Learning State Selection for Reconfigurable Antennas: A Multi-Armed Bandit Approach

Nikhil Gulati, *Member, IEEE*, and Kapil R. Dandekar, *Senior Member, IEEE*

***Abstract*—Reconfigurable antennas are capable of dynamically re-shaping their radiation patterns in response to the needs of a wireless link or a network. In order to utilize the benefits of reconfigurable antennas, selecting an optimal antenna state for communication is essential and depends on the availability of full channel state information for all the available antenna states. We consider the problem of reconfigurable antenna state selection in a single user MIMO system. We first formulate the state selection as a multi-armed bandit problem that aims to optimize arbitrary link quality metrics. We then show that by using online learning under a multi-armed bandit framework, a sequential decision policy can be employed to learn optimal antenna states without instantaneous full CSI and without *a priori* knowledge of wireless channel statistics. Our objective is to devise an adaptive state selection technique when the channels corresponding to all the states are not directly observable and compare our results against the case of a known model or genie with full information. We evaluate the performance of the proposed antenna state selection technique by identifying key link quality metrics and using measured channels in a 2 × 2 MIMO OFDM system. We show that the proposed technique maximizes long term link performance with reduced channel training frequency.**

***Index Terms*—Beamsteering, cognitive radio, MIMO, multi-armed bandit, OFDM, online learning, reconfigurable antennas.**

I. Introduction

Reconfigurable antenna technology has gained a lot of attention in recent years for applications in wireless communications. Both theoretical and experimental studies have shown that reconfigurable antennas can offer additional performance gains in multiple input multiple output (MIMO) systems by increasing channel capacity [2], [3]. These have also been shown to perform well in low SNR regimes [4]. Gradually making their way into commercial wireless systems [5], [6], these antennas bring two major benefits to traditional multi-element wireless systems. First, the additional degree of freedom to dynamically alter radiation patterns enable MIMO systems to adapt to physical link conditions. This adaptation leads to enhanced performance and can provide robustness to varying channel conditions. Second, these antennas provide space and cost benefits either by incorporating multiple elements in a single physical device [7] or by reducing the number of RF chains [8], [9]. Beyond single user MIMO systems, reconfigurable antennas have more recently been used for improving advanced interference management techniques such as interference alignment (IA) [10]–[12]. With the continued interest in developing practical cognitive radios having greater autonomy, learning, and inference capabilities, the integration of reconfigurable antennas in cognitive radios along with inteligent algorithms to control them, is going to play a significant role.

One of the key challenge to effectively use the reconfigurability offered by these antennas and integrate them in practical wireless systems, is to select an optimal radiation state[1] (in terms of capacity, SNR, diversity etc.) among all the available states for a wireless transceiver in a given wireless environment. There are two fundamental challenges to achieve this goal: (*i*) the requirement of additional channel state information (CSI) corresponding to each state for each transmitter-receiver pair; (*ii*) The amount and the frequency of the channel training required and the associated overhead. These challenges become even more difficult to overcome when reconfigurable antennas are employed at both the ends of the RF link thus creating a large search space in order to find an optimal radiation state for communication. Moreover, the effect of node mobility, changes in physical antenna orientation, and the dynamic nature of the wireless channel can render previously found "optimal" states suboptimal over time.

Translating the benefits of reconfigurable antennas into a practical realizable MIMO system is thus a highly challenging task. Existing antenna state selection techniques (see Section II-B) have primarily relied on the availability of perfect instantaneous CSI coupled with modified periodic pilot based training to perform state selection. Besides the performance loss caused by imperfect or incomplete CSI, the additional problem of changing the data frame to enable additional channel training render these approaches impractical for use in systems such as IEEE 802.11x and IEEE 802.16 as the number of available states increase.

In this work, instead of relying on the availability of full, instantaneous CSI and periodic channel training, we ask the following questions:

Manuscript received October 15, 2012; revised February 06, 2013; accepted April 23, 2013. Date of publication August 02, 2013; date of current version February 27, 2014. This work was supported in part by NSF Grant #0916480. Preliminary version of this work appeared in IEEE RWS, 2012.

The authors are with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104 USA (e-mail: nikhil.gulati@drexel.edu; dandekar@coe.drexel.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAP.2013.2276414

[1] State refers to either the radiation pattern selected at the receiver or the transmitter or a combination of radiation patterns selected at the receiver and transmitter respectively.

- Can a reconfigurable antenna system learn to select an optimal radiation state without the availability of instantaneous full CSI and periodic training while optimizing an arbitrary link quality metric?
- Can a reconfigurable antenna system also adapt to varying channel conditions without extensive channel training and parameter optimization?

Towards addressing these challenges, we present an online learning framework for reconfigurable antenna state selection, based on the theory of multi-armed bandit. We model a system where each transmitter-receiver pair (i.e., a link) develops a sequential decision policy to maximize the link throughput. For each decision, the system receives some reward from the environment which is assumed to be an i.i.d random process with an arbitrary distribution. It is also assumed that mean reward for each link is unknown and is to be determined online by the learning process. Specifically, we provide the following contributions in this work:

1) We first present a learning framework to learn the optimal state selection based on the theory of multi-armed bandit, in order to maximize system performance.
2) We identify key link quality metrics which can be used with the learning framework to assess the long-term performance of the system.
3) We implement and evaluate the learning algorithm in a practical IEEE 802.11x based single user MIMO-OFDM system to evaluate the performance over measured wireless channels.

The rest of our paper is organized as follows: In Section II, we provide a background on reconfigurable antennas and multi-armed bandit theory along with related work on both the topics. Section III describes the MIMO system model and explains the multi-armed bandit formulation for reconfigurable antenna state selection. In Section IV, we describe the selection policies and the reward metrics used to evaluate the performance of the proposed schemes. In Section V, we provide a description of experimental setup, hardware components, and the implementation of the learning policies. Section VI provides detailed performance analysis and empirical evaluation followed by the conclusion in Section VII.

II. Background and Related Work

A. Background

Pattern Reconfigurable Antennas: With the introduction of reconfigurable antennas, there was a departure from the notion that antennas can only have fixed radiation characteristics. Reconfigurable antennas are capable of changing the operating characteristics of the radiating elements through either electrical, mechanical or other means. Over the last ten years, research has primarily been focused on designing reconfigurable antennas with the ability to dynamically change either frequency [13], radiation pattern [3] and polarization [14] or the combination of one of these properties. Reconfigurability is based on the fact that the change in the current distribution in the antenna structure effects the spatial distribution of radiation from the antenna element [15]. Current distribution in the reconfigurable antenna can be modified by mechanical/structural changes, material variations or using electronic components like PIN diodes [16] and MEMS switches [17]. In this paper, we focus only on the pattern reconfigurable antennas and using the pattern diversity offered by them.

Multi-Armed Bandit Theory: The multi-armed bandit problem (MAB) is a fundamental mathematical framework for learning unknown variables. It embodies the well known *exploitation vs. exploration* trade off seen in reinforcement learning and limited feedback learning. In the classic formulation [18]–[20], there are N independent arms with a single player, playing arm $i(i = 1, \dots N)$. The trade-off involves choosing an arm with the highest expected payoff using current knowledge and exploring the other arms to acquire more knowledge about the expected pay-offs of the rest of the arms. On each play of a single arm, the player receives a random reward. The goal is to design a policy to play one arm at each time sequentially to maximize the total expected reward in the long run. In the seminal work of Gittins [21], it was shown that the optimal policy of MAB with Bayesian formulation where reward distributions are known, has a simple index structure, i.e., Gittins index. Within the non-Bayesian framework, Lai and Robbins [18] provided a performance measure of an arm selection policy referred to as *regret or cost of learning*. Regret is defined as the difference in the expected reward gained by always selecting the optimal choice and the reward obtained by a given policy. Since the best arm cannot always be identified in most cases using a finite number of prior observations, the player will always have to keep learning. Due to the continuous learning process, the player will make mistakes which will grow the regret over time. The goal of the learning policies under multi-armed bandit framework is to keep the regret as low as possible and bound the growth of regret over time.

B. Related Work

1) Pattern Reconfigurable Antennas and Beam Selection: Though the architecture of pattern reconfigurable antennas and their performance in MIMO systems have received significant attention, there are only a handful of studies which focus on practical control algorithms and optimal state selection techniques [1]. In [22], the authors estimate the channel response for each antenna state at the transmitter and receiver using pilot based training and select an optimal state combination. They theoretically show the effect of channel estimation error on link capacity and power consumption. Since, their technique relies on the availability of full CSI at every time slot, a change in the data frame is required to enable additional channel training which can lead to a loss of capacity as the number of antenna states increase. The authors in [23], provide a selection technique based on second order channel statistics and average SNR information for state selection without changing the OFDM data frame. This technique maximizes the average system performance over long run. Their selection technique relies on building offline lookup tables for a given SNR and power angular spread for a given environment and is not adaptive in nature. Periodic exhaustive training techniques with reduced overhead is presented in [24] where the authors highlight the effect of channel training frequency on the capacity and the bit-error rate (BER) of a MIMO system. In order to reduce the overhead of exhaustive training for all the beam combinations, the authors make assumptions on the prior availability of statistics of the channel in order to eliminate sub-optimal beams. But in their work it is not clear how to re-introduce the excluded states which may become optimal over time due to channel variations. Though some of these techniques were successful in showing the benefits of multi-beam selection and motivated the need for a selection algorithm, none solved the challenges mentioned above. Moreover, all the schemes mentioned above provide only simulated results without any practical implementation and assume the availability of full CSI from the periodic channel training. In our work, we neither assume full CSI for all states at every time slot nor do we perform periodic exhaustive training. Our approach does not require a change in the frame structure and can still adapt on a per packet basis.

2) Multi-Armed Bandit (MAB) for Cognitive Radio: Stochastic online learning via multi-armed bandit has gained significant attention for applications in opportunistic and dynamic spectrum access. The authors in [25] applied the multi-armed bandit formulation to the secondary user channel access problem in cognitive radio network. Later, a restless MAB formulation for opportunistic multi-channel sensing for secondary user network was presented and evaluated in [26]. Further in [27], a combinatorial version of MAB was proposed for a network with multiple primary and secondary users taking into account the collisions among the secondary users. Distributed channel allocation among multiple secondary users was further studied and proposed in [28] and [29]. Another application of multi-armed bandit for cognitive radio was proposed in [30] for adaptive modulation and coding. The application of online learning for cognitive radio for dynamic spectrum access can be enhanced by either using pattern reconfigurable antennas to avoid interference using spatial means or can be combined with spectrum sensing where a secondary user can opportunistically select a radiation state once an unoccupied channel is found using channel sensing.

Notation: We use capital bold letters to denote matrices and small bold letters for vectors. $\mathbf{H}^{-1}$, $\mathbf{H}^{\dagger}$ and $\mathbf{H}^{T}$ denote the matrix inverse, Hermitian and transpose operation respectively. $\|\mathbf{H}\|_F$ represents the Frobenius norm of $\mathbf{H}$ respectively. The $\mathrm{d} \times \mathrm{d}$ identity matrix is represented by $\mathbf{I}_d$.

III. System Model and Bandit Formulation

A. MIMO System Model with Reconfigurable Antennas

Consider a point to point MIMO link with M pattern reconfigurable antennas at the transmitter (Tx) and N pattern reconfigurable antennas at the receiver (Rx). We assume that reconfigurable antennas at the transmitter have $\mathcal{J}$ such unique states and the antennas at the receiver are capable of switching between $\mathcal{K}$ such states. The degree of correlation between all the resulting $\mathcal{J} \times \mathcal{K}$ combination of channel realizations is governed by the physical structure of the reconfigurable antenna. We employ the V-BLAST [31] architecture for transmission of the spatial multiplexed input symbol $\mathbf{x}$ ($\in \mathbb{C}^{M\times 1}$) (i.e., each antenna element carries an independent stream of information). Further, in order to approximate the channel as having flat fading, we employ OFDM as the multicarrier modulation scheme. Under such a setting, the received signal is given as

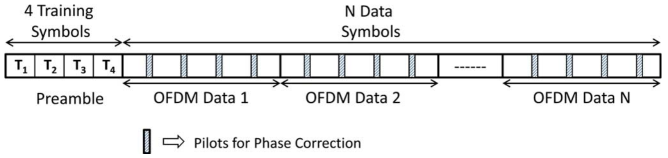


Fig. 1. OFDM frame structure. Preamble is loaded with 4 training symbols for channel training of a single selected antenna state.

$$\mathbf{y}(f) = \mathbf{H}_{k,j}(f)\mathbf{x}(f) + \mathbf{n}(f) \tag{1}$$

where f denotes the OFDM subcarrier index, k, j represent the antenna state of the reconfigurable antenna selected at the receiver and transmitter respectively, $\mathbf{y}$ is the $N \times 1$ received signal vector, $\mathbf{H}_{k,j}$ is the $N \times M$ MIMO channel between the transmitter and the receiver, $\mathbf{x}$ is $M \times 1$ and $\mathbf{n}$ represents the $N \times 1$ vector of complex zero mean Gaussian noise with covariance matrix $\mathbb{E}[\mathbf{n}\mathbf{n}^{\dagger}] = \sigma^2 \mathbf{I}_N$. For brevity, we will drop the symbols k, j and f. The input vector $\mathbf{x}$ is subject to an average power constraint, $\mathbb{E}[Tr(\mathbf{x}\mathbf{x}^{\dagger})] = P$ with total power equally distributed across the input streams, i.e. the input covariance matrix is given by $\mathbf{Q} = (P/M)\mathbf{I}_M$. The set containing all the combinations of states of the reconfigurable antenna at the transmitter and receiver will be represented by the vector $\Omega = \{j \times k : j \in \mathcal{J}, k \in \mathcal{K}\}$.

Frame Structure and Channel Estimation: To enable frequency domain estimation of MIMO channel coefficients at the receiver, every OFDM frame carries four training symbols in the preamble, two for each antenna element as shown in Fig. 1. Each transmit antenna is trained orthogonally in time, i.e., while the training symbols are transmitted from one transmit antenna, the other transmit antenna stays silent. The received training symbols are then used to estimate the MIMO channel using a least squares estimator. Let $\mathbf{t}_1$ and $\mathbf{t}_2$ represent the vector of training symbols and $\mathbf{T}$ represent the matrix of training symbols, i.e. $\mathbf{T} = [\mathbf{t}_1 \ \mathbf{t}_2]$. Then, the least square estimate of a single tap channel is given as

$$\mathbf{H}_{\mathrm{LS}} = \mathbf{Y}\mathbf{T}^{\dagger}\left(\mathbf{T}\mathbf{T}^{\dagger}\right)^{-1} \tag{2}$$

where $\mathbf{Y}$ represents the matrix of received signal vectors corresponding to both the training symbols. It should be noted that (2) is evaluated for every OFDM subcarrier f. In addition to the channel estimates obtained via (2), each OFDM symbol carries 4 evenly placed pilot tones, which are used for phase correction at the receiver.

$$\mathbf{H}_{\mathrm{phase-corrected}} = \mathbf{H}_{\mathrm{LS}} e^{-j\phi_i} \tag{3}$$

where ϕ_i represents the average phase of the 4 pilot tones for the ith OFDM symbol in a frame. These frequency domain channel estimates are then used to carry out channel equalization using zero forcing-successive interference cancelation (ZF-SIC) equalizer. To further enhance the performance of ZF-SIC, symbols are decoded based on optimal SNR ordering and then finally combined using maximal ratio combining (MRC). We estimate the channel for a single antenna state using the OFDM frame structure described above. Therefore, this frame structure

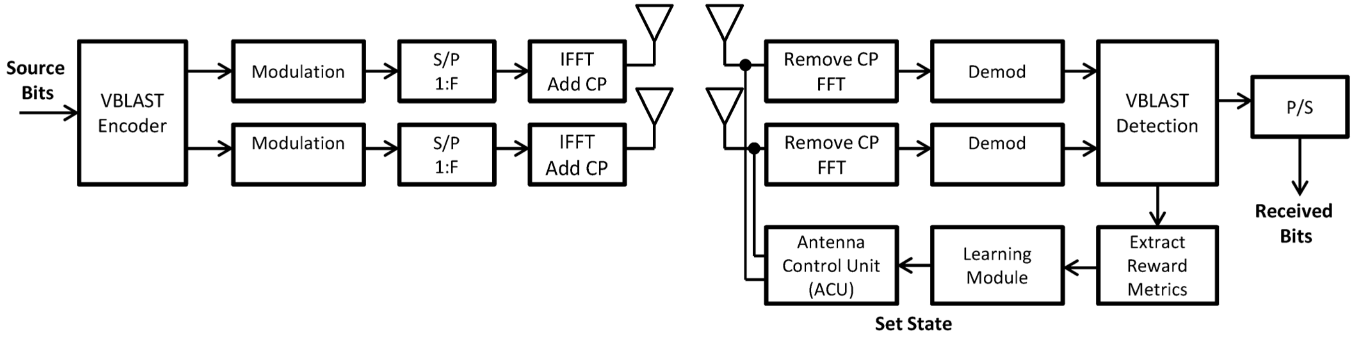


Fig. 2. System diagram of 2 × 2 MIMO OFDM system with reconfigurable antennas and the receiver employing learning module for antenna state selection.

does not change as the number of antenna states are increased. Once the channel estimates are available for the selected antenna state at each time slot, the reward metrics are extracted and used as input to the learning model to select the antenna state for next time slot as shown in Fig. 2.

B. Bandit Formulation for Antenna State Selection

Our work is influenced by the formulation in [20] where arms have non-negative rewards that are i.i.d over time with an arbitrary un-parametrized distribution. We consider the set up where there is a single transmitter and L wireless receivers and both the transmitter and the receiver employ pattern reconfigurable antennas. The receivers can select from K available antenna states and the state at the transmitter is fixed which reduces the problem to selecting an antenna state only at the receiver where each receiver can select state i independently. It can be shown that this framework can be easily extended to the case where the antenna state at the transmitter is not fixed and state selection is also performed for the transmitter. In that case an antenna state will refer to a combination of the radiation state j at the transmitter and radiation state i at the receiver. In the context of multi-armed bandit a radiation state i is interchangeably referred to as an arm i. The decision is made at every time slot (packet) n to select the antenna state to be used for the next reception. If a receiver selects a state i and assuming the transmission is successful, an instantaneous random reward is achieved which we denote as $R_i(n)$. This reward is assumed to evolve as an i.i.d random process and the mean of this random process is unknown to the receiver. Without loss of generality, we normalize $R_i(n) \in [0,1]$. When a receiver selects an antenna state i, the reward $R_i(n)$ is only observed by that receiver and the decision is made only based on the locally observed history. Also, the reward is only observed for the selected state i and not for the other states. In other words, a receiver receives channel state information for only the selected radiation state at a given time slot and acquires no new information about the other available radiation states. In this way our proposed technique differs from the other techniques in the literature as it does not rely on the availability of instantaneous CSI for all the radiation states at each time slot.

We represent the unknown mean for a given state i as $\bar{x}_i = E[R_i]$. The collection of these means for all states is then represented as $\bar{X} = E\{\bar{x}_i, 1 \leq i \leq K\}$. We further define the deterministic policy $\pi(n)$ at each time serving as a mapping between the reward history $\{R_k\}_{k=1}^{n-1}$ and the vector of antenna states $r(n)$ to be selected at time slot n where receiver l selects antenna state $r_l(n)$. The goal is to design policies for this multi-armed bandit problem that perform well with respect to *regret*. Intuitively, the regret should be low for a policy and there should be some guarantees on the growth of regret over time. Formally, the regret of a policy after n plays is given by (4).

$$\mu^* n - \sum_{i=1}^{K} \mu_i E\left[T_i(n)\right] \tag{4}$$

where

$$\mu^* = \max\nolimits_{1\leq i\leq K} \mu_i. \tag{5}$$

μ^* is the average reward of the optimal antenna arm, μ_i is the average reward for arm i, n is the number of total trials. $E[\cdot]$ is the expectation operator and $T_i(n)$ is the number of times arm i has been sampled up to time slot n. It has been shown in [18] that the minimum rate at which regret grows is of logarithmic order under certain regularity conditions. The authors established that for some families of reward distributions there are policies that can satisfy

$$E\left[T_i(n)\right] \leq \left(\frac{1}{D\left(\mu_i \| \mu^*\right)} + o(1)\right) \ln(n) \tag{6}$$

where $o(1) \rightarrow 0$ as $n \rightarrow \infty$ and

$$D\left(\mu \| \mu^*\right) \equiv \int \mu_i \ln \frac{\mu_i}{\mu_*} \tag{7}$$

is the Kullback-Leibler divergence between the reward density μ_i of a suboptimal arm i and the reward density of the optimal machine μ^*. Therefore, over infinite horizon, the optimal arm is played exponentially more often than any other arm.

IV. Policies and Reward

A. Selection Policies

A learning policy for antenna state selection (also referred to interchangeably as selection technique) must overcome certain challenges. We identify such challenges below.

1) Optimal antenna state for each wireless link (between a single transmitter and a receiver location) is unknown *a priori*. Moreover, each wireless link may have a different optimal state. A selection technique should be able to learn and find the optimal state for a given link.

2) For a given wireless link, there might be several states which are near optimal over time based on channel conditions and multipath propagation. A selection technique should provide a policy to balance between exploiting a known successful state and exploring other available states without excessive retraining to account for the dynamic behavior of the channel.
3) For the purpose of real-time implementation in a practical wireless system, a selection technique must be computationally efficient and employ simple metrics which can be extracted from the channel without large overhead or extensive feedback data.

Most of the learning policies for the non-Bayesian multi-armed bandit problem in the literature, works by associating an index called *upper confidence index* to each arm. The calculation of such an index relies on the entire sequence of rewards obtained up to a point from selecting a given arm. The computed index for each arm is then used as an estimate for the corresponding reward expectations and is used to select the arm with highest index.

We base the antenna state selection technique on the deterministic policy UCB1 and its variants as given in [20].

1) UCB1—Selection Policy: To implement the UCB1 policy, each receiver stores and updates two variables; the average of all the instantaneous reward values observed for state i up to the current packet n denoted as $\bar{R}_i(n)$ (sample mean) and the number of times antenna state i has been selected up to the current packet n, denoted as $n_i(n)$. The two quantities $\bar{R}_i(n)$ and $n_i(n)$ are updated using the following update rule:

$$\bar{R}_i(n) = \begin{cases} \frac{\bar{R}_i(n-1)n_i(n-1)+R_i(n)}{n_i(n-1)+1} & \text{if state } i \text{ is selected} \\ \bar{R}_i(n-1) & \text{else} \end{cases} \tag{8}$$

$$n_i(n) = \begin{cases} n_i(n-1)+1 & \text{if state } i \text{ is selected} \\ n_i(n-1) & \text{else.} \end{cases} \tag{9}$$

The UCB1 policy as shown in Algorithm 1, first begins by selecting each antenna state at least once and $\bar{R}_i(n)$ and $n_i(n)$ are then updated using 8 and 9. Once the initialization is completed, the policy selects the state that maximizes the criteria on line 6. From line 6, it can be seen that the index of the policy is the sum of two terms. The first term is simply the current estimated average reward. The second term is the size of the one-sided confidence interval of the estimated average reward within which the true expected value of the mean falls with a very high probability. The estimate of the average reward improves and the confidence interval size reduces as the number of times an antenna state is selected increases. Eventually, the estimated average reward reaches as close as possible to the true mean. The size of the confidence interval also governs the index of the arm for future exploration.

For an instance, consider a two-armed bandit problem shown in Fig. 3. In order to select between the two arms, if only the average reward $\bar{R}_1$ and $\bar{R}_2$ are considered, clearly arm 1 will be selected. However, this will ignore the confidence in the other arm in order to explore higher pay-offs. By adding the size of the confidence interval to the index term, even the though the $\bar{R}_2 < \bar{R}_1$, arm 2 is selected for further exploration instead of arm 1. Since the confidence interval size depends on the number of times an arm has been played as well as the total number of trials, an arm which has been played only a few times and has estimated average reward near to an optimal, will be chosen to be played.

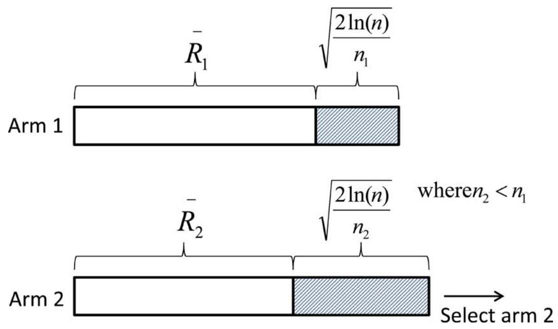


Fig. 3. Illustration of UCB1 Policy for a 2-armed bandit problem.

Algorithm 1 UCB1 Policy, Auer *et al.* [20]

```
1: // Initialization
2: $n_i, \bar{R}_i \leftarrow 0$
3: Select each antenna state at least once and update $n_i$, $\bar{R}_i$ accordingly.
4: // Main Loop
5: while 1 do
6:     Select antenna state $i$ that maximizes $\bar{R}_i + \sqrt{2\ln(n)/n_i}$
7:     Update $n_i, \bar{R}_i$ for antenna state $i$
8: end while
```

The UCB1 policy has an expected regret of at most [20]

$$\left[8 \sum_{i:\mu_i<\mu^*} \frac{\ln n}{\Delta_i}\right] + \left(1+\frac{\pi^2}{3}\right)\left(\sum_{i:\mu_i<\mu^*} \Delta_i\right) \tag{10}$$

where $\Delta_i = \mu^* - \mu_i$

2) UCB1-Tuned Policy: For practical implementations, it has been shown that by replacing the upper confidence bound of UCB1 with a different bound to account for the variance in the reward yields better results [20]. In UCB1-Tuned, the exploration term is given as

$$\sqrt{\frac{\ln(n)}{n_i} \min\left\{\frac{1}{4}, V_i(n_i)\right\}} \tag{11}$$

where V_i is defined as

$$V_i(s) \equiv \left(\frac{1}{s}\sum R_{i,s}^2\right) - \bar{R}_{i,s}^2 + \sqrt{\frac{2ln(t)}{s}} \tag{12}$$

when state i has been selected s times during the first t time slots. Therefore, the UCB1-Tuned policy can now be written as Algorithm II. Even though UCB1-Tuned policy has been shown to work experimentally well, there are no proofs available for the regret bounds in the literature. We expect this policy to work best for the scenarios where the link quality metrics have large variance due to dynamic channel variations.

Algorithm 2 UCB1-Tuned Policy, Auer *et al.* [20]

```
// Initialization
$n_i, \bar{R}_i \leftarrow 0$
Select each antenna state at least once and update $n_i, \bar{R}_i$ accordingly.
// Main Loop
**while** 1 **do**
    Select antenna state $i$ that maximizes $\bar{R}_i + \sqrt{(\ln(n)/n_i)\min\{1/4, V_i(n_i)\}}$
    Update $n_i, \bar{R}_i$ for antenna state $i$
**end while**
```

3) UCB1-Normal Policy: The two UCB1 polices described above made no assumptions on the reward distributions. In the UCB1-Normal policy, it is assumed that the reward follows a normal distribution with unknown mean and variance. This is a special case and it is shown in [20] that the policy in Algorithm 3 achieves logarithmic regret uniformly over n time slots. The index associated with each arm is still calculated based on the one sided confidence interval of the average reward, but since the reward distribution is known, instead of using the Chernoff-Hoeffding bound, to compute the index, the sample variance is used as an estimate of the unknown variance. Thus, the UCB1 policy can be modified as shown in Algorithm 3 and the highest index is calculated using line 8.

Algorithm 3 UCB1-Normal Policy, Auer *et al.* [20]

```
1: // Initialization
2: $n_i, \bar{R}_i \leftarrow 0$
3: // Main Loop
4: **while** 1 **do**
5:     Select the antenna state which has not been selected at least 8 logn times
6:     Otherwise select the antenna state $i$ that maximizes
7:
8:     $\bar{R}_i + \sqrt{16((q_i - n_i\bar{R}_i{}^2)/(n_i - 1))(\ln(n-1)/n_i)}$
9:
10:    Update $n_i, \bar{R}_i$ for antenna state $i$
11: **end while**
```

where q_i is the sum of squared rewards for state i.

UCB1-Normal policy has an expected regret of at most

$$256(\log n)\left(\sum_{i:\mu_i<\mu^*}\frac{\sigma_i^2}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{2} + 8\log n\right)\left(\sum_{i=1}^{K}\Delta_i\right). \tag{13}$$

B. Reward Metrics

In this section, we discuss the link quality metrics that we use as instantaneous reward for the selection policies described above. The selection of reward metrics is dependent on the specific system implementation and based on the desired objective, the system designer can identify a relevant reward metric. In this paper, we evaluate two commonly used link quality metrics for MIMO systems.

1) Post-Processing SNR (PPSNR): We first use post-processing SNR as the reward metric to perform the antenna state selection. PPSNR can be defined as the inverse of the error vector magnitude (EVM) [32], [33]. As true SNR is not easily available for MIMO systems on a per-packet basis, PPSNR can be used to approximate the received SNR and can be used as the link quality metric. EVM for a MIMO spatial multiplexed system is defined as the squared symbol estimation error calculated after the MIMO decoding of the spatial streams. At every OFDM packet reception, we calculate the PPSNR for all the subcarriers and separately for all the spatial streams. Then, the instantaneous reward $R_i(n)$ for time slot (packet) n can be calculated as

$$R_i(n) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{F}\sum_{f=1}^{F}\mathrm{PPSNR}_f^{[s]} \tag{14}$$

where F is the number of subcarriers and S is the number of spatial streams.

2) Demmel Condition Number (DCN): For MIMO systems employing spatial multiplexing (SM) technique, the separate antenna streams on which data is modulated are required to be as uncorrelated as possible for achieving maximum capacity. The correlation among the spatial streams is influenced by the propagation effects such as multipath propagation and the amount of scattering. In MIMO systems equipped with reconfigurable antennas, the additional degree of freedom to select the radiation states can potentially reduce the correlation between the spatial streams. Previously, regular condition number or its reciprocal have been used to evaluate the quality of MIMO channel matrix. However, motivated by results in [34], we use the Demmel condition number as it can be related to a sufficient condition for multiplexing to be better than diversity. If the Demmel condition number is high, it represents high correlation between the streams. We calculate the Demmel condition number per subcarrier as

$$\kappa_f = \frac{\|H_f\|_F^2}{\lambda_k} \tag{15}$$

where $\|\cdot\|_F$ is the Frobenius norm of the channel and λ_k is the smallest eigenvalue of the MIMO channel matrix H_f. We then calculate instantaneous reward $R_i(n)$ as

$$R_i(n) = \frac{1}{\kappa} \tag{16}$$

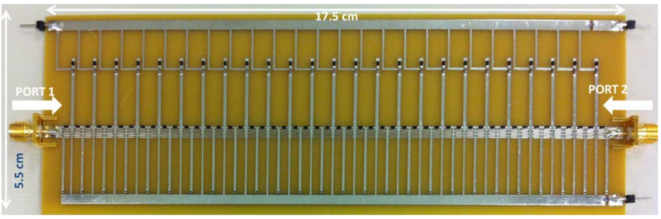


Fig. 4. Two port reconfigurable leaky wave antenna [7].

where

$$\kappa = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{F}\sum_{f=1}^{F}\kappa_f^{[s]}. \tag{17}$$

V. Experimental Setup

A. Pattern Reconfigurable Leaky Wave Antennas

For our experiments, we use the reconfigurable leaky wave antenna (RLWA) which is a two port antenna designed to electronically steer two highly directional independent beams over a wide angular range. Initially proposed by the authors in [7], the prototype shown in Fig. 4 is a composite right/left-handed leaky wave antenna composed of 25 cascaded metamaterial unit cells [35] loaded with varactor diodes. In order to achieve the CRLH behavior, a unit cell is implemented by inserting an artificial series capacitance and a shunt inductance into a conventional microstrip line using an interdigital capacitor and a shorted stub. Further, there are two varactor diodes (D_S) in parallel with the microstrip series interdigital capacitor and one varactor diode (D_{SH}) is placed in series with the shunt inductor. The application of various combinations of bias voltages "S" and "SH" to the independent bias networks, controls the beam direction allowing for symmetrical steering of the two radiation beams at the two ports over a 140° range. As the two ports are located on the same antenna structure, it is used as a two-element array in a MIMO setup. The radiation states were selected so that all the states have approximately similar measured gain with as low pattern correlation as possible. All the radiation states at the two ports are matched for a target return loss of 10 dB with the isolation between the two ports being higher than 10 dB.

Though the antenna in [35] is ideally capable of switching between an infinite number of radiation states, in order to characterize the effect of beam direction on the efficacy of a wireless system with RLWAs deployed at both ends of a link, a subset of 5 states was selected to allow the beam to steer over a range of 140° in the azimuthal plane. Fig. 5 shows the measured radiation patterns for the selected states and their corresponding bias voltages are shown in Table I.

B. Measurement Setup

In our experiments we make use of the wireless open access research platform (WARP), an FPGA-based software defined radio testbed and WARPLab, the software development environment used to control WARP nodes from MATLAB [36]. Four WARP nodes were distributed throughout the fifth floor of the Drexel University Bossone Research Center as shown in Fig. 7. This setup allowed us to capture performance for both line-of-sight (LOS) and non-line-of-sight (NLOS) links and also cover a wide SNR regime. As shown in Fig. 6, the PPSNR varies significantly from location to location and for each antenna state. When node 1 is active, it means that node 1 is transmitting and the other three nodes are receiving. Once all the channels are collected for node 1, the measurement controller makes node 2 as the transmitting node and other remaining nodes receive. This process is repeated until all four nodes have finished transmitting and the data is collected. Since we have 4 nodes and at a given time 1 node transmits and 3 nodes receive, we have a set of total of 12 links each corresponding to a unique transmitter-receiver combination. For certain links, there are more than one antenna states which are near optimal thus making the state selection more challenging. By using

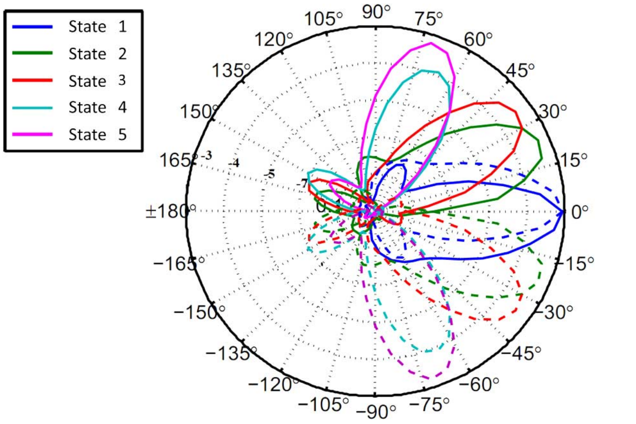


Fig. 5. Measured radiation patterns for port 1 & 2(Gain is shown in dB and is ≈ -3 dB).

TABLE I
Main Radiation Characteristics of Five Antenna States

Index	Bias Voltage(V)	Gain(dB)	Direction(expected)
1	S = 38 SH = 5	-3	0°
2	S = 5 SH = 30	-3	18°
3	S = 10 SH = 5	-3.2	36°
4	S = 2 SH = 10	-3.4	60°
5	S = 2 SH = 2	-3	72°

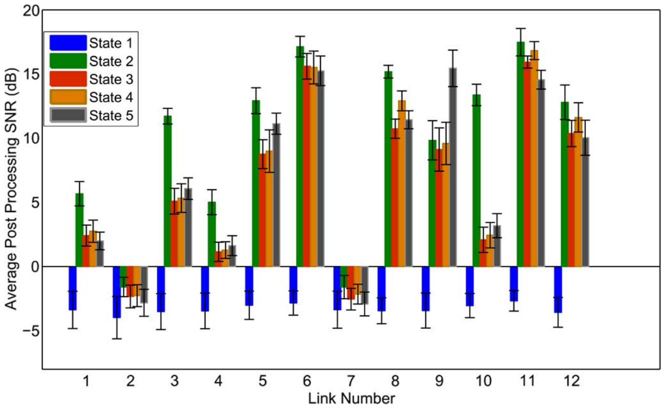


Fig. 6. Average received Post Processing SNR for 5 antenna states at the receiver measured for 12 links. Link 1–3 is when node 1 is active, Link 4–6 is when node 2 is active, Link 7–9 is when node 3 is active and Link 10–12 is when node 4 is active. Standard deviation is also shown for all links and antenna states.

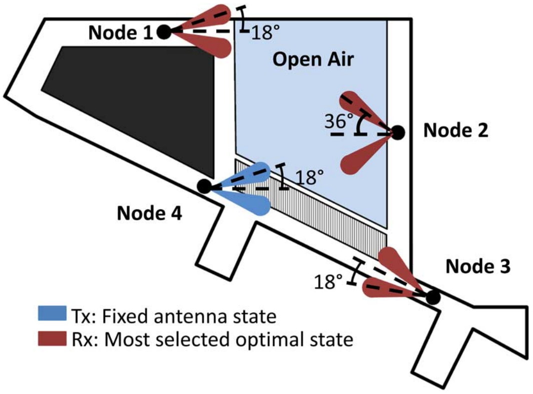


Fig. 7. Node positions on the 5th floor of the Drexel University Bossone Research Center.

WARPLab, each of the nodes were centrally controlled for the synchronization of the transmission and reception process and to provide control over the antenna states selected at each of the nodes. Although, the nodes were controlled centrally for data collection purposes, the learning algorithm was decentralized. Specifically, no information during the learning process was shared with the transmitter.

The performance of the RLWA was evaluated in a 2 × 2 MIMO system with spatial multiplexing as the transmission technique [31]. The implemented communication system followed an OFDM PHY as described in the IEEE 802.11n standard with total 64 subcarriers. 48 subscribers were used for loading data symbols, 4 for Carrier Frequency Offset correction and 12 left blank (i.e., $F = 48$, $S = 2$). For collecting channel realizations, we use a broadcast scheme where each designated WARP node transmitter broadcasts packets modulated using BPSK. For each packet transmission, the receiver nodes stored channel estimates and extracted the reward metrics as described above. Furthermore, the antenna states for each receiver node were switched after each packet until all 5 possible antenna states between the transmitter and receivers were tested. This process was repeated until 200 channel realizations were stored for all the state combinations and for each node acting as a transmitter. The beam directions in Fig. 7 corresponds to the optimal state selected most often at each of the receivers when node 4 was transmitting. The selection policies described in Section III-B are online learning policies but we note that we collected the channel realizations corresponding to each state and evaluated the algorithm in post-processing. This is essential in order to benchmark the performance of different selection policies under the same channel conditions and to make sure that channel conditions do not bias performance results.

VI. Performance Analysis and Results

We evaluate the performance of the proposed online selection policies using the measurement setup described above. We compare these polices with three policies 1) Genie Policy: In this policy it is assumed that the true mean rewards for all the antenna states are known *a priori* and a genie always selects the optimal antenna state. This closely represents the ideal case where instantaneous full CSI corresponding to all the states are available to select the optimal antenna state. 2) Exhaustive Search with Periodic Training (ESPT): In this selection scheme, all the antenna states are selected periodically in sequence and the corresponding channel estimates are then used to select the optimal state. The frequency and the amount of the training is fixed and are given by τ and ϕ. Since, we do not change the frame structure to enable periodic training for all antenna states, this represents an alternative where all antenna states are periodically selected both for channel training and sending data. This can be viewed as the process of consecutive exploration and exploitation, except that the duration of the exploration and exploitation is fixed and exploration occurs uniformly across all the states 3) Random Selection: In this selection scheme, at each time slot, one antenna state is randomly selected with uniform probability from the available K states.

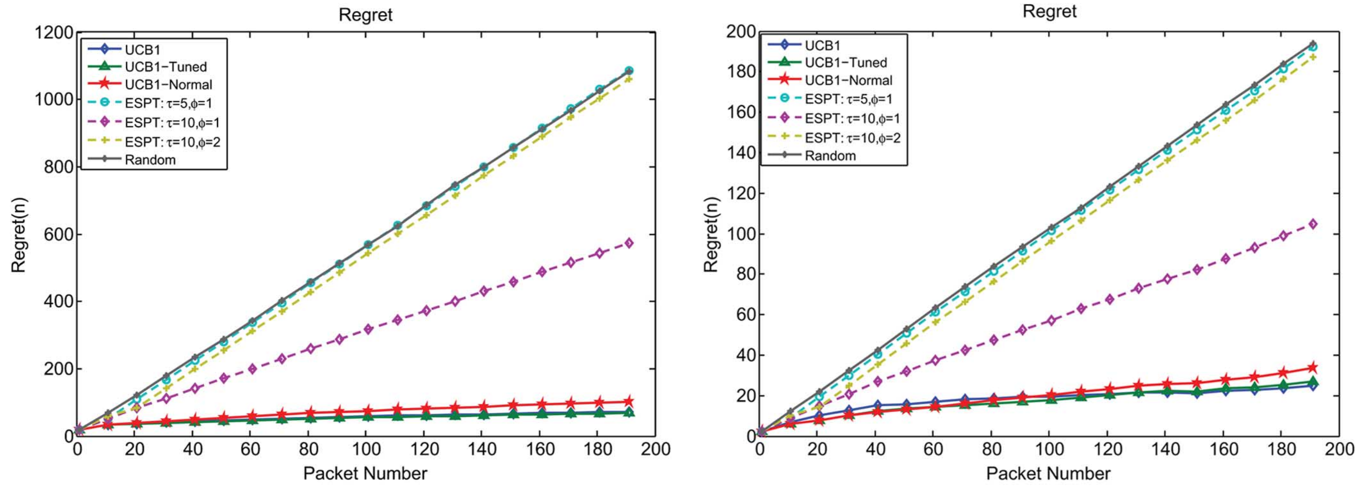
A. Regret Analysis

The goal of the learning policies is to maximize the long-term average reward, but in order to analyze the cost of learning, the regret of a system is an essential metric. Regret is a finer performance criteria as compared to long-term average reward and it indicates the growth of the loss incurred due to learning without complete state information. In Fig. 8(a) and (b), we show the regret with respect to the number of packets for all the selection policies corresponding to two different reward metrics respectively. The regret is averaged across all the transmitter and receiver locations shown in Fig. 7. It can be seen that the UCB1 policy and its variants have a sublinear regret as compared to the ESPT and Random selection. Further, it can be seen that varying the parameters of ESPT has direct impact on the regret of the system. ESPT with $\tau = 10$ and $\phi = 1$ shows minimum regret among other ESPT policies, where a single round of periodic training was performed every 10 OFDM symbols. If there are many suboptimal states and exhaustive training is performed with high frequency ($\tau = 5$, $\phi = 1$), suboptimal states will be selected much more often thus incurring higher regret. Also, if the frequency is kept constant and the amount of training is increased in order to get a better estimate of the reward ($\tau = 10$, $\phi = 2$), it will reduce the rate of regret but it will still be linear. This indicates the trade-off between the amount and the frequency of the channel training which makes tuning the parameters τ and ϕ very challenging for a given environment. On the other hand, UCB1 polices do not require parameter tuning and inherently adjusts the exploration rate based on the acquired estimate of the average reward and the associated confidence in the estimate.

Further in the Fig. 9(a) and (b), we show the percentage of time optimal state is selected by a policy up to time slot n. When using instantaneous PPSNR, UCB1-Tuned policy outperforms other policies, selecting the optimal state up to 90% of the time.

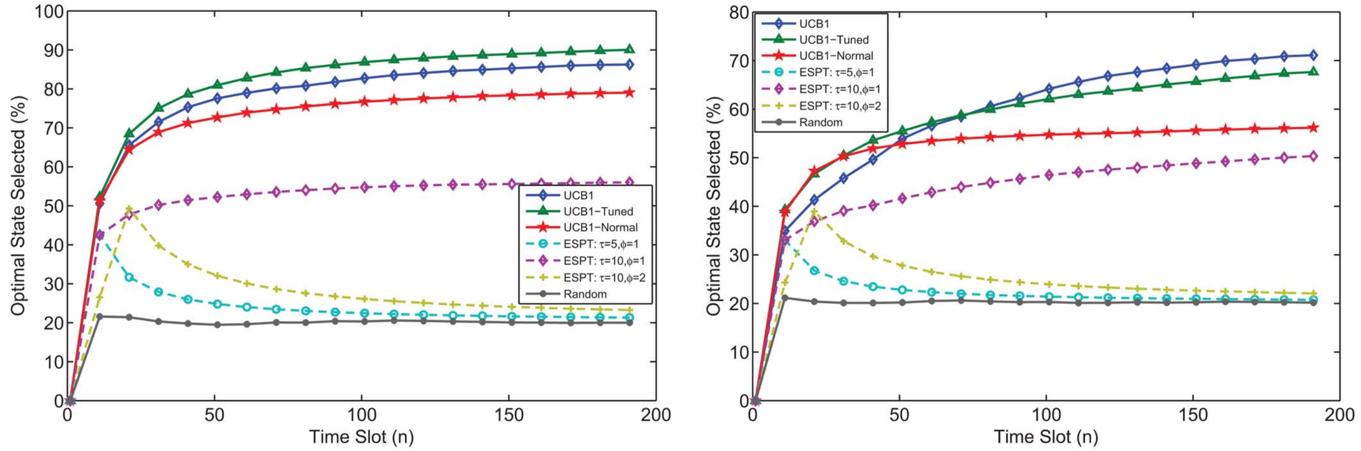
While UCB1 policy performs closer to UCB1-Tuned, the UCB1-Normal performs suboptimally due to the assumption that the reward is normally distributed. As shown in [32], PPSNR is chi-squared distributed, therefore the UCB1-Normal policy is not expected to perform as well as the rest of the UCB1 policies. Further, as shown in the case of regret, performance of the ESPT policy varies as τ and ϕ are varied, where in this case, best ESPT policy selected the optimal state only up to 50%.

In Fig. 10, we show the impact of increasing the number of antenna states on regret. As the number of antenna states



(a) (b)

Fig. 8. Regret $\bar{R}(n)$. (a) Regret vs time slot (n), $\mathrm{K} = 5$, Reward: PPSNR; (b) Regret vs time slot (n), $\mathrm{K} = 5$, Reward: DCN.



(a) (b)

Fig. 9. Percentage of the optimal state is selected. (a) $\mathrm{K} = 5$, Reward: PPSNR; (b) $\mathrm{K} = 5$, Reward: DCN.

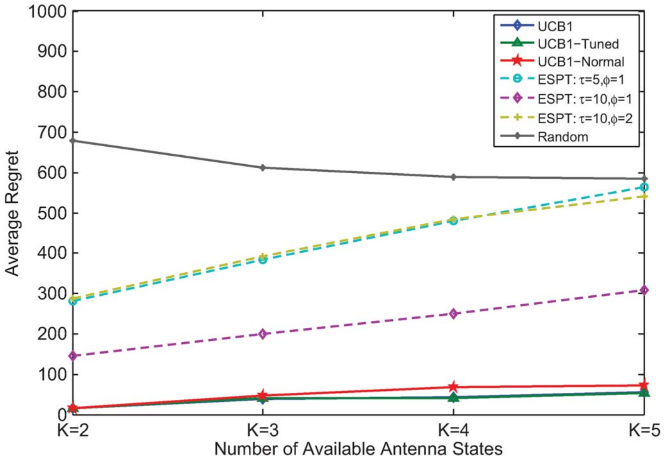


Fig. 10. Regret $\bar{R}(n)$ vs Number of Available Antenna States available at each receiver.

increase, the regret for ESPT polices increase with higher leading constant while the regret for UCB1 policies increase only slightly. More interesting is the fact that the regret for random policy decreases as the number of available states to select, increases. This decrease is due to the fact that the random policy explores uniformly and if there are more than one near optimal states, the regret will decrease. The regret for the random policy will eventually grow in a situation where there is only one constantly optimal state and the rest of the states are sub optimal as the probability of sampling sub optimal states will increase.

B. Average PPSNR

We first study the impact of increased pattern diversity on the average PPSNR. In Fig. 12, it can be seen that as the number of antenna states at the receiver is increased, a higher gain in average PPSNR is achieved. There is a 25% improvement achieved by learning policies in average PPSNR across all the links when the number of antenna states is increased from 2 to 5. On the other hand, the gain in average PPSNR achieved by the best ESPT policy is only 10% while the worst case ESPT policy shows negative improvement. This is due to the fact that, as the number of antenna states are increased, performing exhaustive search periodically negates diversity benefits due to increased training overhead. Further, the random policy shows higher gain since the random policy explores more frequently to find the optimal antenna state.

We further study the improvement in the average PPSNR by evaluating the gain in the average PPSNR for both the reward

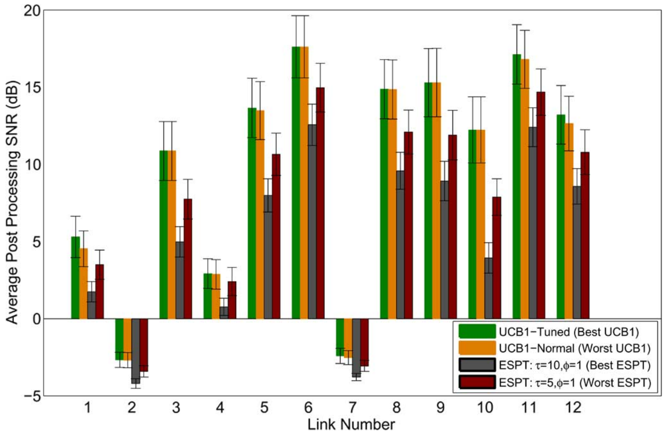


Fig. 11. Average received Post Processing SNR for best and worst UCB and ESPT policies for 12 links. Link 1–3 is when node 1 is active, Link 4–6 is when node 2 is active, Link 7–9 is when node 3 is active and Link 10–12 is when node 4 is active. Standard deviation is also shown for all links and antenna states.

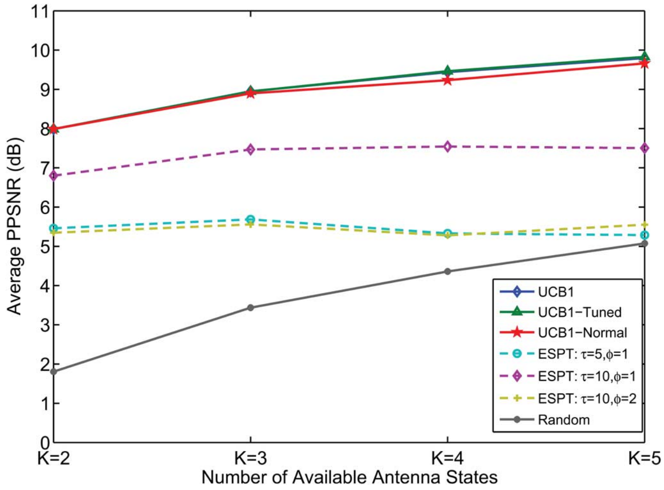


Fig. 12. Gain in average PPSNR (dB) vs number of available antenna states available at each receiver.

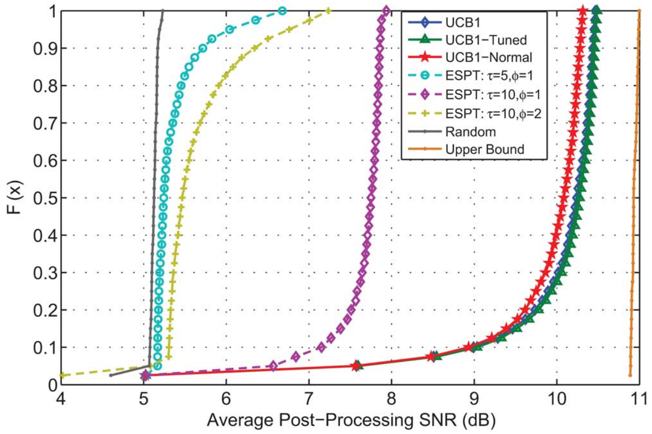


Fig. 13. Empirical CDF of Post-Processing SNR (PPSNR) (db) averaged across all links. Reward: Instantaneous PPSNR.

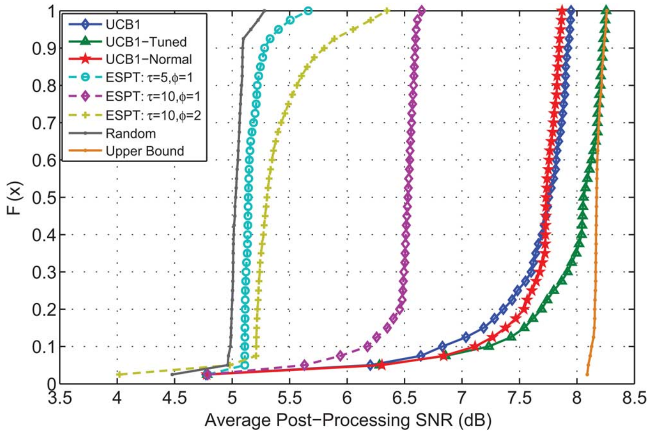


Fig. 14. Empirical CDF of Post-Processing SNR (PPSNR) (db) averaged across all links. Reward: Instantaneous DCN.

metrics; instantaneous PPSNR and Demmel Condition Number (DCN). In Fig. 11, we compare the performance of best case and worst UCB1 policies with respective ESPT policies. It can be seen that the percentage improvement in average PPSNR is significant between the worst case UCB1 and ESPT policies. Also, the two UCB1 polices have only marginal performance difference. In Figs. 13 and 14, we show the empirical CDF of average PPSNR for all the selection policies averaged across all the links in the network for the two reward metrics respectively. Additionally, we also show the average PPSNR achieved by the genie policy as a reference ideal case which defines the upper bound. It can be seen from Fig. 13 that all UCB1 policies have better performance than the rest of the policies.

The average PPSNR for all UCB1 policies lie within 1 dB of the best case scenario which is shown as upper bound. The relative performance of the ESPT policies follow a similar trend as seen above. The best performing UCB1 policy achieves 2.3 dB higher average PPSNR providing 31% improvement over the best ESPT ($\tau = 10$, $\phi = 1$) policy while the worst UCB1 policy achieves 4.4 dB higher average PPSNR providing 82% improvement over the worst case ESPT ($\tau = 5$, $\phi = 1$).

Further, in Fig. 14, when instantaneous DCN is selected as the reward metric, the overall gain in average PPSNR is less than the case when instantaneous PPSNR is selected as the reward metric. This difference in average PPSNR indicates that the optimal state which provides the most correlation between the streams may not always provide the best PPSNR, since received PPSNR also depends on the power in the channel and is influenced by propagation effect such as fading. In this case, the best performing UCB1 policy achieves 1.4 dB higher average PPSNR providing 22% improvement over the best ESPT ($\tau = 10$, $\phi = 1$) policy while the worst UCB1 policy achieves 2.4 dB higher average PPSNR providing 46% improvement over worst case ESPT ($\tau = 5$, $\phi = 1$).

C. Sum Throughput

We further assess the impact of improved PPSNR on the throughput of the network. To analyze throughput improvement, we perform a simple discrete rate look up table based adaptive modulation and coding (AMC) technique. The fixed look up table is defined by SNR ranges obtained by using the

TABLE II
AMC SCHEMES WITH DATA RATES (bps/Hz)

Index	Modulation(M)	Coding Rate (r)	Data rate (bps/Hz)
1	BPSK	1/2	0.5
2	4-QAM	1/2	1
3	4-QAM	3/4	1.5
4	16-QAM	1/2	2
5	16-QAM	3/4	3
6	64-QAM	2/3	4
7	64-QAM	3/4	4.5

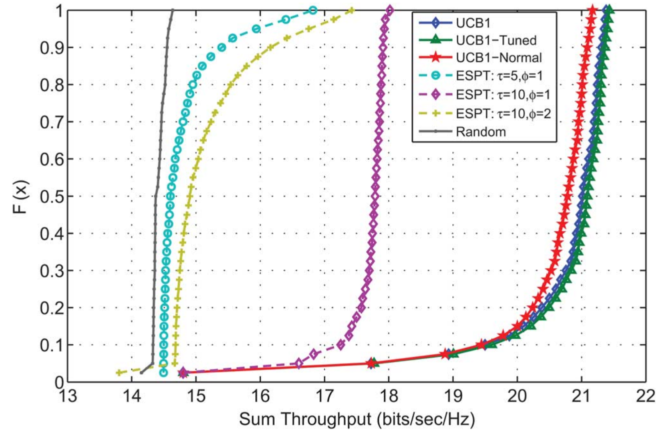


Fig. 15. Empirical CDF of Sum throughput (bits/sec/Hz) across all links. Reward: Instantaneous PPSNR.

upper bound expression for symbol error probability in AWGN channels [37]

$$P_{\sqrt{M}} \approx 2\left(1-\frac{1}{\sqrt{M}}\right)\mathcal{Q}\left(\sqrt{\frac{3m}{(M-1)}\frac{E_b}{N_0}\frac{1}{r}}\right) \le \mathcal{E} \quad (18)$$

where

$$\mathcal{E} = 2\exp\left(\frac{3m}{(M-1)}\frac{E_b}{N_0}\frac{1}{r}\right) \quad (19)$$

where M is the constellation order, r is the coding rate, E_b/N_0 is the SNR per bit, and $m = log_2(M)$. The selected AMC scheme is used for all the subcarriers. Based on the measured received PPSNR of each antenna state, we calculate the throughput each link would achieve when using the AMC schemes shown in Table II.

In Figs. 15 and 16, we show the empirical CDF of calculated sum throughput using the AMC scheme described above for two reward metrics respectively. The PPSNR improvement realized by using the learning policies allow each receiver to select an appropriate AMC scheme to maximize link throughput, thereby improving the sum throughput of the network. As shown in Fig. 15, the best performing UCB1 policies achieve 17% improvement over best case ESPT policy, while the worst case UCB1 policy achieves 38% improvement over worst case ESPT policy.

For the scenario where DCN is used as reward metric UCB1 policy achieves 8% and 13% improvement over best case and worst case respectively.

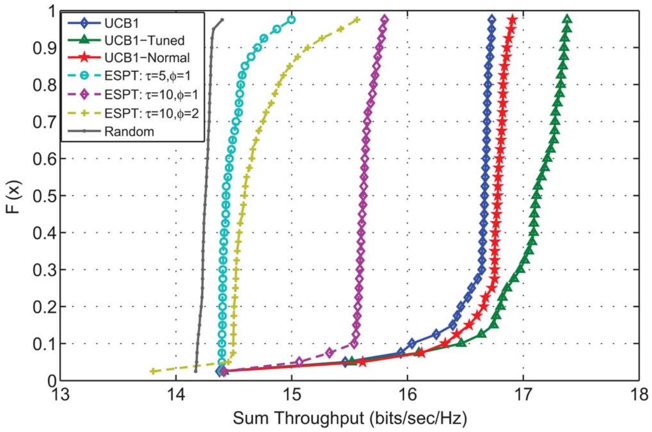


Fig. 16. Empirical CDF of Sum throughput (bits/sec/Hz) across all links. Reward: Instantaneous DCN.

VII. CONCLUSION

We have proposed a novel online learning based antenna state selection technique and have shown that wireless systems employing reconfigurable antennas can benefit from such technique. The proposed selection technique allows the optimal state selection without requiring instantaneous CSI for all the antenna states and does not require modification to OFDM frame for periodic training. This leads to reduced overhead of channel training. The performance of the proposed selection technique is empirically evaluated in a practical wireless system covering wide SNR range and both LOS and NLOS links. We show the impact of available antenna states on system regret and long time average reward. A relative average PPSNR gain and corresponding throughput gain is achieved and the performance is compared to the ideal selection technique utilizing instantaneous full CSI. Future work will involve devising new learning policies which utilize multiple reward metrics at the same time for sequential decision making. In addition, investigating the application of antenna state selection in conjunction with optimal channel selection in a cognitive radio network or multi-user MIMO network are possible future directions.

ACKNOWLEDGMENT

The authors wish to acknowledge K. Wanuga, D. Gonzalez, and R. Bahl for their valuable suggestions and feedback.

REFERENCES

[1] N. Gulati, D. Gonzalez, and K. Dandekar, “Learning algorithm for reconfigurable antenna state selection,” in *Proc. IEEE RWS*, 2012, pp. 31–34.

[2] J. Boerman and J. Bernhard, “Performance study of pattern reconfigurable antennas in MIMO communication systems,” *IEEE Trans. Antennas Propag.*, vol. 56, no. 1, pp. 231–236, 2008.

[3] D. Piazza, N. Kirsch, A. Forenza, R. Heath, and K. Dandekar, “Design and evaluation of a reconfigurable antenna array for MIMO systems,” *IEEE Trans. Antennas Propag.*, vol. 56, no. 3, pp. 869–881, 2008.

[4] A. Sayeed and V. Raghavan, “Maximizing MIMO capacity in sparse multipath with reconfigurable antenna arrays,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 156–166, 2007.

[5] Y. Jung, “Dual-band reconfigurable antenna for base-station applications,” *Electron Lett.*, vol. 46, no. 3, pp. 195–196, 4, 2010.

[6] Adant [Online]. Available: http://www.adant.com/

[7] D. Piazza, M. D’Amico, and K. Dandekar, “Two port reconfigurable CRLH leaky wave antenna with improved impedance matching and beam tuning,” in *Proc. IEEE EuCAP*, 2009, pp. 2046–2049.

[8] C. Sukumar, H. Eslami, A. Eltawil, and B. Cetiner, "Link performance improvement using reconfigurable multiantenna systems," *IEEE Antennas Wireless Propag. Lett.*, vol. 8, pp. 873–876, 2009.

[9] J. Kountouriotis, D. Piazza, K. Dandekar, M. D'Amico, and C. Guardiani, "Performance analysis of a reconfigurable antenna system for MIMO communications," in *Proc. EUCAP*, 2011, pp. 543–547.

[10] R. Bahl, N. Gulati, K. Dandekar, and D. Jaggard, "Impact of reconfigurable antennas on interference alignment over measured channels," in *Proc. IEEE GLOBECOM Workshops*, 2012, pp. 557–562.

[11] T. Gou, C. Wang, and S. Jafar, "Aiming perfectly in the dark-blind interference alignment through staggered antenna switching," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2734–2744, Jun. 2011.

[12] L. Ke and Z. Wang, "Degrees of freedom regions of two-user MIMO Z and full interference channels: The benefit of reconfigurable antennas," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3766–3779, Jun. 2012.

[13] G. Huff, J. Feng, S. Zhang, and J. Bernhard, "A novel radiation pattern and frequency reconfigurable single turn square spiral microstrip antenna," *IEEE Microw. Wireless Compon. Lett.*, vol. 13, no. 2, pp. 57–59, 2003.

[14] M. Fries, M. Grani, and R. Vahldieck, "A reconfigurable slot antenna with switchable polarization," *IEEE Microw. Wireless Compon. Lett.*, vol. 13, no. 11, pp. 490–492, 2003.

[15] J. Bernhard, "Reconfigurable antennas," *Synthesis Lectures on Antennas*, vol. 2, no. 1, pp. 1–66, 2007.

[16] D. Piazza, P. Mookiah, M. D'Amico, and K. Dandekar, "Two port reconfigurable circular patch antenna for MIMO systems," in *Proc. IET EuCAP*, 2007, pp. 1–7.

[17] D. Anagnostou, G. Zheng, M. Chryssomallis, J. Lyke, G. Ponchak, J. Papapolymerou, and C. Christodoulou, "Design, fabrication, measurements of an RF-MEMS-based self-similar reconfigurable antenna," *IEEE Trans. Antennas Propag.*, vol. 54, no. 2, pp. 422–432, 2006.

[18] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[19] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: I.I.D rewards," *IEEE Trans Automatic Contr.*, vol. 32, no. 11, pp. 968–976, 1987.

[20] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[21] J. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Stat. Soc. Ser. B (Methodological)*, pp. 148–177, 1979.

[22] A. Grau, H. Jafarkhani, and F. De Flaviis, "A reconfigurable multiple-input multiple-output communication system," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1719–1733, 2008.

[23] D. Piazza, J. Kountouriotis, M. D'Amico, and K. Dandekar, "A technique for antenna configuration selection for reconfigurable circular patch arrays," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1456–1467, 2009.

[24] H. Eslami, C. Sukumar, D. Rodrigo, S. Mopidevi, A. Eltawil, L. Jofre, and B. Cetiner, "Reduced overhead training for multi reconfigurable antennas with beam-tilting capability," *IEEE Trans. Wireless Commun.*, vol. 9, no. 12, pp. 3810–3821, 2010.

[25] L. Lai, H. El Gamal, H. Jiang, and H. Poor, "Cognitive medium access: Exploration, exploitation, competition," *IEEE Trans. Mobile Comput.*, vol. 10, no. 2, pp. 239–253, 2011.

[26] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, 2008.

[27] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. IEEE Symp. New Frontiers in Dynamic Spectrum*, 2010, pp. 1–9.

[28] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *Proc. IEEE GLOBECOM*, 2011, pp. 1–6.

[29] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, 2010.

[30] H. Volos and R. Buehrer, "Cognitive engine design for link adaptation: An application to multi-antenna systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2902–2913, 2010.

[31] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-blast: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. IEEE URSI*, pp. 295–300.

[32] R. Shafik, S. Rahman, R. Islam, and N. Ashraf, "On the error vector magnitude as a performance metric and comparative analysis," in *Proc. IEEE ICEC*, 2006, pp. 27–31.

[33] H. Arslan and H. Mahmoud, "Error vector magnitude to SNR conversion for nondata-aided receivers," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2694–2704, 2009.

[34] R. Heath and A. Paulraj, "Switching between diversity and multiplexing in MIMO systems," *IEEE Trans. Commun.*, vol. 53, no. 6, pp. 962–968, 2005.

[35] D. Piazza, M. D'Amico, and K. Dandekar, "Performance improvement of a wideband MIMO system by using two-port RLWA," *IEEE Antennas Wireless Propag. Lett.*, vol. 8, pp. 830–834, 2009.

[36] Rice University WARP Project [Online]. Available: http://warp.rice.edu

[37] A. Goldsmith*, Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

Nikhil Gulati received the B.Eng. degree in electronics instrumentation and control from the University of Rajasthan, India, in 2005 and the M.S. degree specializing in control theory and robotics from Drexel University, Philadelphia, PA, USA, in 2007, where he conducted research on sensor networks, autonomous systems and real-time control and where he is pursuing his Ph.D. degree in electrical and computer engineering.

He worked as a Software Engineer from 2007–2010 with Drexel University. His research is focused on developing adaptive algorithms for cognitive radios employing electrical reconfigurable antennas. He also works on applying benefits of electrically reconfigurable antennas to Interference Alignment and Physical Layer Security.

Kapil R. Dandekar received the B.S. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 1997 with specializations in communications and signal processing, applied electrophysics, and computer engineering and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin, TX, USA, in 1998 and 2001, respectively.

In 1992, he worked at the U.S. Naval Observatory and from 1993–1997, he worked at the U.S. Naval Research Laboratory. In 2001, Dandekar joined the Electrical and Computer Engineering Department at Drexel University in Philadelphia, PA, USA. He is currently a Professor and the Director of the Drexel Wireless Systems Laboratory (DWSL). DWSL has been supported by the U.S. National Science Foundation, Army CERDEC, National Security Agency, Office of Naval Research, and private industry. His current research interests involve MIMO ad hoc networks, reconfigurable antennas, free space optical communications, ultrasonic communications, and sensor networks. He has published articles in several journals including IEEE Transactions on Antennas and Propagation, IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, and IEE Electron Lett..