:

1) Mobility parameters: When vehicles move through a crossroad, the probability of going straight is 0.5, turning left is 0.25, and turning right is 0.25. We set the average vehicle speed as 40 km/h and the vehicle density as 10 vehicles per kilometer.

2) Array and band parameters: We assume the number of antennas at the base station and the user are $N_{\text{BS}} = 32$ and $N_{\text{UE}} = 16$ at mmWave and $\underline{N}_{\text{BS}} = 4$ and $\underline{N}_{\text{UE}} = 4$ at sub-6 GHz. The number of streams are $N_{\text{S}} = \underline{N}_{\text{S}} = 4$ and the number of RF chain are $N_{\text{BS,RF}} = 8$ at mmWave. We assume a uniform linear array (ULA) with half-wavelength spacing used at mmWave. We assume the mmWave and sub-6 GHz arrays are co-located and aligned. Note that the aligned arrays imply that the physical line-of-sight (LOS) between the base station and the user is invariant over the sub-6 GHz and mmWave bands. We select $K = 256$ OFDM subcarriers at mmWave and $\underline{K} = 32$ subcarriers at the sub-6 GHz band. The sub-6 GHz band has 150 MHz bandwidth and the mmWave band has 850 MHz bandwidth [45].

3) Beam management parameters: In the mmWave band, we apply beam management with $M_{\text{SS}} = 1$ and $N_{\text{SS}} = 4$. We assume single bit limited feedback and set $\kappa_{\text{channel}} = \underline{\kappa}_{\text{channel}} = 1$. We assume that a discrete Fourier transform (DFT) codebook is employed at mmWave and the Type-I PMI codebook is used at sub-6 GHz.

B. Performance metric and baseline policies

We evaluate the cumulative rate as specified in (20). We approximate the ensemble mean by averaging over 1,000 channel instances generated by SUMO and QuaDRiGa. For the performance of the learning-based policy, either DRL-based or HRL-based, we measure the average of the last 20 iterations out of the $M = 200$ total iterations to represent the converged reward. We use two hidden layers with 400 nodes in the network structure of the actor and critic networks. For both upper and lower-level policies, we use the target update value $\eta = 0.005$, actor learning rate of 0.0001, and critic learning rate of 0.001. The choice of the number of iterations and the parameters of the neural network structure was made to ensure a reasonable algorithm runtime within several hours. Exploring the potential for increased rewards under less stringent computational resource constraints is an avenue for future investigation.

We compare the proposed HRL-based algorithm to three baseline policies:

- **Genie-aided** policy: This algorithm has perfect knowledge of the channel on both the mmWave and sub-6 GHz bands. Subsequently, this policy chooses the data transmission action with the correct frequency band and the best beam indices. Thus, the performance achieved by the genie-aided policy represents the theoretical upper limit of the system.
- **Three-threshold** policy: This algorithm applies DRL using threshold-based actions. The spectral efficiency feedback is compared to the learned thresholds to either perform band switching, digital beam training, analog beam training, or data transmission. The second threshold is masked when the sub-6 GHz band is selected.
- **Greedy** policy: This algorithm chooses an action in each iteration following the genie-aided policy while being restricted to mmWave. This policy

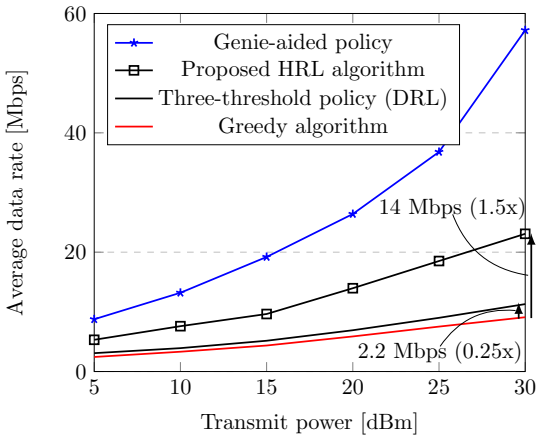

Fig. 3. Average data rate versus transmit power for (i) the genie-aided policy, (ii) the proposed HRL-based policy, (iii) DRL approach using threshold-based action, and (iv) the policy that only use the mmWave band. The distinctive beam management procedures between bands causes the DRL-based heuristic to incrementally improve over the greedy policy. Employing hierarchy between the band assignment and beam management leads to further improvement in the achieved data rate by resolving the nonstationary actions.

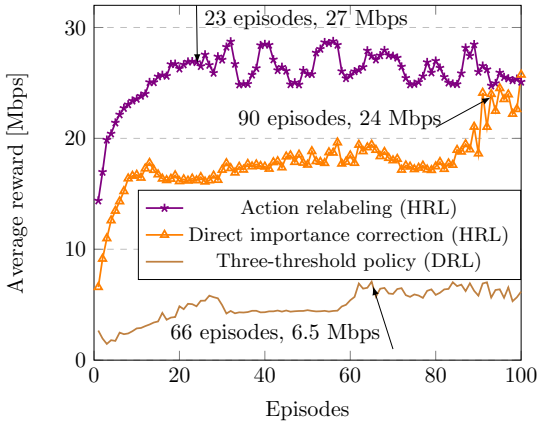

Fig. 4. Reward convergence across learning episodes compared among the HRL algorithm with action relabeling, HRL algorithm with direct importance correction, and DRL algorithm with no hierarchy. Both HRL algorithms exhibit rapid average reward increase over the DRL approach. Action relabeling aids in faster convergence, while direct importance correlation minimizes reward deviation.

represents the performance that can be achieved with beam tracking and alignment alone, without the aid of a sub-6 GHz band.

C. Numerical results and discussion

Fig. 3 shows the average data rate versus transmit power, ranging over 5 dBm to 30 dBm. The proposed band assignment and beam management algorithm based on HRL outperforms the traditional DRL-based heuristic. At a high transmit power of 30 dBm, the HRL-based algorithm shows a 2.7-fold improvement over the greedy method in contrast to the DRL-based heuristic getting 0.25-fold gain over the greedy baseline. This suggests that the HRL-based method effectively learns the policy by decomposing the joint band assignment and beam management, unlike the DRL approach, which struggles with the nonstationary action between the sub-6 GHz and mmWave band.

Fig. 4 displays a comparison of the achieved data rate over 100 training episodes between the proposed HRL-based algorithm and the traditional RL algorithm as a baseline. Additionally, we implement direct importance correction as a baseline to examine its impact on the algorithms' performance. The results demonstrate that both HRL algorithms outperform the DRL approach, exhibiting a substantial increase in average reward. We interpret that the faster increase in episodic reward is due to the enhanced sample efficiency of HRL-based methods, thanks to the extended horizon in the experience replay. Among the different off-policy correction methods, action relabeling promotes faster convergence, while direct importance correction results in less deviation of reward. The DRL-based method takes around 60 episodes to converge at approximately 6.5 Mbps, whereas the HRL algorithms can achieve up to 27 Mbps. Notably, the importance-based action relabeling leads to the fastest convergence in approximately 20 episodes, while the direct importance correction method takes around 90 episodes to achieve more than 24 Mbps. We observed hours of runtime using a simulation environment with a GTX 1080 GPU to achieve the 27 Mbps of the HRL algorithm throughout 20 episodes. Still, base station deployments typically last for tens of years. This indicates that the investment of time in training is justified by the long-term performance benefits.

Fig. 5 displays a comparison of the achieved data rate over 100 training episodes between HRL algorithms with different approaches to set the upper-level policy period. The proposed method based on round-skipping that adapts the upper-level period shows the best convergence behavior, converging to 27 Mbps around 20 episodes. While moderately short fixed upper-level period shows convergence around 20 episodes as well, the achieved reward drops down below 20 Mbps. When the fixed upper-level period is excessive, matching the analog beam training overhead, the performance is comparable to the vanilla DRL approach taking over 40 episode to converge to the reward below 10 Mbps. This highlights the importance of adaptively adjusting the upper-policy trajectory sampling period in achieving better perfor-

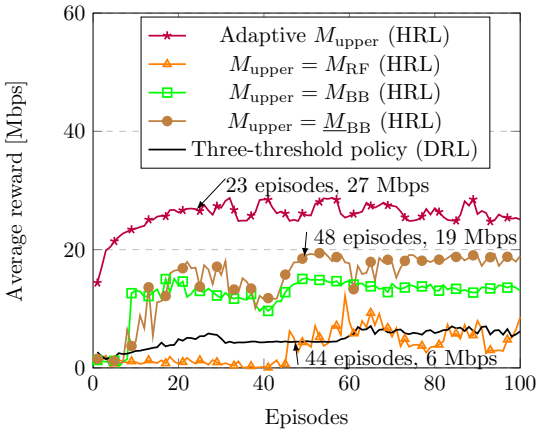

Fig. 5. Convergence behavior over the period of the upper-level policy. The proposed method that uses round-skipping outperforms the baselines that have a fixed period of upper-policy trajectory sampling. The performance of the fixed period policies can deteriorate down to the vanilla DRL appraoch, losing the benefits of using the hierarchical structured learning.

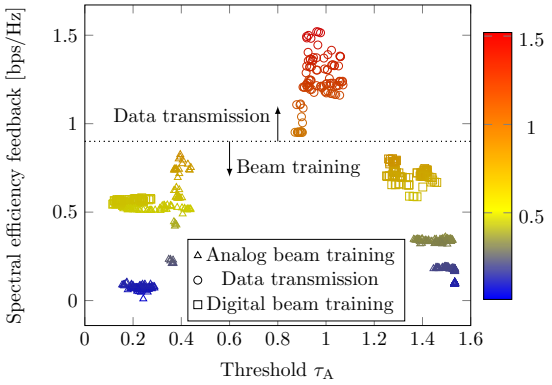

Fig. 6. Beam training behavior per the learned threshold and spectral efficiency feedback at the mmWave band. Three clusters are observed: data transmission, analog beam training, and digital beam training. Clusters form based on spectral efficiency feedback and threshold levels, with high spectral efficiency for data transmission, moderate spectral efficiency and high thresholds for digital beam training, and low spectral efficiency for analog beam training.

mance in the HRL setting, where round-skipping is an effective way to adjust the upper-policy period.

Fig. 6 shows the threshold and spectral efficiency feedback datapoints of the proposed HRL-based algorithm at the mmWave band during episodes 1-40 of the learning phase. The data points are color-coded in a heatmap style, ranging from red to blue to represent high to low spectral efficiency feedback. Clusters are evident, with data transmission occurring when spectral efficiency feedback is over 0.9 bps/Hz and the threshold is between 0.8 and 1.2. Beam training occurs when the feedback deteriorates, with a high threshold indicating digital beam training and a low threshold indicating analog beam training. The cluster formation indicates that the threshold-based action in the HRL algorithm enables efficient beam management per spectral efficiency feedback.

Fig. 7 shows the average data rate per the number of quantization bits of the RVQ codebook used in the digital effective channel feedback. The range of the bits is selected from 1 through 11. Increasing the quantization codebook bits from 1 shows an increase in the average data rate since the digital effective channel feedback will become more accurate. The increase in the average data rate continues up to the quantization bits of 5 for the three-threshold policy and 8 for the proposed HRL-based method. We interpret that the DRL approach is bound to the number of SS blocks $N_{\text{SS}} = 4$ whereas the HRL-based method can further benefit from the accurate digital effective channel to increase the achievable rate. Moreover, the proposed HRL-based algorithm outperforms the greedy approach over the codebook quantization bits ranging from 1 through 11, where for the three-threshold policy the quantization codebook is preferred within 4 to 8 to outperform the greedy baseline. The HRL-based method exhibits a limitation, resulting in a higher data rate loss with lower RVQ codebook quantization bits compared to the DRL-based heuristic. The robustness of HRL-based methods would need careful consideration, especially in scenarios where the feedback from the receiver can be erroneous [46].

Fig. 8 shows the average data rate achievable per vehicle density, ranging from 10 to 40 vehicles per kilometer in the SUMO simulation, under different QuaDRiGa scenarios. The solid lines represent the performance of the policies under the 3GPP-UMi LOS scenario, while the dotted lines depict the performance of the policies under the 3GPP-UMi NLOS scenario. As the vehicle density increases, resulting in a higher likelihood of blockages, the achievable data rate decreases. However, our observation reveals that the exploitation of the LOS channel in the HRL-based method experiences a comparatively lesser performance loss in contrast to the NLOS scenario. The performance of the proposed HRL algorithm outperforms the vanilla DRL approach and the greedy algorithm under the LOS scenario over the increasing vehicle density, where we observe the comparable degradation of policy performance. The proposed method also outperforms the baselines under the NLOS scenario, but the performance degradation is more severe under the NLOS scenario over the increasing vehicle density. This observation may be due to the fact that the Type-1 codebook in the sub-6 GHz band is designed

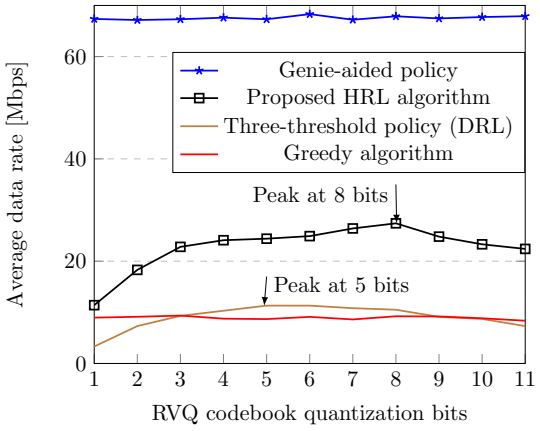

Fig. 7. Achievable rate per the number of bits used in the digital effective channel quantization. Learning-based methods show a peaked curve in performance, with data rate increasing at low quantization due to more accurate channel estimates, but decreasing at excessive quantization due to significant overhead. The proposed HRL-based method outperforms the baseline DRL-based method, with the HRL-based method saturating at 8 bits and the DRL-based method at 5 bits.

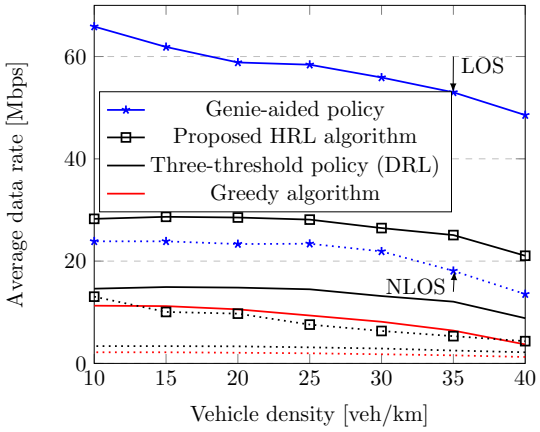

Fig. 8. Achievable rate per different vehicle densities. The solid line shows the 3GPP-UMi LOS scenario and the dotted line shows the 3GPP-UMi NLOS scenario. Increased vehicle density in urban scenarios can reduce the achievable data rate due to more frequent blockages, but the HRL-based method can exploit the LOS channel to achieve milder performance loss compared to the NLOS scenario.

for LOS conditions to enjoy the short beam training overhead while maintaining high data rate. An interesting future direction in this regard would be to consider a Type-2 codebook with more sophisticated precoder computation that accounts for the multipath channel in the sub-6 GHz band.

VII. Conclusions and Future Work

Exploiting multiple band characteristics will be a major approach addressing challenges in wireless networks, including mobility and blockage, while avoiding their associated drawbacks. We formulated the joint band assignment and beam management problem in wireless networks operating on FR1 and FR2. We devised an MDP that introduces hierarchy between the band assignment and beam management to avoid nonstationary action space. The numerical evaluation based on QuaDRiGa-generated channel showed that the proposed HRL-based method improves over traditional DRL approaches. This suggests that the introduction of hierarchy is an effective approach addressing the complex problem of joint band assignment and beam management. For future work, the extension to multi-user scenario is an interesting direction that may require queuing theory to resolve conflicts between users with the same preferred band. The robustness of the RL methods also needs further investigation to ensure their suitability in scenarios where feedback from the receiver might be erroneous.

References

[1] J. Zhao, S. Ni, L. Yang, Z. Zhang, Y. Gong, and X. You, "Multiband cooperation for 5G HetNets: A promising network paradigm," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 85–93, Dec. 2019.

[2] L. Yan, H. Ding, L. Zhang, J. Liu, X. Fang, Y. Fang, M. Xiao, and X. Huang, "Machine learning-based handovers for sub-6 GHz and mmWave integrated vehicular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4873–4885, Oct. 2019.

[3] X. Lin, "An overview of 5G advanced evolution in 3GPP release 18," *IEEE Commun. Standards Mag.*, vol. 6, no. 3, pp. 77–83, 2022.

[4] F. J. Martin-Vega, M. C. Aguayo-Torres, G. Gomez, J. T. Entrambasaguas, and T. Q. Duong, "Key technologies, modeling approaches, and challenges for millimeter-wave vehicular communications," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 28–35, 2018.

[5] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 1st Quart. 2019.

[6] Z. L. Fazliu, F. Malandrino, C. F. Chiasserini, and A. Nordio, "A belief propagation solution for beam coordination in mmWave vehicular networks," *IEEE Trans. Wireless Commun.*, Dec. 2022.

[7] Y. Li, S. K. Jayaweera, M. Bkassiny, and C. Ghosh, "Learning-aided sub-band selection algorithms for spectrum sensing in wide-band cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2012–2024, Apr. 2014.

[8] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent reinforcement learning based cognitive anti-jamming," in *Proc. IEEE Wireless Commun. Netw. Conf.* IEEE, Mar. 2017, pp. 1–6.

[9] M. Najla, P. Mach, and Z. Becvar, "Deep learning for selection between RF and VLC bands in device-to-device communication," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1763–1767, Oct. 2020.

[10] D. Burghal, R. Wang, A. Alghafis, and A. F. Molisch, "Supervised ML solution for band assignment in dual-band systems with omnidirectional and directional antennas," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, Sep. 2022.

[11] D. Kim, M. R. Castellanos, and R. W. Heath, "Joint relay selection and beam management based on deep reinforcement learning for millimeter wave vehicular communication," *IEEE Trans. Veh. Technol. early access*, May 2023, doi: 10.1109/TVT.2023.3274763.

[12] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2487–2500, Dec. 2018.

[13] M. Hussain and N. Michelusi, "Learning and adaptation for millimeter-wave beam tracking and training: a dual timescale variational framework," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 37–53, Jan. 2022.

[14] J. Tan, L. Zhang, Y.-C. Liang, and D. Niyato, "Intelligent sharing for LTE and WiFi systems in unlicensed bands: A deep reinforcement learning approach," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2793–2808, May 2020.

[15] A. Alwarafy, M. Abdallah, B. S. Çiftler, A. Al-Fuqaha, and M. Hamdi, "The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 322–365, Feb. 2022.

[16] G. Pang, W. Liu, Y. Li, and B. Vucetic, "DRL-based resource allocation in remote state estimation," *IEEE Trans. Wireless Commun.*, vol. 22, no. 7, Jul. 2023.

[17] L. Liu, J. Feng, X. Mu, Q. Pei, D. Lan, and M. Xiao, "Asynchronous deep reinforcement learning for collaborative task computing and on-demand resource allocation in vehicular edge computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15 513 – 15 526, Mar. 2023.

[18] T.-W. Weng, K. D. Dvijotham, J. Uesato, K. Xiao, S. Gowal, R. Stanforth, and P. Kohli, "Toward evaluating robustness of deep reinforcement learning with continuous control," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2020.

[19] V. H. L. Lopes, C. V. Nahum, R. M. Dreifuerst, P. Batista, A. Klautau, K. V. Cardoso, and R. W. Heath, "Deep reinforcement learning-based scheduling for multiband massive MIMO," *IEEE Access*, vol. 10, pp. 125 509–125 525, Dec. 2022.

[20] M. Eppe, C. Gumbsch, M. Kerzel, P. D. Nguyen, M. V. Butz, and S. Wermter, "Intelligent problem-solving as integrated hierarchical reinforcement learning," *Nat. Mach. Intell.*, vol. 4, no. 1, pp. 11–20, Jan. 2022.

[21] O. Nachum, S. S. Gu, H. Lee, and S. Levine, "Data-efficient hierarchical reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 3303–3313.

[22] Y. Geng, E. Liu, R. Wang, and Y. Liu, "Hierarchical reinforcement learning for relay selection and power optimization in two-hop cooperative relay network," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 171–184, Jan. 2022.

[23] S. Liu, J. Wu, and J. He, "Dynamic multichannel sensing in cognitive radio: Hierarchical reinforcement learning," *IEEE Access*, vol. 9, pp. 25 473–25 481, Feb. 2021.

[24] T. Ren, J. Niu, B. Dai, X. Liu, Z. Hu, M. Xu, and M. Guizani, "Enabling efficient scheduling in large-scale UAV-assisted mobile-edge computing via hierarchical reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7095–7109, May 2022.

[25] R. W. Heath Jr and A. Lozano, *Foundations of MIMO communication*. Cambridge, U.K.: Cambridge University Press, 2018.

[26] X. Sun, C. Qi, and G. Y. Li, "Beam training and allocation for multiuser millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1041–1053, Feb. 2019.

[27] C. Wang, E. K. Au, R. D. Murch, W. H. Mow, R. S. Cheng, and V. Lau, "On the performance of the MIMO zero-forcing receiver in the presence of channel estimation error," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 805–810, Mar. 2007.

[28] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[29] D. R. Brown III and D. J. Love, "On the performance of MIMO nullforming with random vector quantization limited feedback," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2884–2893, May 2014.

[30] *NR; Physical layer procedures for data*, Standard 3GPP TS 38.214 V15.6.2 Jun. 2019. [Online]. Available: https://www.3gpp.org/DynaReport/38214.htm

[31] H. Lee, H. Choi, H. Kim, S. Kim, C. Jang, Y. Choi, and J. Choi, "Downlink channel reconstruction for spatial multiplexing in massive MIMO systems," *IEEE Tran. Wireless Commun.*, vol. 20, no. 9, pp. 6154–6166, Sep. 2021.

[32] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart. 2019.

[33] A. Kanervisto, C. Scheller, and V. Hautamäki, "Action space shaping in deep reinforcement learning," in *Proc. IEEE Conf. Games*. IEEE, Aug. 2020, pp. 479–486.

[34] S. Huang and S. Ontañón, "A closer look at invalid action masking in policy gradient algorithms," in *Proc. Int. Florida Artif. Intell. Res. Soc*, May 2022.

[35] A. Jain, N. Kosaka, K.-M. Kim, and J. J. Lim, "Know your action set: Learning action relations for reinforcement learning," in *Int. Conf. Learn. Represent.*, May 2021.

[36] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Int. Conf. Learn. Representations*, 2016.

[37] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, vol. 80, Jul. 2018, pp. 1587–1596.

[38] J. P. Dickerson, K. A. Sankararaman, A. Srinivasan, and P. Xu, "Allocation problems in ride-sharing platforms: Online matching with offline reusable resources," *ACM Trans. Econ. Comput.*, vol. 9, no. 3, pp. 1–17, Jun. 2021.

[39] S. Basu, O. Papadigenopoulos, C. Caramanis, and S. Shakkottai, "Contextual blocking bandits," in *Int. Conf. Artif. Intell. Stat.* PMLR, Apr. 2021, pp. 271–279.

[40] Y. Ma, Z. Wang, H. Yang, and L. Yang, "Artificial intelligence applications in the development of autonomous vehicles: a survey," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 2, pp. 315–329, Mar. 2020.

[41] F. Rasheed, K.-L. A. Yau, R. M. Noor, C. Wu, and Y.-C. Low, "Deep reinforcement learning for traffic signal control: A review," *IEEE Access*, vol. 8, pp. 208 016–208 044, Nov. 2020.

[42] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO-Simulation of Urban MObility," *Int. J. Adv. Syst. Meas.*, vol. 5, no. 3-4, Dec. 2012.

[43] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Mar. 2014.

[44] *Study on 3D Channel Model for LTE*, Standard 3GPP TR 36.873 V12.5.0 Jun. 2017. [Online]. Available: https://www.3gpp.org/DynaReport/36873.htm

[45] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.

[46] Z. Yin, N. Cheng, Y. Hui, W. Wang, L. Zhao, K. Aldubaikhy, and A. Alqasir, "Multi-domain Resource Multiplexing Based Secure Transmission for Satellite-Assisted IoT: AO-SCA Approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7319–7330, Nov. 2023.