Robust Reinforcement Learning-based Wald-type Detector for Massive MIMO Radar

Aya Mostafa Ahmed, Stefano Fortunati, Aydin Sezgin, Maria S Greco, Fulvio Gini

Robust Reinforcement Learning-based Wald-type Detector for Massive MIMO Radar

Aya Mostafa Ahmed[1], Stefano Fortunati[2], Aydin Sezgin[1], Maria S. Greco[3], and Fulvio Gini[3]

[1]Institute of Digital Communication Systems, Ruhr University Bochum, Germany
[2]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes & DR2I-IPSA, France
[3]Dipartimento di Ingegneria dell'Informazione, Università di Pisa, Italy
[1]Email: aya.mostafaibrahimahmad@rub.de

***Abstract*—The two basic performance indices characterizing the multi-target detection task in a radar system are the probability of false alarm (P_{FA}) and the probability of detection P_D. It is well-known that, when the disturbance model (i.e., clutter and noise) is perfectly known, the Neyman-Pearson (NP) detector provides the *best* decision strategy, i.e., the detector that maximizes the P_D, while keeping a constant P_{FA}. However, in practical scenarios, the a priori knowledge of the statistical model of the disturbance is rarely available. In this paper we investigate the robustness of a reinforcement learning (RL) based Wald-type test to guarantee reliable detection performance even without knowledge of the disturbance distribution. Specifically, the constant false alarm Rate (CFAR) property is obtained by applying tools from misspecified asymptotic statistics, while the P_D is maximized by exploiting an RL-based scheme.**

***Index Terms*—Cognitive Radar, Reinforcement Learning, Massive MIMO, robust statistics, Wald test.**

I. Introduction

The main idea underlying cognitive radars (CR) is that a radar can enhance its performance by continuously sensing the environment by means of an active feedback between the transmitter and receiver modules. In CR schemes, this feedback is usually implemented through Bayesian filtering [1]. However, this might require some prior information about the environment, which is hardly achieved in practice especially in dynamic environments. In order to overcome this possible limitation, reinforcement learning (RL) approaches can be deployed. RL procedures are characterized by the presence of an *agent* that seeks to attain a certain goal by means of a sequence of *decisions* taken by learning through *trial-error* interactions with the unknown environment [2]. The agent assesses those decisions on the basis of its current *state* and the *reward*. RL procedures have been already exploited in radar detection, for example in [3], where deep RL schemes are adopted to implement an "end-to-end" single target detection. Specifically, the authors use a neural network to approximate the decision statistic. However, no statistical guarantees are given on the overall detection performance under a variable disturbance distribution. In our recent paper [4], we proposed a novel approach to combine a robust Wald-type test, derived for Massive MIMO (MMIMO) radar system in [5], with a RL-based procedure. The aim was to maximize the detection performance of the resulting algorithm in an unknown environment. However, the algorithm was only tested against a specific unknown disturbance distribution. The goal of the present paper is then to verify the overall robustness of the joint RL/Wald-type detector. To this end, extensive investigations have been performed to check its effectiveness against different (unknown) disturbance distributions by using P_{FA} and P_D as performance metrics. More specifically, the CFAR property and the power of the test (the P_D) are assessed for *i)* different levels of disturbance spikiness and *ii)* for different model orders. The numerical results support the robustness property of the joint RL/Wald test with respect to the unknown disturbance model without sacrificing its statistical power.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 287022738 TRR 196 (S03)

The work of S. Fortunati has been partially supported by DGA under grant ANR-17-ASTR-0015.

Profs. Gini's and Greco's work supported by the Italian Ministry of Education and Research (MIUR) within the framework of the CrossLab Project (Departments of Excellence program).

II. Problem Formulation

Consider a colocated MIMO radar with N_T transmit and N_R receive antennas and a point-like target, located at an agle θ. We assume that the radar cross section (RCS) is constant over all the receiving elements. The transmit and receive steering vectors are denoted by $\mathbf{a}_T(\theta)$ and $\mathbf{a}_R(\theta)$, where:

$$\mathbf{a}_T(\theta) = [1, e^{j2\pi\nu}, \dots, e^{j2\pi(N_t-1)\nu}]^T, \tag{1}$$

and $\mathbf{a}_R(\theta)$ is defined similarly. Note that $\nu \triangleq \frac{df}{c}\sin(\theta)$ where f is the carrier frequency and c is the speed of light. We assume a uniform linear array (ULA) with inter-element spacing $d = \lambda/2$ for both the transmitter and the receiver. The baseband representation of the received signal at continuous time t is defined as [6], [7]

$$\mathbb{C}^{N_r} \ni \mathbf{z}(t) = \alpha \mathbf{a}_R(\theta)\mathbf{a}_T^T(\theta)\mathbf{x}(t-\tau) + \hat{\mathbf{c}}(t) \tag{2}$$

where $\alpha \in \mathbb{C}$ accounts for the target RCS and the two-way path loss and τ represents the time delay due to the target position with respect to the radar. The random disturbance vector is denoted as $\hat{\mathbf{c}}(t) \in \mathbb{C}^{N_R}$. The transmit signal $\mathbf{x}(t) \in \mathbb{C}^{N_T}$ is composed of a linear combination of independent orthonormal signals $\mathbf{\Phi}(t) \in \mathbb{C}^{N_T}$, such that $\mathbf{x}(t) = \mathbf{W}\mathbf{\Phi}(t)$, where $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{N_T}]^T \in \mathbb{C}^{N_T \times N_T}$ indicates a beamforming

weight matrix satisfying $\mathrm{tr}\{\mathbf{W}\mathbf{W}^H\} = P_T$, where P_T is the total transmit power. After standard matched filtering, the signal at the output of the receiver is given by:

$$\mathbb{C}^N \ni \mathbf{y} = \mathrm{vec}\left(\int_0^T \mathbf{z}(t)\mathbf{\Phi}^H(t-\tau)dt\right) = \alpha\mathbf{h}(\theta) + \mathbf{c}, \quad (3)$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operator and $N = N_R N_T$ is the number of virtual spatial channels and

$$\mathbf{h}(\theta) = (\mathbf{W}^T\mathbf{a}_T(\theta)) \otimes \mathbf{a}_R(\theta), \quad (4)$$

where $\otimes$ is the Kronecker product. The spatially colored disturbance vector is denoted by $\mathbf{c} = \mathrm{vec}(\mathbf{C})$ where $\mathbf{C} = \int_0^T \hat{\mathbf{c}}(t)\mathbf{\Phi}^H(t-\tau)dt$ is disturbance matrix at the output of the matched filter. It is worth mentioning that the statistical description of the disturbance vector $\mathbf{c}$ is usually unknown, hence its accurate modeling is a challenging task in practice [8], [9]. Even if some simplistic disturbance models have been adopted in literature, their a priori adoption may lead to a *misspecification problem* [10] causing a performance drop in real-word scenario.

To minimize the risk of model mismatch, a very weak statistical assumption on the disturbance is made here [5, A1]:

A1 *The disturbance is a realization of a discrete-time, circular, complex random process with a polynomial decay of its autocorrelation function.*

Note that this assumption is weak enough to include most practical disturbance models such as autoregressive (AR), autoregressive moving average (ARMA) or general correlated Compound-Gaussian model [5].

A. Detection Problem

The received signal in (3) is processed by a bank of spatial filters. Each filter is tuned to a specific angle range l, where the radar field of view is divided into L separate discrete angle bins each at θ_l. It is assumed that each angle bin l contains only one target and the system transmits in total K pulses such that $k \in \{1, \ldots, K\}$. For a single angle bin l, the radar detection can be cast in terms of the following hypothesis testing problem:

$$\begin{aligned} H_0: &\quad \mathbf{y}_l^k = \mathbf{c}_l^k \qquad k = 1, \ldots, K \\ H_1: &\quad \mathbf{y}_l^k = \alpha_l^k\mathbf{h}_l^k + \mathbf{c}_l^k \qquad k = 1, \ldots, K, \end{aligned} \quad (5)$$

As previously mentioned, the disturbance entries of $\mathbf{c}_l^k$ are sampled from complex random process, satisfying Assumption A1. Furthermore, the disturbance covariance matrix $\mathbf{\Gamma} = \mathbb{E}\{(\mathbf{c}_l^k)(\mathbf{c}_l^k)^H\}$ is assumed to be unknown. The disturbance statistics can vary in time and space. The targets can also change over time. In particular, we allow to change from one pulse to the other: *i)* the number of targets; *ii)* their spatial frequencies; *iii)* their signal-to-noise ratio (SNR). Then, we consider a *single snapshot scenario* and consequently the detection is performed *per pulse*. To discriminate between H_0 and H_1, we implement the test statistic for the k^{th}pulse as $\Lambda\left(\mathbf{y}_l^k\right) \underset{H_0}{\overset{H1}{\gtrless}} \lambda_\Lambda$. Conventional model-based test statistics such as the *generalized likelihood ratio test (GLRT)*, are generally adopted in the radar literature. However, GLRT-like schemes can not be directly applied to our model, since they require a priori information about the disturbance probability density function (pdf). In our work, to avoid the risk of running into a misspecified scenario, we do not assume any functional form of the pdf of $\mathbf{c}_l^k$ in (5).

In order to handle the detection problem in (5) under the extremely general and weak assumption A1, the following robust Wald-type detector has been deployed [5]:

$$\Lambda_{l,\mathsf{RW}}^k = \frac{2|(\mathbf{h}_l^k)^H\mathbf{y}_l^k|^2}{(\mathbf{h}_l^k)^H\widehat{\mathbf{\Gamma}}\mathbf{h}_l^k}, \quad (6)$$

where $\widehat{\mathbf{\Gamma}}$ is the estimate of the unknown $\mathbf{\Gamma}$ [5]. Specifically, it can be shown that, if Assumption A1 holds true, this Wald-type detector satisfies the following asymptotic (i.e., $N \to \infty$) relations:

$$\Lambda_{l,\mathsf{RW}}^k\left(\mathbf{y}_{l,g}^k|H_0\right) \underset{N_T N_R\to\infty}{\overset{d}{\sim}} \chi_2^2\left(0\right), \quad (7)$$

$$\Lambda_{l,\mathsf{RW}}^k\left(\mathbf{y}_{l,g}^k|H_1\right) \underset{N_T N_R\to\infty}{\overset{d}{\sim}} \chi_2^2\left(\zeta\right), \quad (8)$$

where $\zeta = 2|\alpha|^2\frac{\|\mathbf{h}\|^4}{\mathbf{h}^H\mathbf{\Gamma}\mathbf{h}}$. [1] These asymptotic properties allow to choose the detection threshold λ_Λ that is able to guarantee a pre-assigned P_{FA} *irrespective of the unknown pdf of the disturbance*. In particular, λ_Λ can be obtained as:

$$\lambda_\Lambda = H_{\chi_2^2}^{-1}(1 - P_{FA}), \quad (9)$$

in which $H_{\chi_2^2}^{-1}(\cdot)$ is the inverse of the cumulative distribution function (cdf) of a χ_2^2 random variable. Moreover, from (8), a closed form expression for P_D can be obtained as:

$$P_\mathsf{D}\left(\lambda\right) \to_{N\to\infty} Q_1\left(\sqrt{\zeta}, \sqrt{\lambda}\right), \quad (10)$$

where $Q_1\left(\cdot,\cdot\right)$ is first order *Marcum Q function* [11].

An important remark is in order here. While the asymptotic distribution of $\Lambda_{l,\mathsf{RW}}^k(\mathbf{y}_{l,g}^k|H_0)$ does not depend on the beamforming matrix $\mathbf{W}$, the asymptotic distribution of $\Lambda_{l,\mathsf{RW}}^k(\mathbf{y}_{l,g}^k|H_1)$ does through the dependence on $\mathbf{W}$ of the vector $\mathbf{h}$ in the non-centrality parameter ζ. This fact is of crucial importance since it provides the theoretical guarantee that it is possible to implement a RL-based algorithm capable of enhancing the detection performance of the above-mentioned Wald-type detector while keeping the CFAR property.

III. RL-BASED MMIMO COGNITIVE RADAR

RL is a machine learning technique which enables a certain agent to achieve an assigned goal through learning the surrounding environment by trial and error. The agent gets a continous feedback from the environment based on the actions it takes. Consequently, the agent evaluates its *action* a_k using two types of information: *state* s_k and *reward* r_k. In our detection problem, the agent is the MIMO radar with an assigned goal to detect multiple targets within unknown disturbance [4].

[1] Further details about the calculation of $\widehat{\mathbf{\Gamma}}$ and the asymptotic distribution of $\Lambda_{l,\mathsf{RW}}^k$ are provided in [5]

A. The set of states

A state s_k in a RL problem defines the current status of the unknown environment. In our problem, the state space $\mathcal{S}$, is defined in terms of the statistic $\Lambda^k_{l,\mathsf{RW}}$ in (6). In particular, a new statistic $\bar{\Lambda}^k_l$ is defined such that:

$$\bar{\Lambda}^k_l = \begin{cases} 1 & \Lambda^k_{l,\mathsf{RW}} > \lambda_\Lambda \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Hence, $\bar{\Lambda}^k_l$ indicates if a certain angle bin l at time k contains a target or not. Therefore, s_k can be described as the total number of angle bins containing a target at specific time k:

$$s_k = \sum_{l=1}^{L} \bar{\Lambda}^k_l. \tag{12}$$

Consequently, the set of states is $\mathcal{S} = \{0, \ldots, M\}$, where M is the maximum number of targets that can be detected.

B. The set of actions

The MIMO radar, i.e., agent, at every time k chooses a certain action a_k from a set of available actions $\mathcal{A}$ based on s_k. An action is generally defined by two main tasks: candidate angle bins selection and beamforming. In particular, based on the environmental state, the agent selects the corresponding angle bins that most likely contain targets. Subsequently, the agent optimizes the beamformer matrix $\mathbf{W}$ to focus the beampattern towards the direction of those bins.
Therefore, $a_k \in \mathcal{A} = \{\Theta_i | i \in \{0, 1, \ldots, M\}\}$, where the set of i candidate angle bins is $\Theta_i = \{\hat{\theta}_1, \ldots, \hat{\theta}_i\}$ and $\hat{\theta}$ is the estimated angle bin of the target. Θ_i is defined based on the highest i values of $\Lambda^k_{l,\mathsf{RW}}$ in (6). As previously mentioned, the agent utilizes this acquired information to optimize $\mathbf{W}$ towards the desired angle bins Θ_i. This is done by focusing the transmit power towards Θ_i, hence the optimization problem is formulated as maximizing the minimum of the beampattern. In more details, the optimization problem is cast as:

$$\begin{aligned} \max_{\mathbf{W}} \min_{j \in \mathcal{T}_i} \{\mathbf{a}_T^T(\hat{\theta}_j) \mathbf{W} \mathbf{W}^H \mathbf{a}_T^*(\hat{\theta}_j)\} \\ \text{s.t. } \mathrm{tr}(\mathbf{W}\mathbf{W}^H) = P_T, \end{aligned} \tag{13}$$

where $\mathcal{T}_i = \{1, \ldots, i\}$ and $\hat{\theta}_j \in \Theta_i$. This problem is solved using iterative inner convex approximations algorithm [4].

C. The reward

The reward is defined as the enviromental feedback which defines how well the agent is doing at a certain step k. The agent's main goal is to maximize the total cumulative reward function [2]. In our specific application, the agent's goal is to detect all the targets without assuming any prior information about the environment, (i.e., number of targets and disturbance statistics are unknown). The radar agent continuously explores changes in the environment in real time, and modifies its actions accordingly, i.e., optimizing the beamformers. To achieve this specific goal, the reward is defined in terms of the estimated $\hat{P}^k_{D_l}$ that can be calculated in a closed form asymptotically, i.e., $N \to \infty$ as

$$\hat{P}^k_{D_l} = Q_1\left(\sqrt{\hat{\varsigma}^k_l}, \sqrt{\lambda_\Lambda}\right), \tag{14}$$

$$\hat{\varsigma}^k_l = 2|\hat{\alpha}^k_l|^2 \frac{\left\|\mathbf{h}^k_l\right\|^4}{(\mathbf{h}^k_l)^H \widehat{\mathbf{\Gamma}}_l \mathbf{h}^k_l}, \tag{15}$$

$$\hat{\alpha} = \frac{(\mathbf{h}^k_l)^H \mathbf{y}^k_l}{||\mathbf{h}^k_l||}. \tag{16}$$

The reward for each time step k is given as:

$$r_{k+1} = \sum_{l=1}^{s_k} \hat{P}^k_{D_l} - \sum_{j=1}^{L-s_k} \hat{P}^k_{D_j}. \tag{17}$$

In particular, the reward consists of two components, a negative and a positive reward. The positive one is a summation of $\hat{P}^k_{D_l}$ over all s_k, which means it is summed over all the bins that most likely contain a target. On the contrary, the negative reward is summed over the bins that do not. The best case scenario occurs when there is a target in every bin such that $s_k = L$, as this means that the decision statistic $\Lambda^k_{l,\mathsf{RW}} > \lambda, \forall l$ (i.e., L targets are detected).

D. SARSA algorithm and target detection

SARSA is an acronym for *state-action-reward-state-action* sequence. In more details, in SARSA the sequence s_k,a_k,r_{k+1},s_{k+1} and a_{k+1} is used to update the Q-function at each time k [12]. The Q function is defined as the expected cumulative reward starting from state s_k and taking action a_k following a certain policy π. The radar agent in our problem continously updates a state-action matrix $\mathbf{Q} \in \mathbb{R}^{(M+1)\times(M+1)}$ of elements $Q(s_k, a_k)$. The matrix is first initialized with zeros, then updated based on the Q function after the execution of a certain action. The *Q-function* is chosen according to the following update rule [2]

$$\begin{aligned} Q(s_k, a_k) \leftarrow & Q(s_k, a_k) + \\ & \alpha\left(r_{k+1} + \gamma Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k)\right). \end{aligned} \tag{18}$$

$\alpha \in [0, 1]$ is the learning rate controlling how much the new experiences override the old ones. Note that γ is the discount factor which controls the impact of future rewards.
The agent follows a certain policy π to determine which action should be taken. In our algorithm, en ϵ-greedy policy is employed to define a_k through defining the size of Θ_i (i.e., i). The optimal action $a_{\mathsf{opt}} \triangleq \arg\max_{a \in A} \mathbf{Q}(s_{k+1}, a)$ is chosen with a probability of $1 - \epsilon$, while another random action a_{rnd} (excluding a_{opt}) is chosen with a probability of ϵ. The algorithm steps are explained in Alg. 1.

IV. Simulation results

In our simulations, we consider a total of $L = 21$ angle bins, where the angle grid is expressed in terms of the spatial frequency $\boldsymbol{\nu} = [-0.5 : 0.5]$. Furthermore, the disturbance vector $\mathbf{c}^k_l$ is modeled as circular complex $\mathrm{AR}(n)$ process

Algorithm 1 SARSA

Initialize $\mathbf{Q} = \mathbf{0}_M$, $s_0 = 1$, $a_0 = 1$, $K = 50$ and $\mathbf{W}_k = \mathbf{I}$
repeat for each time step k:
  Take action a_k by using $\mathbf{W}_k$ as beamforming matrix
  Acquire the received signal $\mathbf{y}_l^k$, $\forall\ l = 1, \ldots, L$
  Calculate s_{k+1} from (12) and r_{k+1} as in (17)
  Choose action a_{k+1} with ϵ greedy, identify Θ_i and $\mathcal{T}_i$
  Update $Q(s_k, a_k)$ as in (18)
  $s_k \leftarrow s_{k+1}$;$a_k \leftarrow a_{k+1}$
  if $s_{k+1} \neq 0$ **then**
    Solve for $\mathbf{W}_{k+1}$ in (13)
  else
    $\mathbf{W}_k = \mathbf{I}$
until Observation time ends

[5] $c_n = \sum_{i=1}^{n} \rho_i c_{n-i} + w_n,\ \ n \in (-\infty, \infty)$, driven by independent, identically t-distributed (i.i.d.) innovations w_n whose variance is σ_w^2 and pdf p_w is defined as:

$$p_w(w_n) = \frac{\lambda}{\sigma_w^2}\left(\frac{\lambda}{\xi}\right)^{\lambda}\left(\frac{\lambda}{\xi} + \frac{|w_n|^2}{\sigma_w^2}\right)^{-(\lambda+1)}, \qquad (19)$$

where $\xi = \lambda/\left(\sigma_w^2(\lambda - 1)\right)$ is a scale parameter, while the shape parameter $\lambda \in (1, \infty)$ controls the non-Gaussianity of w_n. Specifically, p_w is a heavy tailed pdf with highly non-Gaussian behavior when $\lambda \to 1$. On the contrary, if $\lambda \to \infty$, then p_w collapses into a Gaussian distribution. In order to test the robustness of our algorithm, we analyze its performance against different disturbance scenarios characterizing harsh environments. We compare the performance of our RL-based waveform matrix selection scheme against omnidirectional transmission with equal power allocation. In the latter case, orthonormal waveforms are transmitted and the total power is equally divided across all antennas under the constraint $P_t = 1$. In the following three different scenarios are analyzed.

A. Varying N

In this scenario, the parameters of the innovation process are chosen to be $\lambda = 2$ and $\sigma_w^2 = 1$. Furthermore, the normalized power spectral density (PSD) of the AR disturbance is modeled as in [5]

$$S(\nu) \triangleq \sigma_w^2 \left| 1 - \sum_{n=1}^{p} \rho_n e^{-j2\pi n\nu} \right|^{-2}, \qquad (20)$$

with $p = 6$ as the order of the AR process, while the coefficient vector ρ is

$$\rho = [0.5e^{-j2\pi 0.4}, 0.6e^{-j2\pi 0.2}, 0.7e^{-j2\pi 0}, 0.4e^{-j2\pi 0.1}, \quad (21)$$
$$0.5e^{-j2\pi 0.3}, 0.6e^{-j2\pi 0.35}]^T.$$

Hence, the disturbance power is distributed across the whole spatial frequency range. Four targets are generated at $\nu = \{-0.2, 0, 0.2, 0.3\} \subset \boldsymbol{\nu}$, with SNR $= [-5\text{dB}, -8\text{dB}, -10\text{dB}, -9\text{dB}]$, respectively. Fig. 1 shows the $\hat{P}_D$ for the target at $\nu = 0.3$ as a function of the virtual spatial channels N for a pre-assigned $P_{FA} = 10^{-4}$. The detection of this target might be a hard task since it is masked within a clutter peak. Furthermore, it suffers from very low SNR. However, our algorithm can successfully detect the target as $N \to 10^4$ (i.e. $N_T = 100$), in contrast to the omnidirectional approach. In addition, we can see that the estimated $\hat{P}_D$ of the RL algorithm through multiple Monte Carlo runs agrees with the theoretical nominal one provided in (10).

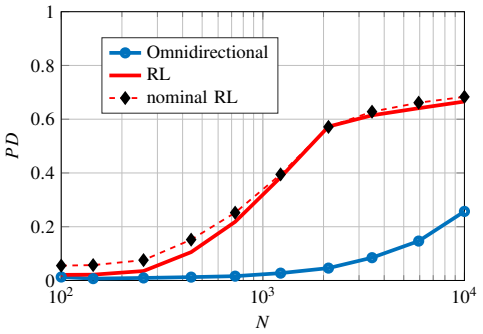

Figure 1: $\hat{P}_D$ at $P_{\mathsf{FA}} = 10^{-4}$ across N

B. Varying λ

In this scenario, we asses the robustness of our algorithm against different levels of non-Gaussianity of the disturbance. We choose $N = 10^4$ and $P_{FA} = 10^{-4}$. Fig. 2 shows the $\hat{P}_D$ as a function of the non-Gaussianity parameter λ. The results show a constant $\hat{P}_D$ for target at $\nu = 0.3$ across different values of λ. This proves that the algorithm has a robust and constant superior behavior compared to the omnidirectional approach. In addition, as expected, the estimated $\hat{P}_D$ matches perfectly with the nominal theoretical one provided in (10).

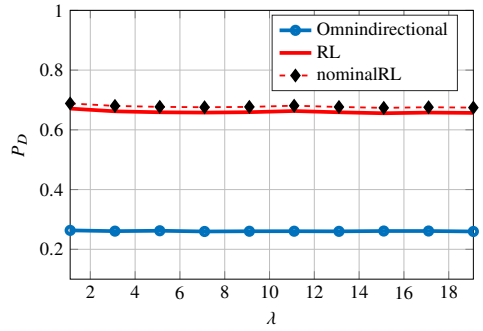

Figure 2: $\hat{P}_D$ at $P_{\mathsf{FA}} = 10^{-4}$ and $N = 10^4$.

In Fig. 3, the CFARness of the algorithm is assessed against the disturbance spikiness. Fig. 3 shows that our RL algorithm provides a constant P_{FA} across λ, similar to the omnidirectional approach. Both algorithms achieve the nominal $\bar{P}_{FA} = 10^{-4}$. This proves the theoretical results in (8), which indicates that the CFAR property is always (asymptotically) achieved using the Wald-type statistic $\Lambda_{l,\mathsf{RW}}^k$ irrespective of the specific waveform matrix $\mathbf{W}$. This is a consequence of (8) that shows that, under H_0, $\Lambda_{l,\mathsf{RW}}^k(\mathbf{y}_{l,g}^k|H_0)$ is distributed

as a central chi-squared χ^2_2 random variable regardless of $\mathbf{W}$.

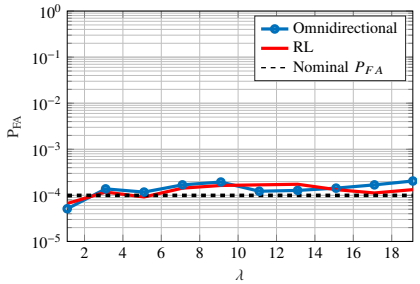

Figure 3: P_{FA} at $N = N_T N_R = 10^4$ across λ.

C. Varying AR(p)

The robustness of the RL algorithm is further validated across more general disturbance models. In this scenario, the $\hat{P}_D$ is evaluated across many orders of the autoregressive process (AR). Specifically, p varies as $p \in [1, \ldots, 10]$. The magnitude of ρ_n in (20) is chosen from $[0.8, 0.7, 0.7, 0.6, 0.6, 0.4, 0.4, 0.5, 0.5, 0.3]$, while the corresponding spatial frequency is selected from $[0, 0.1, -0.1, 0.2, -0.2, 0.1, -0.1, 0.4, -0.4, 0.5]$. For instance, if $p = 1$, then $\rho = 0.8e^{-j2\pi 0}$, while if $p = 2$, then $\rho = \left[0.7e^{-j2\pi - 0.1}, 0.8e^{-j2\pi 0}, 0.7e^{-j2\pi 0.1}\right]$. Fig. 4 shows the probability of detection of the target at $\nu = 0$. Note that at $\nu = 0$, there is always a disturbance peak, regardless of the value of p. Despite that, the $\hat{P}_D$ of this target using our algorithm is constantly higher compared to the omnidirectional case, no matter the order of the AR. It can be noticed a slight drop in the case of AR(1), $p = 1$, as all the disturbance energy in this case is focused on the target at $\nu = 0$, while an AR($p > 1$) will spread it all over multiple spatial frequency points. Again here, the estimated $\hat{P}_D$ agrees with the theoretical nominal $\hat{P}_D$ in (10).

Finally, Fig. 5 shows that the CFAR property with respect to the order p is satisfied for both the proposed RL-based and the omnidirectional algorithms. Again, this represent a numerical validation of the theoretical result that the CFAR property is satisfies using the Wald statistic $\Lambda^k_{l,\mathsf{RW}}$ in any disturbance statistics independent of $\mathbf{W}$.

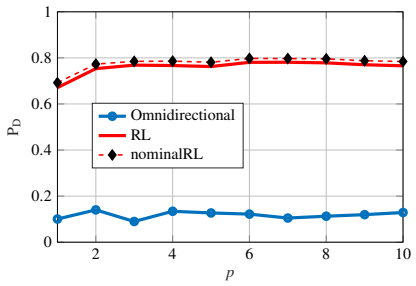

Figure 4: $\hat{P}_D$ at $P_{\mathsf{FA}} = 10^{-4}$ across several p

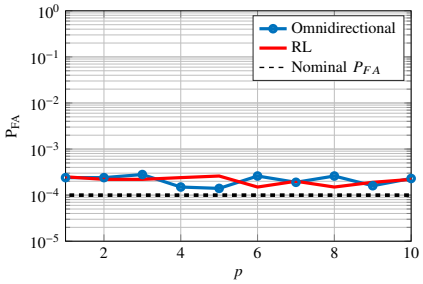

Figure 5: P_{FA} at $N = N_T N_R = 10^4$ across p.

V. Conclusion

In this paper, we investigated the robustness of the multi-target RL-based Wald-type detector proposed in [4], [5]. The performance of the algorithm has been assessed for various unknown disturbance models. The main results is that the RL-based Wald-type detector is able to achieve the CFAR property with respect to a wide range of (unknown) disturbance models. At the same time, the RL-based waveform selection scheme will provide the detector with a remarkable increase of its P_D while keeping the CFAR property. Last but not the least, the estimated P_D obtained by using the RL-based scheme is in agreement with the theoretical closed form expression provided in [5].

References

[1] S. Haykin, Y. Xue, and P. Setoodeh, "Cognitive radar: Step toward bridging the gap between neuroscience and engineering," *Proceedings of the IEEE*, vol. 100, no. 11, pp. 3102–3130, Nov 2012.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[3] W. Jiang, A. M. Haimovich, and O. Simeone, "End-to-end learning of waveform generation and detection for radar systems," in *53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1672–1676.

[4] A. M. Ahmed, A. A. Ahmad, S. Fortunati, A. Sezgin, M. S. Greco, and F. Gini, "A reinforcement learning based approach for multi-target detection in massive MIMO radar," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–1, 2021.

[5] S. Fortunati, L. Sanguinetti, F. Gini, M. S. Greco, and B. Himed, "Massive MIMO radar for target detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 859–871, 2020.

[6] B. Friedlander, "On transmit beamforming for MIMO radar," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3376–3388, October 2012.

[7] J. Li and P. Stoica, "MIMO radar with colocated antennas," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 106–114, Sep. 2007.

[8] K. J. Sangston, F. Gini, M. V. Greco, and A. Farina, "Structures for radar detection in compound Gaussian clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 35, no. 2, pp. 445–458, 1999.

[9] K. J. Sangston, F. Gini, and M. S. Greco, "Coherent radar target detection in heavy-tailed compound-Gaussian clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 1, pp. 64–77, 2012.

[10] S. Fortunati, F. Gini, M. S. Greco, and C. D. Richmond, "Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142–157, Nov 2017.

[11] A. H. Nuttall, "Some integrals involving the (q sub m)-function," 1974.

[12] D. Poole and A. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, 2nd ed. Cambridge, UK: Cambridge University Press, 2017. [Online]. Available: http://artint.info/2e/html/ArtInt2e.html