HRL-based Joint Band Assignment and Beam Management in 4G/5G Multi-Band Networks

Dohyun Kim
Inst. of Comput. Technol.
Seoul National University
Seoul 08826, South Korea
dohyun.p.kim@snu.ac.kr

Miguel R. Castellanos
Dept. of Elect. and Comput. Eng.
North Carolina State University
Raleigh, NC 27606, United States
mrcastel@ncsu.edu

Robert W. Heath Jr.
Dept. of Elect. and Comput. Eng.
University of California, San Diego
La Jolla, CA 92093, United States
rwheathjr@ucsd.edu

***Abstract*—Incorporating the sub-6 GHz band into 5G networks can improve data rates by leveraging the benefits of propagation across frequency ranges. Sequential decision making algorithms such as deep reinforcement learning (DRL) can adaptively select a band over time to take full advantage of the multi-band operation. The distinctive beam management procedure between the sub-6 GHz band and the millimeter wave (mmWave) band, though, pose a sample efficiency challenge for DRL algorithms. In this paper, we use hierarchical reinforcement learning (HRL) to divide and conquer the joint band assignment and beam management problem. The proposed HRL-based method uses rate feedback for intermittent band determination and frequent beam management mode decisions. We show with numerical evaluation that the proposed algorithm achieves a quicker increase in data rate compared to baselines and identify off-policy correction methods as a key factor for this enhancement.**

I. Introduction

Multi-band operation in 5G networks can enhance both data rates from the mmWave band and link resiliency from the sub-6 GHz band [1]. While allowing simultaneous usage of bands in a single time slot offers greater data rate potential, a sophisticated scheduling algorithm involving high radio-frequency (RF) complexity may be required. Band assignment, which refers to the selection of the operating band over time, alleviates the complexity of multi-band operation and can be suitable for user devices with low RF processing capability [2].

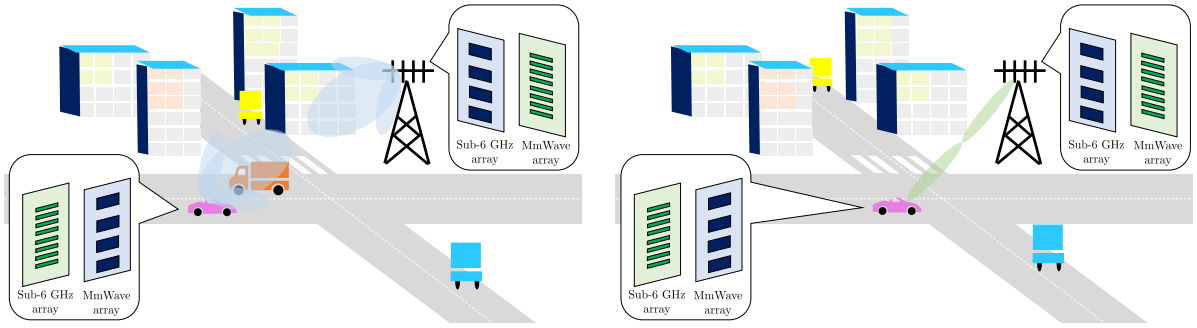
While several solutions have been proposed for band assignment [2] (and references therein) they have not incorporated the overhead of beam management, which can be a significant bottleneck in achieving high data rates. The formulation of joint band assignment and beam management is required, since evaluating a band involves beam management distinctive across bands.

DRL algorithms are well known in addressing the exploration-exploitation tradeoff in resource allocation problems [3]. Beam management in multi-band wireless networks, though, can be challenging based on traditional DRL approaches because beam training can only be performed in one band at a time and the sample efficiency in each band will be low. HRL is a recent advancement of DRL that introduces hierarchy in the learning process [4]. HRL is a viable approach for the joint band assignment and beam management problem, as it separates band assignment and beam management, improving sample efficiency while accommodating distinct beam management procedures across bands.

In this paper, we propose an HRL-based algorithm that leverages rate feedback to determine the operating band and when to perform beam training. We assume 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. We further assume perfect rate feedback from the user to the base station without quantization or overhead. 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 rate feedback and two adaptive thresholds determined by the lower-level policy. 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 rate 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.

II. System model and problem formulation

We assume a downlink scenario in a multi-band MIMO-OFDM wireless network, as in Fig. 1, 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 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



(a) (b)

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 line-of-sight is available.

deployed. When the system uses the mmWave band, the system uses a bandwidth B with K subcarriers. When the system uses the sub-6 GHz band, the system operates over a bandwidth $\underline{B}$ with $\underline{K}$ subcarriers. Hereinafter, we underline the sub-6 GHz parameters.

Each node employs a fully connected hybrid beamforming architecture in the mmWave band and digital beamforming architecture in the sub-6 GHz band. Fully connected hybrid architecture is selected for simplicity and we plan to extend to partially connected hybrid architecture for future work. In the mmWave band, we denote N_{BS} as the number of antennas and $N_{\text{BS,RF}}$ as the number of RF chains at the base station, N_{UE} as the number of antennas and $N_{\text{UE,RF}}$ as the number of RF chains at the user, and N_{S} as the number of data streams. We denote $\mathbf{F}_{\text{BB}}[k,m]$ as the $N_{\text{BS,RF}} \times N_{\text{S}}$ frequency-selective baseband precoder, $\mathbf{F}_{\text{RF}}[m]$ as the $N_{\text{BS}} \times N_{\text{BS,RF}}$ frequency-flat RF precoder, $\mathbf{W}_{\text{RF}}[m]$ as the $N_{\text{UE}} \times N_{\text{UE,RF}}$ frequency-flat RF combiner, and $\mathbf{W}_{\text{BB}}[k,m]$ as the $N_{\text{UE,RF}} \times N_{\text{S}}$ frequency-selective baseband combiner. 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 $\|\mathbf{F}_{\text{RF}}[k,m]\mathbf{F}_{\text{BB}}[k,m]\|_F^2 = N_{\text{S}}$. No other hardware-related constraints are assumed.

We further assume a time-varying frequency-selective $N_{\text{UE}} \times N_{\text{BS}}$ channel matrix $\mathbf{H}[k,m]$, where k, m are the subcarrier and time index. We denote $G[m]$ as the large-scale fading, $\mathbf{n}[k,m]$ as the identically distributed (IID) noise following the distribution $\mathcal{N}_C(0,\sigma_{\text{n}}^2)$, and $\mathsf{s}[k,m]$ as the symbol vector with $\mathbb{E}[|\mathsf{s}[k,m]|^2] = 1$. The end-to-end input-to-output relation in the mmWave band is

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

We use the instantaneous spectral efficiency [5] averaged over the subcarriers, denoted as $S[m,\mathbf{H}[k,m]]$, in defining the performance metric later in (3).

The user measures the instantaneous spectral efficiency and feedback the rate estimate of current band and beam, denoted as the beam measurement $S_{\text{UE}}[m,\mathbf{H}[k,m]]$, to the base station. We assume the system uses the beam measurement to determine the best band and beamformers. The beam measurement at each time horizon m can be written as

$$\begin{aligned}S_{\text{UE}}[m;\mathbf{H}[k,m]] = &\frac{1}{K}\sum_{k=1}^{K}\log_2(1+\text{SNR}_{\text{eff}} \\ &\times |\mathbf{W}_{\text{RF}}^*[m]\mathbf{H}[k,m]\mathbf{F}_{\text{RF}}[m]|^2), \end{aligned} \quad (2)$$

where SNR_{eff} is the effective SNR accounting for the MMSE channel estimator under a rectangular Doppler spectrum [5, Sec. 4.8].

We assume a greedy approach to configure the analog beamformers $\mathbf{F}_{\text{RF}}[m]$ and $\mathbf{W}_{\text{RF}}[m]$ for simplicity, where distinct beams are used for separate RF chains to achieve spatial multiplexing gain [6]. To subsequently determine the digital beamformers $\mathbf{F}_{\text{BB}}[k,m]$ and $\mathbf{W}_{\text{BB}}[k,m]$, we assume the digital effective channel is fed back from the user to the base station via the random vector quantization (RVQ) codebook [7]. In the sub-6 GHz band, we presume Type-1 precoding matrix indicator (PMI) codebook is employed and the PMI feedback indicates the PMI table index, which includes both candidate precoders and the channel quantization [8].

The beam training overhead M_{BT} varies over frequency bands as listed in Table I in decreasing order of length. The mmWave analog beam training overhead depends on the number N_{SS} of synchronization signal (SS) blocks per burst and periodicity M_{SS} between two SS burst exchangements [9]. The sub-6 GHz beam training overhead depends on the size ν_{PMI} of the PMI codebook and the number $\underline{\kappa}_{\text{channel}}$ of bits that can be sent

TABLE I: Closed-form expressions of beam training overhead

Beam training type	Overhead
MmWave analog	$M_{\text{SS}}\lceil \nu_{\text{BS}}\ \nu_{\text{UE}}/N_{\text{SS}} \rceil$
Sub-6 GHz	$\lceil \log_2 \nu_{\text{PMI}}/\underline{\kappa}_{\text{channel}} \rceil$
MmWave digital	$\lceil \kappa_{\text{RVQ}}/\kappa_{\text{channel}} \rceil$

through the sub-6 GHz feedback channel over a single time slot. The mmWave digital beam training overhead depends on the number κ_{RVQ} of quantization bits of the RVQ codebook and the number κ_{channel} of bits sent through the feedback channel per time slot.

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 data transmission and $n_{\text{mode}}[m] = 0$ to indicate beam training. The system's data rate, which is the performance metric of interest, can be written as

$$R[m] = (1 - b[m])\underline{B}\ \underline{S}[m] + b[m]B\ S[m]. \quad (3)$$

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

$$\max_{\{\mathcal{A}[m]\}} \sum_{m=1}^{M} c(\mathcal{A}[m])R[m]. \quad (4)$$

RL is a well known approach for solving optimization problems like (4) as in [3] and references therein. Furthermore, the distinct beam management procedures in the mmWave band and sub-6 GHz band suggest that exploiting an hierarchical structure in decision making can further improve the learning algorithm. In this regard, we propose a HRL-based algorithm in Section III.

III. HRL-BASED ALGORITHM FOR JOINT BAND ASSIGNMENT AND BEAM MANAGEMENT

HRL algorithms build upon DRL algorithms, which aim to find the policy that maximizes the cumulative reward by training neural networks. The key difference of HRL algorithms to traditional DRL 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 DRL framework, the policy of the agent maps a state $\mathcal{T}$ to an action $\mathcal{A}$ and receive a reward. HRL algorithms, depicted in Fig. 2, extend the framework to consist the upper-level policy μ^{upper} and the lower-level policy μ^{lower} [4]. 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 $\mathcal{A}$. The environment provides the extrinsic reward r_{E} to the upper-level policy, whereas the intrinsic reward r_{I} is given to the lower-level policy by the upper-level policy.

In HRL, the upper-level policy provides its action or the goal to the lower-level policy per periods M_{upper}. To adaptively determine M_{upper}, we propose the use of round skipping, inspired by bandit algorithms [10]. Round skipping ensures a short default period while avoiding unnecessary goals to the lower-level policy. 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 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 incorporates the selected beamformers and feedback used throughout the beam management procedures as discussed in Section II. The state can be written as

$$\mathcal{T}[m] = \Big\{\mathbf{F}_{\text{RF}}[m], \mathbf{W}_{\text{RF}}[m], S_{\text{UE}}[m], \big\{\hat{\mathbf{H}}[k,m]\big\}_{k=1}^{K}, \big\{\underline{\mathbf{F}}_{\text{BB}}[k,m]\big\}_{k=1}^{\underline{K}}, \big\{\mathbf{P}[k,m]\big\}_{k=1}^{\underline{K}}\Big\}. \quad (5)$$

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

2) Goal: The goal $g[m]$ corresponds to the band of operation and set to $g[m] = b[m]$.

3) Action: The action space consist of two continuous variables

$$\mathcal{A}[m] = \{\tau_{\text{A}}[m], \tau_{\text{D}}[m]\}. \quad (6)$$

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, 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.

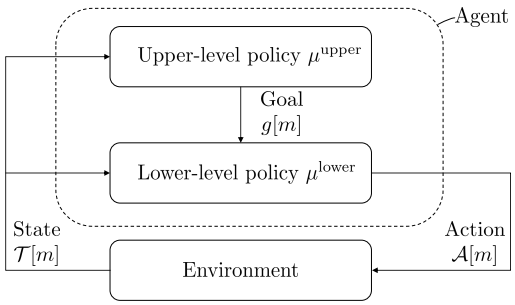


Fig. 2: Hierarchy between the upper-level and lower-level policy in the HRL framework. The upper-level policy generates goals as its action, which is inputted to the lower-level policy to determine the action interacting with the environment.

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

$$r_{\text{I}}(\mathcal{T}[m], g[m], \mathcal{A}[m]) = c(\mathcal{A}[m])R[m]. \tag{7}$$

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

$$r_{\text{E}}[m] = \frac{1}{M_{\text{upper}}} \sum_{m'}^{m'+M_{\text{upper}}-1} r_{\text{I}}[m'], \tag{8}$$

We use deep deterministic policy gradient (DDPG) [11] 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 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 $(\mathcal{T}[m], g[m], \mathcal{A}[m], r_{\text{I}}[m], \mathcal{T}[m+1])$. The update of the neural networks for the lower-level policy incorporates the goals in the typical DDPG update. 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 and the lower-level $\boldsymbol{\theta}^{\text{A,ON}}$ is updated with the policy gradient, and the target networks are slowly updated from the online networks [4]. We denote such group of updates as *Update*($\boldsymbol{\theta}^{\text{A,ON}}_{\text{lower}}, \boldsymbol{\theta}^{\text{C,ON}}_{\text{lower}}, \boldsymbol{\theta}^{\text{A,TAR}}_{\text{lower}}, \boldsymbol{\theta}^{\text{C,TAR}}_{\text{lower}}; \mathcal{D}_{\text{lower}}, \xi$) in Algorithm 1. The upper-level policy involves a transition as a tuple of aggregated state, goal, action, and extrinsic reward over the horizon window of length M_{upper} The neural network update for uppper-level policy is similarly performed using *Update*($\boldsymbol{\theta}^{\text{A,ON}}_{\text{upper}}, \boldsymbol{\theta}^{\text{C,ON}}_{\text{upper}}, \boldsymbol{\theta}^{\text{A,TAR}}_{\text{upper}}, \boldsymbol{\theta}^{\text{C,TAR}}_{\text{upper}}; \mathcal{D}_{\text{upper}}, \xi$).

When updating the upper-level $\boldsymbol{\theta}^{\text{C,ON}}$, an off-policy correction is required to address the varying μ^{lower} in a single upper-level trajectory. We apply the direct importance correction and goal correction based on importance relabling as in [4]. Later in the experiments, we use each off-policy correction methods as baselines.

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 $M_{\text{upper}} > M_{\text{RF}}$. For completeness, the pseudocode is given in Algorithm 1.

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*, Batch sample size ξ
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
    Set aggregated state as $\mathcal{T}^{\text{agg}}[m] = \mathcal{T}[m' : m' + M_{\text{upper}} - 1]$
    Set goal as according to importance relabling
    Set reward as $\sum r_{\text{E}}[m']$
    Get successor state $\mathcal{T}[m + M_{\text{upper}}]$
    Store upper-level transition in $\mathcal{D}_{\text{upper}}$
    Update($\boldsymbol{\theta}^{\text{A,ON}}_{\text{upper}}, \boldsymbol{\theta}^{\text{C,ON}}_{\text{upper}}, \boldsymbol{\theta}^{\text{A,TAR}}_{\text{upper}}, \boldsymbol{\theta}^{\text{C,TAR}}_{\text{upper}}; \mathcal{D}_{\text{upper}}, \xi$)
    Update $b[m + M_{\text{upper}}]$
  end if
  Select lower-level action $\mathcal{A}[m]$ according to $\boldsymbol{\theta}^{\text{A,ON}}_{\text{lower}}$ and exploration noise distribution $\mathcal{N}$
  Set reward $r_{\text{I}}[m]$ as in (7)
  Update $n_{\text{mode}}[m+1]$
  Get successor state $\mathcal{T}[m+1]$
  Store transition $(\mathcal{T}[m], g[m], \mathcal{A}[m], r_{\text{I}}[m], \mathcal{T}[m+1])$ in $\mathcal{D}_{\text{lower}}$
  Update($\boldsymbol{\theta}^{\text{A,ON}}_{\text{lower}}, \boldsymbol{\theta}^{\text{C,ON}}_{\text{lower}}, \boldsymbol{\theta}^{\text{A,TAR}}_{\text{lower}}, \boldsymbol{\theta}^{\text{C,TAR}}_{\text{lower}}; \mathcal{D}_{\text{lower}}, \xi$)
end for

IV. Numerical results

In this section, we assess the HRL algorithm on a realistic multi-band wireless network. We outline simulation parameters, baselines, and analyze the results.

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) [12]. We set the average vehicle speed as 40 km/h and the vehicle density as 10 vehicles per kilometer. Among the simulated vehicles, we 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 [13]. 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 [14].

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. 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 [15]. 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. Baseline policies and numerical evaluation

We evaluate the cumulative rate as specified in (3). 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 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 represents the performance that can be achieved with beam tracking and alignment alone, without the aid of a sub-6 GHz band.

Fig. 3 shows the average data rate versus transmit power, ranging over 5 dBm to 30 dBm. The proposed HRL-based algorithm 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 DRL 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. 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

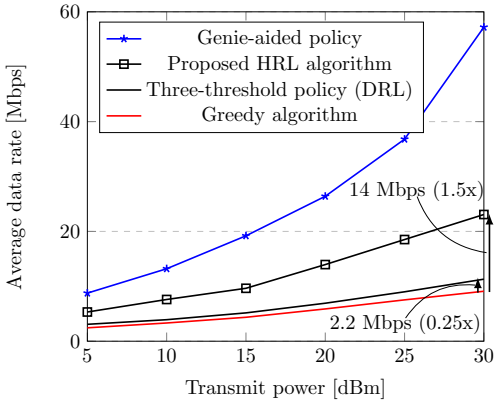


Fig. 3: Illustration of average data rate versus transmit power for the proposed HRL-based algorithm to baselines. The distinctive beam management procedures between bands causes the DRL-based heuristic to incrementally improve over the greedy policy. Employing hierarchy between the band assignment and beam management leads to further improvement in the achieved data rate by resolving the nonstationary actions.

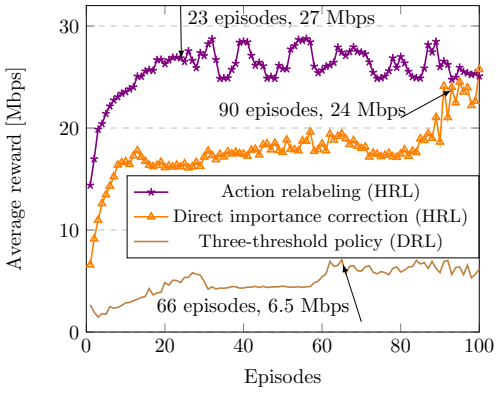


Fig. 4: Illustration of reward convergence per episodes for HRL with action relabeling, HRL with direct correction, and DRL without hierarchy. Both HRL methods show steeper average reward increase compared to DRL. Action relabeling aids faster convergence, while direct correction reduces reward deviation.

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.

V. Conclusions

In this paper we formulated the joint band assignment and beam management problem in 5G networks operating on the sub-6 GHz band and mmWave band. We devised an MDP and a corresponding HRL algorithm that assigns bands followed up by beam management. The numerical evaluation based on QuaDRiGa-generated channel showed that the proposed HRL-based method achieves 1.5-fold data rate gain compared to the traditional DRL baselines. Furthermore, numerical results on episodic reward demonstrate that off-policy correction is a key enabler of the fast reward gains achieved by the proposed HRL-based algorithm.

VI. Acknowledgments

The authors would like to acknowledge support in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program, the National Science Foundation under grant nos. NSF-ECCS-2153698, NSF-CCF-2225555, NSF-CNS-2147955, the National Research Foundation of Korea (NRF) grant, and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (Nos. RS-2023-00222663 and 2018-0-00581).

References

[1] L. Yan *et al.*, "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.

[2] 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.

[3] N. C. Luong *et al.*, "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.

[4] O. Nachum *et al.*, "Data-efficient hierarchical reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 3303–3313.

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

[6] X. Sun *et al.*, "Beam training and allocation for multiuser millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1041–1053, Feb. 2019.

[7] A. Alkhateeb *et al.*, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[8] *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

[9] D. Kim *et al.*, "Joint relay selection and beam management based on deep reinforcement learning for millimeter wave vehicular communication," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13 067–13 080, Oct. 2023.

[10] J. P. Dickerson *et al.*, "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.

[11] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[12] D. Krajzewicz *et al.*, "Recent development and applications of SUMO-Simulation of Urban MObility," *Int. J. Adv. Syst. Meas.*, vol. 5, no. 3-4, Dec. 2012.

[13] S. Jaeckel *et al.*, "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.

[14] *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

[15] A. Ali *et al.*, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.