LEARNING ALGORITHMS FOR ENERGY-EFFICIENT MIMO ANTENNA SUBSET SELECTION: MULTI-ARMED BANDIT FRAMEWORK

Amitav Mukherjee

Nokia Research Center
Berkeley, CA 94704, USA
amitav.mukherjee@nokia.com

Ari Hottinen

Nokia Research Center
P.O.Box 407, FI-00045 NOKIA GROUP, Finland
ari.hottinen@nokia.com

ABSTRACT

The use of multiple antennas in mobile devices provides enhanced data rates at the cost of increased power consumption. The stochastic nature of the wireless propagation medium and random variations in the utilization and operating environment of the device makes it difficult to estimate and predict wireless channels and power consumption levels. Therefore, we investigate a robust antenna subset selection policy where the power-normalized throughput is assumed to be drawn from an unknown distribution with unknown mean. At each time instant, the transceiver decides upon the active antenna subset based on observations of the outcomes of previous choices, with the objective being to identify the optimal antenna subset which maximizes the power-normalized throughput. In this work, we present a sequential learning scheme to achieve this based on the theory of multi-armed bandits. Simulations verify that the proposed novel method that accounts for dependent arms outperforms a naïve approach designed for independent arms in terms of regret.

***Index Terms*—** Antenna selection, energy efficiency, learning, multi-armed bandit.

1. INTRODUCTION

An increasingly standard approach towards achieving high data-rate wireless communications is to deploy multiple antennas at the mobile terminal, as reflected in current and forthcoming cellular radio and WLAN standards. The application of multiple-input multiple-output (MIMO) transmission techniques such as spatial multiplexing and beamforming enable significantly greater spectral efficiencies or reliability on both the downlink and uplink [1]. However, the use of multiple antennas at the mobile terminal comes with a cost of increased RF circuit power consumption, since each antenna RF chain is associated with a multiplicity of RF analog components such as power amplifiers, filters, mixers, and ADC/DACs. Since the mobile device is generally a battery-powered device, the indiscriminate use of multiple antennas (equivalently, RF chains) will degrade the device life time and lead to an earlier exhaustion of the battery lifetime. Thus, energy efficiency of MIMO systems has been studied intensively in the literature [2]- [4].

A potentially more energy-efficient approach would be to dynamically activate a subset of the available antenna chains so as to balance the data rate with the device power consumption. Roughly speaking, this is achieved via receive antenna subset selection on the downlink, and transmit antenna subset selection at the transceiver on the uplink [5]. However, the stochastic nature of the wireless propagation medium and random variations in the utilization and operating environment of the device makes it difficult to estimate and predict wireless channels and power consumption levels. Therefore, we investigate a robust antenna subset selection policy where the power-normalized throughput is assumed to be drawn from an unknown distribution with unknown mean. At each time instant, the transceiver decides upon the active antenna subset based on observations of the outcomes of previous choices, with the objective being to identify the optimal antenna subset which maximizes the power-normalized throughput. In this work, we present a non-parametric sequential learning scheme to achieve this based on the theory of *multi-armed bandits* (MABs).

To our best knowledge, such tools have not been applied previously to resource optimization problems in MIMO systems with incomplete side information. A MAB approach to antenna reconfiguration without analysis was given in [6]. A different line of work in the domain of cognitive radio opportunistic spectrum access that has received a lot of attention recently considers dynamic decisions by a single secondary user when the underlying primary user behavior on each channel is a two-state Markov chain. This can be formulated as a partially observable Markov decision process (POMDP), and when the channels are independent, as a special class of POMDP known as restless bandits [7]- [9]. Furthermore, while [7]- [9] feature MABs with independent arms, we will see that our case specializes to the less commonly studied MAB with dependencies across arms.

The remainder of this work is organized as follows. Section 2 introduces the multi-antenna transceiver system model. The theory of multi-armed bandit problems and their application to antenna subset selection along with performance bounds is presented in Section 3. A numerical example comparing the proposed and existing schemes is given in Section 4, and we conclude in Section 5.

2. SYSTEM MODEL

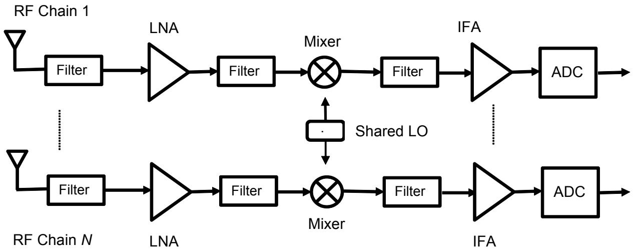

Fig. 1. Multiple receive RF chains (N total) in a multi-antenna transceiver.

The multiple-antenna system model is described in this section. Consider a transceiver with N antennas, each with an associated RF chain. At a given time interval, the transceiver activates a subset of M antennas, $M \leq N$, either on the uplink or downlink in order to optimize a chosen performance metric. Let $\mathcal{S}$ represent the set of all possible antenna subsets, with cardinality

$$|\mathcal{S}| = \sum_{i=1}^{N} \left(\begin{array}{c} N \\ i \end{array} \right).$$

Considering for example *receive antenna subset selection* as in Fig. 1, the baseband signal received by the transceiver from a T-antenna source can be written as

$$\mathbf{y} = \mathbf{H}_s \mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{H}_s \in \mathbb{C}^{M \times T}$ is the channel matrix corresponding to an arbitrary antenna subset $s \in \mathcal{S}$ of cardinality K_s, $\mathbf{x} \in \mathbb{C}^{T \times 1}$ is the information signal, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Z})$ is additive colored Gaussian noise with covariance matrix $\mathbf{Z}$. Let $E\left\{\mathbf{x}\mathbf{x}^H\right\} = \mathbf{Q}$ represent the transmit covariance matrix of the source.

The i^{th} row $\mathbf{h}_{i,s}$ of the matrix $\mathbf{H}_s$ represents the vector $\mathbf{h}_{i,s}$ of channel coefficients between the T transmit antennas and receive antenna i:

$$\mathbf{H}_s = \left[\begin{array}{cccc} \mathbf{h}_{1,s}^T & \mathbf{h}_{2,s}^T & \ldots & \mathbf{h}_{K_s,s}^T \end{array} \right]^T \tag{2}$$

We can define the throughput or goodput as

$$G_s = (1 - \epsilon_s) R_t \tag{3}$$

where R_t is the packet transmission rate of the source and ϵ_s is the packet error rate associated with the chosen antenna subset. The error probability ϵ_s is a function of the signal-to-noise ratio at the transceiver. For reliable detection, the maximum packet transmission rate R_t must be bounded by the information-theoretic capacity defined as [5]

$$C = \log_2 \left| \mathbf{I} + \mathbf{H}_s \mathbf{Q} \mathbf{H}_s^H \mathbf{Z}^{-1} \right|$$

The corresponding transceiver power consumption is

$$P_s = P_b + \sum_{k \in s} P_k \tag{4}$$

where P_b is the shared baseband processing power consumption and P_k is the power consumed by RF chain k in the set s, both of which may also be stochastic. Therefore, a performance metric that captures both data rate and energy efficiency is the throughput normalized by the power consumption:

$$T_s = \frac{G_s}{P_s}. \tag{5}$$

For conventional throughput or rate maximization, it is optimal to always use all antennas. However, this is not necessarily true when the device power consumption is included in the performance criterion [2]- [4].

The stochastic nature of the wireless propagation medium and presence of co-channel interference (reflected in the assumption of colored noise) makes it difficult to perfectly estimate $\mathbf{H}_s$ at the receiver. Furthermore, random variations in the utilization and operating environment of the transceiver hinders the prediction of the exact power consumption. Therefore, we construct a robust antenna subset selection policy where the 'reward' T_s is drawn from an unknown distribution with unknown mean. At each time instant, the transceiver decides upon the active antenna subset based on past observations of the reward obtained from previous choices, with the objective being to identify the optimal antenna subset that maximizes the average power-normalized throughput. In the next section, we present a non-Bayesian sequential learning scheme to achieve this based on the theory of multi-armed bandits.

3. MULTI-ARMED BANDIT FRAMEWORK

We now consider an abstraction of the antenna selection problem as follows. We first introduce some relevant background on the theory of multi-armed bandits, and then specialize to the specific problem considered in this work.

In the classic multi-armed bandit problem with K choices or 'arms', a player must decide which one of the K arms to play at each step in a sequence of trials so as to maximize the long-term reward [7]. Every time he plays an arm, he receives a reward. The structure of the reward for each arm is unknown to the player *a priori*, but in most prior work the reward has been assumed to be independently drawn from a fixed (but unknown) distribution. The reward distribution in general differs from one arm to another, therefore the player must use all his past actions and observations to essentially learn the quality of these arms (in terms of their expected reward) so he can keep playing the best arm.

The multi-armed bandit embodies the classic trade-off between exploration and exploitation. This is because the player needs to sufficiently explore all arms so as to minimize the likelihood of settling upon an inferior one erroneously believed to be optimal. On the other hand, the player needs to avoid spending too much time exploring the arms and maximize the time spent playing the optimal arm. In view of the above, the performance of a decision policy is typically measured by the notion of *regret*, which is defined as the loss in expected reward compared to that yielded by an 'genie' or ideal policy with *a priori* knowledge of the reward distributions of each arm.

We now formalize the application of the MAB technique to the energy-efficient antenna subset selection problem. Consider a MAB where each of the $|\mathcal{S}|$ possible antenna subsets is mapped to a virtual arm. The reward $X_k(n)$ for arm k at time n is drawn from an arbitrary unknown distribution with unknown mean θ_k, which necessitates a non-Bayesian learning approach. The rewards are allowed to be dependent across arms, but are independent over time. Define $\theta^* = \max(\theta_1, \ldots, \theta_{|\mathcal{S}|})$, and

$$\Delta_k = \theta^* - \theta_k.$$

The regret obtained from any selection policy π is defined as

$$R^{\pi}\left(n\right) = n\theta^* - \sum_{j=1}^{K} \theta_j \mathcal{E}\left\{T_j\left(n\right)\right\} \tag{6}$$

where $\mathcal{E}(\cdot)$ denotes expectation and $T_j\left(n\right)$ is the cumulative number of times arm j is played up to time n.

It is clear that the presence of the same antenna(s) in multiple subsets introduces dependencies across the rewards obtained from the associated arms, unlike the independent arms assumed in [7]- [10]. For example, the rewards obtained from antenna subsets $s_1 = \{1, 3, 4, 5\}$ and $s_2 = \{1, 2, 4, 5\}$ are highly correlated. In the limited prior work on MABs with dependent arms, Mersereau *et al.* [11] consider a setting where the expected reward is defined as a linear function of an random variable, and the prior distribution is known, which does not apply in our case. Gai *et al.* [12, 13] consider a combinatorial cognitive radio channel allocation problem with an arm defined as a specific user-to-channel map. The same user-to-channel allocation present in different maps introduces dependencies across arms in their model, but the key difference is that they assume the reward from each individual user-to-channel allocation is known, whereas we assume only the cumulative reward for the *entire arm* (antenna subset) is observed.

3.1. Naïve Approach

In the first approach to the problem, it is possible to naïvely reuse the existing solution in [10] where the arms are considered to be independent. Define the counters $\hat{Y}_k, d_k$, that track the sample mean of the rewards obtained from each arm, and the number of times a particular arm is played, respectively, which are updated at time n as

$$\hat{Y}_k\left(n\right) = \begin{cases} \frac{\hat{Y}_k(n-1)+X_s(n)}{d_k(n-1)+1} & \text{if arm } k \text{ is played} \\ \hat{\theta}_k\left(n-1\right) & \text{else} \end{cases} \tag{7}$$

$$d_k\left(n\right) = \begin{cases} d_k\left(n-1\right)+1 & \text{if arm } k \text{ is played} \\ d_k\left(n-1\right) & \text{else} \end{cases} \tag{8}$$

At time n, the naïve approach plays the arm that maximizes $\hat{Y}_k + \sqrt{\frac{2\ln n}{d_k}}$. This policy has a regret upper-bounded as [10]

$$R^{\pi}\left(n\right) \leq \sum_{k:\theta_j \leq \theta^*} \frac{8\ln n}{\Delta_j} + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{k:\theta_j \leq \theta^*} \Delta_j\right). \tag{9}$$

The naïve approach ignores dependencies across arms and their associated rewards, which is likely to be suboptimal. Therefore, this motivates a more sophisticated approach to the antenna subset selection problem that accounts for correlated arms in an effort to more quickly identify the optimal subset.

3.2. Proposed Approach

In the proposed approach, instead of tracking the rewards for each arm, we record an approximation for the proportional reward gained by each constituent antenna within an arm. Define the counters $\hat{\theta}_i, d_i$ that track the sample mean of the rewards obtained from each antenna, and the number of times a particular antenna is played, respectively. These metrics are then updated after each play at time n as follows:

$$\hat{\theta}_i\left(n\right) = \begin{cases} \frac{\hat{\theta}_i(n-1)+X_s(n)/|s|}{d_i(n-1)+1} & \text{if } i \in s \\ \hat{\theta}_i\left(n-1\right) & \text{else} \end{cases} \tag{10}$$

$$d_i\left(n\right) = \begin{cases} d_i\left(n-1\right)+1 & \text{if } i \in s \\ d_i\left(n-1\right) & \text{else} \end{cases} \tag{11}$$

The estimated sample mean of the reward due to each individual antenna is computed by normalizing the cumulative reward observed for the arm by the number of antennas comprising the arm. The idea is that antennas that do not contribute towards higher rewards are de-weighted over time and played with decreasing frequency.

The proposed antenna subset selection policy is described in Algorithm 3.2.1, where at each decision epoch, an exhaustive search is carried out to determine the optimal constituent antennas. Since the number of antennas is generally less than 8 in practical systems, the exhaustive search does not entail high computational complexity.

Algorithm 3.2.1 Antenna Subset Selection Policy

```
Require: $\hat{\theta}_i = 0, d_i = 0, i = 1, \ldots, N$.
  // INITIALIZATION
  Play each arm once.
  Update $\hat{\theta}_i, d_i$.
  // MAIN LOOP
  while 1 do
    $n = n + 1$;
    Play arm that solves
        $\arg\max_{s\in\mathcal{S}} \sum_{i\in s} \hat{\theta}_i + \sqrt{\frac{(N+1)\ln n}{\sum_{i\in s} d_i}}$
    Update $\hat{\theta}_i, d_i$.
  end while
```

Following the arguments in [10]- [13], the regret accrued from the proposed policy can be upper-bounded as

$$R^{\pi}(n) \leq \sum_{k:\theta_j \leq \theta^*} \frac{4N^2 \ln n}{\Delta_j} + N^2\left(1 + \frac{\pi^2}{3}\right)\left(\sum_{k:\theta_j \leq \theta^*} \Delta_j\right). \tag{12}$$

4. SIMULATION RESULTS

From the discussion in Sec. 2, numerical simulations require the distribution of the normalized throughput T_s in (5) for some choice of channel fading environment. For example, under the commonly-used Rayleigh fading assumption, the MIMO channel matrix $\mathbf{H}_s$ in (2) is composed of zero-mean unit variance complex Gaussian variables, and the corresponding transmission rate and throughput are functions of the random matrix $\mathbf{H}_s$. However, there does not appear to be a known distribution in the literature for the circuit power consumption P_s, which stymies efforts to characterize the statistics of T_s.

Therefore, to verify the advantage of the proposed MAB scheme, we consider a toy example with $N = 4$ antennas and the possible subsets are mapped to 15 arms. The rewards of each arm follow a Bernoulli process that is i.i.d. over time, with respective means

$$\begin{aligned} \boldsymbol{\theta} &= \{0.48, 0.44, 0.64, 0.70, 0.75, 0.27, 0.67, 0.65, \ldots \\ &\quad 0.16, 0.11, 0.49, 0.95, 0.34, 0.58, 0.22\}, \end{aligned} \tag{13}$$

where we recall that the reward distribution and means are unknown *a priori* to the transceiver. Here, θ_1 corresponds to the arm comprising the single antenna $\{1\}$, and so forth up to θ_{15} which corresponds to antenna subset $\{1, 2, 3, 4\}$. Therefore, in this scenario the 12^{th} antenna subset $s = \{1, 3, 4\}$ with mean reward $\theta_{12} = 0.95$ is set as the optimal choice in terms of reward.

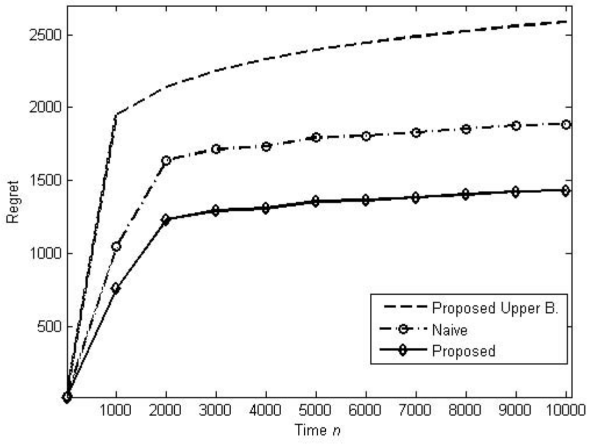

Fig. 2. Regret versus time for N=4.

In Fig. 2, we compare the upper bound and actual regret obtained from the naïve and proposed MAB solutions for 10000 plays averaged over 100 runs. The theoretical upper bound is observed to be somewhat loose. Nevertheless, the proposed MAB method is seen to clearly outperform the naïve approach designed for independent arms in terms of regret. An equivalent interpretation of this figure is that under the proposed method, the optimal antenna subset is located earlier and played more frequently.

5. CONCLUSION

The use of multiple antennas in mobile devices enables enhanced data rates at the cost of increased power consumption. The stochastic nature of the wireless propagation medium and random variations in the utilization and operating environment of the device makes it difficult to estimate and predict wireless channels and power consumption levels. Therefore, we investigate a robust antenna subset selection policy where the power-normalized throughput is assumed to be drawn from an unknown distribution with unknown mean. At each time instant, the transceiver decides upon the active antenna subset based on past observations of the reward obtained from previous choices, with the objective being the maximization of the average power-normalized throughput. In this work, we present a non-parametric sequential learning scheme to achieve this based on the theory of multi-armed bandits. Simulations verify that the proposed novel method that accounts for dependent arms outperforms a naïve approach designed for independent arms in terms of regret. For future work, a more precise statistical characterization of the power-normalized throughput and comparison with a parametric learning scheme would be of interest.

6. REFERENCES

[1] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*, Prentice-Hall, June 2010.

[2] S. G. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE J. Sel. Areas. Commun.*, vol. 22, no. 6, pp. 1089-1098, Aug. 2004.

[3] H. Kim, C.-B. Chae, G. Veciana, and R. W. Heath, "A cross-layer approach to energy efficiency for adaptive MIMO systems exploiting spare capacity," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4264-4273, Aug. 2009.

[4] H. Yu, L. Zhong, and A. Sabharwal, "Power management of MIMO network interfaces on mobile systems," to appear, *IEEE Trans. VLSI Syst.*. Available on *Early Access*.

[5] A. F. Molisch, M. Z. Win, Y.-S. Choi, and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1759-1772, July 2005.

[6] N. Gulati, D. Gonzalez, and K. R. Dandekar, "Learning algorithm for reconfigurable antenna state selection," in *Proc. Radio and Wireless Symposium (RWS)*, 2012.

[7] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2053-2071, May 2008.

[8] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multichannel opportunistic access: structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431-5440, Dec. 2008.

[9] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040-4050, Sep. 2009.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, 47, pp. 235-256, 2002.

[11] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, "A structured multiarmed bandit problem and the greedy policy," *IEEE Trans. Auto. Control*, vol. 54, no. 12, pp. 2787-2802, 2009.

[12] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," *Proc. IEEE DySPAN*, 2010.

[13] Y. Gai, B. Krishnamachari, and M. Liu, "On the combinatorial multi-armed bandit problem with Markovian rewards," *Proc. IEEE Global Communications Conference (GLOBECOM)*, 2011.