www.czasopisma.pan.pl PAN www.journals.pan.pl

INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2021, VOL. 67, NO. 1, PP. 95-102
Manuscript received May 13, 2020; revised January, 2021. DOI: 10.24425/ijet.2021.135949

Millimeter Wave Beamforming Training: A Reinforcement Learning Approach

Ehab Mahmoud Mohamed

***Abstract*—Beamforming training (BT) is considered as an essential process to accomplish the communications in the millimeter wave (mmWave) band, i.e., 30 ~ 300 GHz. This process aims to find out the best transmit/receive antenna beams to compensate the impairments of the mmWave channel and successfully establish the mmWave link. Typically, the mmWave BT process is highly-time consuming affecting the overall throughput and energy consumption of the mmWave link establishment. In this paper, a machine learning (ML) approach, specifically reinforcement learning (RL), is utilized for enabling the mmWave BT process by modeling it as a multi-armed bandit (MAB) problem with the aim of maximizing the long-term throughput of the constructed mmWave link. Based on this formulation, MAB algorithms such as upper confidence bound (UCB), Thompson sampling (TS), epsilon-greedy (e-greedy), are utilized to address the problem and accomplish the mmWave BT process. Numerical simulations confirm the superior performance of the proposed MAB approach over the existing mmWave BT techniques.**

***Keywords*—millimeter wave, beamforming training, multi-armed bandit, reinforcement learning**

I. Introduction

FIFTH (5G) and beyond 5G (B5G) wireless communications aim to support variety of intensive data rate applications ranging from virtual and augmented reality to internet of things (IoTs), unmanned aerial vehicles (UAVs) and wearables [1]. Immigration towards higher frequency bands, e.g., millimeter wave (mmWave) and Terahertz (THz) bands, seems to be an attractive solution due their large available spectrum [2,3]. MmWave band (30 ~ 300 GHz), which is the main concern of this paper, has a large swath of unlicensed spectrum that can support 5G and B5G massive data requirements [4,5]. However, due to its highly operating frequency, mmWave channel is fragile in nature compared to the conventional microwave band, i.e., sub 6 GHz band [6]. High propagation path loss is expected at the mmWave band reaching 28 dB loss compared to the 5 GHz microwave band [7]. Moreover, mmWave band suffers from high susceptibility to shadowing even human/wall shadowing can extremely degrade the quality of the mmWave link [8]. To overcome this tough channel impairments, directional communication using steerable antenna arrays is typically used for establishing the mmWave communication links. Thus, antenna beamforming is typically exploited for constructing the mmWave link between transmitter (TX) and receiver (RX) [9,10]. Due to the massive multi-input multi-output (MIMO) antenna arrays used at mmWave TX/ RX, analogue beamforming using antenna phase shifters is commonly used to accomplish the beamforming process [9,10]. Moreover, hybrid (analogue/digital) precoding can be used to increase the radio frequency (RF) chains of the mmWave transceiver system [11]. Beamforming training (BT) is defined as the process of finding out the best TX/RX beam directions maximizing the achievable data rate of the mmWave link. Variety of mmWave BT strategies can be found in literature aiming to maximize the achievable data rate of the mmWave link while reducing the BT overhead [12]. MmWave standards, e.g., IEEE 802.11ad WiGig standard [13], suggested the use of exhaustive search (EX) BT [13]; by which all available TX/RX beam combinations are examined and the best beam pair is selected for constructing the WiGig link. EX BT has the maximum data rate while it has the highest BT overhead as well. This motivates the design of efficient mmWave BT schemes that emulates the best data rate attained by the EX BT while highly overcoming its substantial BT overhead.

Recently, machine learning (ML) is considered as a talented approach that can address a lot of 5G and B5G challenges and optimize the network performance by the means of sophisticated learning and autonomous decision-making processes [14,15]. Specifically, future networks will be in a great need to learn devices characteristics as well as human behaviors for optimizing the system performance by taking advantage of the powerful smart handheld devices available nowadays [14,15]. Broadly speaking, ML algorithms can be categorized into three main categories: namely, supervised learning, un-supervised learning and reinforcement learning [14,15]. In supervised learning, ML algorithms are used to map the labeled outputs with their corresponding labeled inputs using either regression or classification techniques. Towards that, variety of regressions and classification techniques were proposed in literature for supervised ML such as K-nearest neighbor (KNN), support vector machine (SVM) and different neural networks (NNs) architectures [14,15]. These algorithms are used to model the underlying system and then predicting the outputs for new inputs. Supervised ML algorithms can be applied for estimating/predicting radio parameters associated with massive MIMO channels, spectrum sensing in cognitive radio systems, modulation detection and classification, etc., [14,15]. In un-supervised learning, only the labeled inputs are available for the ML algorithms and the task of the algorithms is to identify the hidden patterns in the labeled input data. Towards that, variety of clustering techniques were utilized in the un-supervised ML. Un-supervised ML algorithms can be

E. M. Mohamed is with 1) Electrical Engineering Dept., College of Engineering, Prince Sattam Bin Abdulaziz University, Wadi Aldwaser 11991, Saudi Arabia and 2) Electrical Engineering Dept., Faculty of Engineering, Aswan University, Aswan 81542, Egypt. (e-mail: ehab_mahmoud@aswu.edu.eg).

applied for users' behavior learning and classification, resource allocation and association, optimal cells deployment, etc., [14,15]. In reinforcement learning (RL), an agent is interacting proactively with the environment with the aim of maximizing its designated long-term reward in trial and error fashion. The main challenge of the RL algorithms is to resolve the tradeoff between manipulating the current selection and discovering new selections, which is formally denoted as exploitation-exploration dilemma. Q-learning and multi-armed bandit (MAB) are considered as the most famous RL algorithms. RL can be applied for base station/ relay/channel online selections, access and handover decision making, power control, etc., [14,15].

In this paper, RL will be utilized to address the problem of mmWave BT by considering it as a MAB problem. In this MAB formulation, the mmWave transceiver system will act as the agent which aims to maximize its long-term reward, i.e., the average throughput in this case. This agent will interact with the environment by selecting a different beam setting at each time and obtain its corresponding reward. Based on the achieved rewards, the agent tries to compromise the exploitation-exploration tradeoff, i.e., either exploiting the best beam direction so-far or exploring new ones. Towards that, three MAB algorithms, namely upper confidence bound (UCB), Thompson sampling (TS), and epsilon greedy (e-greedy) will be investigated to address the MAB based mmWave BT problem. Due to the use of one beam direction at a time while maximizing the achievable throughput, very low BT overhead is consumed by the proposed MAB based BT. This in turns gets the proposed BT scheme has better long-term average throughput performance compared to the existing mmWave BT techniques while reducing its energy consumption as well.

The main contributions of this paper can be summarized as follows:

- MmWave BT is formulated as a MAB problem with the mmWave transceiver acting as the agent trying to maximize its long-term average throughput via interplaying through the available beam directions.
- Three main MAB algorithms; namely UCB, TS and e-greedy are utilized to address this problem and iteratively selects the beam direction maximizing the long-term average throughput.
- Numerical simulations are conducted to prove the effectiveness of the proposed MAB based mmWave BT over the baseline EX BT with respect to the obtained throughput and energy efficiency. Moreover, the convergence analysis of the proposed MAB based BT algorithms is investigated.

The rest of this paper is constructed as follows, Section II gives the literature review, and Section III gives the system model of mmWave BT. Section IV formulates the mmWave BT as a MAB problem and gives the suggested three MAB algorithms to address it. Performance evaluations are given in Section V followed by the conclusion in Section VI.

II. Literature Review

The existing mmWave BT schemes can be divided into 1) Without mmWave channel estimation and 2) With mmWave channel estimation [12]. In the first category, the BT process is done by testing the whole/partial beam space of the mmWave transceiver for obtaining the best beam direction maximizing the achievable data rate. EX BT, adaptive beam search BT, numerical search BT, and location-based BT are types of this category [13], [16,20]. While the EX BT tests all available beam space [13], the adaptive beam search BT uses multi-level BT strategy [16]. Although adaptive beam search BT highly relaxes the BT overhead required by the EX BT, it suffers from low coverage due to the use of wider beams at the earlier stages of the BT process [14]. In the numerical search BT, numerical algorithms such as Rosenbrock or Tabu algorithms [17,18] are used to find out the best beam direction starting with a randomly selected beam. In the location-based BT, the location of the mmWave transceiver is utilized to narrow the number of searched beams to be that only expected to cover the mmWave device at its current location. Different localization techniques with different localization errors were investigated in the location-based BT schemes, such as GPS, Wi-Fi, Li-Fi localization techniques [19,20]. In the second category, mmWave channel estimation is used to firstly estimate the mmWave channel, then the beamformer is adjusted based on the estimated channel coefficients, i.e., angle of departures (AoDs) and angle of arrivals (AoAs). In this regard, compressive sensing (CS) was extensively used to exploit the sparsity inherent in the low scattering mmWave channel for estimating its coefficients, i.e., path gains, AoDs, and AoAs. Based on the estimated channel coefficients, adaptive beam search was proposed by the authors in [11] to optimize the hybrid precoding construction. However, this scheme still suffers from low coverage due to the use of the adaptive beam searching mechanism. To further reduce the complexity of the channel estimation-based BT, localization was used by the authors in [9,19] to further reduce the complexity of the constructed CS matrices and hence reducing the complexity of the BT process. The main drawback of these conventional BT techniques is that the BT process should be re-performed either using the whole beam space or a sub-set of it at every scheduled beacon frame even in case of beam refinement. This results in highly increasing the BT overhead especially when using too sharp beams with a large beam space. The large BT overhead of the conventional BT techniques can be efficiently overcome if learning is introduced to the BT process. That is, the mmWave transceiver can learn from its previous BT interactions with the environment to enhance its future beam selections. Recently, ML especially MAB algorithms attract researchers to apply them to reduce the complexity of the mmWave BT process as given in [21,23]. The authors in [21] proposed linear TS for constructing the mmWave beamformer based on Kalman filter sparse Bayesian learning (KSBL) channel estimator. Despite the novelty of the algorithm, the TS algorithm is performed based on estimating the mmWave channel, which highly increases the complexity of the algorithm. However, in the proposed MAB based BT, no prior channel estimation is needed, and only the achievable data rate is required. This contributes in highly reducing the complexity of the BT process. In [22], the authors considered the mmWave beamforming problem as an adversarial MAB problem and proposed an exponential weight

(EXP3) algorithm with one-bit feedback to address it. The main drawback of this work is the use of indirect, i.e., binary, feedback, which degrades the performance of the beamformer. Instead, in this paper, we will use the whole achievable data rate of the beamformer, which can be measured by the TX side

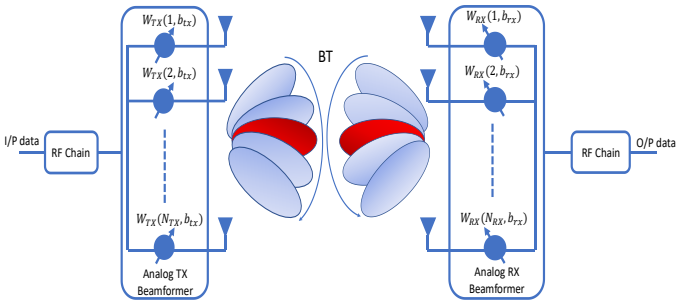

Fig.1. Block diagram of mmWave analog BT

without the need for feedback. In [23], the authors proposed to use the UCB algorithm for position aided coarse and fine levels BT. Also, the algorithm makes use of a small offline database for the purpose of initialization. Yet, in the proposed MAB setting, a single common MAB is designed to flexibly work with neither pre-knowledge of the environment nor users' positions. Also, not only UCB is used but also TS and e-greedy are adopted as well.

III. System Model

Herein, the system model of mmWave analog BT will be explain in detail in addition to the used mmWave channel model. Also, the BT optimization problem will be formulated.

A. MmWave Beamforming and Channel Model

Fig.1 shows the system model of mmWave analog BT, in which N_{TX} and N_{RX} antenna elements are used by the TX and RX respectively. In Fig.1, $W(n,b)$ is the antenna weight vector (AWV) of antenna element n required for beam steering the direction b, where N_{TX} (N_{RX}) and B_{TX} (B_{RX}) are the total number of TX (RX) antenna elements and beam directions, respectively. TX and RX codebooks, i.e., $\boldsymbol{W}_{TX}$ or $\boldsymbol{W}_{RX}$, support a variety of antenna array geometries and offer flexibility in terms of the number, size and the spacing between antenna elements. For phased array antennas, the columns of the codebook matrix specify the discrete phase shifts applied to individual antenna elements to form the beam in a certain direction. Conventional codebook designs such as that proposed by WiGig standards [24] are based on AWVs drawn from the following equation:

$$\mathrm{W}(n,b) = \jmath^{\,floor\left\{\frac{n\times mod\left(b+\left(\frac{B}{2}\right),\ B\right)}{\frac{B}{4}}\right\}},$$
$$\jmath=\sqrt{-1},\ 1\le b\le B,\quad 1\le n\le N \qquad (1)$$

This codebook design has drawbacks of low beamforming efficiency and vulnerability of phase shifting errors. Thus, several codebook designs were proposed in literature to overcome these drawbacks as given in [25,26]. Moreover, novel codebook designs were given in [11,19] by utilizing the quantization angles covered by each beam direction.

Based on $\boldsymbol{W}_{TX}$ and $\boldsymbol{W}_{RX}$, the received signal y for a given signal x can be expressed as:

$$y=\boldsymbol{W}_{RX}^{H}(:,b_{rx})\boldsymbol{H}\boldsymbol{W}_{TX}(:,b_{tx})x+\boldsymbol{W}_{RX}(:,b_{rx})\boldsymbol{n} \qquad (2)$$

where $\boldsymbol{W}_{TX}(:,b_{tx})$ and $\boldsymbol{W}_{RX}(:,b_{rx})$ are the TX and RX AWVs of lengths $N_{TX}\times1$ and $N_{RX}\times1$ in the directions of b_{tx} and b_{rx}, i.e., corresponding to the columns b_{tx} and b_{rx} in $\boldsymbol{W}_{TX}$ and $\boldsymbol{W}_{RX}$, respectively. H denotes the Hermitian transpose, and $\boldsymbol{n}$ is the zero mean additive white gaussian noise term (AWGN) vector of length $N_{RX}\times1$. $\boldsymbol{H}$ is the $N_{RX}\times N_{TX}$ channel matrix, which can be represented as:

$$\mathbf{H}=\frac{1}{\sigma}\sum_{\ell=1}^{L}\chi_\ell\mathbf{T}_{RX}(\Phi_\ell)\mathbf{T}_{TX}^{H}(\Theta_\ell), \qquad (3)$$

where $1\le\ell\le L$ indicates the number of channel paths and L is the total number of multi-paths, and σ is the average path loss depending on the separation distance between the mmWave TX and RX and the path loss exponent. χ_ℓ is the path gain of channel path ℓ, and $\Theta_\ell\in[0,2\pi]$ and $\Phi_\ell\in[0,2\pi]$ are the AoD and AoA of path ℓ. $\mathbf{T}_{TX}(\Theta_\ell)$ and $\mathbf{T}_{RX}(\Phi_\ell)$ are the array responses of both TX and RX, which can be expressed as:

$$\mathbf{T}_{TX}(\Theta_\ell)=\left[1,e^{j\frac{2\pi}{\lambda}d\sin(\Theta_\ell)},..,e^{j(N_{TX}-1)\frac{2\pi}{\lambda}d\sin(\Theta_\ell)}\right]^{T}, \qquad (4)$$

$$\mathbf{T}_{RX}(\Phi_\ell)=\left[1,e^{j\frac{2\pi}{\lambda}d\sin(\Phi_\ell)},..,e^{j(N_{RX}-1)\frac{2\pi}{\lambda}d\sin(\Phi_\ell)}\right]^{T}, \qquad (5)$$

where λ is the signal wavelength and d is the separation distance between TX and RX.

B. Optimization Problem Formulation of MmWave BT

The main goal of the mmWave BT is to maximize the long-term average throughput in bit per second (bps), i.e., finding out the optimal TX/RX beam directions $b_{tx}^{\star}$ and $b_{rx}^{\star}$ maximizing the achievable data rate while using the lowest BT overhead. This can be formulated as follows:

$$(b_{tx}^{\star},b_{rx}^{\star})=$$
$$\arg\max\ \omega\left(\frac{T_D\log2\left(1+\frac{|\boldsymbol{W}_{RX}^{H}(:,b_{rx})\boldsymbol{H}\boldsymbol{W}_{TX}(:,b_{tx})|^2}{N_0}\right)}{KT_{BT}+T_D}\right),$$

s.t.

$$b_{tx}\in\phi_{B_{TX}},\ b_{rx}\in\phi_{B_{RX}},\boldsymbol{W}_{TX}\in\boldsymbol{\phi}_{c_{TX}},\boldsymbol{W}_{RX}\in\boldsymbol{\phi}_{c_{RX}} \qquad (6)$$

where $\phi_{B_{TX}}$, $\phi_{B_{RX}}$ and $\boldsymbol{\phi}_{c_{TX}}$, $\boldsymbol{\phi}_{c_{RX}}$ are the beam spaces and codebook spaces of the TX, RX respectively, and ω is the used bandwidth. K is the total number of TX/RX beam pairs used in the BT process, and T_D and T_{BT} are time durations of data transmission and BT, respectively. The EX BT can find the optimal $(b_{tx}^{\star},b_{rx}^{\star})$ beam pair, which maximizes the achievable data rate, i.e., $\log2\left(1+\frac{|\boldsymbol{W}_{RX}^{H}(:,b_{rx})\boldsymbol{H}\boldsymbol{W}_{TX}(:,b_{tx})|^2}{N_0}\right)$, but using the highest value of K, which is equal to $K=|\phi_{B_{TX}}||\phi_{B_{RX}}|$ beam pairs. Several BT designs can be found in literature trying to find out the beam directions that can achieve the data rate of the

EX BT while using lower K value than that used by the EX BT. This results in enhancing the overall throughput and energy consumption over that obtained using EX BT. The ideal mmWave BT scheme is that can achieve the maximum data rate obtained by the EX BT by using just one beam pair in the BT process, i.e., $K = 1$.

IV. Proposed MAB Based BT Algorithms

In this section, we will consider the optimization problem given in (6) as a MAB problem with the mmWave transceiver acts as the agent and the arms are the TX/RX beam pairs. Thus, the problem will be solved in time bases using MAB algorithms like UCB, TS, and e-greedy with the aim of maximizing the long-term average throughput. The MAB algorithms will select only a single beam pair at a time, i.e., every beacon frame, thus $K = 1$ is always satisfied. Based on the historical data rates achieved by the previously selected beam pairs up to time t and by considering the exploitation- exploration trade-off addressed by these algorithms, a beam pair will be selected for the beacon frame at time t.

A. Proposed UCB Based MmWave BT

The UCB deals with the exploitation-exploration trade-off very effectively. In which, the exploitation term is represented by the average rewards obtained by the played arms so far, while the exploration term is represented by how many times these arms were played so far. Thus, the algorithm is based on maximizing the confidence of the chosen arm by decreasing the un-certainty. The inputs for the proposed algorithm are the precoding matrices $\boldsymbol{W}_{TX}$ and $\boldsymbol{W}_{RX}$, and the total cardinality of the beams space, i.e., $M = |\phi_{B_{TX}}||\phi_{B_{RX}}|$. The algorithm is initialized by selecting each beam pair once for M beacons and calculating their corresponding rewards as follows:

$$\Upsilon_{m,t} = \frac{T_D \log 2\left(1+\frac{\left|W_{RX}^{H}(:,b_{rx,m,t}) H W_{TX}(:,b_{tx,m,t})\right|^2}{N_0}\right)}{KT_{BT}+T_D}, 1 \le m,t \le M \quad (7)$$

where $(b_{tx,m,t}, b_{rx,m,t})$ is the selected beam pair m, $1 \le m \le M$ at time t. After initialization, the algorithm is running using the equation of the UCB based beam pair selection as follows:

$$(b_{tx}, b_{rx})_{m,t} = \arg \max_{1 \le m \le M} \left(\overline{\Upsilon}_{m,t} + \sqrt{\frac{2\ln(t)}{s_{m,t}}} \right), \; M+1 \le t \le T \quad (8)$$

where $(b_{tx}, b_{rx})_{m,t}$ is the selected beam pair m at time t, and T is the total horizon duration. $s_{m,t}$ and $\overline{\Upsilon}_{m,t}$ are the total number of times a beam pair m is selected and its achievable average throughput up to time t. In (8), $\overline{\Upsilon}_{m,t}$ and $\sqrt{\frac{2\ln(t)}{s_{m,t}}}$ represent the exploitation and the exploration terms, respectively. Thus, the UCB based BT tries to compromise between exploiting the beam pair having the maximum average throughput or exploring new beam pairs that have lower values of $s_{m,t}$. Fig. 2 summarizes the proposed UCB based mmWave BT algorithm.

B. Proposed TS Based MmWave BT

TS is a Bayesian algorithm that tries to build a probabilistic model for the reward obtained by each arm. That is, the collected rewards are used to construct posterior distributions and then selects arms randomly in a way that the drawing probability of each arm matches the probability of the particular arm being optimal. In detail, the TS algorithm samples the constructed posterior distributions of the arms' rewards and then selects the arm having the maximum sample to play. For the underlaying mmWave BT problem, the attained throughput (the reward), can be modeled as a normal distribution. Thus, we will make use of the model given in [27,28], where the posterior 2

Algorithm: UCB based mmWave BT

Inputs: $\boldsymbol{W}_{TX}$ and $\boldsymbol{W}_{RX}$,

Initialize: each Tx/RX beam pair, i.e., $(b_{tx}, b_{rx})_m$, $1 \le m \le M$, will be selected once, and their corresponding $\Upsilon_{m,t}$ are evaluated

For $t = M+1:T$

1. Draw a beam pair and obtain the reward $\Upsilon_{m,t}$

$$(b_{tx}, b_{rx})_{m,t} = \arg \max_{1 \le m \le M} \left(\overline{\Upsilon}_{m,t} + \sqrt{\frac{2\ln(t)}{s_{m,t}}} \right)$$

2. $s_{m,t} = s_{m,t} + 1$
3. $\overline{\Upsilon}_{m,t} = \frac{1}{s_{m,t}} \sum_{j=1}^{s_{m,t}} \Upsilon_{m,j}$

END For

Fig.2. Proposed UCB based mmWave BT algorithm

Algorithm: TS based mmWave BT

Inputs: $\boldsymbol{W}_{TX}$ and $\boldsymbol{W}_{RX}$,

Initialize: $\overline{\Upsilon}_{m,t} = 0$, $s_{m,t} = 0$

For $t = 1:T$

Sample $\Pi_{m,t}$, $1 \le m \le M$, from normal distributions $\mathcal{N}(\overline{\Upsilon}_{m,t}, \alpha_{m,t}^2)$

1. Draw a beam pair and obtain the reward $\Upsilon_{m,t}$

$$(b_{tx}, b_{rx})_{m,t} = \arg \max_{1 \le m \le M} (\Pi_{m,t})$$

2. $s_{m,t} = s_{m,t} + 1$
3. $\overline{\Upsilon}_{m,t} = \frac{1}{s_{m,t}} \sum_{j=1}^{s_{m,t}} \Upsilon_{m,j}$

END For

Fig.3. Proposed TS based mmWave BT algorithm

distribution of the throughput of beam pair m comes from $\mathcal{N}(\overline{\Upsilon}_{m,t}, \alpha_{m,t}^2)$. In this model, the mean and the variance of the normal distribution are $\overline{\Upsilon}_{m,t} = \frac{1}{s_{m,t}} \sum_{j=1}^{s_{m,t}} \Upsilon_{m,j}$ and $\alpha_{m,t}^2 = \frac{1}{s_{m,t}+1}$, respectively. In TS, at every time t, i.e., at every beacon frame, a sample $\Pi_{m,t}$ is taken from each constructed posterior distribution and the beam pair having the maximum sample value will be played based on the following equation:

$$(b_{tx}, b_{rx})_{m,t} = \arg \max_{1 \le m \le M} (\Pi_{m,t}), \quad 1 \le t \le T \quad (9)$$

Fig.3 gives the proposed TS based mmWave BT algorithm.

C. Proposed e-greedy based MmWave BT

E-greedy is considered as the simplest MAB algorithm when dealing with the exploitation-exploration dilemma for arm selection. Simply, the arms exploitation is done with probability $1 - \varepsilon$ and the arms exploration is done with probability ε, where ε is a system design parameter. Thus, in the proposed beam pair

selection, at a time t, the beam pair having the highest average throughput, i.e., the highest $\overline{\Upsilon}_{m,t}$, will be selected with probability $1-\varepsilon$; otherwise random beam pair is drawn from uniform random distribution, i.e., $\mathcal{U}(1,M)$, as shown in the proposed e-greedy based mmWave BT algorithm given in Fig.4.

V. Performance Evaluations

In this section, performance evaluations are conducted to prove the effectiveness of the proposed MAB based mmWave BT algorithms over the baseline EX BT and some of the existing approaches. The EX BT is selected as a benchmark approach because it has the optimal data rate performance.

Algorithm: e-greedy based mmWave BT

Inputs: $\boldsymbol{W}_{TX}$ and $\boldsymbol{W}_{RX}$, ε

Initialize: $\overline{\Upsilon}_{m,t}=0$, $s_{m,t}=0$

For $t=1:T$

1. Draw a beam pair and obtain the reward $\Upsilon_{m,t}$

$$(b_{tx},b_{rx})_{m,t}=\begin{cases}\arg\max\limits_{1\le m\le M}\left(\overline{\Upsilon}_{m,t}\right) & \text{with probability } 1-\varepsilon\\ \mathcal{U}(1,M) & \text{with probability } \varepsilon\end{cases}$$

2. $s_{m,t}=s_{m,t}+1$
3. $\overline{\Upsilon}_{m,t}=\frac{1}{s_{m,t}}\sum_{j=1}^{s_{m,t}}\Upsilon_{m,j}$

END For

Fig.4. Proposed e-greedy based mmWave BT algorithm

A. Simulation Parameters

For realistic considerations, ray tracing is used to construct the mmWave channel in the conducted simulations. Fig. 5 shows the used ray tracing indoor study area of dimension $30\times15\times4$ m^3, where the mmWave AP operating at 60 GHz is attached at the ceiling and the mmWave user equipment (UE) is uniformly dropped inside the room area at a height of 0.75 m. Three mmWave paths are assumed with one line of sight (LOS) path and other non-LOS paths, other important parameters are given in Table I. Also, downlink transmission is assumed where the TX is the mmWave AP and the RX is the mmWave UE.

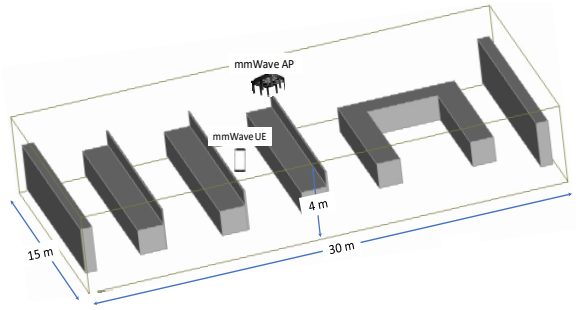

Fig.5. Ray tracing indoor study area

B. MmWave 3D Beamforming

For the mmWave link, the mmWave AP is assumed to use 3D beamforming while the UE is using omni-directional antenna pattern, i.e., B_{RX} = 1. For the 3D beamforming, the azimuth coverage angle of the mmWave AP, ϑ_{azm} is divided into a number of beam tiers, which is equal to $N_{\text{tier}}=\frac{\vartheta_{\text{azm}}}{\vartheta_{-3\text{dB}}}$, where $\vartheta_{-3\text{dB}}$ is the 3D beamwidth. Then, the total number of beams is equal to $B_{TX}=1+\frac{6N_{\text{tier}}(N_{\text{tier}}+1)}{2}$ [10]. For example, using $\vartheta_{-3\text{dB}}=30^{\circ}$ and $\vartheta_{\text{azm}}=85^{\circ}$, as given in [10] and used in the simulation setting in Table I, then $N_{\text{tier}}\approx3$ and $B_{TX}=36$ beams.

C. Simulation Results

In the conducted simulations, the performances of the MAB base BT algorithms will be compared with the performance of the EX BT in terms of average throughput in Gbps, i.e., $\omega\overline{\Upsilon}_m$, and energy efficiency in Gbps/mJ, where the energy efficiency is calculated using the following equation

$$\gamma_E=\frac{\omega\overline{\Upsilon}_m}{P_t(KT_{BT}+T_D)}. \tag{10}$$

Fig. 6 shows the average throughput comparisons against the 3D beamwidth at LOS blocking probability of 0. As shown in this figure, as the 3D beamwidth is increased, the average throughputs of the MAB based BT schemes are decreased due to the decrease in the beamforming gain and hence the achievable data rate. TS shows the best performance due its inherent Bayesian functionality, and e-greedy has the worst performance among the MAB algorithms. However, all MAB based BT algorithms have better average throughput performance than the EX BT. This is due to the lower number of beams used in each BT step which is equal to one beam per a beacon time. It interesting to note that the average throughput of the EX BT is highly decreased when the 3D beamwidth is equal to 10° due to the large number of trained beams which is equal to 270 beams. However, as the 3D beamwidth is increased, the BT overhead is decreased which increases the average throughput of the EX BT till reaching a point where the beamforming gain is highly decreased resulting in decreasing the average throughput again as shown in Fig. 6. From Fig.6, using $\vartheta_{-3\text{dB}}=10^{\circ}$, about 4.5, 4.37, and 3.3 increase in average throughputs are obtained using TS, UCB and e-greedy based BT over using EX BT. However, using $\vartheta_{-3\text{dB}}=60^{\circ}$, these values are decreased to 1.24, 1.2 and 1.18., respectively.

Table I
Simulation Parameters

Parameter	Value
ω	2.16 GHz [13]
T_D	1 msec [10]
T_{BT}	23 usec [13]
3D Beamwidth ($\vartheta_{-3\text{dB}}$)	$10^{\circ}, 20^{\circ}, 30^{\circ}, 40^{\circ}, 50^{\circ}, 60^{\circ}$
LOS blocking probability	0, 0.2, 0.4, 0.6, 0.8
TX power (P_t)	10 dBm
ε	0.1 [28]
T	2000
Azimuth coverage angle (ϑ_{azm})	85° [10]

Fig. 7 shows γ_E of the compared schemes against $\vartheta_{-3\text{dB}}$. The MAB based BT shows superior γ_E performances over the EX BT at all values of $\vartheta_{-3\text{dB}}$. This comes from the lower BT energy consumptions of the MAB based BT schemes compared to the EX BT due to the high decrease in the K value in (10). In

consequence, this results in highly increasing the numerator of (10) and highly decreasing its dominator as well. Using $\vartheta_{-3\mathrm{dB}} = 10^{\circ}$, about 1.5e+3, 1.48e+3, and 1.12e+3 increase in γ_E are obtained using TS, UCB and e-greedy based BT over using EX BT. However, using $\vartheta_{-3\mathrm{dB}} = 60^{\circ}$, about 35.42, 35.3 and 31.3 increase in the average throughputs are obtained.

Fig. 8 shows the average throughput comparisons against the LOS blocking probability using $\vartheta_{-3\mathrm{dB}} = 20^{\circ}$. As shown by this figure, the average throughputs of the compared BT schemes are decreasing with the increase of the LOS blocking probability. This is due to the low channel gains of the non-LOS paths, which results in decreasing the achievable data rate of all BT schemes. However, the proposed MAB based BT algorithms have better average throughput performances over the EX BT for all tested values of LOS blocking probability. This comes from the lower BT overhead of the proposed schemes. Also, it is interesting to note that the rate of decrease in the average throughputs of the MAB based BT algorithms are comparable to that belongs to the EX BT. This means that the MAB based schemes can withstand the harsh blockage environment

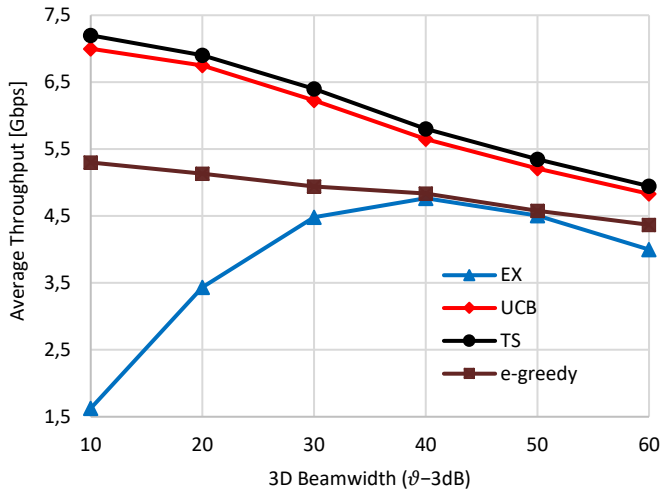

Fig.6. Average throughput comparisons against 3D beamwidth

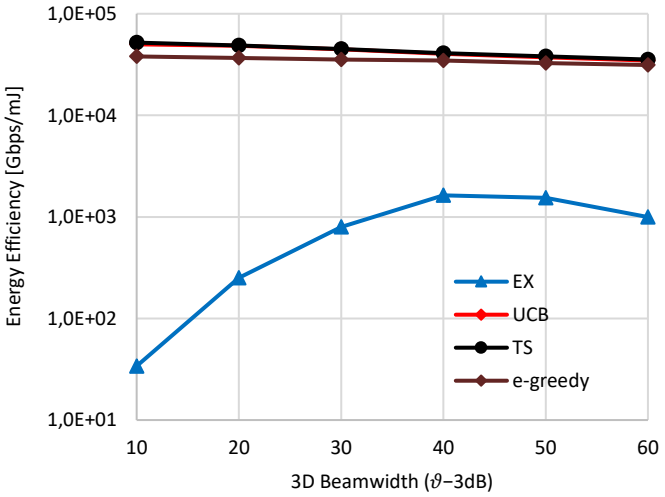

Fig.7. Energy efficiency comparisons against 3D beamwidth

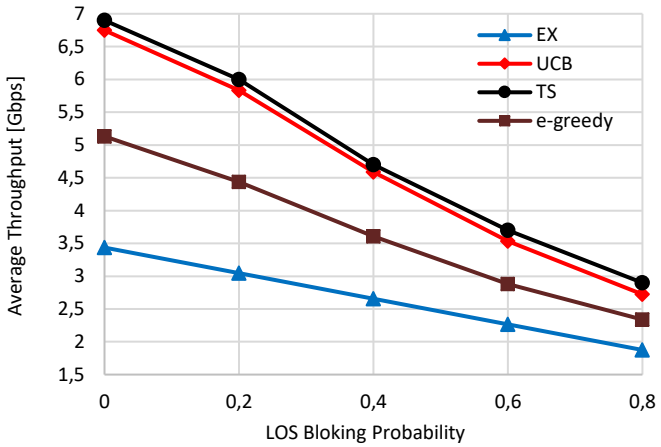

Fig.8. Average throughput comparisons against LOS blocking probability

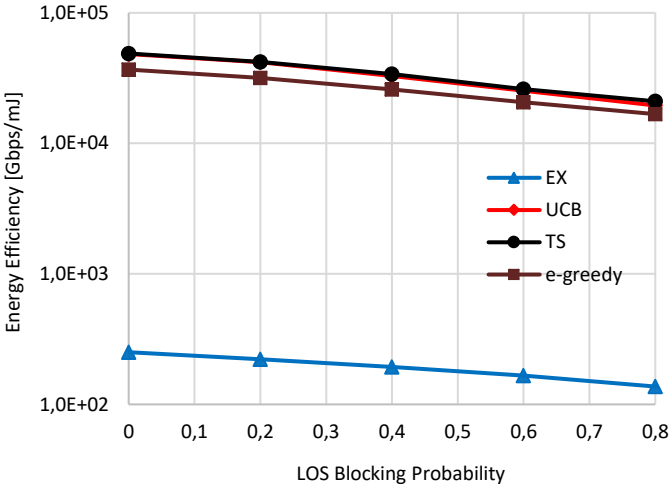

Fig.9. Energy efficiency comparisons against LOS blocking probability

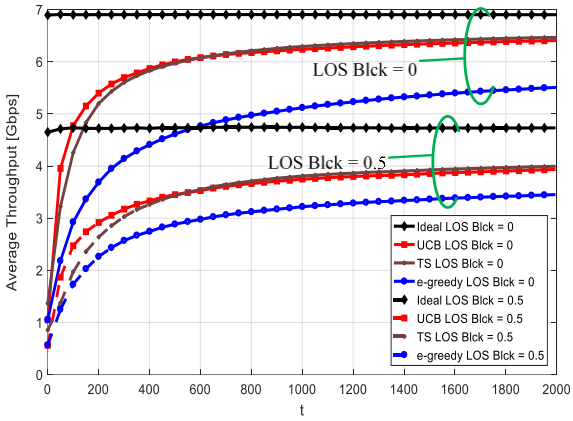

Fig.10. Average throughput convergence rate using 3D beamwidth of 20°

comparable to exhaustively searching all available beam space. At LOS blocking of 0, about 2, 1.96, and 1.5 increase in average throughputs are obtained using the proposed TS, UCB and e-greedy BT algorithms over using the EX BT. However, at LOS blocking of 0.8, these values are decreased to be 1.54, 1.45 and 1.24, respectively.

Fig. 9 shows γ_E comparisons against the LOS blocking probability. Again, the proposed MAB based BT algorithms demonstrate superior performances over the EX BT at all values of LOS blocking probability. Likewise, γ_E of the MAB based BT algorithms are decreasing with a comparable rate with that of the EX BT. At zero blocking, about 195, 195 and 146.8 increase in γ_E are obtained using the proposed TS, UCB and e-greedy BT algorithms over using the EX BT. However, at LOS blocking of 0.8, about 153.3, 153, and 121.9 increase in γ_E are obtained.

Fig. 10 shows the average throughput convergence rate of the proposed MAB based BT schemes using $\vartheta_{-3\mathrm{dB}}$ of 20° and at LOS blocking of 0 and 0.5 against the horizon. Moreover, the ideal average throughput performance is shown in Fig. 10, which corresponds to the BT scheme that can achieve the maximum data rate obtained by the EX BT using just one beam pair in the BT process, i.e., using $K = 1$. This ideal performance is used as an upper limit performance of the MAB based BT schemes. As shown by Fig. 10, all MAB based BT schemes converge towards the ideal average throughput performance. At LOS blocking of 0, about 95 % of the ideal performance can be achieved by TS and UCB based BT while about 80 % is achieved by the e-greedy BT scheme. However, at LOS blocking of 0.5, about 83.3 % of the ideal performance is achieved by TS and UCB based BT while about 73% is achieved by the e-greedy based BT.

Compared to the existing mmWave BT techniques, in [20], the authors proved that 8 and 26 beam pairs should be used in the BT process when using the high accurate Li-Fi and Wi-Fi localization-based BT to obtain 95% of the data rate obtained by the EX BT. This emphasizes the high superior performance of the proposed MAB based BT over the high accurate localization-based BT techniques such as Li-Fi based localization. Also, to obtain 95% of the data rate achieved by the EX BT, the schemes given in [9], [11] and [19], which are based on CS based mmWave channel estimation, need almost 961, 94 and 64 beam switchings respectively, as stated in [19]. Also, numerical search BT needs high number of beam switchings to achieve 95% of the data rate achieved by the EX BT as given in [17].

VI. Conclusion

In this paper, a reinforcement learning approach is introduced to address the crucial problem of mmWave BT by considering the problem as a multi-armed bandit problem. In this formulation, the mmWave transceiver was acting as the agent, the candidate TX/RX beam pairs are the arms and the attained average throughputs are the corresponding rewards. Based on this formulation, three MAB algorithms were adopted to address mmWave BT. The proposed MAB based BT techniques employed only one beam pair for BT at a time. Then, based on the historical performance of the operated beams, new beam pair selection is decided by the MAB algorithms at each beacon time. Thus, very low overhead is consumed by the BT process, which highly improves the average throughput and energy efficiency of the proposed MAB based BT techniques over the baseline EX BT under different scenarios. Moreover, the proposed MAB BT schemes converged to the optimal performance with high percentages especially at zero blocking. Furthermore, the proposed scheme showed superior performance over the location-based BT even using high accurate localization technique in addition to the CS channel estimation-based BT techniques.

References

[1] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu, "Future of Ultra-Dense Networks Beyond 5G: Harnessing Heterogeneous Moving Cells" *IEEE Communications Magazine*, vol. 57, no. 16, pp. 86 – 92, 2019.

[2] T. S. Rappaport, et al., "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond" *IEEE ACCESS*, vol. 7, pp. 78729 – 78757, 2019.

[3] Z. Chen, et al., "A Survey on Terahertz Communications," *China Communications*, vol. 16, no. 2 pp. 1673-5547, 2019.

[4] E. M. Mohamed, M. A. Abdelghany, and M Zareei "An Efficient Paradigm for Multiband WiGig D2D Networks," *IEEE ACCESS*, vol. 7, pp. 70032-70045, 2019.

[5] E. M. Mohamed, et al., "Relay Probing for Millimeter Wave Multi-Hop D2D Networks," *IEEE ACEESS*, vol. 8, pp. 30560 – 30574, 2020.

[6] T. S. Rappaport, et al., "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. on Antenn. and Propag.*, vol. 61, no. 4, pp. 1850-1859, 2013.

[7] T. S. Rappaport, et al., "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models," *IEEE Trans. on Antenn. and Propag.*, vol. 65, no. 12, pp. 6213-6230, 2017.

[8] T. Bai, R. Vaze, and R. W. Heath, Jr., "Analysis of Blockage Effects on Urban Cellular Networks," *IEEE Transc. On Wirele. Commun*., vol. 13, no. 9, pp. 5070-5083, 2014.

[9] A. Abdelreheem, E. M. Mohamed, and H. Esmaiel, "Location-based millimeter wave multi-level beamforming using compressive sensing," *IEEE Commun. Lett*., vol. 22, pp. 185-188, 2018.

[10] E. M. Mohamed, K. Sakaguchi, and S. Sampei, "Wi-Fi coordinated WiGig concurrent transmissions in random access scenarios," *IEEE Trans. on Vehi. Techn*., vol. 66, no. 11, pp. 10357-10371, 2017.

[11] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selec.Topics in Signal Process*., vol. 8, pp. 831-846, 2014.

[12] I. Ahmed, et al., "A survey on hybrid beamforming techniques in 5g: Architecture and system model perspectives," *IEEE Communic. Surv. Tut*., vol. 20, no. 4, pp. 3060–3097, Fourthquarter 2018.

[13] IEEE 802.11ad Standard, "Enhancements for very high throughput in the 60 GHz band," ed, Dec. 2012.

[14] C. Jiang, et al, "Machine Learning Paradigms for Next-Generation Wireless Networks" *IEEE Wireless Communications*, vol. 24, no.2, pp. 98 – 105, 2017.

[15] J. Wang, et al., "Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks" *IEEE Communic. Surv. Tut*. (early access) 2020.

[16] Hur, S., Kim, T., Love, D. J., *et al*.: 'Millimeter wave beamforming for wireless backhaul and access in small cell networks', *IEEE Trans. on Commun*., vol. 61, no. 10, pp. 4391–4403. 2013.

[17] B. Li, Z. Zhou, H. Zhang, and A. Nallanathan, "Efficient beamforming training for 60-GHz millimeter-wave communications: A novel numerical optimization framework," *IEEE Trans. Veh. Technol*., vol. 63, no. 2, pp. 703–717, 2014.

[18] Gao, X., Dai, L., Yuen, C., Wang, Z.: 'Turbo-like beamforming based on tabu search algorithm for millimeter-wave massive MIMO systems', *IEEE Trans. On Vehi. Techn*., vol. 65, no.7, pp. 5731–5737, 2016.

[19] A. Abdelreheem, E. M. Mohamed, H. Esmaiel, "Adaptive location-based millimetre wave beamforming using compressive sensing based channel estimation," *IET Communications*, vol. 13, no. 9, pp. 1287-1296, 2019.

[20] A. M. Nor and E. M. Mohamed, "Li-Fi Positioning for Efficient Millimeter Wave Beamforming Training in Indoor Environment," *Mobile Networks and Applications*, vol. 24, no.2, pp. 517-531, 2019.

[21] M. B. Booth, V. Suresh, N. Michelusi, and D. J. Love, "Multi-Armed Bandit Beam Alignment and Tracking for Mobile Millimeter Wave Communications," IEEE Commun. Lett., vol. 23, no. 7, pp.1244-1248, 2019.

[22] I. Chafaa, E. V. Belmega, and M. Debbah, "Adversarial multi-armed bandit for mmWave beam alignment with one-bit feedback," in *Proc. 12th EAI Int. Conf. Perform. Eval. Methodol. Tools*, pp. 23–30, 2019.

[23] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, Jr., "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30507–30526, 2019.

[24] IEEE 802.15.3c Part 15.3: 'Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rat Wireless Personal Area Neworks (WPANS) Amendment', 2009.

[25] Wang, J., Lan, Z., Sum, C., et al 'Beamforming codebook design and performance evaluation for 60GHz wideband WPANs'. *Proc. IEEE Vehi. Techn*., Anchorage, 2009, pp. 1-6.

[26] Zou, W., Cui, Z., Li, B., Zhou, Z., Hu, Y, 'Beamforming codebook design and performance evaluation for 60GHz wireless communication'. *Proc. International Symposium on Communications & Information Technologies (ISCIT)*, 2011, pp. 30-35.

[27] S. Agrawal, and N. Goyal, "Further optimal regret bounds for thompson sampling," in *Artificial Intelligence and Statistics*, pp. 99-107, 2013.

[28] F. Wilhelmi, "Collaborative spatial reuse in wireless networks via selfish multi-armed bandits," in Ad Hoc Networks, vol. 88, pp. 129-141, 2019.