Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Pulkkinen, Petteri; Aittomäki, Tuomas; Koivunen, Visa

Reinforcement learning based transmitter-receiver selection for distributed MIMO radars

Published in:
2020 IEEE International Radar Conference, RADAR 2020

DOI:
10.1109/RADAR42522.2020.9114644

Published: 01/04/2020

Document Version
Peer reviewed version

Please cite the original version:
Pulkkinen, P., Aittomäki, T., & Koivunen, V. (2020). Reinforcement learning based transmitter-receiver selection for distributed MIMO radars. In *2020 IEEE International Radar Conference, RADAR 2020* (pp. 1040-1045). [09114644] IEEE. https://doi.org/10.1109/RADAR42522.2020.9114644

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Reinforcement Learning Based Transmitter-Receiver Selection for Distributed MIMO Radars

Petteri Pulkkinen, Tuomas Aittomäki, and Visa Koivunen
Department of Signal Processing and Acoustics,
Aalto University
PO Box 15400, FI-00076 Aalto, Finland

***Abstract*—Active transmitter-receiver (TX-RX) subset selection facilitates efficient resource use and adaptation to varying target and propagation environments in distributed multiple-input multiple-output (MIMO) radar systems. The problem has been addressed in the literature and objective functions related to radar tasks that depend on the signal-to-interference-plus-noise ratio (SINR) have been proposed. The SINR values observed at the receivers can be estimated assuming a particular propagation environment and target models. In this paper, a novel machine learning approach is proposed in which no such assumptions are needed. We formulate the TX-RX subset selection as a multi-armed bandit (MAB) problem and further extend it to the combinatorial MAB framework. A variety of reinforcement learning algorithms developed for the MAB problem are employed to learn the optimal subset in real-time. It is shown that such algorithms can be effectively used for the TX-RX subset selection problem even in non-stationary scenarios.**

***Index Terms*—reinforcement learning, multi-armed bandits, subset selection, distributed MIMO radar**

I. Introduction

A distributed multiple-input and multiple-output (MIMO) radar is a system in which multiple transmitters and receivers are deployed in different locations [1]. The widely distributed antennas increase spatial diversity which can be utilized for improved target localization, parameter estimation and detecting low-observable targets. In a MIMO radar, the used waveforms are designed jointly and the waveforms are typically orthogonal. Therefore, each transmitter-receiver (TX-RX) pair forms an independent channel for illuminating and observing the targets. Increasing the number of channels provides additional spatial diversity at the cost of a larger amount of data to be processed and higher power consumption [2]–[4]. In addition, active transmitters may expose their location which might be undesired in certain applications. Strategies for TX-RX subset selection are studied to preserve spatial diversity and reduce the costs at the same time.

The TX-RX subset selection problem, especially for target localization, has been previously addressed in the literature [2], [3]. In these studies, the signal-to-interference-plus-noise ratio (SINR) is assumed to be known or estimated. It is possible to estimate the SINR based on assumed particular propagation environments and target models, but a real-world radar environment may deviate from the assumed model and consequently lead to performance degradation. Another way to form the estimates is to probe the different channels for a sufficient number of times to learn their state. The probing needs to be efficient since the channels are continuously evolving and the probing will impact the radar performance because it requires performing extra tasks in addition to the main operation of the radar. Therefore, it is necessary to decide when to exploit the TX-RX subset currently providing the best payoff or explore other subsets that may or may not provide even higher SINR levels. This is the exploration-exploitation trade-off.

A multi-armed bandit (MAB) framework provides policies for selecting actions to balance the exploration and exploitation trade-off while maximizing the employed reward function [5]. In the literature, such a framework is effectively used for example for opportunistic spectrum access, in which the secondary user must choose a frequency band in a way that does not cause interference to the primary user [6]. Moreover, authors in [7] and [8] use a combinatorial MAB approach as a robust way for selecting the MIMO antenna subset for maximizing throughput in communication systems. Extensions from combinatorial multi-armed bandits are used to address the problem with large combinatorial action space.

This paper proposes a machine learning approach for the active TX-RX subset selection in the distributed MIMO radar context. It is based on the same principle as in the papers [7] and [8]. However, the approach is generalized for any index-based policies and non-stationary environments. The generalized approach is simulated in a radar environment and the performance of several different MAB algorithms is compared.

II. Problem formulation

Assume a distributed MIMO radar system that consists of N receivers and M transmitters. The radar system is constrained to use a subset of the transmitters and the receivers at the same time in order to save resources such as power and reduce the probability of being detected by an adversary. Each TX-RX pair is considered as a channel and overall there are $K = NM$ channels. The number of channels in a subset is $K_S = N_S M_S$, where both, the number of receivers N_S and the number of transmitters M_S in a subset are constrained with an equality constraint. Selecting receiver $n \in \{1, 2, ..., N\}$ and transmitter $m \in \{1, 2, ..., M\}$ is indicated with vectors $\boldsymbol{\delta}_{\text{rx}}$ and $\boldsymbol{\delta}_{\text{tx}}$ where

$$\delta_{\text{rx}_n} = \begin{cases} 1, \text{ when receiver } n \text{ is included in the set} \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

$$\delta_{\text{tx}_m} = \begin{cases} 1, \text{ when transmitter } m \text{ is included in the set} \\ 0, \text{ otherwise.} \end{cases} \tag{2}$$

Radar receivers are subject to both unintentional and intentional interference when observing target returns. The quality of the received signal is typically characterized by signal-to-interference-plus-noise ratio (SINR).

In this paper, the measured signal power ρ_{nm} at receiver n from transmitter m via the target is modeled as a stochastic process that may be non-stationary. Non-stationary character of ρ_{nm} can stem from the target moving in the environment, and the movement causes variability to path losses and the radar cross-section (RCS). Moreover, unintentional and intentional interference levels might change in time further contributing to the non-stationary behavior of the system. It is assumed that sufficiently accurate estimates of the noise power σ^2_{th} and the interference power $\sigma^2_{\text{int}_n}$ at receiver n are available. Therefore, we can express the channel SINR on a linear scale as

$$\Gamma_{nm} = \frac{\mathbb{E}\left[\rho_{nm}\right] - (\sigma^2_{\text{th}} + \sigma^2_{\text{int}_n})}{\sigma^2_{\text{th}} + \sigma^2_{\text{int}_n}}, \tag{3}$$

where the noise power and the interference power is subtracted from the received signal power and $\mathbb{E}\left[\rho_{nm}\right] \geq \sigma^2_{\text{th}} + \sigma^2_{\text{int}_n}$. Note that the SINR is time-dependent because random variable ρ_{nm} is non-stationary.

Performance in target detection and target localization depend on the SINR level [2]–[4]. Therefore, a reward function for reinforcement learning is proposed that depends on vector $\mathbf{\Gamma}$ that consists of all the SINR values Γ_{nm}. The following reward function is used for the employed MAB learning

$$r(\boldsymbol{\delta}_{\text{rx}}, \boldsymbol{\delta}_{\text{tx}} | \mathbf{\Gamma}) = \frac{1}{N_S M_S} \sum_{n=1}^{N} \sum_{m=1}^{M} \delta_{\text{rx}_n} \delta_{\text{tx}_m} \Gamma_{nm}, \tag{4}$$

which is a mean of SINR values for the selected channels on a linear scale. Furthermore, the reward maximization problem can be formulated as

$$\begin{aligned} &\max\nolimits_{\boldsymbol{\delta}_{\text{rx}}, \boldsymbol{\delta}_{\text{tx}}} r(\boldsymbol{\delta}_{\text{rx}}, \boldsymbol{\delta}_{\text{tx}} | \mathbf{\Gamma}) \\ \text{s.t.} \quad &\begin{cases} \sum_{n=1}^{N} \delta_{\text{rx}_n} = N_S \\ \sum_{m=1}^{M} \delta_{\text{tx}_m} = M_S. \end{cases} \end{aligned} \tag{5}$$

It is possible to find an optimal solution for the equation (5) if $\mathbf{\Gamma}$ is known. However, $\mathbf{\Gamma}$ is not known since we can only obtain the measurements ρ_{nm}. Furthermore, it might not be possible to probe all the channels at the same time. Therefore, $\mathbf{\Gamma}$ needs to be estimated from the received signal. By probing the different channels, an estimate $\widehat{\mathbf{\Gamma}}$ of the channel SINR values can be formed. An approximate solution for the objective (5) is found when the estimate $\widehat{\mathbf{\Gamma}}$ is used to condition the reward function (4). Different subsets must be explored sufficiently large number of times to reinforce the estimate $\widehat{\mathbf{\Gamma}}$ to identify the subset of TX-RX pairs yielding the highest SINR values.

III. Multi-armed bandits

The MAB problem is a sequential decision-making problem in which an agent needs to decide on an action from several competing actions [9]. The word arm originates from a slot machine that has a pull lever, and the pull lever is called an arm. Pulling an arm of a slot machine changes the state according to the Markov chain model and gives a reward based on a certain probability distribution. In the MAB problem, the agent needs to sequentially decide from many arms which arm to pull to maximize the cumulative reward. The strategy of how the agent chooses the arms is called a policy. The specific formulation considered in this paper is a stochastic multi-armed bandit, in which each arm is associated with one state.

Usually, the performance of a policy is measured with regret. The regret quantifies the cost of learning by measuring how much reward agent has missed from the cumulative reward. In this paper, the performance for policy π is measured with normalized regret

$$R_\pi(T) = \sum_{t=1}^{T} \left(1 - \frac{\mu_t^\pi}{\mu_t^*}\right), \tag{6}$$

where T is the time horizon, μ_t^π is the expected reward for the policy π at time instant t and μ_t^* is the expected reward for the optimal arm at time instant t. Moreover, the normalized regret describes how large proportion of the optimal reward is missed at each iteration, which is useful when the rewards are non-stationary. Sublinear regret as a function of T indicates that the agent has made choices in previous time instances that improve the future choices.

To maximize the cumulative reward, the agent needs to find the arm which gives the highest expected reward. This means that the agent needs to decide when to explore different arms to possibly identify those with higher expected rewards and when to keep pulling the arm with currently known highest expected reward. An estimate for the expected reward is updated each time when the agent has pulled an arm. The update equation for each arm can be written as

$$q_{t+1}(a) = q_t(a) + \alpha \left(r_t + q_t(a)\right), \tag{7}$$

where r_t is the received reward from selecting the arm a, α is a step size and $q_t(a)$ is called an action-value [9]. For stationary rewards, α can be set to t^{-1}, so that (7) calculates the empirical mean. For non-stationary rewards, the parameter α needs to be constant so that old rewards have a lower weight than more recent ones [9].

In practice, a policy is realized as an algorithm. We briefly review five different MAB algorithms which are used in the simulations. The selected algorithms are widely used to solve MAB problems. These algorithms are

1) ϵ-greedy [9],
2) Upper Confidence Bound (UCB1) [9], [10],
3) Kullback Leibler Upper Confidence Bound (KL-UCB) [11],
4) Thompson sampling [12], [13], and

5) Recency-Based Exploration (RBE) [14], [15].

An analysis for stationary and non-stationary reward distributions can be found in the references. Also, other algorithms and more extensive discussion on the stochastic multi-armed bandits can be found in [5].

In the ϵ-greedy algorithm the probability of choosing a random arm is $0 < \epsilon \ll 1$ and this probability remains constant or decreases slowly in time. When not selecting an arm randomly, the algorithm chooses the arm which has the highest action-value. The UCB1 and KL-UCB are policies which are based on deriving an upper confidence bound for the expected rewards and the arm with highest bound is selected. Thompson sampling is a Bayesian algorithm where the posterior probability distributions for the expected rewards are formed from the collected rewards. Then the posterior distributions are sampled at each time instant and the arm with the highest sample is selected. Finally, the RBE is based on defining an exploration term to support choosing arms which have not been explored recently. In addition, the term is constructed in a way that the agent prefers arms which have the highest action-values.

Most of the MAB algorithms such as UCB1, KL-UCB, Thompson sampling and RBE are index-based polices that calculate a quantity, called an index, for each arm. The index captures the uncertainty on the action-value $q_t(a)$ and emphasizes exploration for those arms that might have desirable expected rewards. Typically, the index is constructed in a way that the arm with highest index is selected.

Another way to solve a MAB problem is to divide the exploration and the exploitation into two distinct phases. This division can be fully deterministic or random. Deterministic algorithms divide the exploration and the exploitation phases into blocks of specific lengths [5]. Random algorithms, such as ϵ-greedy, explore different arms with some probability and otherwise they exploit the arm with currently highest action-value.

IV. Multi-armed bandit problem formulation

Classical multi-armed bandits have a finite set of arms and a single arm is selected at each time instant. Therefore, the TX-RX selection problem in distributed MIMO radars could be formulated as follows. An arm could be a subset of TX-RX pairs and the reward is calculated using the reward function (4). However, number of the subsets is $\binom{N}{N_S}\binom{M}{M_S}$ which grows exponentially. This could make the problem unsolvable with the MAB approach because there is no time to explore all the arms in a given time horizon. The amount of time which is available for learning the different channels depends on the dynamic nature of targets and the propagation scenario. For example, if a target appears or disappears or does abrupt maneuvers, then rapid changes in the reward distributions will happen. On the other hand, smooth changes take place when a target moves on a smooth trajectory and changes its orientation gradually.

The problem with exponentially increasing number of arms can be avoided by reformulating the MAB model so that each arm represents SINR for each channel and the agent can choose multiple arms at each time instant. The arms are chosen to maximize the reward function and satisfy the constraints. When this formulation is used, the action-values introduced in section III are the channel SINR estimates $\widehat{\mathbf{\Gamma}}$. The total number of arms is reduced to NM and the agent can choose $N_S M_S$ arms at each iteration. The formulation is known as the combinatorial multi-armed bandits, in which the subset of arms is called a super arm [16]. The reformulation is possible because the reward function (5) is an increasing function of the radar channel SINR values and each arm is independent since waveforms from different transmitters do not interfere with each other if orthogonal waveforms are used.

Usually in the combinatorial MAB problem, the agent does not know the mapping from the arm rewards to the super arm rewards. However, here the reward function (4) is known which enables us to use MAB algorithms from the classical MAB problem. Authors in [7] and [8] use a similar approach for MIMO antenna selection in mobile communications to maximize throughput. The approach in [7] uses the UCB-1 algorithm and the reward statistics are constant. While, authors in [8] use Thompson sampling and both, non-stationary and stationary rewards are considered.

In this paper we generalize the approaches in [7] and [8] to any index-based algorithm and non-stationary reward distributions. The proposed algorithm for solving the combinatorial MAB problem with known mapping from the arm rewards to the super arm rewards is shown in Algorithm 1. On line 2 of the Algorithm 1 any index-based MAB algorithm can be used to find the indexes, and on line 3 any optimization method can be used to find the super arm. The indexes calculated by a MAB algorithm will ensure that the exploration and exploitation trade-off is balanced well.

The principles of the Algorithm 1 can be also used for the random policies by using the SINR estimates $\widehat{\mathbf{\Gamma}}$ instead of the indexes. In addition, random policies explore different subsets by selecting them randomly. However, the exploration is not as efficient as with index-based policies, because the fact that the super arm rewards are a function of the arm rewards is not utilized.

Algorithm 1: Proposed generalized algorithm

```
1  while not end of the time horizon do
2      calculate indexes for all arms;
3      find the super arm using the indexes;
4      pull the super arm;
5      if non-stationary rewards then
6          discount rewards for all arms;
7          discount exploration parameters for all arms;
8      end
9      observe the arm rewards;
10     update the arm rewards;
11     update the exploration parameters;
12 end
```

V. Simulation Setup

A. System Configuration

The simulated MIMO radar system consists of $N = 6$ receivers and $M = 4$ transmitters. A subset with $N_S = 3$ receivers and $M_S = 2$ transmitters is selected, so that six out of 24 possible channels are used at any time instance. It is possible to use exhaustive search to find the super arm because there are only 120 different subsets. The simulation environment is visualized in Fig.1.

B. Scattering Model

The target illumination angle and the scattering angle are usually different for each TX-RX pair in distributed MIMO radars. Therefore, the RCS model depends on the angles to the receiver and the transmitter. The dependency on the angles for receiver n and transmitter m is denoted by ψ_{nm} which is a product between two scattering coefficients that are taken from the monostatic target RCS model at the illumination angle and the scattering angle. The simplistic monostatic RCS model used in simulations is expressed by a simple sum of cosine functions $0.064 \cdot |2.5\cos(\theta) + 7\cos(2\theta) + 3\cos(3\theta) + 3\cos(4\theta)|$, where θ is the backscatter angle.

The target RCS fluctuation is modeled based on the Swerling I model. Hence, the power loss of the target fluctuation c is modeled by the exponential distribution with the scale parameter equal to one, and c remains constant between two subsequent time instances.

C. Propagation Environment Model

The environment model includes path losses, thermal noise, and external interference. The path loss $L_{nm} = d_{\text{tx}}^{-2} d_{\text{rx}}^{-2}$ is a product between reciprocal of targets' squared distance to the transmitter and the receiver. The thermal noise power $\sigma_{\text{th}}^2 = 0.001$ is constant in time and the same for all the channels. Also, the interference power $\boldsymbol{\sigma}_{\text{int}}^2 = [0.1, 0.9, 0.3, 0.1, 0.2, 0.4]^T$ is constant in time and $\sigma_{\text{int}_n}^2$ is the interference power at receiver n.

D. Rewards

The agent aims to find the super arm which has the highest reward based on the equation (4). To create the arm rewards, we define an instantaneous SINR measure γ_{nm} that satisfies $\mathbb{E}[\gamma_{nm}] = \Gamma_{nm}$ where Γ_{nm} is the channel SINR. The value γ_{nm} is calculated using the equation (3) where the expectation is replaced with the power measurement ρ_{nm}. To simplify the simulations, the target scattering coefficient c remains constant through a single measurement period and the stochasticity of the noise and the interference powers in the measurements are approximated to be negligible. Therefore, the rewards for each arm are simulated by

$$\gamma_{nm} \sim \text{Exp}\left(\Gamma_{nm}\right), \tag{8}$$

which is the exponential distribution with mean of Γ_{nm}. The channel SINR Γ_{nm} on a linear scale for receiver n and transmitter m is calculated from the models defined in Sections V-A, V-B, and V-C as follows

$$\Gamma_{nm} = \frac{L_{nm}\psi_{nm}p_m}{\sigma_{\text{th}}^2 + \sigma_{\text{int}_n}^2}, \tag{9}$$

where the transmit power $p_m = 1$ is constant in time and equal for each transmitter.

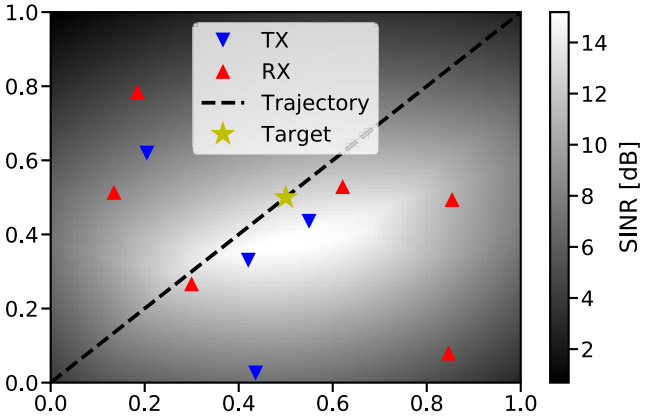
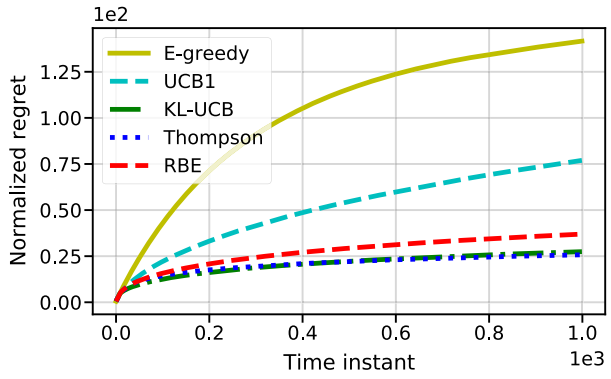


Fig. 1: Simulation setup. The target position is shown for the stationary case and the trajectory for the non-stationary case. The heat map indicates the mean channel SINR at every position when the target scattering coefficient is excluded and all the channels are active.

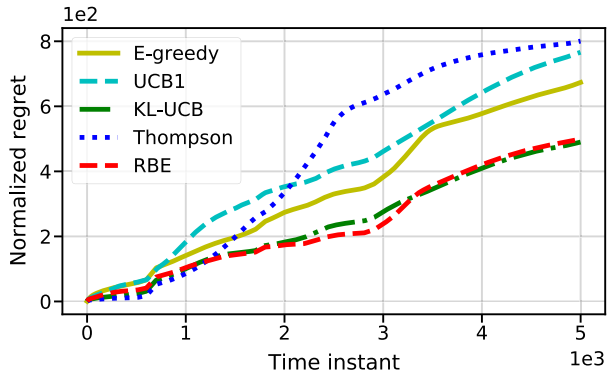
VI. Simulation Results

Two different scenarios are studied in the simulation examples. In the stationary scenario, there is a target that remains stationary at position $(0.5, 0.5)$. In the non-stationary scenario, a target moves from $(0, 0)$ to $(1, 1)$ on a linear trajectory with a constant velocity. All other parameters for the environment remain the same in both scenarios. The five algorithms that were briefly reviewed in section III are used in the simulations. In addition, a simple method that selects the closest subset at each time instant is used to compare the MAB algorithms to a more conventional exploration-free method. The performance of the different MAB algorithms were evaluated using Monte Carlo simulations with 1000 iterations.

The MAB algorithms are compared between each other in terms of regret. The regret through the simulation period is shown in Fig 2. Also, box plots of the regrets are shown in Fig 3. The comparison between the exploration-free method and MAB algorithms is performed by comparing expectation of the achieved rewards through the simulation period. The achieved rewards for two well-functioning MAB algorithms, the worst MAB algorithm and the exploration-free method are shown in Fig. 4. The overall results show that the MAB algorithms improve the performance through the time horizon and outperform the exploration-free method in stationary and non-stationary cases. Moreover, the RBE algorithm stands out with excellent reliability and regret performance in both cases.

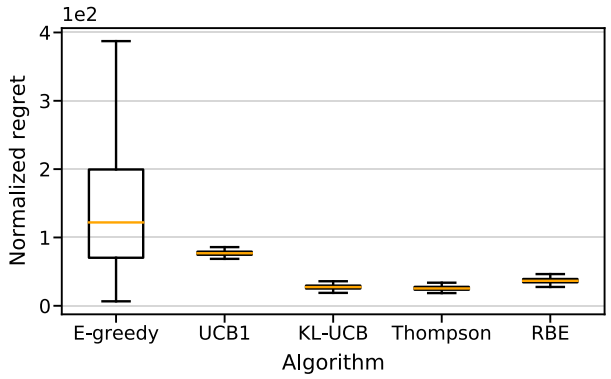


(a) Stationary case.

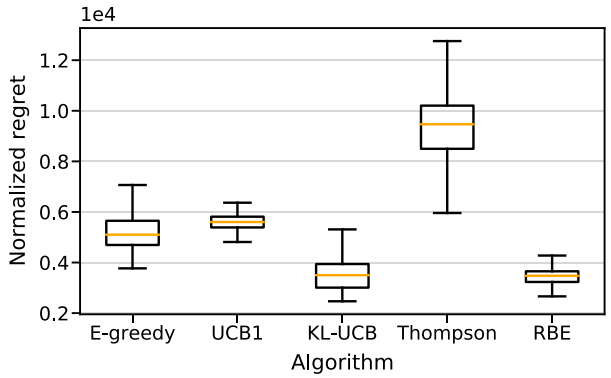


(b) Non-stationary case.

Fig. 2: The normalized regret at each time instant. RBE and KL-UCB perfom well in the both cases.



(a) Stationary case.



(b) Non-stationary case.

Fig. 3: The normalized regret for the whole time horizon compared between different simulation runs. RBE and KL-UCB obtain excellent results in both stationary and non-stationary scenarios.

A. *Stationary target*

In the stationary case, the discount factor can be set to $\alpha = t^{-1}$. The value of ϵ for ϵ-greedy exploration is set to 0.1, which means that it explores random subsets 10% of the time. Lower ϵ can sometimes result in lower regret, but the variance of the regret may increase. The value 0.1 was found empirically in this simulation. The other algorithms do not have any tunable parameters. The time horizon is set to 1000 which is sufficiently long for finding the optimal TX-RX configuration.

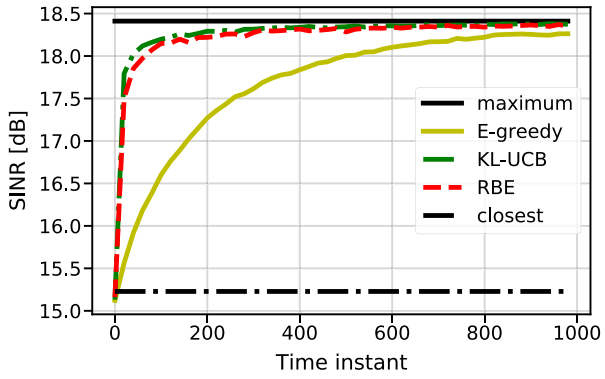
The stationary target implies that the expected rewards of the arms do not change in time. Therefore, it is possible to achieve a logarithmic regret [5]. From Fig.2a it can be observed that all algorithms other than ϵ-greedy can achieve the sublinear regret. Algorithms like Thompson sampling and KL-UCB which require knowledge about the reward distribution have excellent performance in these simulations. The ϵ-greedy algorithm has the highest regret and it is visible from Fig.3a that such random policies have a high variance. Algorithms with higher variance make the learning less reliable even though they can some times find the optimal action faster than more reliable algorithms.

Fig.4*a* visualizes the achievable reward at each time instant. The main difference between the MAB algorithms is the time taken to find the optimal super arm. The gap between the maximum achievable reward and achieved expected reward will decrease as a function of time for algorithms that achieve sublinear regret. Therefore, ϵ-greedy with constant exploration probability will eventually have a constant gap between maximum reward and achievable expected reward. It can be observed that the MAB algorithms perform much better than the exploration-free method.

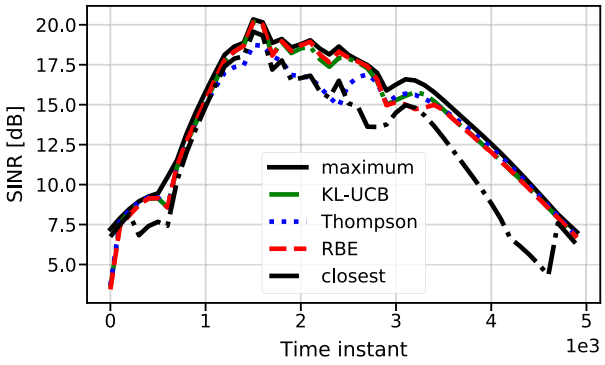
B. *Non-stationary target*

The algorithms adapt to the non-stationary conditions by using the discounted action-value and discounted exploration parameters. The discount factor is set to $\alpha = 0.998$. Also, each element of $\widehat{\mathbf{\Gamma}}$ is divided by the maximum value before calculating the index. This ensures that enough exploration is done at each time instant even if the scale of the rewards changes over time. The value of ϵ for ϵ-greedy algorithm is kept at 0.1. The asymptotic optimality can not be achieved since the exploration term will never vanish completely. However, the considered algorithms have differences in their exploration efficiencies, as can be observed from the different regrets in Fig.2b.

In Fig. 2b it can be seen that KL-UCB and RBE achieve the lowest regret in the non-stationary case. Thompson sampling



(a) Stationary case.



(b) Non-stationary case.

Fig. 4: Reward at each time instant. The reward is a mean of the channel SINRs in a linear scale for the selected subset. The MAB algorithms are compared to a simple exploration-free method in which the subset of the receivers and the transmitters closest to the target is selected. In average, most of the MAB algorithms achieve a reward close to the maximum reward.

does not adapt as well for the non-stationary case even if it performed quite well in the stationary case. The ϵ-greedy algorithm has lower regret than Thompson sampling but it still performs worse than the other algorithms. The Fig.3b demonstrates that RBE algorithm has a very small regret with low variance. Also, the median performance is better than with any other of the used algorithms. Hence it is promising algorithm for the radar problem at hand. The other algorithms which performed well in stationary reward scenario have poorer performance in non-stationary reward case. The changes in reward distributions will force the agent to switch the arm if the action-value for the arm under exploitation becomes lower than the other action-values. Therefore, in case of non-stationary target, the ϵ-greedy algorithm achieves a smaller variance on regret than with stationary targets.

The achieved rewards are compared in Fig.4b. All MAB algorithms perform on average better than the exploration-free method through the whole simulation period. Also, it is visible that the MAB algorithms can reach a SINR value close to the maximum reward at most of the time instances. Moreover, the performance gap between MAB algorithms is not as significant as in the stationary scenario.

VII. Conclusions

The transmitter-receiver subset selection problem, in which the channel SINR values are unknown, was considered for the distributed MIMO radars. Since only a subset of the channels can be selected at the same time, it was shown that such problems have to deal with the exploration and exploitation trade-off to identify the optimal subset without degrading the radar performance. The problem was formulated as the combinatorial multi-arm bandit problem and reinforcement learning algorithm was proposed to solve the problem. It was shown that reinforcement learning can be effectively used to continuously improve the subset selections and outperform the proposed exploration-free method.

References

[1] A. M. Haimovich, R. S. Blum, and L. J. Cimini, "MIMO radar with widely separated antennas," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 116–129, 2008.

[2] H. Godrich, A. Petropulu, and H. V. Poor, "Antenna subset selection in distributed multiple-radar architectures: A knapsack problem formulation," in *19th European Signal Processing Conference*, Aug. 2011, pp. 1693–1697.

[3] B. Sun, H. Chen, D. Yang, and X. Li, "Antenna selection and placement analysis of MIMO radar networks for target localization," *International Journal of Distributed Sensor Networks*, vol. 10, no. 5, May 2014.

[4] T. Aittomäki, H. Godrich, H. V. Poor, and V. Koivunen, "Resource allocation for target detection in distributed MIMO radars," in *Conference Record of the 45th Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov 2011, pp. 873–877.

[5] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2019, Draft of 27th June, Revision: 8b22.

[6] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, December 2008.

[7] A. Mukherjee and A. Hottinen, "Learning algorithms for energy-efficient MIMO antenna subset selection: Multi-armed bandit framework," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 659–663.

[8] Z. Kuai, T. Wang, and S. Wang, "Transmit antenna selection in massive MIMO systems: An online learning framework," in *IEEE/CIC International Conference on Communications in China (ICCC)*, Aug 2019, pp. 496–501.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[10] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," *arXiv e-prints*, p. arXiv:0805.3415, May 2008.

[11] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," *arXiv e-prints*, p. arXiv:1102.2490, Feb 2011.

[12] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*, 2012, pp. 39.1–39.26.

[13] V. Raj and S. Kalyani, "Taming non-stationary bandits: A Bayesian approach," in *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[14] J. Oksanen and V. Koivunen, "An order optimal policy for exploiting idle spectrum in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1214–1227, March 2015.

[15] ——, "Learning spectrum opportunities in non-stationary radio environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2447–2451.

[16] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *arXiv e-prints*, p. arXiv:1407.8339, Jul 2014.