Fundamental Limits for ISAC - Asymptotics in massive MIMO sensing systems

Stefano Fortunati, Francesco Lisi, Aya Mostafa Ibrahim Ahmed, Aydin Sezgin, Maria Sabrina Greco, Fulvio Gini

Fundamental Limits for ISAC - Asymptotics in massive MIMO sensing systems

Stefano Fortunati and Francesco Lisi and Aya Mostafa Ibrahim Ahmed and Aydin Sezgin and Maria Sabrina Greco and Fulvio Gini

Abstract Asymptotic analysis is a common tool in statistics aiming at investigating the properties of an inference methodology as the number of observations grows to infinity. Even if the asymptotic regime cannot be achieved in real-world scenarios, its practical usefulness has been proved in an uncountable number of engineering applications. In the contest of ISAC, one of the brightest example is the Massive Multiple-Input-Multiple-Output (MMIMO) communications framework. The breakthrough brought by the MMIMO systems was in showing that, as the number of antenna elements grows to infinity, linear combining and precoding algorithms can mitigate the interference even in the presence of a partial knowledge of the communication channel. Inspired by this fundamental result, in this chapter we show that the massive (asymptotic) paradigm can bring essential benefits also in radar systems. In particular, we considered a co-located MIMO radar having a massive number of virtual spatial antenna channels. We focus on the target detection problem by showing that the massive regime allows for the derivation of a cognitive, robust, reinforcement learning (RL)-based, Wald-type test that guarantees certain performance regardless of the unknown statistical characterization of the disturbance. As

Stefano Fortunati
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France and DR2I-IPSA, 94200, Ivry sur Seine, France, e-mail: stefano.fortunati@centralesupelec.fr

Francesco Lisi
Università di Pisa e-mail: francesco.lisi@phd.unipi.it

Aya Mostafa Ibrahim Ahmed
Ruhr-Universität Bochum e-mail: aya.mostafaibrahimahmad@rub.de

Aydin Sezgin
Ruhr-Universität Bochum e-mail: Aydin.Sezgin@rub.de

Maria Sabrina Greco
Università di Pisa e-mail: maria.greco@unipi.it

Fulvio Gini
Università di Pisa e-mail: fulvio.gini@unipi.it

concluding remarks, some explorative idea on a massive integrated communication/sensing system will be provided.

1 Introduction

A key feature enabled by sixth generation (6G) systems is the integration of sensing within the communication capabilities. It is foreseen that 6G will empower novel sensing applications as in the *smart cities* paradigm, which includes intelligent vehicular networks, robot and drone tracking, enhanced emergency call localization, personal radar and location aware communications in general [29]. To support this huge variety of potential applications, an innovative *integrated sensing and communication (ISAC)* strategy has to be put in place. The simplest form of ISAC is cohabitation, where both systems share the same resources (i.e., time, frequency, space) while interfering on each other [22]. Normally, in this case each system develops a co-existence strategy of resource sharing which can be non-cooperative or cooperative for example [22, 19, 20, 12]. A more advanced form of ISAC is the co-design, which implies the implementation of a fully integrated system with fully shared transmitters, largely shared receivers and joint transmitted waveforms [13, 14]. 6G systems provide the perfect framework to realize ISAC. In fact, thanks to the high frequency and large bandwidth offered within millimeter wave (mmW) and sub-Terahertz (THz) bands, it is suitable to increase the users data rate from the communication side while improving the range and Doppler resolution from a remote sensing side. Furthermore, 6G may become a fast catalyst for the realization of massive multiple input multiple output (MMIMO) systems due to the small antenna size. As a matter of fact, current 5G commercial solutions can even realize up to 64 fully digital transceivers[6]. It is envisioned that 6G systems will be capable of employing much more antennas, in the order of hundreds, within the footprint of few square centimeters [3].

Let us focus now on the strict link between the ISAC goal and the MMIMO paradigm. The concept of MMIMO was first introduced in [21] as a solution to the shortcomings of conventional multi-user MIMO. From a communication perspective, the MMIMO system can bring huge advantages in terms of spectral efficiency, since the latter increases monotonically with the number of antennas. Moreover, it has been shown that MMIMO can also mitigate spatial interference even in the presence of imperfect channel state information (CSI) [6]. At the same time, from the sensing perspective, large scale antenna arrays offer a significant spatial resolution capability. Specifically, a MMIMO system can steer a pencil-like beam towards a direction of interest, hence improving target detection and related parameter estimation. However, we can go far beyond that in term of advantages for remote sensing tasks. For example, the potential benefits that large virtual spatial channels can bring into the radar detection problem has been recently investigated in [7]. Specifically, it has been proved that a massive number of virtual spatially antenna channels N allows for the derivation of robust detection algorithms, which in turn guarantee certain sta-

tistical performance regardless the, generally unknown, disturbance model. More formally, in [7], it has been shown that the constant false alarm property (CFAR) of a derived Wald-type detector can be achieved under a wide variety of disturbance (i.e., clutter plus noise) statistical models as $N \to \infty$. Along with the CFAR property, the optimization of the Probability of Detection (P_D) is a crucial task for a sensing system and it can be fulfilled by means of a suitable waveform design strategy [11]. Moreover, in the ISAC framework, the selected waveforms have to fulfill, *at the same time*, the maximization of both the sensing and communication metrics in an environment whose statistical characterization is generally unknown and variable with time. The complexity of this kind of scenarios has motivated the integration of artificial intelligence (AI), machine and reinforcement learning (ML and RL) techniques in more classical statistical-oriented signal processing frameworks [4].

Motivated by the evidence that a cross-fertilization between classical asymptotic statistics and learning approaches may bring a breakthrough in the unsolved ISAC challenges, for instance mutual interference, this chapter describes the successful exploitation of this hybrid cognitive approach in the MMIMO sensing context. In particular, building upon recent works [1, 18], we show that an RL based algorithm for waveform selection can be fruitfully combined with outcomes on robust asymptotic statistics to develop a fully adaptive and data-driven detection algorithm for MMIMO radar systems that may achieve good performance without relying on any a priori knowledge of the environment. More specifically, in Sec. 2 the signal and disturbance models of the MMIMO system will be presented along with the robust Wald-type test. As we will see, asymptotic statistics provide us with the tools needed to investigate the statistical properties of the derived test as the number of virtual antenna goes to infinity. In order to make the chapter as self-contained as possible, Sec. 3 collects the main concepts about Reinforcement Learning, Markov Decision Problem and related learning algorithms. The core of this chapter is Sec. 4, where the RL methodologies, previously introduced, will be applied to the MMIMO detection problem in order to derive an adaptive RL-based waveform selection algorithm aiming at maximizing the P_D of the system. Finally, Sec. 5 concludes the chapter with some considerations about the extension of the proposed joint asymptotic statistics-RL methodologies to the general ISAC framework of MMIMO communication-sensing systems.

2 Massive MIMO radar system description

This section is dedicated to a detailed description of a massive MIMO (MMIMO) radar system and to the presentation of the main advantages over classical MIMO system. Specifically, in Sec. 2.1 a general signal model for a co-located MIMO radar is presented. Then, Sec. 2.2 describes a robust Wald-type test for radar detection that satisfies the constant false alarm rate (CFAR) property when the radar operates in the MMIMO regime.

2.1 Signal model

Suppose that a point-like target, with angular position θ_0, is within the radar's detection range, the baseband signal at the receiving array can be modelled as [8], [9]

$$\boldsymbol{y}(t) = \alpha \boldsymbol{a}_R(\theta_0)\boldsymbol{a}_T^T(\theta_0)\boldsymbol{s}(t-\tau)e^{j\omega_d t} + \boldsymbol{n}(t), \quad t \in [0, T) \tag{1}$$

where $\boldsymbol{y}(t) \in \mathbb{C}^{N_R}$ is the receiving array output vector, $\boldsymbol{s}(t) \in \mathbb{C}^{N_T}$ is the vector of transmitted signals and $\boldsymbol{n}(t) \in \mathbb{C}^{N_R}$ is the disturbance at the receiver. The scalar $\alpha \in \mathbb{C}$ is a parameter that accounts for the two-way path loss and the radar cross section, which is supposed to be the same for all the transmitter-receiver pairs since we are dealing with a co-located radar system, $\tau \in \mathbb{R}$ and $\omega_d \in \mathbb{R}$ are the delay and doppler shift of the target. Let us assume an Uniformly Linear Array (ULA) configuration for both the transmitter and the receiver, so that $\boldsymbol{a}_T(\theta_0) = [1, e^{j2\pi\nu}, \ldots, e^{j2\pi(N_T-1)\nu}]^T \in \mathbb{C}^{N_T}$ and $\boldsymbol{a}_R(\theta_0) = [1, e^{j2\pi\nu}, \ldots, e^{j2\pi(N_R-1)\nu}]^T \in \mathbb{C}^{N_R}$ are the transmit and receive steering vectors, with $\nu \triangleq \frac{d}{\lambda}\sin(\theta_0)$. In the rest of the chapter, d is always assumed equal to $\lambda/2$. The transmitted signals vector $\boldsymbol{s}(t)$ can be expressed as:

$$\boldsymbol{s}(t) = \boldsymbol{W}\boldsymbol{\Phi}(t), \tag{2}$$

where $\boldsymbol{W} \in \mathbb{C}^{N_T \times N_T}$ is the *weighting matrix* used to shape the transmit beampattern as better detailed below. The term $\boldsymbol{\Phi}(t) \in \mathbb{C}^{N_T}$ is a vector of quasi-orthonormal waveforms, i.e.

$$\int_0^T \Phi_i(t-\tau')\Phi_j^*(t-\tau)dt \cong \delta_{ij}, \ \forall \tau', \tau. \tag{3}$$

where δ_{ij} is the Kronecker delta. Note that this orthogonality condition is assumed to be true for all the possible delays τ', τ. This stringent requirement will be removed in Sec. 2.2. At the receiver the signal is processed as:

$$\begin{aligned}
\boldsymbol{Y}(n,h) &\triangleq \int_0^T \boldsymbol{y}(t)\boldsymbol{\Phi}^H(t-n\Delta t)e^{-jh\Delta\omega t}dt && (4)\\
&= \int_0^T \left(\alpha \boldsymbol{a}_R(\theta_0)\boldsymbol{a}_T^T(\theta_0)\boldsymbol{W}\boldsymbol{\Phi}(t-\tau)e^{j\omega_d t} + \boldsymbol{n}(t)\right)\boldsymbol{\Phi}^H(t-n\Delta t)e^{-jh\Delta\omega t}dt \\
&= \alpha \boldsymbol{a}_R(\theta_0)\boldsymbol{a}_T^T(\theta_0)\boldsymbol{W}\boldsymbol{R}_{\boldsymbol{\Phi}}(n,h) + \boldsymbol{C}(n,h), && (5)
\end{aligned}$$

with $n = 1, \ldots, L_\tau$ and $h = 1, \ldots, L_\omega$, where:

$$\boldsymbol{R}_{\boldsymbol{\Phi}}(n,h) \triangleq \int_0^T \boldsymbol{\Phi}(t-\tau)\boldsymbol{\Phi}^H(t-n\Delta t)e^{j(\omega_d - h\Delta\omega)t}dt, \tag{6}$$

$$\boldsymbol{C}(n,h) \triangleq \int_0^T \boldsymbol{n}(t)\boldsymbol{\Phi}^H(t-n\Delta t)e^{-jh\Delta\omega t}dt. \tag{7}$$

Recasting the problem in vectorial form we have:

$$\boldsymbol{y}(n,h) \triangleq \mathrm{vec}\big(\boldsymbol{Y}(n,h)\big) = \alpha \boldsymbol{h}(n,h) + \boldsymbol{c}(n,h) \in \mathbb{C}^{N\times 1} \tag{8}$$

with $N \triangleq N_T N_R$ and

$$\boldsymbol{h}(n,h) \triangleq \big(\boldsymbol{R}_{\boldsymbol{\Phi}}^T(n,h) \otimes \boldsymbol{I}_{N_R}\big)\big(\boldsymbol{W}^T \boldsymbol{a}_T(\theta_0) \otimes \boldsymbol{a}_R(\theta_0)\big) \tag{9}$$

where $\boldsymbol{c}(n,h) \triangleq \mathrm{vec}\big(\boldsymbol{C}(n,h)\big)$. Note that, in radar terminology, the measurement vector $\boldsymbol{y}(n,h)$ is usually called *snapshot*. Under the hypothesis that $\boldsymbol{n}(t)$ is a zero-mean wide sense stationary process, i.e $\mathbb{E}\{\boldsymbol{n}(t)\} = \boldsymbol{0}$ and $\mathbb{E}\left\{\boldsymbol{n}(t_1)\boldsymbol{n}(t_2)^H\right\} = \boldsymbol{\Sigma}(t_1 - t_2)$, then $\mathbb{E}\{\boldsymbol{c}(n,h)\} = \boldsymbol{0}$ and consequently, as shown in [7]:

$$\boldsymbol{\Gamma}(n,h) \triangleq E\{\boldsymbol{c}(n,h)\boldsymbol{c}^H(n,h)\} \tag{10}$$

$$= \int_0^T \int_0^T [\boldsymbol{\Phi}^*(t_1 - n\Delta t)\boldsymbol{\Phi}^T(t_2 - n\Delta t) \otimes \boldsymbol{\Sigma}(t_1 - t_2)] e^{-jh\Delta\omega(t_1 - t_2)} dt_1 dt_2. \tag{11}$$

In the following sections, we assume that the samples $\bar{n}$ and $\bar{h}$ are chosen in such a way that $\bar{n}\Delta t = \tau$ and $\bar{h}\Delta\omega = \omega_d$, so using (3), (9) becomes

$$\boldsymbol{h}(\bar{n},\bar{h}) = \boldsymbol{W}^T \boldsymbol{a}_T(\theta_0) \otimes \boldsymbol{a}_R(\theta_0) \equiv \boldsymbol{h}, \tag{12}$$

and the indexes are omitted for ease of notation. It is worth underlining that most of the existing works assume that the noise samples are uncorrelated in both spatial and temporal domains [16], i.e. $\boldsymbol{\Gamma}(n,h) \equiv \boldsymbol{\Gamma} = \sigma^2 \boldsymbol{I}$. As we will extensively discuss in the next subsection, in [7] this assumption has been greatly relaxed to take into account a huge variety of realistic disturbance statistics.
To conclude, let's recall the *transmitter beampattern* of a MIMO radar defined as [10]:

$$BP(\theta) = \boldsymbol{a}_T^T(\theta) \boldsymbol{R}_{\boldsymbol{W}} \boldsymbol{a}_T^*(\theta), \tag{13}$$

where $\boldsymbol{R}_{\boldsymbol{W}} \triangleq \boldsymbol{W}\boldsymbol{W}^H$ and $\boldsymbol{W}$ is he weighting matrix defined in eq. (2).

The beampattern corresponds to the normalised power density [10] of the electromagnetic field under the hypothesis of ideal isotropically radiating elements, thus we can compute the transmitted power as $P_T \triangleq tr\{\boldsymbol{R}_{\boldsymbol{W}}\}$.

Eq. (13) shows that by choosing the *weighting matrix* $\boldsymbol{W}$, and consequently the *correlation matrix* $\boldsymbol{R}_{\boldsymbol{W}}$, one can shape the transmitting beampattern. Thus the choice of the transmitted waveforms plays a crucial role in the design of the MIMO radar systems as amply discussed in [11, Ch. 4] and [9], among many others. Just to cite the limit cases, an omnidirectional beampattern can be obtained by selecting orthogonal waveforms, i.e. $\boldsymbol{W} = \boldsymbol{W}_{ort} \triangleq \sqrt{\frac{P_T}{N_T}} \boldsymbol{I}_{N_T}$. On the contrary, if we want to focus all the transmitter power in the direction $\bar{\theta}$, we should select $\boldsymbol{W} = \sqrt{\frac{P_T}{N_T}} \boldsymbol{a}_T^*(\bar{\theta}) \boldsymbol{a}_T^T(\bar{\theta})$.

2.2 A robust Wald-type test for target detection

After having introduced the signal model for a MMIMO radar system, we will now focus our attention on the target detection problem. Remarkably, as we will see in the following, the massive regime allows us to overcome most of the unrealistic assumptions that are generally made in the radar detection literature.

Starting from the characterization of the received signal given in eq. (8), the *single-snapshot* target detection problem for MMIMO system can be cast as:

$$\mathcal{H}_0 : \boldsymbol{y} = \boldsymbol{c} \tag{14}$$

$$\mathcal{H}_1 : \boldsymbol{y} = \alpha \boldsymbol{h} + \boldsymbol{c}. \tag{15}$$

A huge amount of literature can be found on the implementation of a decision statistic, i.e. a *detector*, to discriminate between the null hypothesis $\mathcal{H}_0$ (target absence) against the alternative $\mathcal{H}_1$ (target present). However, most of the existing works rely on the following simplifying assumptions (see, among many others [5], [17] and [16, Ch. 4] for the MIMO case.):

- availability of a sufficiently large number of *independent* and *identically distributed* (i.i.d.) measurement vectors (or snapshots),
- the disturbance is assumed to be Gaussian distributed with diagonal (or block-diagonal) covariance matrices,
- the waveform matrix $\boldsymbol{\Phi}(t)$ is assumed to be perfectly orthogonal.

These three assumptions are rather unrealistic and generally violated in practice [8]. For this reason, by exploiting the advantages resulting from a massive number of virtual antenna channels, in [7] it was proposed to overcome the previous two simplified assumptions by considering a *single snapshot* scenario and by imposing only the following:

Assumption A1

The disturbance $\boldsymbol{c}$ is a realization of a discrete-time, complex circular and possibly non-Gaussian random process $\{c_n, \forall n\}$ with a polynomial decay of its autocorrelation function, that is $r_c[m] \triangleq \mathbb{E}\{c_n c^*_{n-m}\} = O\big(|m|^{-\gamma}\big), m \in \mathbb{Z}, \gamma > \varrho/(\varrho-1), \varrho > 1.$

Note that Assumption A1 is weak enough to be satisfied by, for example, all the Gaussian and non-Gaussian Autoregressive-Moving Average (ARMA) models of arbitrary order and by all the Compound Gaussian models [7]. Remarkably, under Assumption A1, the target detection problem in (14) can be solved by a Wald-type test without the need of any a priori knowledge on the statistical distribution of the disturbance vector $\boldsymbol{c}$. Briefly, to build this Wald-type statistic, only an asymptotically normal, $\sqrt{N}$-consistent estimator of α, together with its asymptotic error covariance, are required. Under A1, the estimator is simply given by the linear least square (LLS) estimator [7]:

$$\hat{\alpha} = \boldsymbol{h}^H \boldsymbol{y} / ||\boldsymbol{h}||^2, \tag{16}$$

whose asymptotic error covariance is $\mathbb{E}\{|\hat{\alpha}-\alpha|^2\} = \frac{\boldsymbol{h}^H \boldsymbol{\Gamma} \boldsymbol{h}}{||\boldsymbol{h}||^4}$ where $\boldsymbol{\Gamma}$ is the disturbance covariance matrix defined in eq. (10). Then, a Wald-test statistic can be derived as:

$$\Lambda(\boldsymbol{y}) = \frac{2|\boldsymbol{h}^H \boldsymbol{y}|^2}{\boldsymbol{h}^H \hat{\boldsymbol{\Gamma}} \boldsymbol{h}}, \tag{17}$$

where $\hat{\boldsymbol{\Gamma}}$ is the following $\sqrt{N}$-consistent estimate of $\boldsymbol{\Gamma}$:

$$[\hat{\boldsymbol{\Gamma}}]_{i,j} = \begin{cases} \hat{c}_i \hat{c}_j^*, & |i-j| \leq l, \\ 0, & |i-j| > l, \end{cases} \tag{18}$$

where $\hat{\boldsymbol{c}} = \boldsymbol{y} - \hat{\alpha}\boldsymbol{h}$ and l is the truncation lag that must grow with N, but more slowly than $N^{1/3}$. Further considerations on the choice of the truncation lag l can be found in [27, Theorem 6.20]. Moreover, [7, Theorem 3], states that the derived Wald-test statistic is asymptotically distributed as:

$$\Lambda(\boldsymbol{y}|\mathcal{H}_0) \underset{N\to\infty}{\sim} \chi_2^2(0), \tag{19}$$

$$\Lambda(\boldsymbol{y}|\mathcal{H}_1) \underset{N\to\infty}{\sim}, \chi_2^2(\varsigma) \tag{20}$$

with

$$\varsigma \triangleq 2|\alpha|^2 \frac{||\boldsymbol{h}||^4}{\boldsymbol{h}^H \boldsymbol{\Gamma} \boldsymbol{h}}. \tag{21}$$

To discriminate between the null hypothesis $\mathcal{H}_0$ and its alternative $\mathcal{H}_1$, the value of the decision statistic is compared with a threshold:

$$\Lambda(\boldsymbol{y}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda. \tag{22}$$

It is immediate to verify that, from the Wald's test statistic asymptotic distributions, the closed form asymptotic expressions of the P_{FA} and P_D are [7]:

$$P_{FA}(\lambda) = \Pr\{\Lambda(\boldsymbol{y}) \geq \lambda | \mathcal{H}_0\} = \int_{\lambda}^{+\infty} f_{\Lambda|\mathcal{H}_0}(a) da \underset{N\to\infty}{\to} e^{-\frac{\lambda}{2}}, \tag{23}$$

$$P_D(\lambda) = \Pr\{\Lambda(\boldsymbol{y}) \geq \lambda | \mathcal{H}_1\} = \int_{\lambda}^{+\infty} f_{\Lambda|\mathcal{H}_1}(a) da \underset{N\to\infty}{\to} Q_1(\sqrt{\varsigma}, \sqrt{\lambda}), \tag{24}$$

where $Q_1(\cdot,\cdot)$ is the Marcum Q function of order 1 defined as $Q_1(a,b) \triangleq \int_b^{+\infty} x \exp\left(-(x^2+a^2)/2\right) I_0(ax) dx$, a and b are nonnegative real numbers and $I_0(\cdot)$ is the modified Bessel function of the first kind of zero order.

From the asymptotic expression of $P_{FA}(\lambda)$, the Constant False Alarm Rate (CFAR) property follows immediately. In fact, under Assumption A1 and by inverting the eq. (23), one can determine the required threshold $\bar{\lambda}$ for any desired nominal $\bar{P}_{FA}$ as:

$$\bar{\lambda} = -2\ln\left(\bar{P}_{FA}\right). \tag{25}$$

To summarize, it is worth stressing here that the three biggest advantages of the proposed Wald-type detector are:

- it needs only a single snapshot to extract all the information required to discriminate between $\mathcal{H}_0$ (target absence) and $\mathcal{H}_1$ (target present),
- it satisfies the closed form expression of P_{FA} and P_D, provided in eq. (23) and eq. (24), under any disturbance scenario satisfying Assumption A1, i.e. it is *statistically robust*.
- it satisfies the CFAR property in the *massive MIMO regime*.

We note, in passing, that the previous results about the Wald-type detector hold true for any array geometry, since we are not using any specific structure for the related steering vector.

The last practical question that needs to be addressed is related to the "asymptotic MMIMO regime": how large should the number of virtual spatial antenna channels N be to consider the asymptotic results a good approximation for their finite-sample counterparts? In [7], a preliminary and simulation-based answer has been given. Specifically, it has been shown that the *massive MIMO regime* is achieved if $N \geq 10^4$ for a nominal $\bar{P}_{FA}$ of 10^{-4}. Providing a more general and rigorous answer to this crucial aspect is still an open problem.

Let us now focus our attention on the *power* of the test statistic $\Lambda(\boldsymbol{y})$ in eq. (17), i.e. on its detection performance. In fact, even if the proposed Wald-type detector presents the desirable robustness and CFAR properties over a very large set of disturbance model, nothing can be said about the optimality of the achieved P_D. In particular, according to the closed form expression in (24) and to the fact that $Q_1(\cdot,\cdot)$ is strictly increasing in the first argument, to maximize the P_D we should maximize ς with respect to the only free parameter that is the weighting matrix $\boldsymbol{W}$. The analysis and a possible solution of this optimization problem will be presented in Sec. 4.

We conclude this section with some clarification on the practical radar implementation of target detection problem in (14) and by introducing the useful notation that allows us to integrate the robust Wald-type test in (17) with the Reinforcement Learning-based procedure for the optimal selection of the weighting matrix $\boldsymbol{W}$ presented ahead.

Let us denote with $k \in \mathbb{N}$ the time index corresponding to each transmitted pulse provided in (2). At time instant k, the radar computes the Wald-test statistic in (17) in a predefined fixed grid of angular bins $\Theta = \{\theta_l\}_{l=1}^{L}$, i.e.

$$\Lambda_{k,l} = \frac{2|\boldsymbol{h}_{k-1,l}^H \cdot \boldsymbol{y}_{k,l}|^2}{\boldsymbol{h}_{k-1,l}^H \hat{\boldsymbol{\Gamma}}_{k,l} \boldsymbol{h}_{k-1,l}} \tag{26}$$

with [1]

$$\boldsymbol{h}_{k-1,l} = \boldsymbol{W}_{k-1}^{T} \boldsymbol{a}_T(\theta_l) \otimes \boldsymbol{a}_R(\theta_l), \tag{27}$$

$$[\hat{\boldsymbol{\Gamma}}_{k,l}]_{i,j} = \begin{cases} [\hat{\boldsymbol{c}}_{k,l}]_i [\hat{\boldsymbol{c}}_{k,l}]_j^*, & |i-j| \leq l \\ 0, & |i-j| > l \end{cases}, \tag{28}$$

$$\hat{\boldsymbol{c}}_{k,l} = \boldsymbol{y}_{k,l} - \hat{\alpha}_{k-1,l} \boldsymbol{h}_{k-1,l}, \quad \hat{\alpha}_{k-1,l} = \frac{\boldsymbol{h}_{k-1,l}^H \boldsymbol{y}_{k,l}}{||\boldsymbol{h}_{k-1,l}||^2}. \tag{29}$$

After performing the detection in each angular bin, the radar system should select the weighting matrix $\boldsymbol{W}_k$ to maximize the P_D, starting from some sort of knowledge of the environment. As we will show in the following sections, the weighting matrix selection problem can be perfectly cast in the *reinforcement learning* (RL) framework.

3 Introduction to Reinforcement Learning

As a prerequisite for Sec. 4 and in order to make this chapter self-contained, in the following, a brief introduction to the basic ideas behind reinforcement learning techniques is provided. The interested reader may find exhaustive and in-depth discussions in [24], [25], and [23], just to cite a few.

3.1 Markov Decision Process (MDP)

A Markov Decision Process is a mathematical tool to describe the interaction between a system (in our case, the radar system) and the surrounding environment. In the MDP literature the system that learns and takes decision is referred to as the *agent*, while all the other entities, external to the agent and with whom it interacts, are called *environment*. At each time instant k the agent is in state $s_k \in \mathcal{S}$ and selects an action $a_k \in \mathcal{A}$. The chosen action corresponds to a stimulus for the environment, which responds by sending a reward $r_{k+1} \in \mathcal{R}$. In the meanwhile the agent reaches the new state $s_{k+1} \in \mathcal{S}$. The final objective of the agent is to maximise a *cumulative reward*.

A MDP is completely specified by the following quantities [25]:

- The state space $\mathcal{S}$,
- The action space $\mathcal{A}$,
- The set of possible rewards $\mathcal{R}$,
- The *dynamics* of the process defined as:

[1] Additional considerations of the index k will be done ahead in Sec. 4

$$p(s', r|s, a) \triangleq \Pr\{(s_{k+1} = s') \cap (r_{k+1} = r)|(s_k = s) \cap (a_k = a)\}. \tag{30}$$

We will focus our attention on the subset of *finite, discrete-time MDP with infinite horizon*. A MDP is called finite if the state space $\mathcal{S}$ and action space $\mathcal{A}$ have a finite number of elements. In a MDP with finite horizon the state space $\mathcal{S}$ contains at least one terminal state, so the temporal evolution of the system can be described as a succession of epochs that can have different duration. During each epoch the system evolves from an initial state to a terminal state. Such a description is useful in some applications where a terminal state can be easily identified, such as in the chess game. In an infinite horizon MDP a terminal state can't be identified and the agent continues to interact with the environment endlessly. This is the case for radar detection/weighting matrix selection problem previously introduced.

3.1.1 Policy and *value functions*

A policy is a mapping from each state $s \in \mathcal{S}$ to a probability distribution over the set $\mathcal{A}$. Thus the function $\pi(a|s) \triangleq \Pr\{a_k = a|s_k = s\}$ completely specifies a policy. A deterministic policy is a special case of the previous one, where

$$\pi(a|s) = \begin{cases} 1, & a = f_\pi(s) \\ 0, & a \in \mathcal{A} \setminus \{f_\pi(s)\} \end{cases} \tag{31}$$

then the policy is completely specified by the function $f_\pi : \mathcal{S} \to \mathcal{A}$. Among all the possible policy, the agent should choose the one that will allow it to achieve its final goal. To this end, we need to define an ordering over the set of policies. As we will explain ahead, this ordering will be induced by the *value functions*. Formally, to define the *value functions* for policy π we must introduce the *cumulative reward*

$$G_k \triangleq \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \tag{32}$$

where $\gamma \in [0, 1)$ is called *discount factor*.

The *state value function* for policy π is defined as

$$V_\pi(s) \triangleq \mathbb{E}_\pi \left\{ \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \middle| s_k = s \right\} \tag{33}$$

that corresponds to the expectation of the *cumulative reward* starting from state $s_k = s \in \mathcal{S}$ and following policy π. The expectation is computed over all the states $\{s_{k'} \in \mathcal{S}\}_{k'=k+1}^{+\infty}$, actions $\{a_{k'} \in \mathcal{A}\}_{k'=k}^{+\infty}$ and rewards $\{r_{k'} \in \mathcal{R}\}_{k'=k+1}^{+\infty}$, but we explicit only the dependence on π to simplify the notation.

Similarly the *state-action value function* for policy π is defined as

$$Q_\pi(s,a) \triangleq \mathbb{E}_\pi \left\{ \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \middle| (s_k = s) \cap (a_k = a) \right\}, \tag{34}$$

where the expectation is computed over $\{s_{k'} \in \mathcal{S}\}_{k'=k+1}^{+\infty}$, $\{a_{k'} \in \mathcal{A}\}_{k'=k+1}^{+\infty}$ and $\{r_{k'} \in \mathcal{R}\}_{k'=k+1}^{+\infty}$. The relationship between the *state value function* in (33) and the *state-action value function* in (34) is given by:

$$\begin{aligned} V_\pi(s) &\triangleq \mathbb{E}_\pi \left\{ \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \middle| s_k = s \right\} && (35)\\ &= \sum_{a\in\mathcal{A}} \mathbb{E}_\pi \left\{ \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \middle| (s_k = s) \cap (a_k = a) \right\} \Pr\{a_k = a | s_k = s\} && (36)\\ &= \sum_{a\in\mathcal{A}} Q_\pi(s,a)\pi(a|s) = \mathbb{E}_{a\in\mathcal{A}} \{Q_\pi(s,a)\}. && (37) \end{aligned}$$

From the *state value function* definition one can derive the following equation:

$$\begin{aligned} V_\pi(s) &\triangleq \mathbb{E}_\pi \left\{ \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \middle| s_k = s \right\} && (38)\\ &= \sum_{a\in\mathcal{A}} \mathbb{E}_\pi \left\{ \sum_{h=0}^{+\infty} \gamma^h r_{k+h+1} \middle| (s_k = s) \cap (a_k = a) \right\} \pi(a|s) && (39)\\ &= \sum_{a\in\mathcal{A}} \mathbb{E}_\pi \left\{ r_{k+1} + \gamma \sum_{h=0}^{+\infty} \gamma^h r_{k+h+2} \middle| (s_k = s) \cap (a_k = a) \right\} \pi(a|s) && (40)\\ &= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r|s, a) \left(r + \gamma V_\pi(s')\right). && (41) \end{aligned}$$

This equation is known as the Bellman equation for policy π and since it is valid for all $s \in \mathcal{S}$, it can be written in vectorial form as

$$\boldsymbol{V}_\pi = \boldsymbol{R}_\pi + \gamma \boldsymbol{P}_\pi \boldsymbol{V}_\pi, \tag{42}$$

where the vectors $\boldsymbol{V}_\pi$ and $\boldsymbol{R}_\pi$ and the matrix $\boldsymbol{P}_\pi$ are defined as:

$$\begin{cases} [\boldsymbol{V}_\pi]_i \triangleq V_\pi(s^{(i)}), \\ [\boldsymbol{R}_\pi]_i \triangleq \mathbb{E}_\pi \left\{ r_{k+1} | s_k = s^{(i)} \right\} = \sum\limits_{a\in\mathcal{A}} \pi(a|s^{(i)}) \sum\limits_{s'\in\mathcal{S}} \sum\limits_{r\in\mathcal{R}} p(s', r|s^{(i)}, a) r, \\ [\boldsymbol{P}_\pi]_{i,j} \triangleq \Pr\left\{ s_{k+1} = s^{(j)} | s_k = s^{(i)} \right\} = \sum\limits_{a\in\mathcal{A}} \pi(a|s^{(i)}) \sum\limits_{r\in\mathcal{R}} p(s^{(j)}, r|s^{(i)}, a), \end{cases} \tag{43}$$

where $s^{(i)} \in \mathcal{S}$. If the dynamics (see eq. (30)) of the MDP are known, then eq. (42) can be uniquely solved as:

$$\boldsymbol{V}_\pi = \left(\boldsymbol{I}_{|\mathcal{S}|} - \gamma \boldsymbol{P}_\pi\right)^{-1} \boldsymbol{R}_\pi, \tag{44}$$

where $\boldsymbol{I}_{|\mathcal{S}|}$ is the identity matrix of dimension $|\mathcal{S}| \times |\mathcal{S}|$.

Finally, from eqs. (37) and (41), it follows directly that the Bellman equation for the *state-action value function* can be expressed as:

$$Q_\pi(s,a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) \left(r + \gamma V_\pi(s')\right), \tag{45}$$

and thus, once (42) has been solved, $Q_\pi(s,a)$ can be computed from (45).

3.1.2 Optimal policy

The policy π^* is said to be *optimal* if the following condition is satisfied [23]:

$$V_\pi(s) \leq V_{\pi^*}(s) = \max_{a' \in \mathcal{A}} Q_{\pi^*}(s, a'), \quad \forall s \in \mathcal{S}. \tag{46}$$

It can also be shown that every MDP admits a *deterministic* optimal policy f_{π^*} that satisfies:

$$f_{\pi^*}(s) = \underset{a' \in \mathcal{A}}{\text{argmax}} Q_{\pi^*}(s, a'). \tag{47}$$

Let us now have a look on how the expressions of the state value function and of the state-action value function changes when they are evaluated for the optimal policy. More specifically, we limit ourselves to the case of *deterministic* policies, i.e. those that satisfy the definition (31). For the optimal deterministic policy, the state value function in (35) becomes

$$V_{\pi^*}(s) = \sum_{a \in \mathcal{A}} Q_{\pi^*}(s,a)\pi(a|s) = \max_{a \in \mathcal{A}} Q_{\pi^*}(s,a) \tag{48}$$

$$= \max_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) \left(r + \gamma V_{\pi^*}(s')\right) \right\}, \quad \forall s \in \mathcal{S} \tag{49}$$

that constitutes a system of non linear equations that cannot be solved using eq. (44). Regarding the state-action value function in (45), it becomes:

$$Q_{\pi^*}(s,a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) \left(r + \gamma \max_{a' \in \mathcal{A}} Q_{\pi^*}(s', a')\right) \tag{50}$$

$$= \mathbb{E}_{\pi^*} \left\{ r_{k+1} + \gamma \max_{a' \in \mathcal{A}} Q_{\pi^*}(s_{k+1}, a') \middle| (s_k = s) \cap (a_k = a) \right\}. \tag{51}$$

Eq. (47) tells us that to find the optimal deterministic policy $f_{\pi^*}(s)$, the agent only needs the "optimal" state-action value function $Q_{\pi^*}(s,a)$, while the knowledge of the dynamics in eq. (30) is not strictly required. However, $Q_{\pi^*}(s,a)$ is generally not available a priori to the agent and it needs to be recursively estimated, which will be explained following subsection.

3.2 The SARSA learning algorithm

The learning algorithms are a class of algorithms designed to obtain the optimal policy associated to a specific MDP. As discussed in 3.1.2, if the "optimal" state-action value function $Q_{\pi^*}(s,a)$ were a priori known, the optimal deterministic policy may be obtained by maximizing it as shown in eq. (47). Unfortunately, in most of the real world applications, $Q_{\pi^*}(s,a)$ is not available to the agent. In these cases a "closed-form" solution characterizing the optimal policy does not exist. Nevertheless, it is still possible to find a good approximation of the optimal policy by applying some iterative learning procedure [25]. The Q-learning and the SARSA algorithms are two classic examples of this kind of iterative approximation algorithms. Here, we limit ourselves to discuss the SARSA iteration process, since it is the one that we are going to exploit in the application at hand.

The SARSA algorithm, like the Q-learning one, attempts to estimate eq. (50). The name of the algorithm comes from the succession of operations followed by the algorithm, i.e. State-Action-Reward-State-Action. The system selects an initial state $s_0 \in \mathcal{S}$ and an initial *state-action value function* $Q_0(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$, then proceeds by choosing action $a_0 \in \mathcal{S}$ following a given policy π_0. Then, the system repeats the following steps at each time instant.

SARSA

At each iteration, the SARSA algorithm operates as follows:

1. receive reward r_{k+1} from the environment,
2. reach state s_{k+1},
3. choose action a_{k+1} following a certain policy π_k that depends on $Q_k(s,a)$,
4. update Q as

$$\begin{aligned} Q_{k+1}(s_k,a_k) =& (1-\alpha_{k+1})Q_k(s_k,a_k) + \alpha_{k+1}(r_{k+1} + \gamma Q_k(s_{k+1},a_{k+1})) \\ =& Q_k(s_k,a_k) + \alpha_{k+1}(r_{k+1} + \gamma Q_k(s_{k+1},a_{k+1}) - Q_k(s_k,a_k)). \end{aligned} \tag{52}$$

If the policy π_k guarantees that each couple $(s,a) \in \mathcal{S} \times \mathcal{A}$ is visited infinitely many times and under additional condition on the sequence of scalars $\{\alpha_k\} \in (0,1)$ (see [23] end [15]) then:

$$\lim_{k \to +\infty} Q_k(s,a) = Q_{\pi^*}(s,a). \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{53}$$

How can we assure that this condition holds? This is the focus of the following subsection.

3.2.1 ε-greedy policy

In order to guarantee that each couple $(s, a) \in \mathcal{S} \times \mathcal{A}$ is visited infinitely many times, we can force the agent to "explore" the state-action space, instead of just using the optimal action. Specifically, according to eq. (47), the optimal *greedy* action at the iteration k is given by:

$$a_k^{(greedy)} = \arg \max_{a' \in \mathcal{A}} Q_{\pi_{k-1}}(s_k, a'). \tag{54}$$

Even though choosing the greedy action at each time instant seems a good (optimal) choice, unfortunately it doesn't guarantee that each couple (s, a) is visited infinitely many times. To force the agent to explore, one of the more widely adopted solutions is the ε-greedy policy that consists on selecting the greedy (optimal) action with probability $1 - \varepsilon$ and a random action with probability ε, i.e.

$$\pi_k^{(1)}(s_k) = \begin{cases} a_k^{(greedy)}, & \text{with probability } 1 - \varepsilon \\ \mathbb{U}\left\{\mathcal{A} \setminus \{a_k^{(greedy)}\}\right\}, & \text{with probability } \varepsilon \end{cases} \tag{55}$$

where $\mathbb{U}\{\cdot\}$ denotes a function that selects randomly one of the elements specified by its argument with uniform probability and the apex "(1)" is used to diversify the ε-greedy policy from the two variations that will be introduced in Sec. 4.4. The ε parameter controls the *exploration-exploitation* trade-off: for low values of ε, the agent selects the greedy (optimal) action with high probability (exploitation), while for high values of ε the action is most likely chosen randomly (exploration).

After having introduced the basics of the RL theory, we are now ready to map the radar detection/weighting matrix selection problem, introduced at the end of Sec. 2.2, into this framework and find a solution to it.

4 A RL-based detection algorithm for MMIMO radar

In this section, we merge the asymptotic results on the robust, CFAR, Wald-type detector presented in Sec. 2 with the RL framework briefly summarized in Sec. 3. This cross-fertilization between asymptotic statistics and learning methodologies allows us to present a fully cognitive and robust detection scheme characterized by the CFAR property (with respect to a wide class of disturbance model satisfying Assumption A1) and, at the same time, to maximize the detection performance without the need of any a priori information on the environment [2].

We start this section by mapping the general MDP-related concepts on the MMIMO radar detection problem. We show that, in this specific application, the *state* is related to the number of detected targets, the *action* to the number of angular bins where the system focuses its power, and the *reward* to an estimate of the probability of detection of the targets. Along with the similarity, it is important to highlight that the radar detection problem presents a crucial difference with respect

to the standard RL machinery: the radar environment is *not stationary*. This fundamental aspect will be discussed in Sec. 4.4. Consequently, some modifications aiming at taking into consideration possible non-stationarities, will be discussed and the final resulting algorithm presented. An important remark on the index k is in order now. In Sec. 2, we used the index k to define the discrete time indexing the acquired observations, while in Sec. 3, k was adopted to characterize the iterations of the RL algorithm. In the following, we continue to use the index k for both the above-mentioned quantities since, from an algorithmic point of view, they are basically "the same thing". In fact, as we will amply discuss ahead, the proposed algorithm will perform a new iteration when a new snapshot is acquired.

4.1 The set of the states

The detection procedure starts by applying the Wald-type detector in (26) to all the L angular bins previously defined in Sec. 2.2. Consequently, given the decision statistic at the k^{th} time instant associated to the l^{th} angular bin $\Lambda_{k,l} \in \mathbb{R}$, let us first define the following quantity [1]:

$$\bar{\Lambda}_{k,l} \triangleq \begin{cases} 1, & \text{if } \Lambda_{k,l} \geq \lambda \\ 0, & \text{if } \Lambda_{k,l} < \lambda \end{cases}. \tag{56}$$

In words, $\bar{\Lambda}_{k,l}$ is equal to 1 when the decision statistic is over the threshold λ, otherwise it is equal to 0. Then, the state of the system at time instant k can be defined as $s_k = s^{(i_k)}$ with:

$$i_k \triangleq \min\left\{ \sum_{l=0}^{L-1} \bar{\Lambda}_{k,l}, K \right\} \tag{57}$$

where $\sum_{l=0}^{L-1} \bar{\Lambda}_{k,l}$ is simply the number of target detected at time k. Consequently, the set of all the possible states indexed by eq. (57) constitutes the state space of our MDP:

$$\mathcal{S} \triangleq \left\{ s^{(i)} \right\}_{i=0}^{K} \tag{58}$$

where $K < \infty$ is the maximum number of detectable targets.

4.2 The set of the actions

The set of the actions of the MDP in our application is defined as

$$\mathcal{A} \triangleq \left\{ a^{(j)} \right\}_{j=0}^{K}. \tag{59}$$

Specifically, if the SARSA algorithm selects action $a^{(j_k)}$ at time instant k, then the system focuses its power in the j_k angular bins with the highest decision statistic according to the following procedure. Let $\{l_k^{(n)}\}_{n=1}^{L}$ be the set containing the indexes of the angular bin corresponding to the sequence of the decision statistics at time instant k ordered in a descending way, i.e. $\Lambda_{k,l_k^{(1)}} \geq \Lambda_{k,l_k^{(2)}} \geq ... \geq \Lambda_{k,l_k^{(L)}}$. Moreover, let us define the set:

$$\Omega_k \triangleq \begin{cases} \emptyset\,, & \text{if } j_k = 0 \\ \{l_k^{(n)}\}_{n=1}^{j_k}\,, & \text{if } j_k \neq 0 \end{cases} \tag{60}$$

containing the indexes of the angular bins associated to the j_k highest decision statistics. Then, as discussed in [1, 26], the action made by the system is to choose the weighting matrix $\boldsymbol{W}_k$ by solving the following constrained optimization problem:

$$\boldsymbol{W}_k = \begin{cases} \boldsymbol{W}_{ort} \triangleq \sqrt{\frac{P_{max}}{N_T}} \cdot \boldsymbol{I}_{N_T}\,, & \text{if } \Omega_k = \emptyset \\ \begin{cases} \arg\max\limits_{\boldsymbol{W}} \min\limits_{l \in \Omega_k} \boldsymbol{a}_T^T(\nu_l)\boldsymbol{W}\boldsymbol{W}^H\boldsymbol{a}_T^*(\nu_l) \\ \text{subject to} \quad \text{tr}\{\boldsymbol{W}\boldsymbol{W}^H\} \leq P_{max} \end{cases}, & \text{if } \Omega_k \neq \emptyset \end{cases} \tag{61}$$

where $\boldsymbol{I}_{N_T}$ denotes the $N_T \times N_T$ identity matrix. The choice of the matrix $\boldsymbol{W}_k$ is not univocal since the beampattern depends on the autocorrelation matrix $\boldsymbol{R}_{\boldsymbol{W}} \triangleq \boldsymbol{W}\boldsymbol{W}^H$ and there are infinitely many different $\mathbf{W}$ with the same $\boldsymbol{R}_{\boldsymbol{W}}$. The algorithm chooses one $\mathbf{W}$ among them arbitrarily.

4.3 The reward function

The choice of the reward is a crucial part of every reinforcement learning algorithm and its definition strictly depends on the application at hand. Since in this chapter we are dealing with a radar detection problem, the reward to be maximized should be linked to the P_D of the system.

As already described in Sec. 2.2, the P_D associated to a target in the l^{th} angular bin at time k in the massive MIMO regime can be estimated/approximated as :

$$\hat{P}_{D,k,l} \triangleq Q_1\left(\sqrt{\hat{\zeta}_{k,l}}, \sqrt{\lambda}\right) = Q_1\left(\sqrt{\Lambda_{k,l}}, \sqrt{\lambda}\right), \tag{62}$$

where the non-centrality parameter ς, defined in (21), has been substituted by its consistent estimate: [2]

[2] We suppose that the sequence of actions at time instant k starts when the radar receives the first snapshot. Thus, the received vector at time k is associated to the transmitted waveform at time $k-1$ and so we use $\boldsymbol{h}_{k-1,l}$ and $\alpha_{k-1,l}$ to compute the Wald-test statistic at time k. This is a different notation from the one adopted in [1].

$$\hat{\zeta}_{k,l} \triangleq 2|\hat{\alpha}_{k-1,l}|^2 \frac{||\boldsymbol{h}_{k-1,l}||^4}{\boldsymbol{h}_{k-1,l}^H \hat{\boldsymbol{\Gamma}}_{k,l} \boldsymbol{h}_{k-1,l}} \overset{\text{eq. (29)}}{=} 2\frac{|\boldsymbol{h}_{k-1,l}^H \cdot \boldsymbol{y}_{k,l}|^2}{\boldsymbol{h}_{k-1,l}^H \hat{\boldsymbol{\Gamma}}_{k,l} \boldsymbol{h}_{k-1,l}} = \Lambda_{k,l}. \tag{63}$$

Let us now define the set $\Psi_k \triangleq \left\{l_k^{(n)}\right\}_{n=1}^{K}$ where $l_k^{(n)}$, $\forall n$ are the ordered indexes of the angular bins already defined in subsec. 4.2 and the sets :

$$\Phi_k \triangleq \begin{cases} \emptyset\,, & \text{if } i_k = 0, \\ \left\{l_k^{(n)}\right\}_{n=1}^{i_k}\,, & \text{if } i_k \neq 0, \end{cases} \tag{64}$$

$$\bar{\Phi}_k \triangleq \Psi_k \setminus \Phi_k = \left\{l \in \Psi_k : l \notin \Phi_k\right\}, \tag{65}$$

where i_k is defined in (57). The set Φ_k contains the indexes associated to the angular bins where the decision statistic is over the threshold if these are less then K, otherwise contains the ones associated to the K with the highest decision statistic. The set $\bar{\Phi}_k$ is the complement of Φ_k with respect to Ψ_k.

Following [1], we are finally ready to introduce the reward function as:

$$r_k \triangleq \sum_{l \in \Phi_k} \hat{P}_{D,k,l} - \sum_{l \in \bar{\Phi}_k} \hat{P}_{D,k,l}. \tag{66}$$

Remark. The role of the second "negative term" in eq. (66) can be explained by means of the following example. Suppose that there are three targets present in the environment and that the system selects action $a_k = a^{(3)}$ with Ω_k containing the indexes of the three targets. If the three targets have high SNR, the reward at the next time instant r_{k+1} will be slightly less than 3. Suppose now that the system selects action $a_{k+1} = a^{(5)}$, misses one target due to a miss-detection event, and detects only the two most powerful targets. If the negative term weren't considered then the reward r_{k+2} would be slightly less then 2, while with the negative term it will be decreased with the term $\hat{P}_{D,k+2,l}$ associated to the missed target that can be as high as $Q_1(\sqrt{\lambda}, \sqrt{\lambda}) \approxeq 0.55$ if λ is chosen to guarantee $P_{FA} = 10^{-4}$. The presence of this negative term increases the reward loss between a good choice and a poor choice of the action.

We note, in passing, that the reward function in eq. (66) is an example among the infinitely many others that can be built around the concept of detection probability. As an example, a reward with an additional negative term has been recently proposed in [28]. Determining an optimal (if it exists) reward function is still an open problem and active field of research.

After having introduced the reward function and before discussing the impact of the violation of the *stationarity assumption* of the environment, we summarize here the main step of the proposed weighting matrix selection algorithm.

RL-based, W_k-selection algorithm

Initialization: Let us choose $k = 0$ as the starting point in time.

1. At time $k = 0$, the system is in state $s^{(0)}$ and selects action $a^{(0)}$, so the first transmitted waveform is $\boldsymbol{W}_0 = \boldsymbol{W}_{ort}$,
2. According to analysis in [18], the matrix optimal state-action value function in eq. (50) is initialized as $\boldsymbol{Q}_0 = \boldsymbol{I}_{K+1}$.

At time instant k, once the system collects the set of observations $\{\boldsymbol{y}_{k,l}\}_{l=1}^{L}$, the sequence of instruction executed by the algorithm are the following:

3. compute the Wald-type statistics $\Lambda_{k,l}$ given in eq. (26) for each angular bin in the considered grid $\Theta = \{\theta_l\}_{l=1}^{L}$,
4. compute $\bar{\Lambda}_{k,l}$ as in (56),
5. compute state s_k as described in subsec. 4.1,
6. compute the reward r_k as in (66)
7. select action a_k according to the ε-greedy policy in eq. (55),
8. compute Ω_k and $\boldsymbol{W}_k$ as described in 4.2,
9. update the matrix $\boldsymbol{Q}_k$ according to the SARSA iteration in eq. (52),
10. transmit the new waveforms using the computed weight matrix $\boldsymbol{W}_k$.

From the previous list of algorithmic steps, the roles that the ε-greedy policy selection and the SARSA algorithm play in the recursive evaluation of the weighting matrix $\boldsymbol{W}_k$ it is now clear. However, these two fundamental building blocks of the required learning strategy depend on two crucial hyper-parameters : the ε for the ε-greedy policy in eq. (55) and the α_k for the SARSA updating equation in eq. (52). In the next subsections, an adaptive and data-dependent methodology to select these two parameters will be provided.

4.4 Adaptivity in non-stationarity radar environments

The crucial assumption underlying the theoretical RL framework presented in Sec. 3 is the *stationarity* of the MDP, i.e. the environment has to remain unchanged over time. However, in radar applications, the environment is intrinsically non-stationary since position, Signal to Noise Ratio (SNR) and even the number of the targets may vary over time as well as the disturbance statistics. The first consequence of this non-stationarity is that the concept of *optimal policy* loses its meaning. In fact, the optimal policy depends on the dynamics of the environment: if the dynamics change, the optimal policy will change accordingly. This clearly makes the definition of a *globally optimal policy* impossible in non-stationary MDPs.

If the temporal evolution of the environment is slow enough, a possible way out is to split a non-stationary MDP in a sequence of approximately stationary MDPs. Consequently, a "stationary-based" learning approach may be implemented in each sub-MDP. However, some considerations on how to handle the transition phase between two approximately stationary sub-MDPs is in order. Specifically, we should find out a strategy that makes the adopted learning algorithm, in our case

the SARSA one, more robust to the changes in the environment, that in the radar application at hand are represented by changes in the number of targets, in their SNRs or in their angular positions. In particular, the aspects that we are going to investigate in the following subsections are: 1) definition of a policy tailored on the specific radar environment, 2) adaptive selection of the parameters ε and α for the SARSA learning algorithm.

4.4.1 A new policy for MMIMO radar detection

In this subsection, we present two new policies able to provide significant improvements with respect to the standard ε-greedy policy introduced in eq. (55) of Sec. 3.2.1. Here, we limit ourselves to provide a short operative definition of the two policies, while additional details and in-depth discussions are provided in [18].

1. Quasi $\boldsymbol{\varepsilon}$-greedy policy:

$$\pi_k^{(2)}(s_k) = \begin{cases} \mathbb{U}\left\{\mathcal{A}'(s_k) \setminus \{a_k^{(greedy)}\}\right\}, & \text{w.p.} \quad \varepsilon \\ a_k^{(greedy)}, & \text{w.p.} \quad 1-\varepsilon \end{cases} \tag{67}$$

   where $\mathcal{A}'(s^{(i)}) \triangleq \{a^{(j)}, j = i, ..., K\}$. The aim of this policy is to minimize possible miss-detection events that can be generated during the exploration phase. Specifically, suppose that the system is in state $s^{(i)}$, than the quasi $\boldsymbol{\varepsilon}$-greedy policy avoids that, in the exploration phase, the radar is focusing its power in a number of angular bins less than i, that corresponds to the number of potential detected targets.
2. Quasi $\boldsymbol{\varepsilon}$-greedy policy with target recovery:

$$\pi_k^{(3)}(s_k, s_{k-1}) \triangleq \begin{cases} \arg\max_{a \in \mathcal{A}} Q_k(s_{k-1}, a), & i_k < i_{k-1} \\ \pi_k^{(2)}(s_k), & i_k \geq i_{k-1}. \end{cases} \tag{68}$$

   This policy has been introduced in [18] to handle the *target loss problem*: if, at the time instant k, the radar detects an higher or equal number of targets with respect to the one detected at time $k-1$, then the policy in eq. (68) select the new action according to the quasi ε-greedy policy in eq. (67). On the contrary, if the number of detected targets at time k is smaller than the one at the previous time instant $k-1$, the algorithm tries to recover them as soon as possible by choosing the greedy action associated to the state at the previous time instant s_{k-1}.

It is worth stressing here that the two proposed policies are tailored on the specific non-stationary, radar detection problem at hand. Consequently, no general performance improvement can be claimed for scenarios different from this one.

4.4.2 Adaptive selection of the SARSA hyper-parameters

The setting of the hyper-parameter of a learning algorithm is a critical step in the implementation of a learning algorithm. Despite its direct impact on the overall system performance, this setting is generally made heuristically by the practitioner starting from some previous knowledge of the problem at hand. In non-standard environment, their setting is even more problematic due to the fact that, when the dynamics of environment changes, the original parameter selection is no longer "valid".

The problems related to the choice of the hyper-parameter $\boldsymbol{\varepsilon}$ and $\boldsymbol{\alpha}$ of the SARSA algorithm for our radar detection problem at hand has been discussed first in [1]. Therefore, in [18], a *fully adaptive* and *data-dependent* algorithm able to automatically select both $\boldsymbol{\varepsilon}$ and $\boldsymbol{\alpha}$ in the SARSA algorithm has been proposed and its effectiveness proven through extensive numerical simulation. This adaptive algorithm can be summarized as follows.

Let r_k be the reward in eq. (66) at time instant k. Let us define the sequence $\{d_k,\ \forall k\}$ of real numbers defined as:

$$d_k \triangleq \begin{cases} r_1 & ,\ k = 1 \\ r_k - r_{k-1} & ,\ k \neq 1 \end{cases}. \tag{69}$$

Then, the hyper-parameters and the time instant $k+1$, i.e. ε_{k+1} and α_{k+1}, can be obtained according to the following strategy:

$$x_{k+1} = \begin{cases} \max\{c_1 \cdot x_k\ ,\ x_{min}\}, & |d_k| < \eta_1 \\ \min\{c_2 \cdot x_k\ ,\ x_{max}\}, & \eta_1 < |d_k| < \eta_2 \\ x_{max}, & |d_k| > \eta_2 \end{cases} \tag{70}$$

where x_k corresponds to ε_k or α_k, $c_1 \in (0,1)$ and $c_2 \in (1,+\infty)$. The initial value of x is set to $x_0 = x_{max}$. The values c_1, c_2, η_1 and η_2 are constants, and depend on which parameter (i.e. ε_k or α_k) we are considering. Table 1 lists all their values. Note that x_{k+1} is not updated if the SARSA algorithm was in exploration mode in the two previous time instants, i.e. $k-1$ and k. This choice is motivated by the following observation. If the SARSA algorithm selects a random action (possibly far from the greedy one) at time instant k, the reward r_{k+1} may drop due to some (possibly multiple) miss-detection causing $|d_{k+1}| = |r_{k+1} - r_k|$ to surpass η_1 even though the scenario has not changed. If the algorithm then chooses the correct action at time $k+1$, the reward r_{k+2} rises back to a value around r_k, but $|d_{k+2}| = |r_{k+2} - r_{k+1}|$ is likely to be over η_1 due to the low value of r_{k+1}.

Finally, let us provide some insight behind the choices of the thresholds η_1 and η_2. Clearly, their value are linked to the maximum range of variation of the reward function. Since the one proposed in eq. (66) is basically a linear combination of probabilities of detection, its values will vary (in absolute value) of something close to 1 when the radar misses one target or detects a new one. Consequently, a reasonable choice for η_1 is 0.5. On the same line, η_2 should be chosen large enough to guarantee

Table 1: Values to be used for the adaptive selection of ε and α.

x	x_{min}	x_{max}	c_1	c_2	η_1	η_2
ε	0.1	0.8	0.8	2	0.5	1.8
α	0.2	0.6	0.9	2.5	0.5	1.8

that the value of ε_{k+1} or α_{k+1} is set to its maximum value only when an abrupt change in the scenario happens (i.e. a new target appears or an old one disappears). Some numerical analysis have showed that $\eta_2 \geq 1.8$ is a good choice.

4.5 Simulation results

In the following, a numerical analysis of the previous theoretical results is presented. Specifically, we will start by describing the static and dynamic scenarios used in our simulations. Then, in Sec. 4.5.2, we show that the original policies and the proposed adaptive selection of the SARSA hyper-parameters lead to a remarkable performance improvement. Finally, Sec. 4.5.3 is dedicated to the comparison of the proposed original RL-based beamforming with more classical strategies. As the simulation results show, the RL-based algorithm is the closest one to the performance benchmark with respect to the other competing methods.

4.5.1 The scenarios under investigation

Let us start by introducing the parameters that we keep unchanged in all the settings described below. As in [7, 1], the disturbance vector $\boldsymbol{c}$ in eq. (8) is assumed to be sampled from a complex circular, autoregressive process of order $p = 6$ (AR(6)):

$$c_n = \sum_{i=1}^{p} \rho_i c_{n-i} + w_n, \; n \in (-\infty, +\infty). \tag{71}$$

The innovations $\{w_n, \forall n\}$ are complex t-distributed, independent and identically distributed (i.i.d.) random variables whose pdf is by:

$$f_w(w_k) = \frac{\lambda}{\sigma_w \pi} \left(\frac{\lambda}{\eta}\right)^{\lambda} \left(\frac{\lambda}{\eta} + \frac{|w_k|^2}{\sigma_w^2}\right)^{-(\lambda+1)} \tag{72}$$

with $\lambda = 2$, $\sigma_w^2 = 1$ and $\eta = \lambda/(\sigma_w^2(\lambda - 1))$. The coefficients of the AR(6) process are $\rho_1 = 0.5e^{-j2\pi 0.4}$, $\rho_2 = 0.6e^{-j2\pi 0.2}$, $\rho_3 = 0.7$, $\rho_4 = 0.4e^{j2\pi 0.1}$, $\rho_5 = 0.5e^{j2\pi 0.3}$ and $\rho_6 = 0.6e^{j2\pi 0.35}$ that lead to a Power Spectral Density (PSD) $S_c(\nu) \triangleq \sigma_w^2 \left|1 - \sum_{n=1}^{p} \rho_n e^{-\mathrm{j}2\pi n\nu}\right|^{-2}$ plotted in Fig. 1. Note that, for ULA arrays with $\lambda/2$ inter-element spacing, the *spatial frequency* ν is related to the angular grid by

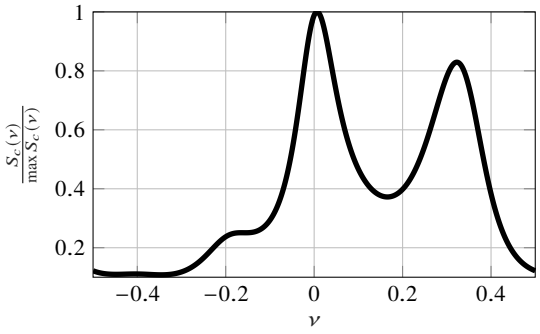


Fig. 1 Normalized power spectrum of the noise process.

$\nu \triangleq \frac{1}{2}\sin(\theta)$. While satisfying Assumption 1, this disturbance model is much more general than the white Gaussian one usually adopted in MIMO radar literature and it allows us to check the claimed robustness property of the Wald-type test proposed in Sec. 2.2. The other parameters that are kept constant in all the scenarios are:

- The nominal P_{FA} is chosen to be equal to 10^{-4}.
- The number of the transmitting and receiving antennas, respectively N_T and N_R, are both equal to 100.
- The angular grid, expressed in terms of spatial frequency ν, is selected as $\left\{\nu_l = -\frac{1}{2} + \frac{l}{L}\right\}_{l=0}^{L-1}$, where $L = 20$.
- The maximum number of targets is assumed to be $K = 5$.
- The maximum transmitted power is $P_{max} = 1$.

We can now proceed with the description of the four considered scenarios consisting of both static and dynamic scenarios. In the dynamic cases the number of targets, their SNRs and positions can vary over time. Note that each target is completely determined by its angular bin index (along with its spatial frequency), SNR and time interval in which it is present in the scenario. All the required information are stored in Table 2, while a short description of each scenario is provided in the following:

- Scenario 1 contains two static targets.
- Scenario 2 contains a single static target.
- Scenario 3 contains two dynamic targets that move towards and then outwards the radar. We characterize this dynamics by assuming that their SNR varies linearly over time during both phases, as shown in Fig. 5a.
- Scenario 4 contains 3 dynamic targets appearing and disappearing: target 1 disappears at time instant 101, target 2 disappears at time instant 301 and target 3 appears at time instant 201, as shown in Table 2.

The interested reader may find the whole code that we implement to carry out the proposed simulative analysis here: `https://github.com/lisifra96/Improved_RL_algorithm_mMIMO_radar`.

Table 2: Target scenarios.

Scenario	Time Interval	Target	Angular Bin	ν	SNR_{dB}
1	[1,300]	1	5	-0.30	-22
		2	17	0.30	-19
2	[1,100]	1	17	0.30	-20
3	[1,200]	1	7	-0.20	variable
		2	15	0.20	(Inset Fig.5a)
4	[1,100]	1	4	-0.35	-18
		2	12	0.05	-19
	[101,200]	2	12	0.05	-19
	[201,300]	2	12	0.05	-19
		3	18	0.35	-22
	[301,400]	3	18	0.35	-22

4.5.2 Validation of the new policy with adaptive hyper-parameters selection

At first, we provide a numerical validation of the original policies, proposed in Sec. 4.4.1, and of the strategy, proposed in Sec. 4.4.2, to select adaptively the SARSA hyper-parameters α and ε. Let us start by comparing in terms of probability of detection (P_D) the "standard" ε-greedy policy with the two original *quasi ε-greedy policy* in eq. (67) and *quasi ε-greedy policy with target recovery* in eq. (68). The comparison is reported in Fig. 2 in terms of probability of detection (P_D) evaluated for the two static targets of the Scenario 1. Since the results are similar, here we show only the results for target 2 (see Table 2). As we can clearly see, both the *quasi ε-greedy policy* and *quasi ε-greedy policy with target recovery* outperform the "standard" ε-greedy one. Specifically, the *quasi ε-greedy policy with target recovery* has the best performances among the three policies.

Let us now move on with the validation of the adaptive selection strategy of the SARSA hyper-parameters. To this end, for the analysis of the ϵ parameter we consider the Scenario 1 with two static targets. In Fig. 3, we compare the P_D of the adaptive ε algorithm with the two "non-adaptive" cases with $\varepsilon = \varepsilon_{min}$ (the *exploitation phase* is dominant) and $\varepsilon = \varepsilon_{max}$ (the *exploration phase* is dominant). The parameter α is kept constant and equal to 0.5. In Fig. 4, we do a similar analysis for the adaptive selection of α by keeping ε constant and equal to 0.5. The adopted set-up is the Scenario 2, where a single static target is present. As the simulation results show, the adaptive strategy has the remarkable ability to combine the positive effects of both *exploration phase* (characterized by high values of ε and α) and of the *exploitation phase* (low values of ε and α). In fact, in the initial transition phase, when the system has to gather information about the surrounding environment the adaptive strategy

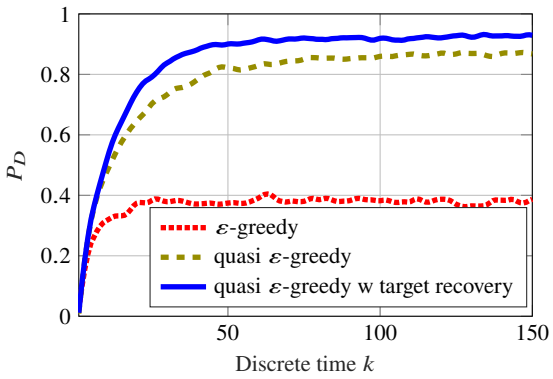


Fig. 2 Policy comparison: P_D of target 2 (Scenario 1).

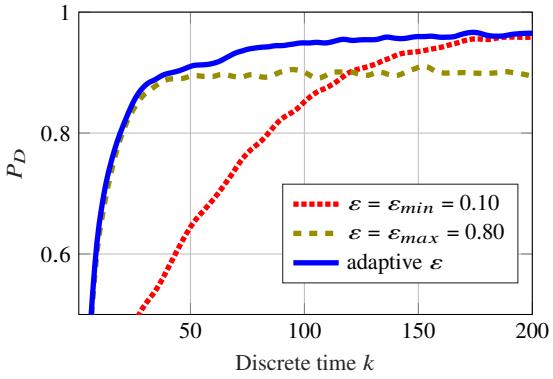


Fig. 3 Adaptive vs static ε: P_D of target 2 (Scenario 1).

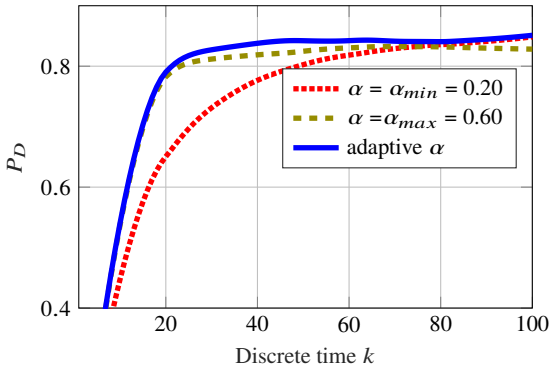


Fig. 4 Adaptive vs static α: P_D of target 1 (Scenario 2).

selects high values of ε and α. Then, their values are gradually reduced in order to fully exploit the acquired knowledge of the environment and maximize the P_D.

4.5.3 Performance comparison: the learning gain

This last subsection is dedicated to the comparison of the RL-based detection algorithm described in Sec. 4 with two "non RL-based"algorithms and a performance benchmark. For the RL-algorithm, we use the SARSA learning strategy with the following setting:

- $\gamma = 0.8$ as in [1].
- Quasi ε-greedy policy with target recovery in eq. (68).
- Adaptive selection of α and ϵ according to eq. (70).

In order to highlight the performance gain that the learning can bring into the MMMIO radar detection problem, we compare the proposed RL-based algorithm with two non RL algorithms:

- **Orthogonal waveform selection**: The radar/agent chooses the orthogonal beamformer at each iteration, i.e.
$$\boldsymbol{W}_k = \boldsymbol{W}_{ort}. \tag{73}$$
This choice is clearly the easiest to implement since it doesn't require any kind of memory. On the other hand the system doesn't take advantage of its high focusing capability to improve its performance.
- **Non RL (NRL) beamforming**: The NRL algorithm exploits the information gathered from the decision statistic to focus the transmitted power. The $\boldsymbol{W}$ matrix is the solution of the optimization problem in (61) with the set Ω_k
$$\Omega_k \triangleq \begin{cases} \emptyset\,, & \text{if } i_k = 0 \\ \{l_k^{(n)}\}_{n=1}^{i_k}\,, & \text{if } i_k \neq 0 \end{cases}. \tag{74}$$
No *state-action-reward* cycle is adopted here. The beamforming depends only on the values of the test statistics.

As upper bound to the maximum achievable detection performance we use a clairvoyant beamforming that select the $\boldsymbol{W}$ matrix as the solution of the optimisation problem in eq. (61), with Ω_k being the set containing the angular bins corresponding to the exact position of the targets.

The above-mentioned algorithms are compared by using the Scenarios 3 and 4. As previously said, Scenario 3 contains two targets characterized by a constant angular position but with a SNR that varies linearly over time, as shown in Fig. 5a. The P_D reached by the orthogonal, NRL and RL beamforming algorithms are compared with the performance benchmark in Fig. 5 for the two targets. It is immediate to verify that the RL-beamformer outperforms the orthogonal and the NRL ones and it is not that far from the (unachievable) upper bound. This proves the effectiveness of the proposed learning strategy in collecting information from the environment and to use them to maximize the P_D. This learning gain also explains the delay in the decay of the P_D (that starts at around $k = 110$) with respect to the decrease of the targets SNR (that begins at $k = 100$). The system in fact exploits the acquired information stored in the Q matrix to contrast the SNR drop.

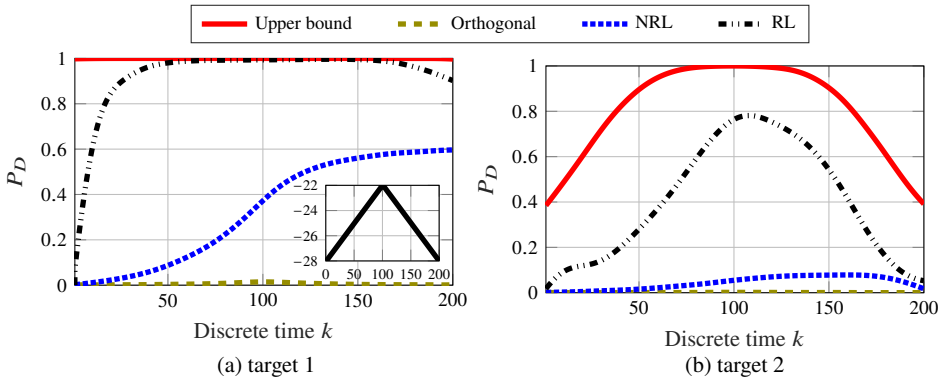


Fig. 5: Probability of detection of the two targets in scenario 3. The inset figure in (a) shows the SNR of both targets expressed in dB.

Finally, let us investigate how the three beamforming algorithms react to a dynamic environment as the one in Scenario 4, where the three considered targets appear and disappear over time. Fig. 6 shows the evolution of the P_D for each target over time. The figures can be easily understood by comparing the progress of the P_D curves with the targets dynamics given in Table 2. In particular, this result highlight the ability of the RL-algorithm to better adapt itself to abrupt changes in the environment. This desirable "adaptivity property" has its roots in the SARSA learning strategy and in the proposed adaptive hyper-parameters selection. In fact, if we compare the targets dynamics in Table 2 (Scenario 4) with the temporal evolution of α_k plotted in Fig. 7, it is immediate to verify that the selection strategy in eq. (70) is perfectly able to detect the abrupt changes and to update the parameters accordingly.

5 Conclusions

In this chapter, we studied the problem of multi-target detection for a MMIMO radar. A hybrid approach combining RL and a robust Wald-type detector has been proposed to solve this problem through cognitively optimizing the transmitted waveform without any prior knowledge of the surrounding environment. Such approach can be exploitable within a non-cooperative ISAC framework, where both systems cause possible unknown interference/disturbance on each other while sharing the same resources. This kind of disturbance can lead to a degradation of the sensing performance due to the low power of the reflected signals from the targets. In addition, further challenges arise if the environment is time varying. In our simulations, the proposed approach was capable of detecting multiple targets with high P_D while maintaining the CFAR property in both static and time varying scenarios even if the target SNR is low or the disturbance is strong. Furthermore, the proposed simulative

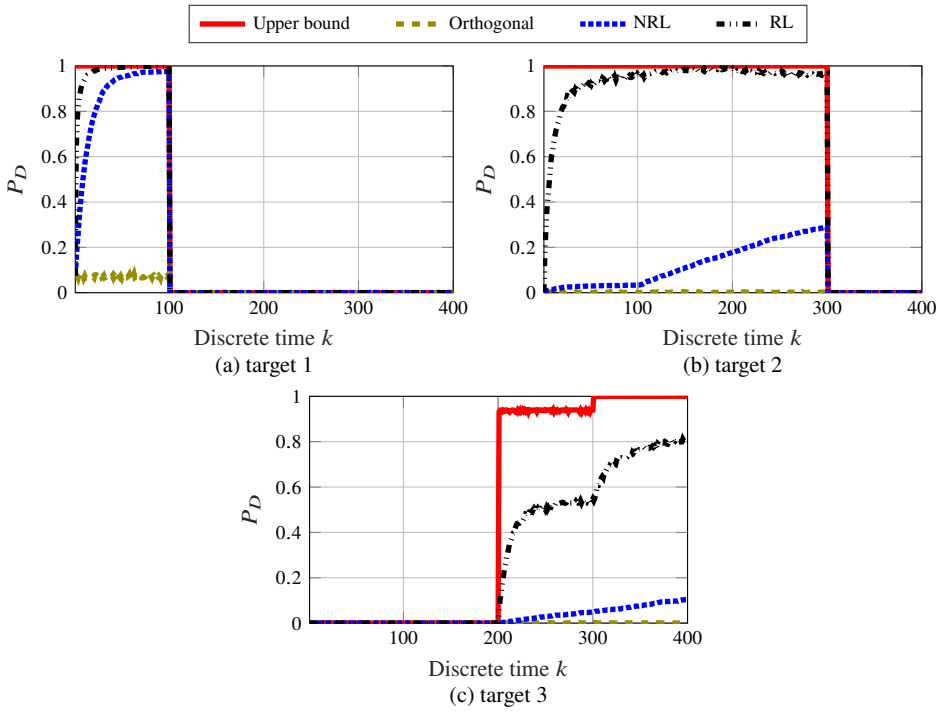
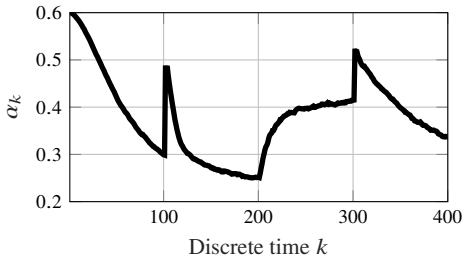


Fig. 6: Probability of detection of the three targets in scenario 4.

Fig. 7 Temporal evolution of α_k in scenario 4.



analysis showed the capability of self-adaptation of the algorithm to abrupt changes in the environment. Benefiting from the waveform diversity and large DoFs of the Massive MIMO radar, the proposed system could be also used for dual functional MIMO radar-communication in the more complex scenario in which the two systems cooperate. In this case, the transmit covariance matrix could be optimized in order to satisfy both radar and communication constraints, and to create a beam, among the others, dedicated to communication purposes. This is an interesting area of ongoing and future research under the ISAC framework.

References

[1] Ahmed AM, Ahmad AA, Fortunati S, Sezgin A, Greco MS, Gini F (2021) A reinforcement learning based approach for multitarget detection in massive mimo radar. IEEE Transactions on Aerospace and Electronic Systems 57(5):2622–2636

[2] Ahmed AM, Fortunati S, Sezgin A, Greco MS, Gini F (2021) Robust reinforcement learning-based wald-type detector for massive mimo radar. In: 2021 29th European Signal Processing Conference (EUSIPCO), pp 846–850

[3] Akyildiz IF, Jornet JM (2016) Realizing ultra-massive MIMO (1024x1024) communication in the (0.06–10) Terahertz band. Nano Communication Networks 8:46–54, electromagnetic Communication in Nano-scale

[4] Ali S, Saad W, Rajatheva N, Chang K, Steinbach D, Sliwa B, Wietfeld C, Mei K, Shiri H, Zepernick HJ, Chu TMC, Ahmad I, Huusko J, Suutala J, Bhadauria S, Bhatia V, Mitra R, Amuru S, Abbas R, Shao B, Capobianco M, Yu G, Claes M, Karvonen T, Chen M, Girnyk M, Malik H (2020) 6G white paper on machine learning in wireless communication networks

[5] Bekkerman I, Tabrikian J (2006) Target detection and localization using MIMO radars and sonars. IEEE Transactions on Signal Processing 54(10):3873–3883

[6] Björnson E, Sanguinetti L, Wymeersch H, Hoydis J, Marzetta TL (2019) Massive MIMO is a reality—What is next? Five promising research directions for antenna arrays. Digital Signal Processing 94:3–2

[7] Fortunati S, Sanguinetti L, Gini F, Greco MS, Himed B (2020) Massive MIMO radar for target detection. IEEE Transactions on Signal Processing 68:859–871

[8] Friedlander B (2012) On signal models for mimo radar. IEEE Transactions on Aerospace and Electronic Systems 48(4):3655–3660

[9] Friedlander B (2012) On transmit beamforming for mimo radar. IEEE Transactions on Aerospace and Electronic Systems 48(4):3376–3388

[10] Fuhrmann DR, San Antonio G (2008) Transmit beamforming for mimo radar systems using signal cross-correlation. IEEE Transactions on Aerospace and Electronic Systems 44(1):171–186

[11] Gini F, De Maio A, Patton L (2012) Waveform design and diversity for advanced radar systems. Institution of engineering and technology London, UK

[12] Grossi E, Lops M, Venturino L (2021) Energy efficiency optimization in radar-communication spectrum sharing. IEEE Transactions on Signal Processing 69:3541–3554

[13] Hassanien A, Amin MG, Zhang YD, Ahmad F (2016) Dual-function radar-communications: Information embedding using sidelobe control and waveform diversity. IEEE Transactions on Signal Processing 64(8):2168–2181

[14] Hassanien A, Aboutanios E, Amin MG, Fabrizio GA (2018) A dual-function mimo radar-communication system via waveform permutation. Digital Signal Processing 83:118–128

[15] Kushner HJ, Clark DS (2012) Stochastic approximation methods for constrained and unconstrained systems, vol 26. Springer Science & Business Media

[16] Li J, Stoica P (2008) MIMO Radar Signal Processing. Hoboken, NJ: Wiley

[17] Li J, Xu L, Stoica P, Forsythe KW, Bliss DW (2008) Range compression and waveform optimization for MIMO radar: A cramé-rao bound based study. IEEE Transactions on Signal Processing 56(1):218–232

[18] Lisi F, Fortunati S, Greco MS, Gini F (2022) Enhancement of a state-of-the-art rl-based detection algorithm for massive MIMO radars. IEEE Transactions on Aerospace and Electronic Systems pp 1–1

[19] Martone A, Amin M (2021) A view on radar and communication systems coexistence and dual functionality in the era of spectrum sensing. Digital Signal Processing 119:103–135

[20] Martone AF, Ranney KI, Sherbondy K, Gallagher KA, Blunt SD (2018) Spectrum allocation for noncooperative radar coexistence. IEEE Transactions on Aerospace and Electronic Systems 54(1):90–105

[21] Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Transactions on Wireless Communications 9(11):3590–3600

[22] Mazahir S, Ahmed S, Alouini MS (2021) A survey on joint communication-radar systems. Frontiers in Communications and Networks 1

[23] Mohri M, Rostamizadeh A, Talwalkar A (2018) Foundations of machine learning, MIT press, chap 17

[24] Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons

[25] Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press

[26] Wang L, Fortunati S, Greco MS, Gini F (2018) Reinforcement learning-based waveform optimization for mimo multi-target detection. In: 2018 52nd Asilomar Conference on Signals, Systems, and Computers, IEEE, pp 1329–1333

[27] White H (1984) Chapter VI - estimating asymptotic covariance matrices. In: White H (ed) Asymptotic Theory for Econometricians, Economic Theory, Econometrics, and Mathematical Economics, Academic Press, San Diego, pp 132 – 161

[28] Zhai W, Wang X, Greco MS, Gini F (2022) Weak target detection in massive mimo radar via an improved reinforcement learning approach. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4993–4997

[29] Zhang JA, Liu F, Masouros C, Heath RW, Feng Z, Zheng L, Petropulu A (2021) An overview of signal processing techniques for joint communication and radar sensing. IEEE Journal of Selected Topics in Signal Processing 15(6):1295–1315