# Weak Target Detection in Massive MIMO Radar via an Improved Reinforcement Learning Approach

Weitong Zhai[1], Xiangrong wang[1], Maria S. Greco[2] and Fulvio Gini[2]
[1] School of Electronic and Information Engineering, Beihang University, Beijing, China
[2]Department of Information Engineering, University of Pisa, Pisa, Italy
Emails: {wtzhai, xrwang}@buaa.edu.cn, {m.greco, f.gini}@iet.unipi.it

***Abstract*—Massive multi-input-multi-output (MMIMO) cognitive radar can enhance the target detection ability in a dynamic environment via a continuous "perception-action" cycle. In our previous work, we proposed a reinforcement learning (RL) based approach for multi-target detection in MMIMO. However, this method shows poor detection performance for weak targets attributed to its imperfect action and reward mechanisms. In this paper, we propose an improved RL based method to enhance the detection probability of weak targets. In the action stage, the transmit power is divided into omni-directional and directional components, the former significantly reduces the missed detection probability of weak targets and the latter improves the detection probability by focusing more power on weak targets. Moreover, the reward mechanism of RL is modified to further improve the detection performance. In addition, the transmit weight matrix is designed by an optimum combination of the beampatterns of all unit orthogonal transmit waveforms, thus greatly reducing the computational complexity. Simulation results are provided to demonstrate the effectiveness of the improved RL based method for weak target detection.**

***Index Terms*—Reinforcement learning, Cognitive radar, Massive MIMO, Weak target detection, Beamforming.**

## I. Introduction

Cognitive radar (CR) can jointly optimize the transmit and receive parameters through a continuous "perception-action" cycle. With the experience accumulation, it can enhance the performance of specific tasks, such as target detection [1]–[3]. Benefiting from the waveform diversity of transmit signals, the MIMO CR can leverage increased degrees of freedom (DoFs) to improve the target detection ability. A common model of target detection is to make some "ad-hoc" assumptions of the disturbance [4], [5], which sometimes might be impractical. Recently, [6] has shown that the MMIMO radar can provide robustness against the unknown disturbance distribution utilizing a robust Wald-type estimator. Capitalizing on this, an RL based approach for multi-target detection in MMIMO radar was proposed in our previous work [7]. This approach works well for strong targets, but poorly for weak targets due to its imperfect action and reward mechanisms. Therefore, in this paper, we propose an improved RL based method in MMIMO radar to enhance the detection ability of weak targets.

The basic principle of RL is that it guides action through the reward obtained by interacting with the environment, with a goal of obtaining the maximum reward [8]. This process can be described as Markov decision processes (MDP). Compared with existing detection approaches that depend on the prior information of dynamic environment [9], [10], RL can eradicate this dependence by perception and learning. The work in [11] applied RL to investigate a target tracking radar that must coexist with a communication system. In [12], RL was used for indoor mapping of UAV. While, [13] examined the radar-communications coexistence problem by modeling the radar environment as an MDP and applying RL to solve the resultant optimization problem. These works either consider the case of single target only or make specific assumptions about the disturbance distribution. In our previous work [7], a fully data-driven RL framework for multi-target detection was proposed, which is applicable in the presence of unkonwn disturbance statistics. However, this approach is not perfect yet, especially the missed detection probability of weak targets is high. The reasons are two-fold: on one hand, the unselected angular bins are completely ignored even those probably containing targets; on the other hand, transmit power is evenly distributed to all possible targets. Therefore, we improve both the action and reward mechanisms of the RL to reduce the missed detection probability, while increasing the detection probability of weak targets simultaneously.

Transmit waveform optimization is a prominent step for cognitive MMIMO radar [14], and can be generally divided into two categories. One is the direct optimization of waveform samples of each antenna [15]–[17]. The other is the optimization of the weighting matrix of different orthogonal signals in the multi-carrier framework [18]–[20]. Semi-definite programming (SDP) is a commonly-used method for both optimizations. Whereas, SDP exhibits high computational complexity especially for the MMIMO case. In [7], a continuous iterative convex optimization method was proposed to design the transmit weight matrix with reduced computational complexity. Nevertheless, a faster waveform design method is necessary in order to further reduce the time of each training. In this paper, we propose a fast convex optimization algorithm by combining the beampatterns of all unit orthogonal transmit signals. One convex optimization is sufficient and no iterations are required. Moreover, the $N_T \times N_T$-dimensional complex weight matrix is downsized to an $N_T$-dimensional real vector variable, which greatly reduces the computational complexity.

The rest of this paper is organized as follows. Section II briefly review the signal model and detection method in [7]. Section III proposes the improved RL method. Section IV proposes the fast transmit weight matrix design based on the combination of beampatterns of all orthogonal signals.

The work by W Zhai and X Wang is supported by National Natural Science Foundation of China under Grant No. 62071021 and No. 61827901.

The work of M.S. Greco and F. Gini has been partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence) of the University of Pisa, laboratory of Industrial Internet of Things (IIoT).

Numerical simulations are provided in Section V and some concluding remarks are reported in Section VI.

## II. Review of Previous Work

### A. Signal model

Consider a colocated MMIMO system consisting of $N_T$ transmit antennas and $N_R$ receive antennas. Both of them are uniform linear arrays (ULA) with an interval of $d_T = d_R = \frac{\lambda}{2}$ ($\lambda$ represents the wavelength).

The received echo from one point-like target at the time instant $t$ can be expressed by [18], [21],

$$\mathbf{x}(t) = \alpha \mathbf{a}_R(\theta)\mathbf{a}_T^T\mathbf{s}(t-\tau)e^{j\omega t} + \mathbf{n}(t),\ t \in [0,T], \quad (1)$$

where $\mathbf{x}(t) \in \mathbb{C}^{N_R}$ is the received data vector, $\mathbf{s}(t) \in \mathbb{C}^{N_T}$ is the transmit signal vector, $\tau$ is the time delay, $\omega$ is the Doppler shift, $\alpha$ is an unknown coefficient representing the radar RCS and two way path loss following Swerling 0, $\mathbf{a}_T(\theta)$ and $\mathbf{a}_R(\theta)$ represent the steering vectors of the transmit and receive array, separately, with $\mathbf{a}_T(\theta) = [1, e^{j\frac{2\pi d_T}{\lambda}sin\theta}, ..., e^{j\frac{2\pi d_T}{\lambda}(N_T-1)sin\theta}]^T$ and $\mathbf{a}_R(\theta)$ defined in the same way, and $\mathbf{n}(t) \in \mathbb{C}^{N_R}$ is the noise plus clutter vector.

Consider the case of multi-carriers, that is, each transmit signal is a linear combination of $N_T$ unit orthogonal signals $\mathbf{s}_o(t) \in \mathbb{C}^{N_T}$,

$$\mathbf{s}(t) = \mathbf{W}\mathbf{s}_o(t),\ ||\mathbf{w}_{(i)}||_2^2 = P_T,\ 1 \le i \le N_T, \quad (2)$$

where $\mathbf{W} \in \mathbb{C}^{N_T \times N_T}$ is the transmit weight matrix and $\mathbf{w}_{(i)}$ represents the $i$th row of $\mathbf{W}$. Constraint $||\mathbf{w}_{(i)}||_2^2 = P_T$ indicates that the transmit power of each antenna is constant. Then the transmit beampattern can be expressed as $B(\theta) = \mathbf{a}_T^T(\theta)\mathbf{R}_W\mathbf{a}_T(\theta)^*$, where $\mathbf{R}_W = \mathbf{W}\mathbf{W}^H$.

After applying the matched filtering, the received signal is transformed into a $N_R \times N_T$-dimensional matrix,

$$\mathbf{Y} = \alpha \mathbf{a}_R(\theta)\mathbf{a}_T^T\mathbf{W}\int_0^T \mathbf{s}_o(t-\tau)\mathbf{s}_o^H(t-\hat{\tau})e^{j(\omega-\hat{\omega})t}dt + \mathbf{C}, \quad (3)$$

where $\mathbf{C} = \int_0^T \mathbf{n}(t)\mathbf{s}_o^H(t-\hat{\tau})e^{-j\hat{\omega}t}dt$. Assuming that the filter matches the model exactly, that is $\hat{\tau} = \tau$ and $\hat{\omega} = \omega$, we have $\int_0^T \mathbf{s}_o(t-\tau)\mathbf{s}_o^H(t-\hat{\tau})e^{j(\omega-\hat{\omega})t}dt = \mathbf{I}$. Substituting it into Eq. (3) and vectorizing matrix $\mathbf{Y}$, we have,

$$\mathbf{y} = \mathrm{vec}(\mathbf{Y}) = \alpha\mathbf{v}(\theta) + \mathbf{c}, \quad (4)$$

where $\mathbf{y} \in \mathbb{C}^{N_T N_R}$ and,

$$\mathbf{v}(\theta) = (\mathbf{W}^T\mathbf{a}_T(\theta)) \otimes \mathbf{a}_R(\theta), \quad (5)$$

where $\otimes$ denotes the Kronecker product, $\mathbf{c}$ is the space disturbance vector. In order to prevent mismatch with the unknown environment, we adopt a very general noise model proposed in [6], as shown in the following assumption,

*Assumption 1: Let $\{c_n : \forall n\}$ be a real and unknown disturbance process, which is a stationary discrete and cyclic complex valued process. It is only assumed that its autocorrelation function $r_C[m] \triangleq \mathbf{E}\{c_n c_{n-m}^*\} = O(|m|^{-\gamma})$ has a polynomial decay.*

### B. Detection problem

We divide the detection area into $L$ discrete angular bins on average, $\{\theta_l | 1 \le l \le L\}$ and consider $K$ transmit pulses. According to Eq. (4), the received signal of the $l$th angular bin and $k$th pulse can be expressed by,

$$\mathbf{y}(l,k) = \alpha(l,k)\mathbf{v}(l,k) + \mathbf{c}(l,k). \quad (6)$$

Then the hypothesis testing problem for the $l$th angular bin is,

$$\begin{aligned} H_0 &: \ \mathbf{y}(l,k) = \mathbf{c}(l,k),\ k = 1, ..., K, \\ H_1 &: \ \mathbf{y}(l,k) = \alpha(l,k)\mathbf{v}(l,k) + \mathbf{c}(l,k),\ k = 1, ..., K. \end{aligned} \quad (7)$$

According to [7], we adopt a robust Wald-type test to each angular bin in one pulse, and the statistic is given by,

$$\Lambda_{\text{RW}}(\mathbf{y}(l,k)) = \frac{2|\hat{\alpha}(l,k)|^2}{\mathbf{v}^H(l,k)\hat{\mathbf{\Gamma}}\mathbf{v}(l,k)}, \quad (8)$$

where $\hat{\alpha}(l,k)$ is the least-square estimator of $\alpha$ and $\hat{\mathbf{\Gamma}}$ is the estimate of the unknown convariance matrix of the disturbance. The expressions of them can be found in [7]. In the following, we abbreviate $\Lambda_{\text{RW}}(\mathbf{y}(l,k))$ as $\Lambda_{l,\text{RW}}^k$ and use $\Lambda_{\text{RW}}$ to represent the general Wald-type statistic. Based on the Wald-type test and Eq. (7), the detection problem can be described as,

$$\Lambda_{\text{RW}} \overset{H_1}{>} \bar{\lambda}, \quad \Lambda_{\text{RW}} \overset{H_0}{<} \bar{\lambda}, \quad (9)$$

where $\bar{\lambda}$ is a preset detection threshold that can satisfy the constant false alarm probability (CFAR) criterion.

If Assumption 1 holds, according to [6], when the number of virtual antennas is large enough, the probability density function (PDF) of $\Lambda_{\text{RW}}$ under $H_0$ and $H_1$ will converge to the chi-square distribution. And the CFAR criterion can be satisfied by setting $\bar{\lambda} = -2\ln P_{\text{FA}}$ ($P_{\text{FA}}$ represents the false alarm probability).

## III. Improved RL-based Approach

In this section, we introduce the target detection based on SARSA (state-action-reward-state-action) in the framework of RL. In particular, we delineate the action and reward of RL to improve the detection performance of weak targets.

In the implementation of SARSA, one training is conducted in one pulse. And for the $k$th training, there are three parameters: states, actions and rewards, which are recorded as $s_k$, $a_k$ and $r_k$ respectively. For our problem, $s_k$ represents the number of targets in the detection area, $a_k$ represents the number of targets pointed by the pre-designed transmit beam. Assuming that the maximum number of possible targets in the detection area is $M$, then the number of possible states and actions are both $M+1$. We define $\mathcal{S} = \{0, 1, ...M\}$ as the set of all states and $\mathcal{A} = \{0, ..., M\}$ as the set of all actions. Then we can define the state-action matrix $\mathbf{Q} \in \mathbb{R}^{(M+1)\times(M+1)}$ [7]. According to the update rule of SARSA, the element $Q(s_k, a_k)$ in the $k$th training is updated by,

$$\begin{aligned} &Q(s_k,a_k) \longleftarrow \\ &\quad Q(s_k,a_k) + \beta(r_{k+1} + \gamma Q(s_{k+1}, a_{k+1}) - Q(s_k,a_k)) \end{aligned} \quad (10)$$

For the $k$th training, $s_k = \sum_{l=1}^{L} \text{sign}(\Lambda_{l,\text{RW}}^k - \bar{\lambda})$ is defined the same as [7]. To enhance the detection ability of weak targets, we improve the updating of $a_k$ and $r_k$ as follows.

### A. Actions

In order to enhance the detection of weak targets, the actions are defined as focusing the transmit beams to the angular bins which may contain targets. Initially, since we don't have any environmental information, we take $\mathbf{W} = \mathbf{I}$ for omni-directional detection. For the subsequent training, we adopt the $\epsilon$-greedy policy to select action, that is,

$$a_k = \begin{cases} \arg\max_{a\in\mathcal{A}} \mathbf{Q}(s_k, a) & \text{with prob. } 1-\epsilon, \\ \text{random action} & \text{with prob. } \epsilon. \end{cases} \tag{11}$$

After that, we design the transmit beampattern according to $a_k$. We select the $a_k$ highest $\Lambda^k_{l,\text{RW}}$ calculated by Eq. (8) and point the beam to the corresponding angular bins. We record the set of these selected angular bins as $\mathbf{\Theta}_k$. So the transmit beampattern design problem can be described as,

$$\max_{\mathbf{W},\delta} \ \delta \tag{12}$$

$$\text{s.t. } \mathbf{a}_T^T(\theta_i)\mathbf{R}_W\mathbf{a}_T(\theta_i)^* \geq \delta_i, \ \theta_i \in \mathbf{\Theta}_k \tag{12a}$$

$$\delta = \min_{1\leq i\leq a_k} \{\delta_i\}, \tag{12b}$$

$$||\mathbf{w}_{(i)}||_2^2 = P_T, 1 \leq i \leq N_T. \tag{12c}$$

If the transmit beampattern is pointed towards the selected angular bins, the power radiated on the other angular bins will be very low. In this case, for the $(k+1)$th training, the estimated $\Lambda_{\text{RW}}$ of the angular bins containing a target may be not statistically larger than those without a target. This can lead to missed detection of some targets in the subsequent trainings, especially for the weak targets. To avoid this, we divide matrix $\mathbf{R}_W$ into two parts, $\mathbf{R}_W = \mathbf{R}_F + P_1\mathbf{I}$, where $\mathbf{R}_F$ is used to focus the beam into the selected angular bins, and $P_1\mathbf{I}$ is the omni-directional detection matrix used to search all angular bins. Note that $P_1\mathbf{I}$ can ensure that Wald-type statistic $\Lambda_{\text{RW}}$ of unselected angular bins with targets are statistically higher than those without targets, thus reducing the missed detection. The constraint (12c) is equivalent to restricting all the diagonal elements of $\mathbf{R}_W$ to be $P_T$. Accordingly, in order to meet the constant power constraint of the transmitter, all the diagonal elements of matrix $\mathbf{R}_F$ should equal to $(P_T - P_1)$.

To further enhance the $P_D$ of the weak targets, the following strategy can be implemented to increase the radiation power on it. Define $\delta_i \in \{\delta, 5\delta\}$, if one selected $\Lambda^k_{l,\text{RW}}$ is less than $\bar{\lambda}/2$, we regard it as a weak target and take the corresponding $\delta_i$ as $5\delta$; otherwise we take $\delta_i$ as $\delta$. A fast algorithm to solve problem (12) will be provided in section IV.

### B. Rewards

According to [6], the detection probability of angular bin $l$ for the $k$th training is given by,

$$\hat{P}^k_{D,l} = Q_1(\sqrt{\zeta^k_l}, \sqrt{\bar{\lambda}}), \ \zeta^k_l = \frac{2|\hat{\alpha}(l,k)|^2||\mathbf{v}(l,k)||^4}{\mathbf{v}^H(l,k)\hat{\Gamma}(l,k)\mathbf{v}(l,k)} \tag{13}$$

where $Q_1(.)$ is the first order Marcum function. Then we defined the reward of the $k$th training as follows,

$$r_k = \sum_{i\in\mathcal{S}_{k+1}} \hat{P}^{k+1}_{D,i} - \sum_{j\in\mathcal{S}-\mathcal{S}_{k+1}} \hat{P}^{k+1}_{D,j} - |a_k - s_{k+1}|, \tag{14}$$

where $\mathcal{S}_{k+1}$ represents the set of $s_{k+1}$ angular bins. Specifically, we introduce term $|a_k - s_{k+1}|$ which denotes the number difference between focused beams and actual targets, so as to penalize the missed detection or false alarm of the selected angular bins.

## IV. Fast optimization of the transmit beampattern

In this section, we provide a fast algorithm to solve Eq. (12). Since we divide $\mathbf{R}_W$ into two parts, the transmit beampattern $B(\theta)$ is also divided into $B_F(\theta)$ and $B_I(\theta)$. As $B_I(\theta)$ is used for omni-directional detection, we only need to optimize $B_F(\theta)$. Obviously, $\mathbf{R}_F$ is a positive definite matrix, so there must exist a matrix $\mathbf{W}_F \in \mathbb{C}^{N_T\times N_T}$ that satisfies $\mathbf{R}_F = \mathbf{W}_F\mathbf{W}_F^H$. And we have that,

$$\begin{aligned} B_F(\theta) &= \mathbf{a}_T^T(\theta)\mathbf{W}_F\mathbf{W}_F^H\mathbf{a}_T^*(\theta) \\ &= \sum_{i=1}^{N_T} \mathbf{a}_T^T(\theta)\mathbf{w}_{F,i}E\{s_{oi}(t)s_{oi}^*(t)\}\mathbf{w}_{F,i}^H\mathbf{a}_T^*(\theta), \end{aligned} \tag{15}$$

where $\mathbf{w}_{F,i}$ represents the $i$th column of $\mathbf{W}_F$ and $s_{oi}(t)$ represents the $i$th element of $\mathbf{s}_o(t)$. We can find that $B_{F,i}(\theta) = \mathbf{a}_T^T(\theta)\mathbf{w}_{F,i}E\{s_{oi}(t)s_{oi}^*(t)\}\mathbf{w}_{F,i}^H\mathbf{a}_T^*(\theta) = |\mathbf{a}_T^T(\theta)\mathbf{w}_{F,i}|^2$ is the beampattern carried by the $i$th orthogonal signal, where $\mathbf{w}_{F,i}$ is the beamforming weight vector. Eq. (15) implies that $B_F(\theta)$ is equivalent to the summation of $N_T$ beampatterns carried by the unit orthogonal signals.

From the analysis above, we can see that Eq. (15) requires to jointly optimize $N_T$ beampatterns, thus the computational complexity is very high. Suppose the beampattern of each orthogonal signal is pre-designed, and then the computational complexity can be greatly reduced by optimally combining these predefined beampatterns. Capitalizing on this insight, we divide the spatial frequency $v = d_T\text{sin}\theta/\lambda$ into a set of uniformly spaced grid points with an interval of $1/N_T$, denoted by $\mathcal{V} = \{v_i = -\frac{1}{2} + \frac{i-1}{N_T} | i = 1, ..., N_T\}$. We steer the beampattern carried by $s_{oi}(t)$ to $v_i$ by applying conventional beamforming, that is $B_{F,i}(\theta) = |\mathbf{a}_T^T(\theta)\mathbf{b}_i|^2$ where $\mathbf{b}_i = [1, e^{-j2\pi v_i}, ..., e^{-j2\pi(N_T-1)v_i}]^T$. Then the synthesis of the total transmit beampattern can be transformed into,

$$\begin{aligned} B_F(\theta) &= \sum_{i=1}^{N_T} r_i|\mathbf{a}_T^T(\theta)\mathbf{b}_i|^2 \\ &= \mathbf{a}_T^T(\theta)\mathbf{B}\text{diag}\{\mathbf{r}\}\mathbf{B}^H\mathbf{a}_T^*(\theta). \end{aligned} \tag{16}$$

where $\mathbf{r} \in \mathbb{R}^{N_T}$ is the weight vector of the predefined beampatterns, $r_i$ is the $i$th element of $\mathbf{r}$, and $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_{N_T}]$ is the beam-space transformation matrix with the property of $\mathbf{B}\mathbf{B}^H = \mathbf{B}^H\mathbf{B} = N_T\mathbf{I}$. Combining Eqs. (15) and (16), we can get $\mathbf{R}_F = \mathbf{B}\text{diag}\{\mathbf{r}\}\mathbf{B}^H$. If we normalize matrix $\mathbf{B}$, this formula can be regarded as the eigenvalue decomposition (EVD) of $\mathbf{R}_F$, and according to the positive definiteness of $\mathbf{R}_F$, the elements of $\mathbf{r}$ should be non-negative real numbers. We can also find that the diagonal elements of $\mathbf{R}_F$ are all the same, which are equal to the $l_1$-norm of $\mathbf{r}$, thus we have $||\mathbf{r}||_1 = P_T - P_1$. And this can ensure the constant power of each transmit antenna. In the following, we are going to formulate the problem in off-grid and on-grid two cases.

### A. Off-grid case

We first consider the general situation, that is, the $L$ discrete angular bins may not locate on the grid point of $v \in \mathcal{V}$.

Combined with Eqs. (12) and (16), the optimization of the total transmit beampattern based on the predefined beampatterns of all orthogonal signals can be described as,

$$\begin{aligned} \max_{\mathbf{r},\delta} \quad & \delta, \\ \text{s.t.} \quad & \mathbf{a}_T^T(\theta_i)\mathbf{B}(\text{diag}(\mathbf{r}))\mathbf{B}^H\mathbf{a}_T(\theta_i)^* \geq \delta_i,\ \theta_i \in \mathbf{\Theta_k}, \\ & \delta = \min_{1\leq i\leq a_k}\{\delta_i\}, \\ & ||\mathbf{r}||_1 = P_T - P_1;\ r_i \geq 0, i = 1, ..., N_T. \end{aligned} \tag{17}$$

Obviously, Eq. (17) is a convex optimization problem, which can be solved directly via off-the-shelf toolbox, such as CVX. Compared with Eq. (12), we transform the problem from the complex domain to the real domain, and transform the variable from a $N_T \times N_T$-dimensional complex matrix $\mathbf{W}$ into a $N_T$-dimensional real vector $\mathbf{r}$, which greatly reduces the computational complexity.

### B. On-grid case

In the special case where all the discrete angular bins are on the grid points, which means $d_T \text{sin}\theta_l/\lambda \in \mathcal{V},\ l = 1, ..., L$. For the $k$th training, we denote the $v$ coordinates of the $a_k$ selected angular bins by $\{v_{k_i}|i = 1, ..., a_k; v_{k_i} \in \mathcal{V}\}$ where $k_i$ represents the $i$th angular bin selected in the $k$th training. The beampattern of the $i$th selected angular bin is given by,

$$B_F(\theta_i) = \mathbf{b}_{k_i}^H\mathbf{B}\text{diag}\{\mathbf{r}\}\mathbf{B}^H\mathbf{b}_{k_i} = N_T^2 r_{k_i}, \tag{18}$$

where $r_{k_i}$ is the $k_i$th element of $\mathbf{r}$. Substituting Eq. (18), we can get the solution of Eq. (17) is equal to,

$$\sum_{i=1}^{a_k} r_{k_i} = P_T - P_1,\ r_{k_i} = \delta_i,\ 1 \leq i \leq a_k. \tag{19}$$

Since $\delta_i \in \{\delta, 5\delta\}$ is preseted, the above problem is a linear equation and can be solved directly.

After obtaining the optimal $\mathbf{r}_o$, the corresponding $\mathbf{R}_W$ can be calculated. By taking EVD to $\mathbf{R}_W$, we can finally get the weight matrix $\mathbf{W}$. Here we provide one representation of $\mathbf{W}$,

$$\mathbf{W}_o = \frac{1}{\sqrt{N_T}}\mathbf{B}\text{diag}\{(\mathbf{r}_o + \frac{P_1}{N_T})^{1/2}\}\mathbf{B}^H. \tag{20}$$

## V. Simulation

In this section, simulation results are provided to verify the effectiveness of the improved RL-based algorithm.

In the first experiment, we compare the $P_D$ of the targets using the improved RL-based algorithm with some existing methods. We consider a harsh environment where the disturbance is generated by the underlying circular model, SOS AR(6), as presented in [6], [22]. An MMIMO radar is consisting of $N_T = 100$ transmit antennas and $N_R = 100$ receive antennas. And the number of virtual antennas is $N = N_R N_T = 10^4$. There are four targets distributed in $v = [-0.4, -0.05, 0.2, 0.45]$ with signal-to-noise ratios (SNR) of $[-5, -5, -15, -15]$dB. The first two targets can be regarded as strong targets, and the latter two are weak targets. The false alarm probability is $P_{\text{FA}} = 10^{-4}$ and $P_T = 0.1(watt)$. For target detection, we calculate the $P_D$ after 100 trainings using the improved RL-based algorithm, RL-based algorithm in [7] and omni-directional method. Then we run 1000 Monte-Carlo simulations and take the average over all runs. The results are shown in Fig. 1. As can be seen from Figs. 1(a) and 1(b), for the improved RL-based algorithm, the detection performance of strong targets is slightly lower than that of the RL algorithm in [7] when the number of antennas is small, as more transmit power is allocated to weak targets. However, from Figs. 1(c) and 1(d), we can see our improved algorithm can significantly enhance the detection performance of weak targets.

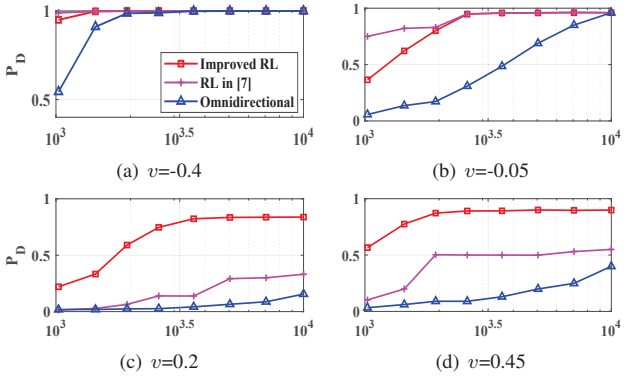


Fig. 1. The $P_D$ versus $N = N_R N_T$ curves for different targets.

In the second experiment, we verify the improved algorithm in a dynamic environment. We consider two scenarios. Scenario 1 is the same as the previous experiment. For scenario 2, there are four targets distributed in $v = [-0.3, -0.2, 0, 0.3]$ with SNRs of $[-5, -5, -5, -15]$dB. Applying the improved RL algorithm and the approach in [7] separately, we conduct 120 trainings totally and after the 60th training, the environment changes from scenario 1 to scenario 2. Similarly, we run 1000 Monte-Carlo simulations and take the average over all runs. The results are given in Fig. 2. We can see that compared with [7], the improved RL algorithm can significantly improve the detection ability to the weak targets in a dynamic environment and requires less training time compared to the method in [7].

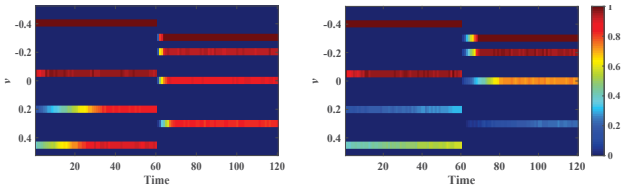


Fig. 2. Detection performance of the (a) improved RL-based algorithm, (b) RL based approach in [7], in dynamic environment.

## VI. Conclusion

In this paper, we proposed an improved RL based method capitalizing on the previous work in [7]. An omni-directional component of the transmit power was added to reduce the missed detection of the angular bins containing a weak target. Then the detection probability of weak targets was further improved by focusing more transmit power on them and changing the reward mechanism of RL. In addition, a fast optimization method to design the transmit weight matrix was proposed to greatly reduce the computational complexity and training time. The simulation results showed the proposed improved RL based method significantly improve the detection ability of weak targets.

## REFERENCES


[1] S. Haykin, "Cognitive radar: a way of the future," in *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 30-40, 2006, doi: 10.1109/MSP.2006.1593335.

[2] S. Haykin, "Cognitive Dynamic Systems," in *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1910-1911, 2006, doi: 10.1109/JPROC.2006.886014.

[3] W. Zhai, X. Wang, S. A. Hamza and M. G. Amin, "Cognitive-Driven Optimization of Sparse Array Transceiver for MIMO Radar Beamforming," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1-6, doi: 10.1109/RadarConf2147009.2021.9455310.

[4] K. J. Sohn, H. Li and B. Himed, "Parametric GLRT for Multichannel Adaptive Signal Detection," in *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5351-5360, 2007, doi: 10.1109/TSP.2007.896068.

[5] A. L. Swindlehurst and P. Stoica, "Maximum likelihood methods in radar array signal processing," in *Proceedings of the IEEE*, vol. 86, no. 2, pp. 421-441, 1998, doi: 10.1109/5.659495.

[6] S. Fortunati, L. Sanguinetti, F. Gini, M. S. Greco and B. Himed, "Massive MIMO Radar for Target Detection," in *IEEE Transactions on Signal Processing*, vol. 68, pp. 859-871, 2020, doi: 10.1109/TSP.2020.2967181.

[7] 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," in *IEEE Transactions on Aerospace and Electronic Systems*, doi: 10.1109/TAES.2021.3061809.

[8] 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

[9] K. L. Bell, C. J. Baker, G. E. Smith, J. T. Johnson and M. Rangaswamy, "Cognitive Radar Framework for Target Detection and Tracking," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 8, pp. 1427-1439, 2015, doi: 10.1109/JSTSP.2015.2465304.

[10] S. Haykin, Y. Xue and P. Setoodeh, "Cognitive Radar: Step Toward Bridging the Gap Between Neuroscience and Engineering," in *Proceedings of the IEEE*, vol. 100, no. 11, pp. 3102-3130, 2012, doi: 10.1109/JPROC.2012.2203089.

[11] E. Selvi, R. M. Buehrer, A. Martone and K. Sherbondy, "Reinforcement Learning for Adaptable Bandwidth Tracking Radars," in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 5, pp. 3904-3921, 2020, doi: 10.1109/TAES.2020.2987443.

[12] A. Guerra, F. Guidi, D. Dardari, and P. M. Djuric, "Reinforcement learning for uav autonomous navigation, mapping and target detection," in *CoRR,* 2020. [Online]. Available: http://arxiv.org/abs/2005.05057

[13] E. Selvi, R. M. Buehrer, A. Martone and K. Sherbondy, "On the use of Markov Decision Processes in cognitive radar: An application to target tracking," in *2018 IEEE Radar Conference (RadarConf18)*, 2018, pp. 0537-0542, doi: 10.1109/RADAR.2018.8378616.

[14] M. S. Greco, F. Gini, P. Stinco and K. Bell, "Cognitive Radars: On the Road to Reality: Progress Thus Far and Possibilities for the Future," in *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 112-125, 2018, doi: 10.1109/MSP.2018.2822847.

[15] O. Aldayel, V. Monga and M. Rangaswamy, "Successive QCQP Refinement for MIMO Radar Waveform Design Under Practical Constraints," in *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3760-3774, 2016, doi: 10.1109/TSP.2016.2552501.

[16] X. Yu, K. Alhujaili, G. Cui and V. Monga, "MIMO Radar Waveform Design in the Presence of Multiple Targets and Practical Constraints," in *IEEE Transactions on Signal Processing*, vol. 68, pp. 1974-1989, 2020, doi: 10.1109/TSP.2020.2979602.

[17] Z. Cheng, Z. He, S. Zhang and J. Li, "Constant Modulus Waveform Design for MIMO Radar Transmit Beampattern," in *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4912-4923, 2017, doi: 10.1109/TSP.2017.2718976.

[18] B. Friedlander, "On Transmit Beamforming for MIMO Radar," in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3376-3388, 2012, doi: 10.1109/TAES.2012.6324717.

[19] G. Hua and S. S. Abeysekera, "MIMO Radar Transmit Beampattern Design With Ripple and Transition Band Control," in *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2963-2974, 2013, doi: 10.1109/TSP.2013.2252173.

[20] Y. Wang, G. Huang and W. Li, "Transmit Beampattern Design in Range and Angle Domains for MIMO Frequency Diverse Array Radar," in *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 1003-1006, 2017, doi: 10.1109/LAWP.2016.2616193.

[21] B. Friedlander, "On Signal Models for MIMO Radar," in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3655-3660, 2012, doi: 10.1109/TAES.2012.6324753.

[22] S. Fortunati, F. Gini and M. S. Greco, "The Misspecified Cramer-Rao Bound and Its Application to Scatter Matrix Estimation in Complex Elliptically Symmetric Distributions," in *IEEE Transactions on Signal Processing*, vol. 64, no. 9, pp. 2387-2399, 2016, doi: 10.1109/TSP.2016.2526961.