REINFORCEMENT LEARNING FOR HYBRID BEAMFORMING IN MILLIMETER WAVE SYSTEMS

REINFORCEMENT LEARNING FOR HYBRID BEAMFORMING IN MILLIMETER WAVE SYSTEMS

Ture Peken Ravi Tandon Tamal Bose
University of Arizona, Electrical and Computer Engineering Dept.
Tucson, AZ, 85721
[turepeken, tandonr, tbose]@email.arizona.edu

ABSTRACT

The use of millimeter waves (mmWave) for next-generation cellular systems is promising due to the large bandwidth available in this band. Beamforming will likely be divided into RF and baseband domains, which is called hybrid beamforming. Precoders can be designed by using a predefined codebook or by choosing beamforming vectors arbitrarily in hybrid beamforming. The computational complexity of finding optimal precoders grows exponentially with the number of RF chains. In this paper, we develop a Q-learning (a form of reinforcement learning) based algorithm to find the precoders jointly. We analyze the complexity of the algorithm as a function of the number of iterations used in the training phase. We compare the spectral efficiency achieved with unconstrained precoding, exhaustive search, and another state-of-art algorithm. Results show that our algorithm provides better spectral efficiency than the state-of-art algorithm and has performance close to that of exhaustive search.

INTRODUCTION

The massive amount of spectrum in the mmWave frequencies is considered as one of the key enablers for next-generation cellular systems [1,2]. Large-scale antenna systems aka massive MIMO systems can be feasible in mmWaves to provide beamforming gains since a large number of antennas can be placed in small spaces at high frequencies [3, 4]. In conventional cellular systems operating in lower frequencies, digital beamforming is implemented to have better control for designing the precoding matrices. However, digital beamforming is not a practical solution in mmWaves due to the complexity, energy consumption, and cost overhead [5]. In particular, each antenna array element requires to be fed with separate transceiver and data converter, which leads to high power consumption and cost. Analog beamforming was proposed to reduce the number of transceivers [6, 7]. With analog beamforming, each transceiver is connected with multiple antennas, and the phase of the transmitted signal at each antenna of the array is controlled by using analog phase shifters. However, analog beamforming also has challenges, for example, the phase shifters allow only quantized phase values for the transmitted signals. Moreover, each transceiver forms a single beam towards a user, so a separate transceiver is required for each user in multiple-user systems. In this case, inter-user interference would be burdensome if the spatial separation between users is not enough [5].

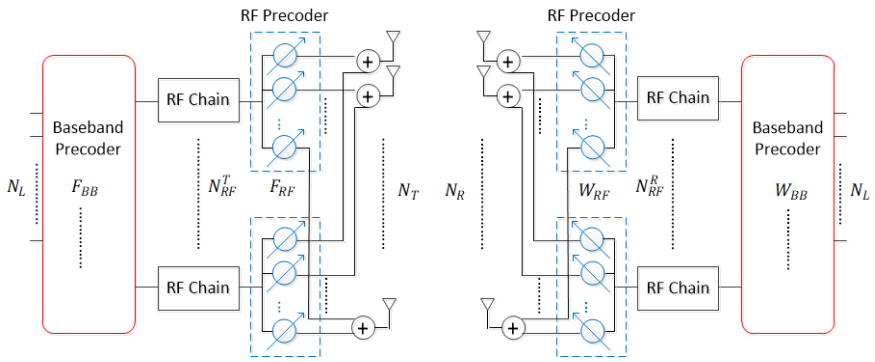


Figure 1: Hybrid beamforming architecture with RF and baseband blocks.

To overcome the bottlenecks of analog and digital beamforming, hybrid beamforming was proposed in [8–10], where digital beamforming is utilized on top of analog beamforming. The sparsity of mmWave channels was exploited in [8] to use the basis pursuit algorithm for developing a hybrid precoding scheme based on complete channel state information (CSI). In [9], a hybrid precoding algorithm, which uses partial CSI, was proposed. Authors of [10] proposed low-complexity channel estimation and hybrid precoding algorithms when both the base station (BS) and the mobile station (MS) are equipped with large antenna arrays.

In hybrid beamforming, RF precoders can be designed based on the optimal (i.e., unconstrained) or sub-optimal (i.e., constrained) solution. In unconstrained solution, the codebook of precoders are not fixed and beam steering directions can be chosen arbitrarily. In the constrained solution (i.e., codebook-based), RF precoders can be selected from some codebook in which beamforming vectors depend on quantized directions. Codebook-based hybrid beamforming design is more practical since it takes hardware limitations into account and it is easier to implement for arbitrary antenna arrays. Therefore, for the scope of this paper, we focus on codebook-based hybrid beamforming. We consider the fully connected architecture (see Figure 1) in which each RF chain is connected to all antennas. When codebook-based hybrid beamforming is considered, optimal RF precoders are found by solving an optimization problem based on a metric (e.g., sum rate, mutual information over the channel, signal-to-noise-ratio (SNR)) over all possible RF precoders that are selected from predefined codebooks. The computational complexity of this optimization problem by using exhaustive search grows exponentially with the number of RF chains at the transmitter and receiver. The number of computations required for the exhaustive search solution is huge even with a moderate number of RF chains. For example, the number of computations for a system with 4 RF chains, and with an RF precoder codebook that consists of 4 beamforming vectors at the transmitter and receiver would be $4^4 \times 4^4 = 65536$. To tackle this challenge, we develop a reinforcement learning based hybrid beamforming algorithm for a mmWave system with large antenna arrays at both the transmitter and receiver. Reinforcement learning has been applied to different problems in mmWave communication [11]. To the best of our knowledge, this is the first work to apply reinforcement learning for hybrid beamforming in mmWaves. The main contributions of our paper are summarized next:

1. We propose a novel reinforcement learning based hybrid beamforming algorithm which applies Q-learning to jointly design RF precoders at the transmitter and receiver.

2. The Q-learning based approach has two phases: a training (learning) phase followed by a testing (precoder selection) phase. In the training phase of our algorithm, Q-table, which essentially captures optimal actions (precoder matrices) for the states (CSI) in the training set, is iteratively learned. In the test phase of our algorithm, optimal precoders are found for a new state by using the learned Q-table. In order to use Q-table for a new state, we find a state in Q-table which has the closest Euclidean distance to the new state. We then select the action for this *closest* state as the action for the new state.

3. We implement our proposed algorithm with both complete and partial CSI and analyze the performance in both scenarios with different training sample sizes.

4. We then analyze the computational complexity of our algorithm as a function of iteration steps and show that a significant reduction in computational complexity is achieved compared to the exhaustive search. The performance of the proposed algorithm in terms of spectral efficiency with imperfect and perfect CSI is shown and compared with existing methods in the literature. According to the results, we improve the spectral efficiency compared to other suboptimal algorithms and achieve a very close performance to the exhaustive search.

System Model: Hybrid Beamforming for mmWaves

We consider a mmWave system given in Figure 1 in which a transmitter with N_T antennas and N_{RF}^T RF chains communicates with a receiver with N_R antennas and N_{RF}^R RF chains. We assume there are N_L data streams. The transmitter applies an $N_{RF}^T \times N_L$ baseband precoder $\mathbf{F}_{BB}$ followed by an $N_T \times N_{RF}^T$ RF precoder $\mathbf{F}_{RF}$. Then, the transmitted signal $\mathbf{x}$ over the mmWave channel can be written as,

$$\mathbf{x} = \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s}, \tag{1}$$

where $\mathbf{s}$ is the $N_L \times 1$ symbol vector. The average total transmit power is denoted as P_L, and the transmitted symbol vector satisfies $\mathbb{E}\left[\mathbf{s}\mathbf{s}^H\right] = \left(\frac{P_L}{N_L}\right)\mathbf{I}_{N_L}$. Analog phase shifters in the RF precoder bring a constraint on the entries of the RF precoder such that the entries of $\mathbf{F}_{RF}$ with constant modulus are normalized to satisfy $|[\mathbf{F}_{RF}]_{i,j}|^2 = N_T^{-1}$, where $|[\mathbf{F}_{RF}]_{i,j}|$ corresponds to the the magnitude of $(i,j)^{th}$ element of $\mathbf{F}_{RF}$. Then, the average total power is satisfied by normalizing $\mathbf{F}_{BB}$ such that $\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2 = N_L$. We denote the mmWave channel between the transmitter and receiver with a $N_R \times N_T$ matrix $\mathbf{H}$. The received signal over N_R antennas of the receiver is given as,

$$\mathbf{r} = \mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{n}, \tag{2}$$

where $\mathbf{n}$ is the noise vector of dimension $N_R \times 1$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Then, the receiver processes signal $\mathbf{r}$ with an $N_R \times N_{RF}^R$ RF precoder $\mathbf{W}_{RF}$ and following that with an $N_{RF}^R \times N_L$ baseband precoder $\mathbf{W}_{BB}$, which leads to the received symbol vector $\mathbf{y}$ of dimension $N_L \times 1$:

$$\mathbf{y} = {\mathbf{W}_{BB}}^H {\mathbf{W}_{RF}}^H \mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + {\mathbf{W}_{BB}}^H {\mathbf{W}_{RF}}^H \mathbf{n}. \tag{3}$$

Various measurements were conducted to model mmWave channel [12, 13]. According to these measurements, mmWave channels have limited scattering, which makes geometric channel model

an appropriate choice. In geometric channel model, each scatterer contributes a single propagation path between the transmitter and receiver. The channel representation based on this model is given as,

$$\mathbf{H} = \sqrt{\frac{N_T N_R}{\rho}} \sum_{s=1}^{S} g_s \mathbf{a}_R(\theta_s) \mathbf{a}_T^H(\phi_s), \tag{4}$$

where S is the number of scatterers, ρ is the average path-loss between the transmitter and receiver, and g_s is the complex gain of the s^{th} path with Rayleigh distribution, i.e., $g_s \sim \mathcal{N}(0, \overline{G})$ for $s = 1, 2, ..., S$. $\overline{G}$ is the average power gain. $\mathbf{a}_T(\phi_s)$ and $\mathbf{a}_R(\theta_s)$ denote the array response vector at the transmitter and receiver. Finally, $\theta_s \in [0, 2\pi]$ and $\phi_s \in [0, 2\pi]$ denote the s^{th} path's azimuth Angle of Arrival (AoA) and Angle of Departure (AoD) of the transmitter and receiver.

Problem Statement - Our main goal is to design optimal hybrid precoders at the transmitter and receiver, $(\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{RF}, \mathbf{W}_{BB})$ which maximize the rate obtained over the mmWave channel under the RF precoder constraints. This problem can be defined as in the following,

$$R = log_2 \left| \mathbf{I}_{N_L} + \frac{P_L}{N_L} \mathbf{W}_{BB}^H \mathbf{W}_{RF}^H \mathbf{H} \mathbf{F}_{RF} \mathbf{F}_{BB} \mathbf{F}_{BB}^H \mathbf{F}_{RF}^H \mathbf{H}^H \mathbf{W}_{RF} \mathbf{W}_{BB} \right|, \tag{5}$$

over all possible RF and baseband precoder matrices $(\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{RF}, \mathbf{W}_{BB})$ for the transmitter and receiver.

For this paper, we use the predefined codebook structure given in [10]. In this codebook structure, each beamforming vector in the codebook is defined in terms of the set of quantized angles. This codebook structure is chosen because it can be easily extended to a codebook with beamforming vectors of different beamwidths. According to this structure, $[\mathbf{F}_{RF}]$ $:, i$, $i = 1, ..., N_{RF}^T$ and $[\mathbf{W}_{RF}]$ $:, j$ for $j = 1, ..., N_{RF}^R$ are selected from predefined codebooks $\mathcal{C}_T$ and $\mathcal{C}_R$, respectively. $\mathcal{C}_T$ consists of N_{Beams}^T beamforming vectors with dimension of $N_T \times 1$, and $\mathcal{C}_R$ consists of N_{Beams}^R beamforming vectors with dimension of $N_R \times 1$. The columns of $\mathcal{C}_T$ and $\mathcal{C}_R$ are chosen such that they satisfy RF beamforming constraints. Elements of beamforming vectors in $\mathcal{C}_T$ and $\mathcal{C}_R$ are represented as quantized phase shifts, where each phase shifter is controlled by an N_q-bit input. In this case, $N_{Beams}^T = N_{Beams}^R = 2^{N_q}$. $n^{th}(m^{th})$ row of the RF precoding matrix at the transmitter (receiver), which corresponds to the phase shifts of the $n^{th}(m^{th})$ antenna of the $\mathbf{F}_{RF}(\mathbf{W}_{RF})$, can be written as $e^{\frac{j2\pi n k_q}{2^{N_q}}}$ or $e^{\frac{j2\pi m k_q}{2^{N_q}}}$ for some $k_q = 0, 1, ..., 2^{N_q} - 1$.

In the next section, we will first briefly review reinforcement learning. We then present our proposed approach of applying reinforcement learning for hybrid beamforming.

Brief Overview of Reinforcement Learning

Reinforcement learning is a machine learning (ML) approach in which an agent learns to choose optimal actions to achieve its goals by observing the states of its environment through an interactive process [14]. The main purpose of an agent in reinforcement learning is to learn a policy that from any initial state, performs actions which maximize the reward accumulated over time. Moreover, the agent's sequence of actions over time affects the distribution of training samples. Therefore,

there is a trade-off between exploration of undiscovered states and actions and exploitation of learned states and actions which generate high rewards. General setting of a reinforcement learning can be summarized as follows. An agent observes the states S of its environment and has a set of actions A that it can choose. After the agent observes its current state $s_t \in S$, it performs an action $a_t \in A$ at time t. Environment responds to the agent's action a_t at state s_t with a reward $r_t = r(s_t, a_t)$ and generates a successor state $s_{t+1} = \delta(s_t, a_t)$. The agent aims to select the policy which gives the maximum cumulative discounted reward over time, i.e., the agent's goal is to learn a policy $\pi : S \rightarrow A$, which maximizes the cumulative value:

$$V^{\pi}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... = \sum_{i=0}^{\infty} \gamma^i r_{t+i}, \tag{6}$$

where $0 \leq \gamma < 1$. Here, γ denotes a discount factor and the future rewards are discounted by this factor exponentially. In this case, an optimal policy π^* which maximizes discounted cumulative reward for all states can be given as,

$$\pi^* = \underset{\pi}{\operatorname{argmax}} V^{\pi}(s), (\forall s). \tag{7}$$

If the agent has a sequence of immediate rewards $r(s_i, a_i)$ for $i = 0, 1, 2, ...$ as the training information, it is easier to learn a numerical evaluation function instead of directly learning the function $\pi^* : S \rightarrow A$. Maximum discounted cumulative reward $V^*(s)$ can be learned to find the optimal policy through the following optimization:

$$\pi^*(s) = \underset{a}{\operatorname{argmax}} \left[r(s, a) + \gamma V^*(\delta(s, a)) \right]. \tag{8}$$

Here, the optimal action to choose in state s is a which maximizes the sum of current reward $r(s, a)$ and the value function $V^*(\cdot)$ of the successor state $\delta(s, a)$, discounted by γ. Since $V^*(\cdot)$ can be used to learn the optimal policy if only the agent knows the reward function r and the state transition function δ exactly, another evaluation function $Q(s, a)$ can be defined for the agent to learn the optimal policy even if it does not have any knowledge about r and δ. The value of $Q(\cdot, \cdot)$ function is the sum of immediate reward obtained after executing action a at state s, and the discounted value of following the optimal policy, i.e.,

$$Q(s, a) = r(s, a) + \gamma V^*(\delta(s, a)). \tag{9}$$

Then, the optimal policy can be rewritten as,

$$\pi^*(s) = \underset{a}{\operatorname{argmax}} Q(s, a). \tag{10}$$

By using the definition of optimal policy π^* as in (10), relationship between $Q(\cdot, \cdot)$ and $V^*(\cdot)$ can be given as,

$$V^*(s) = \max_{a'} Q(s, a'). \tag{11}$$

In this case, (9) can be also rewritten as,

$$Q(s, a) = r(s, a) + \gamma \max_{a'} Q(\delta(s, a), a'). \tag{12}$$

Then, $Q(\cdot, \cdot)$ can be iteratively approximated with Algorithm 1 by using (12). With this Algorithm, the agent estimates $Q(s, a)$ for state-action pair s and a, and the agent's estimate for this state-action pair is denoted by $\hat{Q}(s, a)$.

Algorithm 1 An algorithm for learning Q

Set time to $t = 0$.
for each state-action pair s and a **do**
  Initialize the table entry $\hat{Q}_t(s, a)$ to zero
end for
Observe the current state s at time $t = 0$
while $\hat{Q}_{t+1}(s, a)$ does not converge $Q(s, a)$ for each s, a **do**
  Choose an action a, perform it, and get a reward r
  Observe the new state s'
  Update $\hat{Q}_{t+1}(s, a) \leftarrow r + \gamma \max_{a'} \hat{Q}_t(s', a')$
  Set time to $t = t + 1$
  Set the current state as s' at time $t = t + 1$
end while

Q-Learning Based Hybrid Beamforming

We propose a novel hybrid beamforming algorithm which uses Q-learning to find the optimal RF precoders at the transmitter and receiver. In the setting of our problem, the environment is the mmWave channel, which is described by a continuous state space $\mathcal{H}$. Considering an agent exists in this environment, the agent can perform any of a set of possible actions $\mathcal{V}$, where $\mathcal{V}$ consists of a finite set of RF precoder pairs for the transmitter and receiver. At time t, the agent can select possible RF precoder pair $\mathbf{v}_t = \{(\mathbf{F}_{RF})_t, (\mathbf{W}_{RF})_t\} \in \mathcal{V}$ for the transmitter and receiver in some channel state $\mathbf{H}_t \in \mathcal{H}$. Then, the agent receives a real-valued reward R_t, which is the rate achieved over the mmWave channel at time t.

The Q-learning algorithm given in Algorithm 1 cannot be directly applied to hybrid precoding design problem since the state space has infinite continuous valued elements. Therefore, we propose a modified Q-learning algorithm for hybrid beamforming. In Q-learning based hybrid beamforming algorithm, we assume that the elements of state space, which consists of channel matrices at different time instances, are chosen from a finite training set. In other words, training set consists of N_S channel matrices such that $\mathcal{H} = \{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_{N_S}\}$. Action space $\mathcal{V}$ is a finite set of all possible RF precoder pairs so that it can be used directly in Q-learning algorithm.

Our proposed algorithm consists of two phases. The first phase is the training phase of the algorithm. During training, the table entries $\hat{Q}(\mathbf{H}_i, \mathbf{v}_j)$, where $\mathbf{H}_i \in \mathcal{H}$ and $\mathbf{v}_j \in \mathcal{V}$, are updated by using Algorithm 1. It is also important to decide for a strategy for the agent to choose from all possible RF precoders in channel state $\mathbf{H}$. We use a probabilistic approach for the agent in channel state $\mathbf{H}$ to select a possible pair of RF precoders. In this case, every RF precoder pair would have a nonzero probability to be selected, but RF precoder pairs with higher $\hat{Q}$ values are assigned higher probabilities. We define the probability of selecting a precoder pair $\mathbf{v}_i$, given that the agent is in channel state $\mathbf{H}$ as,

$$P(\mathbf{v}_i|\mathbf{H}) = \frac{c^{\hat{Q}(\mathbf{H},\mathbf{v}_i)}}{\sum_j c^{\hat{Q}(\mathbf{H},\mathbf{v}_j)}}, \tag{13}$$

where $c > 0$ is a constant that determines how strongly the agent would exploit RF precoder pairs with high $\hat{Q}$ values instead of exploring undiscovered RF precoder pairs. Once the learner's estimate $\hat{Q}(\mathbf{H}, \mathbf{v})$ converges to $Q(\mathbf{H}, \mathbf{v})$ for all $\mathbf{H}$ and $\mathbf{v}$ pairs, Q-table can be used to find optimal RF precoders.

Algorithm 2 Training Phase of Q-Learning Based Hybrid Beamforming Algorithm

Input: $\mathcal{V}$,$\mathcal{H}$,P_L,N_L,c,T
Set time to $t = 0$
for each state-action pair $\mathbf{H}$ and $\mathbf{v}$ **do**
  Initialize the table entry $\hat{Q}_t(\mathbf{H}, \mathbf{v})$ to zero
end for
Observe the current state $\mathbf{H}$ at time $t = 0$
for $t = 0, 1, ..., T$ **do**
  Choose $\mathbf{v} = \{(\mathbf{F}_{RF}), (\mathbf{W}_{RF})\}$ from $\mathcal{V}$ with probability $p(\mathbf{v}|\mathbf{H}) = \frac{c^{\hat{Q}(\mathbf{H},\mathbf{v})}}{\sum_{\mathbf{w}} c^{\hat{Q}(\mathbf{H},\mathbf{w})}}$
  $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^*$, $\mathbf{U} = [\mathbf{U}_1\mathbf{U}_2]$, $\mathbf{V} = [\mathbf{V}_1\mathbf{V}_2]$, $\mathbf{U}_1 \in \mathbb{C}^{N_R \times N_L}$, $\mathbf{V}_1 \in \mathbb{C}^{N_T \times N_L}$
  $\mathbf{F}_{opt} = \mathbf{V}_1$ and $\mathbf{W}_{opt} = \mathbf{U}_1$
  $\mathbf{F}_{BB} = (\mathbf{F}_{RF}^* \mathbf{F}_{RF})^{-1} \mathbf{F}_{RF}^* \mathbf{F}_{opt}$ and $\mathbf{W}_{BB} = (\mathbf{W}_{RF}^* \mathbf{W}_{RF})^{-1} \mathbf{W}_{RF}^* \mathbf{W}_{opt}$
  $\mathbf{F}_{BB} = \sqrt{N_L} \frac{\mathbf{F}_{BB}}{\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F}$ and $\mathbf{W}_{BB} = \sqrt{N_L} \frac{\mathbf{W}_{BB}}{\|\mathbf{W}_{RF}\mathbf{W}_{BB}\|_F}$
  Calculate rate R by using Equation (5)
  Observe the new state $\mathbf{H}'$
  Update $\hat{Q}_{t+1}(\mathbf{H}, \mathbf{v}) \leftarrow R + \gamma \max_{\mathbf{v}'} \hat{Q}_t(\mathbf{H}', \mathbf{v}')$
  Set time to $t = t + 1$
  Set the current state as $\mathbf{H}'$ at time $t = t + 1$
end for

In the second phase, optimal RF precoders at the transmitter and receiver are designed by using the table entries $\hat{Q}(\mathbf{H}, \mathbf{v})$ according to given CSI, which can be complete or partial. Let us assume that given CSI is $\mathbf{H}$ at a certain time, the Q-learning based hybrid beamforming algorithm searches over the all states in training set $\mathcal{H} = \{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_{N_S}\}$ and selects the channel $\tilde{\mathbf{H}}$, where $\min_{\tilde{\mathbf{H}} \in \mathcal{H}} \left\|\tilde{\mathbf{H}} - \mathbf{H}\right\|_2$. $\tilde{\mathbf{H}}$ corresponds to a minimum Euclidean distance estimate of the channel $\mathbf{H}$ by one of the elements in the training set. By using the table entries $\hat{Q}(\mathbf{H}, \mathbf{v})$, the RF precoder pair $\mathbf{v}'$ for the transmitter and receiver, which gives the maximum Q-value for the state $\tilde{\mathbf{H}}$, are selected.

Algorithm 3 RF Precoders Selection Phase of Q-Learning Based Hybrid Beamforming Algorithm

Input: $\mathbf{H}$,$\mathcal{H}$,$\mathcal{V}$,$\hat{Q}(\mathbf{H}, \mathbf{v})$,
$\tilde{\mathbf{H}} = \min_{\tilde{\mathbf{H}} \in \mathcal{H}} \left\|\tilde{\mathbf{H}} - \mathbf{H}\right\|_2$
Choose $\mathbf{v} = \{(\mathbf{F}_{RF}), (\mathbf{W}_{RF})\}$, where $\mathbf{v} = \max_{\mathbf{v}} \hat{Q}(\tilde{\mathbf{H}}, \mathbf{v})$
$\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^*$, $\mathbf{U} = [\mathbf{U}_1\mathbf{U}_2]$, $\mathbf{V} = [\mathbf{V}_1\mathbf{V}_2]$, $\mathbf{U}_1 \in \mathbb{C}^{N_R \times N_L}$, $\mathbf{V}_1 \in \mathbb{C}^{N_T \times N_L}$
$\mathbf{F}_{opt} = \mathbf{V}_1$ and $\mathbf{W}_{opt} = \mathbf{U}_1$
$\mathbf{F}_{BB} = (\mathbf{F}_{RF}^* \mathbf{F}_{RF})^{-1} \mathbf{F}_{RF}^* \mathbf{F}_{opt}$ and $\mathbf{W}_{BB} = (\mathbf{W}_{RF}^* \mathbf{W}_{RF})^{-1} \mathbf{W}_{RF}^* \mathbf{W}_{opt}$
$\mathbf{F}_{BB} = \sqrt{N_L} \frac{\mathbf{F}_{BB}}{\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F}$ and $\mathbf{W}_{BB} = \sqrt{N_L} \frac{\mathbf{W}_{BB}}{\|\mathbf{W}_{RF}\mathbf{W}_{BB}\|_F}$
return $\mathbf{F}_{BB}$, $\mathbf{F}_{RF}$, $\mathbf{W}_{BB}$, $\mathbf{W}_{RF}$

Two phases of our proposed algorithm are summarized in Algorithm 2 and Algorithm 3.

Results

In this section, we first give the computational complexity analysis of Q-learning based hybrid beamforming algorithm. Then, we present our Matlab based simulation results.

A. Computational Complexity Analysis of Q-Learning Based Hybrid Beamforming

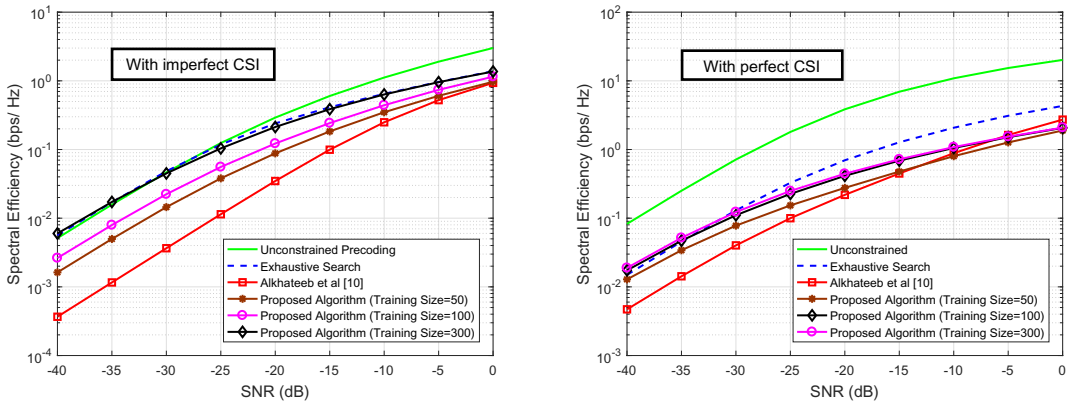
Total computation complexity results from the two phases of Q-learning based hybrid beamforming. Computational complexity[1] of training and optimal RF precoders selection phases are $O(T(N_R(N_T)^2+N_S(N^T_{Beams})^{N^T_{RF}}(N^R_{Beams})^{N^R_{RF}}+(N^T_{RF})^3+N_LN^R_{RF}N_R+N_LN_RN_T+N_LN^T_{RF}N_T+(N_L)^3))$ and $O(N_TN_RN_S+(N^T_{Beams})^{N^T_{RF}}(N^R_{Beams})^{N^R_{RF}}+N_R(N_T)^2+(N^T_{RF})^3+N^T_{RF}N_TN_L+(N^R_{RF})^3+N^R_{RF}N_RN_L)$, respectively. It is important to note that the computational complexity of exhaustive search is $O(((N^T_{Beams})^{N^T_{RF}}(N^R_{Beams})^{N^R_{RF}})(N_LN^R_{RF}N_R+N_LN_RN_T+N_LN^T_{RF}N_T+(N_L)^3+N_R(N_T)^2+(N^T_{RF})^3+N^T_{RF}N_TN_L+(N^R_{RF})^3+N^R_{RF}N_RN_L))$. For $T=100$, $N_S=20$, $N_T=N_R=64$, $N^T_{Beams}=N^R_{Beams}=8$, $N^T_{RF}=N^R_{RF}=4$, $N_L=3$, number of operations for training and optimal RF precoders selection phases are 3.4×10^{10} and 1.7×10^7, respectively. Therefore, the total number of operations for our proposed algorithm is approximately 3.4×10^{10}. With the exhaustive search, 4.7×10^{12} number of operations is required for the same N_T, N_R, N^T_{Beams}, N^R_{Beams}, N^T_{RF}, N^R_{RF}, and N_L. In this case, 138 times reduction is achieved with the proposed algorithm compared to the exhaustive search. The computational complexity of the training phase of the proposed algorithm increases linearly with the number of iterations T. However, this computation cost is compensated with a large number of RF chains at the transmitter and receiver.

B. Simulation Results

In our simulation, we consider a mmWave system with one transmitter-receiver pair. We use the hybrid architecture given in Figure 1. The carrier frequency and the bandwidth of the mmWave system are chosen as 28 GHz and 100 MHz, respectively. The transmitter has $N_T=64$ antennas and 2 RF chains, and the receiver has $N_R=32$ antennas and 2 RF chains. Uniform Linear Arrays (ULAs) are used as the antenna arrays in which the spacing between antennas are $\frac{\lambda}{2}$. The RF phase shifters have quantized phases, and the number of inputs to the phase shifters is $N_q=3$. Size of the space $\mathcal{V}$ equals to 4096. The channel model given in (4) is used in the simulations. In the channel model, $\overline{G}=1$ and the number of paths $S=3$. The AoAs/AoDs of the channel are continuous-valued random variables with uniform distribution $[0,2\pi]$.

Figure 2-a shows the spectral efficiency of the proposed algorithm with different number of training samples $N_S=50$, $N_S=100$, and $N_S=300$ for channel states when SNR increases from -40 dB to 0 dB. Figure 2-a also includes spectral efficiency of the suboptimal solution proposed in [10], exhaustive search, and unconstrained precoding. In Figure 2-a, it is assumed that CSI is incomplete. Therefore, estimated channel matrices are used in the training phase. In RF precoders selection phase of the proposed algorithm, observed channel states in real-time are also the estimated channel matrices. It can be seen in Figure 2-a that unconstrained precoding outperforms all of the other hybrid beamforming algorithms. However, the mmWave system achieves higher spec-

[1]Analysis of the computational complexity of the algorithm is omitted due to space limitation.



(a) Imperfect CSI. (b) Perfect CSI

Figure 2: Spectral efficiency achieved by the proposed algorithm for different training sizes, the method in [10], exhaustive search, and unconstrained precoding.

tral efficiency with our proposed algorithm for all training sizes than the algorithm given in [10]. The spectral efficiency achieved with our algorithm when training size equals to 300 is very close to the exhaustive search based solution.

Figure 2-b shows the spectral efficiency of the proposed algorithm with different number of training samples $N_S = 50$, $N_S = 100$, and $N_S = 300$ for channel states when SNR increases from -40 dB to 0 dB. Results of the proposed algorithm are also compared with the algorithm proposed in [10], exhaustive search, and unconstrained precoding in Figure 2-b. In this case, it is considered that CSI is complete. Unconstrained precoding outperforms all of the other hybrid beamforming algorithms as it is shown in Figure 2-b. The mmWave system achieves higher spectral efficiency with our algorithm for all training sizes than the algorithm given in [10] for smaller values of SNR. However, the algorithm proposed in [10] starts to achieve better spectral efficiency than our algorithm for higher values of SNR. Furthermore, spectral efficiency achieved with our algorithm with an increasing number of training samples is still less than the exhaustive search.

Conclusion

In this paper, we presented a reinforcement learning approach for the hybrid beamforming problem. We give a computational complexity analysis for our proposed algorithm. Finally, we compare the performance of our algorithm in terms of achieved rate over the mmWave channel with unconstrained precoding, exhaustive search, and a suboptimal hybrid beamforming algorithm. We show the results of our algorithm with different sized training sets. We also compare the performance of our algorithm when CSI is perfect and imperfect. It is seen that the performance of the proposed algorithm improves with larger training sets. The results of the proposed algorithm are promising since it gives better spectral efficiency than other suboptimal algorithms and has a very close performance to exhaustive search when the available CSI is not perfect. Future directions include providing a theoretical analysis to show our Euclidean distance-based approach for using Q-table is guaranteed to converge and adapting the proposed algorithm for multi-user scenarios.

Acknowledgments

This work was partially supported by US NSF through grant CNS-1715947. It was also partly supported by the Broadband Wireless Access and Applications Center (BWAC); NSF Award No. 1265960.

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1065–1082, June 2014.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.

[3] A. Abbaspour-Tamijani and K. Sarabandi, "An affordable millimeter-wave beam-steerable antenna using interleaved planar subarrays," *IEEE Transactions on Antennas and Propagation*, vol. 51, pp. 2193–2202, Sep 2003.

[4] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, pp. 366–385, March 2014.

[5] S. Han, C. l. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5g," *IEEE Communications Magazine*, vol. 53, pp. 186–194, January 2015.

[6] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 61, pp. 4391–4403, October 2013.

[7] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace mimo for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Transactions on Antennas and Propagation*, vol. 61, pp. 3814–3827, July 2013.

[8] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave mimo systems," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 1499–1513, March 2014.

[9] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *2013 Information Theory and Applications Workshop (ITA)*, pp. 1–5, Feb 2013.

[10] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. H. Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *CoRR*, vol. abs/1401.7426, 2014.

[11] F. B. Mismar and B. L. Evans, "Deep reinforcement learning for improving downlink mmwave communication performance," *CoRR*, vol. abs/1707.02329, 2017.

[12] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1164–1179, June 2014.

[13] T. S. Rappaport, Y. Qiao, J. I. Tamir, J. N. Murdock, and E. Ben-Dor, "Cellular broadband millimeter wave propagation and angle of arrival for adaptive beam steering systems (invited paper)," in *2012 IEEE Radio and Wireless Symposium*, pp. 151–154, Jan 2012.

[14] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc., 1 ed., 1997.