ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Reinforcement Learning for UAV Autonomous Navigation, Mapping and Target Detection

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Guerra A., Guidi F., Dardari D., Djuric P.M. (2020). Reinforcement Learning for UAV Autonomous Navigation, Mapping and Target Detection. Piscataway, N.J. : IEEE-Institute of Electrical and Electronics Engineers [10.1109/PLANS46316.2020.9110163].

Availability:
This version is available at: https://hdl.handle.net/11585/766174 since: 2020-09-16

Published:
DOI: http://doi.org/10.1109/PLANS46316.2020.9110163

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

12 October 2024

This is the post peer-review accepted manuscript of:

Guerra, A., Guidi, F., Dardari, D., & Djurić, P. M. (2020, April). Reinforcement Learning for UAV Autonomous Navigation, Mapping and Target Detection. In 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS) (pp. 1004-1013). IEEE.

doi: 10.1109/PLANS46316.2020.9110163

The published version is available online at: https://ieeexplore.ieee.org/document/9110163

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

Reinforcement Learning for UAV Autonomous Navigation, Mapping and Target Detection

Anna Guerra
DEI - CNIT
University of Bologna
40136 Bologna, Italy
anna.guerra3@unibo.it

Francesco Guidi
CNR-IEIIT
40136 Bologna, Italy
francesco.guidi@ieiit.cnr.it

Davide Dardari
DEI - CNIT
University of Bologna
40136 Bologna, Italy
davide.dardari@unibo.it

Petar M. Djurić
ECE Department
Stony Brook University
11790 Stony Brook, New York
petar.djuric@stonybrook.edu

***Abstract*—In this paper, we study a joint detection, mapping and navigation problem for a single unmanned aerial vehicle (UAV) equipped with a low complexity radar and flying in an unknown environment. The goal is to optimize its trajectory with the purpose of maximizing the mapping accuracy and, at the same time, to avoid areas where measurements might not be sufficiently informative from the perspective of a target detection. This problem is formulated as a Markov decision process (MDP) where the UAV is an agent that runs either a state estimator for target detection and for environment mapping, and a reinforcement learning (RL) algorithm to infer its own policy of navigation (i.e., the control law). Numerical results show the feasibility of the proposed idea, highlighting the UAV's capability of autonomously exploring areas with high probability of target detection while reconstructing the surrounding environment.**

***Index Terms*—Target Detection, Indoor Mapping, Autonomous Navigation, Reinforcement Learning, Unmanned Aerial Vehicles**

I. Introduction

In recent years, unmanned aerial vehicles (UAVs) have gained more and more autonomy, including the capability of taking decisions based on predicting future possible scenarios and learning from past experiences. For this reason, they are often at the center of artificial intelligence applications for both civilian and military use. They have been adopted to help rescuers during emergency situations (e.g., for care delivery or by providing a privileged point of view in search-and-rescue operations) [1], for the prevention of natural disasters [2], or for increasing the safety of citizens in smart cities [3].

In this context, a new challenge is to enrich such small flying robots with radar capabilities for exploring unknown environments and for tracking targets with a higher accuracy. An additional requirement is to accomplish all this by reducing the time needed to complete the mission [4]. Nowadays, such operations are mainly performed by fusing all the information coming from on-board sensors as, for example, vision-based (camera or video) or inertial sensors [5]. Nevertheless, such technologies fail in scarce visibility conditions or in harsh propagation environments, like indoors. A possible solution to overcome such shortcomings is to embed UAVs with millimeter-wave (mm-wave) radars, which due to the reduced wavelength, can be significantly miniaturized and integrated on board, and can improve the detection and tracking capabilities of targets [6].

An autonomous UAV with radar capabilities can be used for different applications. In particular, the scenario investigated in this paper can fit two different situations: an emergency situation where a single UAV-radar reveals the presence of a target (e.g., a survivor who needs assistance or care delivery) in an unknown environment, and a safety situation where the same single UAV-radar is able to detect a malicious (non-cooperative) user, hidden behind obstacles [7], [8]. In both cases, the single UAV should rapidly decide where to explore the environment, while reconstructing a map of it and guaranteeing a robust detection of targets. In this direction, one possibility is the realization of dynamic radar networks by means of swarms of UAVs [9], [10]. This solution is more suitable for outdoor environments where a swarm of multi-agents can be deployed. An alternative is to enhance a single UAV with reinforcement learning (RL) capabilities for indoor autonomous navigation.

The RL concept has been initially proposed several decades ago with the aim of learning a control policy for maximizing a numerical reward signal [11], [12]. According to this paradigm, an agent (e.g., a UAV) determines which actions yield to the largest expected return while accounting for decisions based not only on the immediate reward but also rewards in the future. Operating like this, the exploring agent is capable to learn a policy while accomplishing its task in a non-myopic way.

The concept of using RL with UAVs or radars is not new. In [13]–[16], RL-based approaches for multi-target detection problems have been proposed through the optimization of MIMO radar waveforms. In [17], the Markov decision process (MDP) has been exploited for dual radar-communication problems, whereas in [18], [19] RL has been adopted for autonomous navigation in unknown environments, and in [20] for UAV trajectory planning. UAV trajectory optimization using deep RL has been also investigated in [21], [22] with the purpose of optimizing the spectral efficiency and the communication performance while using a flying UAV-base station, whereas in [23], for learning policy that optimizes the agent trajectory and the tracking performance.

Most of these approaches, however, avoid the reconstruction of the environment (i.e., indoor mapping) whose knowledge may improve the navigation performance. Moreover, the mis-

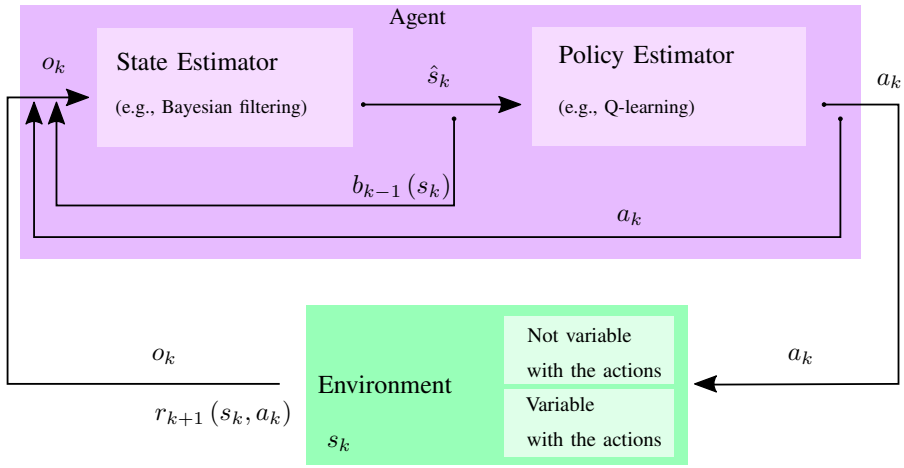

Fig. 1: A belief Markov Decision Process (MDP) considered in the paper with two main blocks for state and control estimation.

sion time constraint is not always taken into account but it is essential for time-critical applications, as in emergency scenarios.

Motivated by this background, in this paper we propose a RL-based approach for UAV autonomous navigation and target detection. Before the navigation step, we consider a prior state estimation step that allows for inference of the map of the environment as well as for the detection of possible targets. Then, the UAV (the agent hereinafter) adopts a RL approach to learn the policy that maximizes the immediate and future rewards, expressed in the form of probability of detection error and mapping accuracy.

The rest of the paper is organized as follows. Sec. II provides a background on MDP, whereas Sec. III and Sec. IV present a description of the considered problem and the proposed solutions, respectively. Sec V reports numerical results, and Sec. VI summarizes the final conclusions.

Notation: A sample space, a random variable (RV) and possible outcomes/values at time instant k are indicated with $\mathcal{X}$, X_k, x_k, respectively. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively; $p(x)$ symbolizes a probability distribution of a random variable x; $p(x|z)$ is the conditional distribution of x given z; $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that $\mathbf{x}$ is distributed according to a Gaussian pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\mathbb{E}\{\cdot\}$ represents the expectation of the argument; $[\cdot]^{\mathsf{T}}$ denotes transpose of the argument.

II. Background on Markov Decision Processes

The problem of learning an optimal policy to be used by an agent when exploring an environment can be formulated as a MDP. Following the same notation as in [11], a MDP is defined by a tuple containing the state space (indicated with $\mathcal{S}$), the action space (i.e., $\mathcal{A}$), the reward space (i.e., $\mathcal{R}$), and the probability of transitioning from one state s_k, at time instant k, to the state s_{k+1} at time $k+1$, defined as

$$p\left(R_{k+1} = r_{k+1}, S_{k+1} = s_{k+1} | S_k = s_k, A_k = a_k\right), \quad (1)$$

satisfying the Markovian property. Notably, the state at time instant k, indicated with S_k, represents the knowledge about the environment available to the agent at time instant k, and can take values $s_k \in \mathcal{S}$. The actions are decided by the agent according to a specific policy given by

$$\pi\left(a_k | s_k\right) \triangleq p\left(A_k = a_k | S_k = s_k\right). \quad (2)$$

The optimal policy is chosen by selecting actions that maximize a state-action value function (or Q-function), defined as

$$Q_\pi\left(s_k, a_k\right) = \mathbb{E}_\pi\left\{\sum_{\ell=0}^{\infty} \gamma^\ell R_{k+\ell+1} \middle| S_k = s_k, A_k = a_k\right\} \quad (3)$$

where $0 \leq \gamma \leq 1$ is a discount rate and, for $\ell = 0$, we have the expected reward at time instant $k+1$, that is [11]

$$\begin{aligned} r_{k+1}(s_k, a_k) &= \mathbb{E}\left[R_{k+1} = r_{k+1} | S_k = s_k, A_k = a_k\right] \\ &= \sum_{r_{k+1} \in \mathcal{R}} r_{k+1} \sum_{s_{k+1} \in \mathcal{S}} p\left(r_{k+1}, s_{k+1} | s_k, a_k\right). \end{aligned} \quad (4)$$

Consequently the problem of finding the optimal policy can be stated as

$$\pi^*\left(a_k | s_k\right) = \arg\max_{a_k} Q_\pi\left(s_k, a_k\right), \quad (5)$$

that can be solved iteratively using RL approaches, e.g., temporal-difference learning [11].

If the state is not known and can be only observed or inferred from noisy measurements, a partially observable Markov decision process (POMDP) or Belief MDP formalism can be used to describe the problem [24], [25]. In order to reduce the processing complexity at the agent, in the next, we will use an approach based on MDP adopting a point estimate of the state instead of its actual value, as it will be detailed in the sequel.

In Fig. 1, a diagram of the system is displayed. The agent is depicted as a block where two estimation processes take place. The first is the "*State Estimator*" which can be implemented using a Bayesian filtering approach, and it provides an

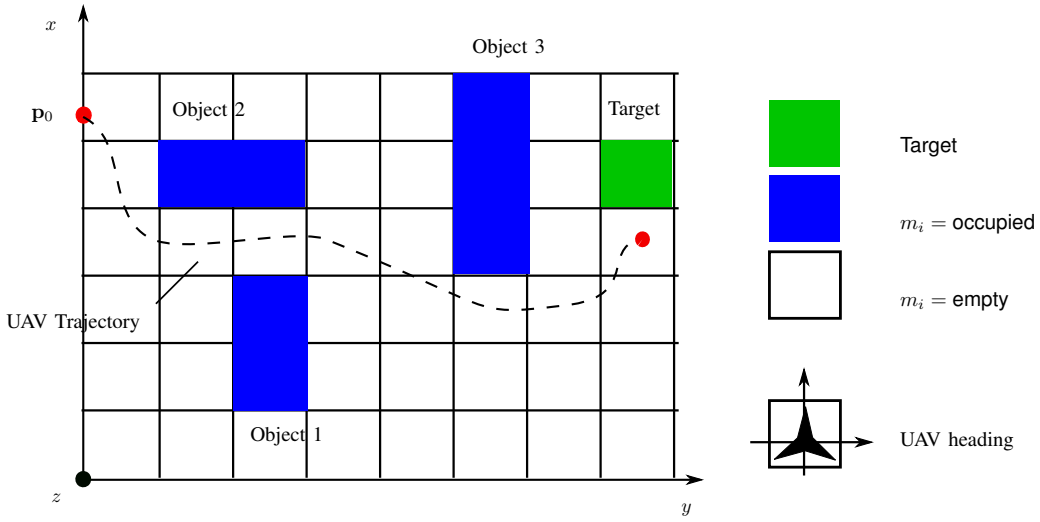

Fig. 2: Considered UAV navigation, mapping and target detection scenario.

estimate $\hat{s}_k$ of the state s_k starting from the previous posterior probability distribution (indicated as $b_{k-1}(s_k)$ in Fig. 1) and the current observation o_k.

The second step is the "*Policy Estimator*", which infers the best action (indicated with a_k) to be taken to maximize the expected return. Once the agent has made a decision, it can evaluate a measure of the goodness of its behaviour by interacting with the environment in its new state. Such a measure, indicated with $r_{k+1}(s_k, a_k)$, could be expressed as a reward or a penalty, and it drives new actions of the agent.

Some actions can change the state (e.g., by modifying the position of the agent in the space), whereas others do not alter the state of the environment (e.g., the presence of obstacles in the surrounding).

III. Problem Statement

We consider here a target detection and mapping problem, performed by a UAV (i.e., the agent), which autonomously navigates an indoor environment. Formally, we state the problem as follows.

$\mathcal{P}_1$: *Detection problem*: Given a fixed maximum time to complete the mission (namely, T_{M}), we aim at minimizing the error in detecting a cooperative target, i.e., minimize the mis-detection and false alarm events. Thus, we define

$$\mathcal{P}_1 : \min P_{\mathsf{e}} = \min \left(P_{01}\, P_1 + P_{10}\, P_0 \right), \tag{6}$$

where $P_{mn} = P\left(\mathcal{D}_m | \mathcal{H}_n\right)$ is the probability of taking the mth decision $\mathcal{D}_m$ when the nth hypothesis $\mathcal{H}_n$ occurs, and $P_m = P\left(\mathcal{H}_m\right)$ is the probability of $\mathcal{H}_m$ being true. In our case, we have

$$\{\mathcal{H}_1, \mathcal{H}_0\} = \{\text{target}, \text{no target}\} \tag{7}$$

$$\{\mathcal{D}_1, \mathcal{D}_0\} = \{\text{decide for } \mathcal{H}_1, \text{decide for } \mathcal{H}_0\}. \tag{8}$$

$\mathcal{P}_2$: *Mapping problem*: Given a fixed maximum time to complete the mission (namely, T_{M}), we aim at minimizing the uncertainty in estimating the map of an unknown environment.

Hereafter, we describe the main assumptions made for solving these two joint problems. First, we consider a grid representation $\mathcal{G} = \{\mathbf{c}_i\}_{i=1}^{N_{\text{cells}}}$ of the environment where each cell of the grid is described by a vector:

$$\mathbf{c}_i = [x_i, y_i, m_i], \quad i = 1, \ldots, N_{\text{cells}}, \tag{9}$$

with N_{cells} being the number of cells with Cartesian coordinates $[x_i, y_i]$.[1] The term m_i represents the *state* of the cell, and we suppose that each cell can have two states, that are $m_i \in \{0, 1\} = \{\text{free}, \text{occupied}\}$. The states are unknown and should be estimated by a mapping process. For example, given the binary nature of the environment, an occupancy grid (OG) algorithm can be used for this purpose as will be explained below. Finally, we consider that the state of the environment does not change with time (static map and target).

Second, the UAV is dynamic and, at each time instant, we define its state as

$$\mathbf{s}_{\mathsf{U},k} = \mathbf{p}_{\mathsf{U},k} = [x_{\mathsf{U},k}, y_{\mathsf{U},k}], \tag{10}$$

with $\mathbf{p}_{\mathsf{U},k} = [x_{\mathsf{U},k}, y_{\mathsf{U},k}] \in \mathcal{G}$ being the true Cartesian position of the UAV at time instant k constrained to lie on the 2D grid.

IV. Target Detection and Mapping

In this section, we describe the main elements characterizing our MDP problem. The objective is to optimize the UAV trajectory considering rewards related to the mapping and detection tasks.

A. State Estimator

The state $\mathbf{s}_k$ is a representation of the system and it consists of three subsets of states:

1) The UAV position, that can be varied by the actions;

[1] Note that a 2D model is used, but its 3D extension is straightforward.

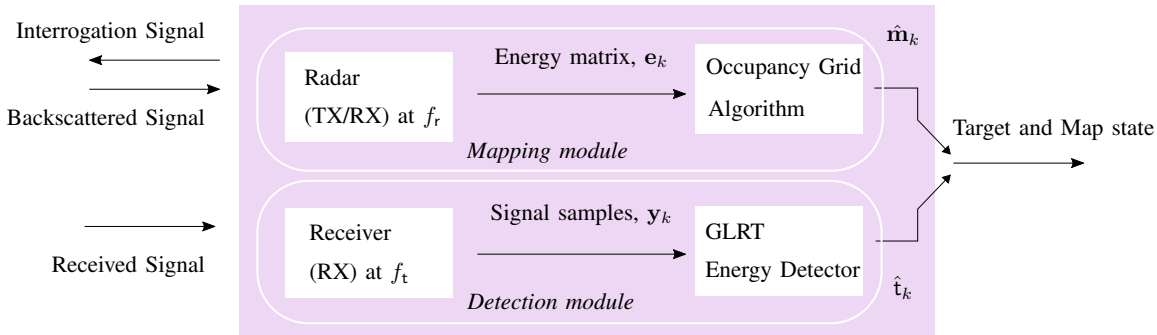

Fig. 3: *State Estimator* at the agent.

2) A binary parameter indicating the presence or absence of a signal source (target) in the environment, which cannot be changed by the actions;
3) The states of each cell, i.e., m_i, (not affected by the actions).

Consequently, we define the state as

$$\begin{aligned}\mathbf{s}_k &= \left[\mathbf{s}_{\mathsf{U},k}, \mathsf{t}, m_1, \ldots, m_i, \ldots, m_{N_{\mathsf{cells}}}\right], \\ &= \left[\mathbf{s}_{\mathsf{U},k}, \mathsf{t}, \mathbf{m}\right], \end{aligned} \tag{11}$$

where $\mathsf{t} \in \{\mathcal{H}_1, \mathcal{H}_0\}$ denotes the presence or not of the target and $\mathbf{m}$ is the true map. The state estimate at time instant k is $\hat{\mathbf{s}}_k = \left[\mathbf{s}_{\mathsf{U},k}, \hat{\mathsf{t}}_k, \hat{\mathbf{m}}_k\right]$ where $\hat{\mathsf{t}}_k \in \{\mathcal{D}_1, \mathcal{D}_0\}$ and $\hat{\mathbf{m}}_k$ are estimated by a detection and mapping modules, respectively, as represented in Fig. 3.

The UAV is equipped with a receiver able to process the signal coming from an active target transmitting at frequency f_t, and with a radar capable of interrogating the environment and operating at a frequency f_r. We also assume that the overall received signals, coming from the target and backscattered by the environment, can be discerned in the frequency domain.

At each time instant k, the UAV performs a scan of the environment by rotating its radar towards different directions in space. In this sense, the possibility of using large antenna arrays at mm-waves allows for a reliable localization [26]–[28] and mapping performance [29] through antennas with limited size [30], [31], and, thus, being apt to be integrated even on small UAVs. To obtain reliable time-of-arrival estimates, one can choose a pulse-based radar. For each steering direction θ_b, where $b = 1, \ldots, N_{\mathsf{rot}}$ is the steering index and N_{rot} is the number of steering directions, the radar emits N_p pulses and collects the related backscattered environment responses. At the same time, the receiver of a detection module receives the signal from the cooperative target, if the target is within the operating range of the RX.

Next, we describe the two estimation processes.

a) Occupancy Grid Mapping: The mapping is performed using energy measurements collected by the radar from each steering direction [29], [32]. The signal received by the radar can be expressed as

$$r(t, \theta_b) = \sum_{n=0}^{N_\mathsf{p}-1} x(t - nT_\mathsf{f}, \theta_b) + n(t), \tag{12}$$

where $x(t, \theta_b)$ is the useful signal acquired when pointing at direction θ_b and $n(t)$ is the additive white Gaussian noise (AWGN) with two-sided power spectral density $N_0/2$.

Successively, the received signal is passed through a band-pass filter with center frequency f_r to eliminate the out-of-band noise, thus producing a filtered signal $y(t, \theta_b)$.

Energy measurements are then computed within a time frame T_f divided into $N_{\mathsf{bins}} = \lfloor T_\mathsf{f}/T_{\mathsf{ED}} \rfloor$ discrete time bins of duration $T_{\mathsf{ED}} \approx 1/W$, with W being the bandwidth of the transmitted signal. Consequently, for each steering direction and for each time bin, the filtered received signal is accumulated over N_p transmitted pulses so that the corresponding final energy value is given by

$$e_{bs} = \sum_{n=0}^{N_\mathsf{p}-1} \int_{(s-1)\,T_{\mathsf{ED}}}^{s\,T_{\mathsf{ED}}} y^2(t + nT_\mathsf{f}, \theta_b)\, dt\,, \tag{13}$$

with $s = 1, 2, \ldots, N_{\mathsf{bins}}$ being the temporal bin index.

Starting from (13) and according to the analysis in [32], the observation vector can be written as[2]

$$\mathbf{e}_k = \left[e_{11,k}, \ldots, e_{bs,k}, \ldots, e_{N_{\mathsf{rot}}\,N_{\mathsf{bins}},k}\right]^\mathsf{T} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{e}_k}, \boldsymbol{\Sigma}_{\mathbf{e}_k}\right), \tag{14}$$

where $\boldsymbol{\mu}_{\mathbf{e}_k}$ is the mean vector and $\boldsymbol{\Sigma}_{\mathbf{e}_k}$ is the measurement covariance matrix whose generic elements are given by [32]

$$\begin{aligned}&\mathbb{E}\left[e_{bs,k}\right] = N_\mathsf{p} \int_{(s-1)\,T_{\mathsf{ED}}}^{s\,T_{\mathsf{ED}}} \tilde{x}^2(t, \theta_b)\, dt + \sigma^2\, N_\mathsf{p}\, T_{\mathsf{ED}} = E_{bs,k} + E_n\,, \\ &\mathrm{var}\left(e_{bs,k}\right) = \sigma^2_{bs,k} = N_0\,\left(2\,E_{bs,k} + E_n\right)\,, \end{aligned} \tag{15}$$

where $N = N_\mathsf{p} N_\mathsf{d}$ is the number of degrees of freedom, with $N_\mathsf{d} = 2WT_{\mathsf{ED}}$, $\sigma^2 = N_0\,W$ is the noise power, and $\tilde{x}(t)$ is the filtered version of $x(t)$. Note that $E_{bs,k}$ depends on the backscattered response of all the map collected when the radar points towards θ_b and, thus, it gathers also the energy

[2]The Gaussian approximation is valid when the number of transmitted pulses is large [33]

contributions coming from all the other spatial directions filtered by the array radiation pattern.

In particular, the considered mapping algorithm accounts for the observation model in (14), and it is based on an extension of the analysis of [34]. Consequently, we have

$$E_{bs,k}(\mathbf{m}) = \sum_{l \in \mathcal{I}(s)} \int_W \frac{L_0(f)\,\rho_l}{(d_{ik})^4} G^2(\tilde{\theta}_l, f)\, df, \tag{16}$$

where $\mathcal{I}(s)$ is the set of cells located at the same discrete distance s from the radar, $L_0(f) = \frac{P_t(f)\,T_{\mathrm{f}}\,c^2}{f^2\,(4\,\pi)^3}$ is the path-loss at the reference distance of 1 meter, $P_t(f)$ is the power spectral density of the transmitted signal, c is the speed of light, $G(\theta)$ is the array gain, $\tilde{\theta}_l = \theta_l - \theta_b$ is the difference between the arrival and steering angles, and ρ_l is the radar cross section (RCS) of the lth cell.

The goal of the UAV is to infer the map of the environment by searching the maximum of the belief given the history of measurements (maximum a-posteriori probability (MAP) estimator), i.e., [35], [36]

$$\hat{\mathbf{m}}_k = \arg\max_{\mathbf{m}} b_k(\mathbf{m}) = \arg\max_{\mathbf{m}} p(\mathbf{m}|\mathbf{e}_{1:k}) \ , \tag{17}$$

where $b_k(\mathbf{m}) = p(\mathbf{m}|\mathbf{e}_{1:k})$ is the posterior of the probability distribution of the map given the set of measurements collected until the discrete time k, i.e., $\mathbf{e}_{1:k}$. According to (17), the mapping problem is described as a maximum a posteriori estimation problem in a high-dimensional space, and, thus, its direct computation is prohibitive. In order to reduce the complexity, instead of computing the joint conditional probability distribution $p(\mathbf{m}|\mathbf{e}_{1:k})$, we operate cell-by-cell as

$$\begin{aligned}\hat{m}_{i,k} &= \arg\max_{m_i} b_k(m_i) = \arg\max_{m_i} p(m_i|\mathbf{e}_{1:k}) \\ &= \arg\max_{m_i} \frac{p(\mathbf{e}_k|m_i)\ b_{k-1}(m_i)}{p(\mathbf{e}_k|\mathbf{e}_{k-1})}.\end{aligned} \tag{18}$$

Moreover, given the binary nature of cells, we can write

$$b_k(m_i = 1) = \frac{p(\mathbf{e}_k|m_i = 1)\, b_{k-1}(m_i = 1)}{p(\mathbf{e}_k|\mathbf{e}_{k-1})} \tag{19}$$

$$\begin{aligned}b_k(m_i = 0) &= 1 - b_k(m_i = 1) \\ &= \frac{p(\mathbf{e}_k|m_i = 0)\, b_{k-1}(m_i = 0)}{p(\mathbf{e}_k|\mathbf{e}_{k-1})}.\end{aligned} \tag{20}$$

Taking the ratio between (19)-(20) and considering a log-odd notation, i.e., $\ell_k(m_i = 1) \triangleq \log\left(\frac{b_k(m_i=1)}{1-b_k(m_i=1)}\right)$, we can obtain

$$\ell_k(m_i = 1) = \log\left(\frac{p(\mathbf{e}_k|m_i = 1)}{p(\mathbf{e}_k|m_i = 0)}\right) + \ell_{k-1}(m_i = 1)\,. \tag{21}$$

In the sequel, we will indicate with $m_{i,1} \triangleq \{m_i = 1\}$ the event for the ith cell of being occupied and with $m_{i,0} \triangleq \{m_i = 0\}$ the opposite event. Recalling the statistical measurement model in (15), we have

$$p(\mathbf{e}_k|m_{i,1}) \propto \exp\left(-\sum_{b=1}^{N_{\mathrm{rot}}}\sum_{s=1}^{N_{\mathrm{bins}}} \frac{(e_{bs,k} - h_{bs,k}(m_{i,1}))^2}{\mathrm{var}(e_{bs,k}|m_{i,1})}\right), \tag{22}$$

$$p(\mathbf{e}_k|m_{i,0}) \propto \exp\left(-\sum_{b=1}^{N_{\mathrm{rot}}}\sum_{s=1}^{N_{\mathrm{bins}}} \frac{(e_{bs,k} - h_{bs,k}(m_{i,0}))^2}{\mathrm{var}(e_{bs,k}|m_{i,0})}\right), \tag{23}$$

where the measurement models $h_{bs,k}(m_{i,1})$ and $h_{bs,k}(m_{i,0})$ are computed using (15) by only considering the contribution of the ith cell as

$$h_{bs,k}(m_{i,1}) = \begin{cases} \int_W \frac{L_0(f)\,\rho_i}{(d_{ik})^4} G^2(\theta_i - \theta_b, f)\, df, & s = s_{i,k} \\ 0, & s \neq s_{i,k} \end{cases}, \tag{24}$$

$$h_{bs,k}(m_{i,0}) = 0, \tag{25}$$

where $s_{i,k} = \left\lfloor \frac{2\,d_{ik}}{c\,T_{\mathrm{ED}}} \right\rfloor$ is the time index where the backscattered signal from the ith cell is expected to arrive. Similarly, $\mathrm{var}(e_{bs,k}|m_{i,1})$ and $\mathrm{var}(e_{bs,k}|m_{i,0})$ can be found by injecting (25) in (15).

In Fig. 4, we present examples of reconstructed maps, using different UAV trajectories and number N of antenna array elements. It is interesting to see how different trajectories result in different mapping accuracies and, thus, one needs a trajectory optimization to enhance the final performance.

b) Target Detection: We now describe the signal processing performed by the detector module of Fig. 3. This module determines if a target is present in the environment. We assume a scenario with unknown deterministic signals (i.e., those transmitted by the target) in AWGN conditions, since multipath is neglected and obstacles can only obstruct the line-of-sight (LOS) component. In addition, we assume that the agent-UAV has knowledge of the target operating bandwidth W. After bandpass filtering over the bandwidth W, the received signal $r(t)$ is sampled at the Nyquist rate, thus obtaining the vector $\mathbf{y} = [y[1], \ldots, y[n], \ldots, y[N]]$, where $N = 2\,T\,W$ is the number of samples,[3] T is the observation time window, and N is an integer. According to the aforementioned definitions, the normalized energy test statistic can be expressed by [37]

$$\frac{2}{N_{0\nu}} \int_0^T [r(t)]^2\, dt \simeq \frac{1}{\sigma_\nu^2} \sum_{n=0}^{N-1} |y[n]|^2\ , \tag{26}$$

where $\sigma_\nu^2 = N_{0\nu}\,W$ represents the noise power and $N_{0\nu}$ is the one-sided noise power spectral density of the receiver of the detector module.

Then, the related discrete time detection problem can be written as

$$\mathcal{H}_0 : \ y[n] = \nu[n], \tag{27}$$

$$\mathcal{H}_1 : \ y[n] = x[n] + \nu[n], \tag{28}$$

[3]In the rest of the manuscript, we assume $N \gg 1$.

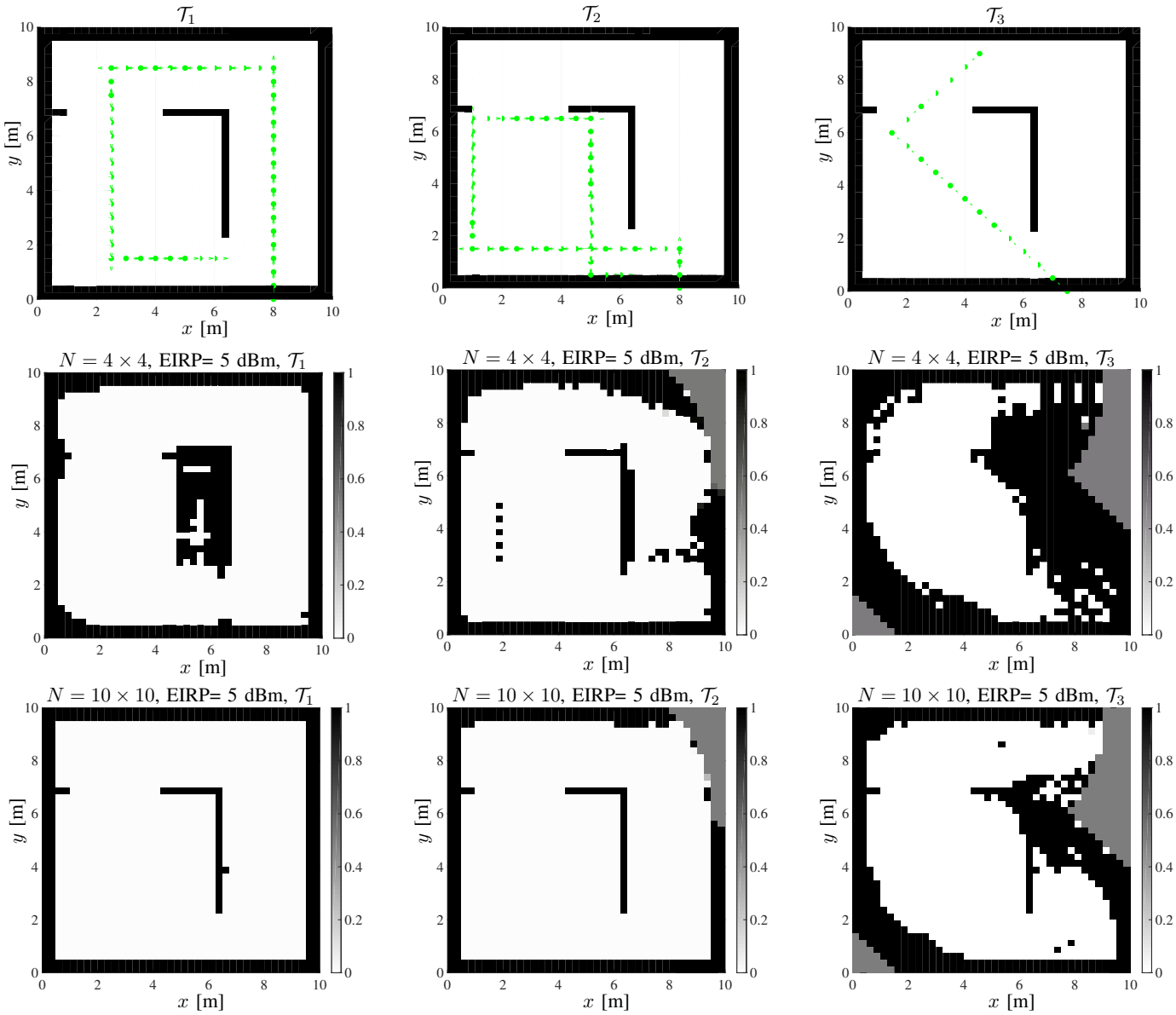

Fig. 4: Reference maps (top) and estimated probabilistic maps for three radar trajectories (namely, $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$) and for $N = 16$ (middle) and $N = 100$ (bottom).

where $x[n]$ and $\nu[n]$ are the nth samples of the low-pass representation of the signal and noise component, respectively, with $\nu[n] \sim \mathcal{N}(0, \sigma_\nu^2)$.

In our case, we consider the normalized energy test [33], [38]

$$\Lambda_{\text{ED}} \triangleq \frac{1}{\sigma_\nu^2} \cdot \sum_{n=1}^{N} |y[n]|^2 \underset{D_0}{\overset{D_1}{\gtrless}} \xi, \tag{29}$$

which represents also the generalized likelihood ratio test (GLRT) when the signal that is being detected is unknown, as is the case here [33], [38]. Notably, in absence of target, we have

$$\Lambda_{\text{ED}} \simeq \frac{1}{\sigma_\nu^2} \cdot \sum_{n=1}^{N} |\nu[n]|^2 , \tag{30}$$

which is distributed according to a central Chi-square distribution with N degrees of freedom. According to the considered model, the probability of false alarm (PFA) is given by [38]

$$P_{\text{FA}} = \frac{\Gamma\left(N, \frac{\xi}{2}\right)}{\Gamma(N)} = \tilde{\Gamma}\left(N, \frac{\xi}{2}\right), \tag{31}$$

where $\Gamma(a, x) = \int_x^\infty x^{a-1} e^{-x} dx$ is the Gamma function, and $\tilde{\Gamma}(\cdot)$ is the Gamma regularized function. According to the Neyman-Pearson criterion, we set the threshold according to

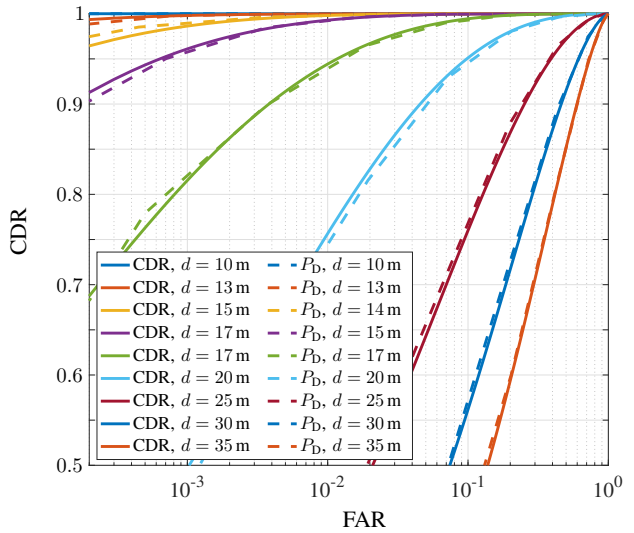

Fig. 5: Continuous line: simulated correct detection rate (CDR) vs. false alarm rate (FAR). Dashed line: theoretical results.

a constraint on the desired $P^{\star}_{\text{FA}}$, so that it can be written by inverting (31) in the form:

$$\xi = 2\left[\text{Inv}\tilde{\Gamma}\left(N, P^{\star}_{\text{FA}}\right)\right], \tag{32}$$

where $\text{Inv}\tilde{\Gamma}(\cdot,\cdot)$ is the inverse gamma regularized function.

According to the defined threshold, the probability of correct detection can be expressed as

$$P_{\text{D}} = \mathcal{Q}_h(\sqrt{\lambda}, \sqrt{\xi}), \tag{33}$$

with $\mathcal{Q}_h$ being the Marcum's $\mathcal{Q}$-function of order $h = N/2$, and $\lambda = \sum_n [x[n]]^2 / \sigma^2_{\nu}$.

B. Actions

In this problem, the actions are the control signals to be applied by the UAV for moving from one cell to another, i.e., $\mathbf{a}_k = \Delta\mathbf{p}_{\text{U},k}$. Since we consider a stationary environment, i.e. $m_{i,k+1} = m_{i,k}$, $\text{t}_{k+1} = \text{t}_k$ $\forall i, k$, here we focus only on the UAV transition dynamics, that is $\mathbf{p}_{\text{U},k+1} = \mathbf{p}_{\text{U},k} + \Delta\mathbf{p}_{\text{U},k}$.

C. Rewards

The expected reward $r(\mathbf{s}_k, \mathbf{a}_k)$ is computed as

$$\begin{aligned} r(\mathbf{s}_k, \mathbf{a}_k) &= \mathbb{E}\left[R_{k+1} = r_{k+1} | \mathbf{s}_k, \mathbf{a}_k\right] \\ &= \sum_{r_{k+1}\in\mathcal{R}} r_{k+1} \sum_{\mathbf{s}_{k+1}\in\mathcal{S}} p\left(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k\right). \end{aligned} \tag{34}$$

In our decision and mapping problem, we define the following rewards:

- r_{map}: Mapping reward. To obtain a quantitative evaluation of the mapping performance, we use the entropy H of the map according to [39]

$$\text{r}_{\text{map},k+1} = \left[H_{k+1|k}(\mathbf{m})\right]^{-1}, \tag{35}$$

where

$$\begin{aligned} H_{k+1|k}(\mathbf{m}) = \sum_{i\in\mathcal{I}} &-b_{k+1|k}(m_{i,1}) \log\left(b_{k+1|k}(m_{i,1})\right) + \\ &- b_{k+1|k}(m_{i,0}) \log\left(b_{k+1|k}(m_{i,0})\right), \end{aligned} \tag{36}$$

represents the entropy of the estimated map from the acquired measurements by time k.

- r_{d}: Detection reward. For each cell, we have

$$\text{r}_{\text{d},k+1} = 1 - P_{\text{e},k+1} \tag{37}$$

where $P_{\text{e},k+1}$ is defined in (6) and is computed for each UAV position, where $P_{10} = P_{\text{FA}}$ and $P_{01} = P_{\text{M}} = 1 - P_{\text{D}}$ are the false alarm and missed detection probabilities, with P_{FA} and P_{D} defined in (31) and (33), and where $P_1 = P(\mathcal{H}_1)$ and $P_0 = P(\mathcal{H}_0)$ are the probabilities of a target in the environment. In our work, we suppose to have the target always present in the environment, so that we have $P_0 = 0$ and $P_1 = 1$. Consequently, (37) becomes $\text{r}_{\text{d},k+1} = P_{\text{D},k+1}$.

Algorithm 1 "Policy Estimator" for inferring π^* [11]

Parameters: Set the discount rate γ
Set the number of future instants (i.e., the horizon) T_{H}
Set the mission time T_{M}
Set the probability of taking a random action ϵ
Initialization: Initialize the map estimate $\hat{\mathbf{m}}_0$;
Set the initial UAV position $\mathbf{s}_{\text{U},0}$;
Initialize the state $\hat{\mathbf{s}}_{k=0} = \left[\mathbf{s}_{\text{U},0}, \hat{\text{t}}_0, \hat{\mathbf{m}}_0\right]$
for *time instant* $k < T_{\text{M}}$ **do**
  Set ϵ as a function of T_{M}, according to (40);
  Generate a random value ϵ_k;
  if $\epsilon_k < \epsilon$ **then**
    Choose a random action $\mathbf{a}_k \in \mathcal{A}$
  else
    For each possible π, evaluate Q_{π} according to the expected rewards $r(\mathbf{s}_k, \mathbf{a}_k)|_{\mathbf{s}_k=\hat{\mathbf{s}}_k}$ until $T_{\text{H}} - 1$;
    Choose an action $\mathbf{a}_k \in \mathcal{A}$ according to (5);
  end
  Agent moves to the new state, , $\mathbf{s}_{\text{U},k+1} = \mathbf{s}_{\text{U},k} + \mathbf{a}_k$;
  Agent acquires measurements (i.e., $\mathbf{e}_k$ and $\mathbf{y}_k$);
  Agent performs state estimation, $\hat{\mathbf{s}}_{k+1}$, using (17) and (30);
end

V. Numerical Results

A. Non Optimized Trajectory - Mapping Results

For the considered case study, we accounted for an EIRP of 5 dBm, a receiver noise figure of 4 dB, and a transmitted pulse with bandwidth of 1 GHz centered at $f_{\text{r}} = 60$ GHz. The number of pulses depends on the fixed scanning time, i.e., $N_{\text{p}} = \lceil T_{\text{obs}}/T_{\text{f}} \rceil$ where $T_{\text{obs}} = T_{\text{scan}}/N_{\text{rot}}$ is the observation time, and $T_{\text{scan}} = 80\,\mu$s is the time needed to perform an entire scan operation. For example, for $N = 4\times4$ ($N = 10\times10$), the steering directions were set to $N_{\text{rot}} = 8$ ($N_{\text{rot}} = 20$) leading to a number of pulses $N_{\text{p}} = 334$ ($N_{\text{p}} = 134$). Figure 4 presents

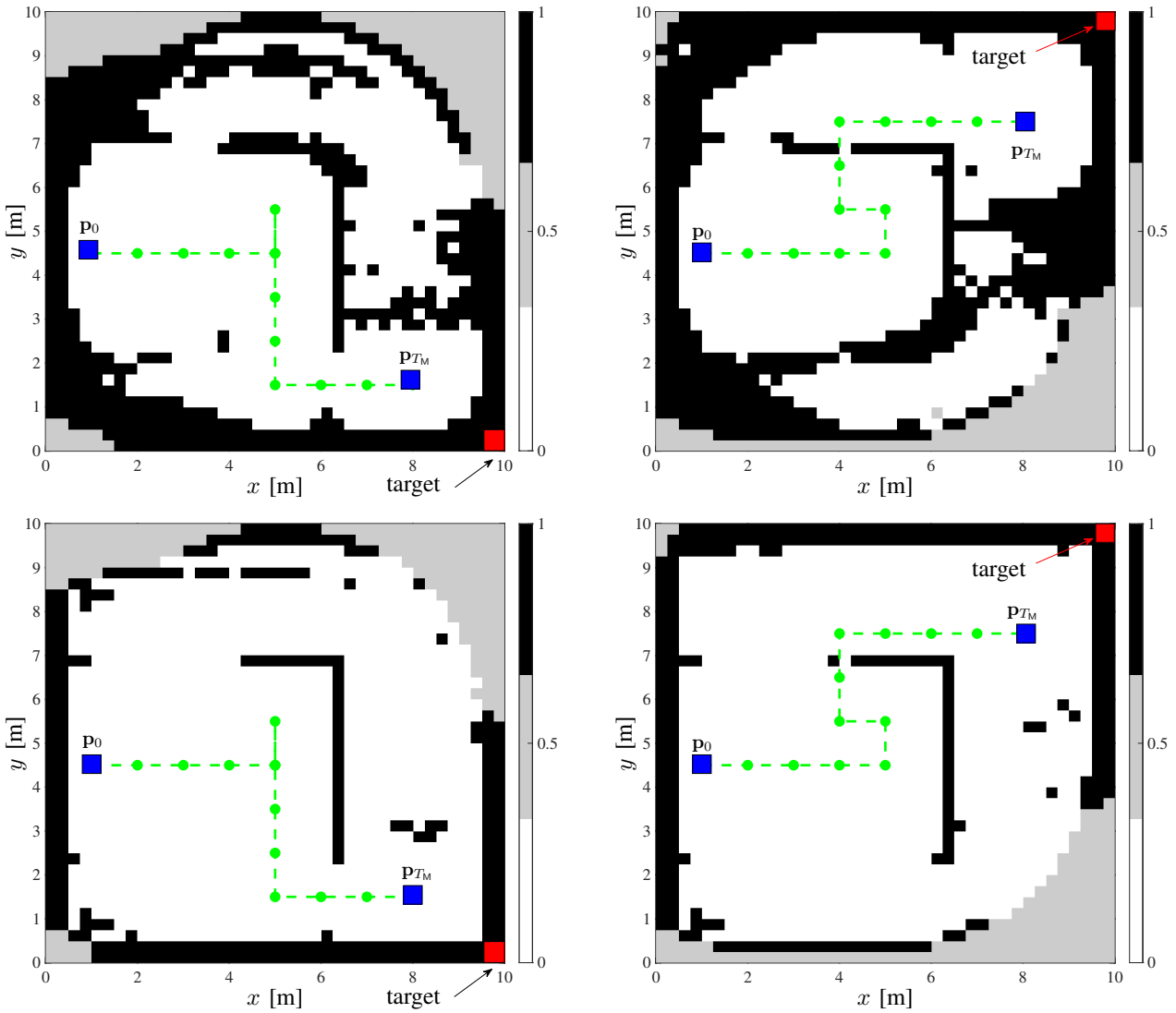

Fig. 6: UAV autonomous navigation for $N = 16$ (top) and $N = 100$ (bottom). Left: target placed on the bottom-right corner. Right: target placed on the bottom-left corner.

the outcomes of the "*State Estimator*" block dedicated to the map reconstruction, i.e., the estimated map $\hat{\mathbf{m}}_k$. The reference maps are shown in Fig. 4-top for different UAV trajectories and mission times, namely $\mathcal{T}_1$-$\mathcal{T}_2$-$\mathcal{T}_3$, whereas the estimated maps are in Fig. 4-middle and Fig. 4-bottom for $N = 4 \times 4$ and $N = 10 \times 10$, respectively. The color scale indicates the probability of occupancy, for example the black cells are occupied with probability equal to 1, whereas the white cells are free. Initially, we suppose to have a complete uncertainty about the map, i.e., $b_0(m_{i,1}) = b_0(m_{i,0}) = 0.5$, $\forall i$. The radar trajectory is depicted with green dots.

As expected, an increased number of antennas, leading to an increased antenna gain and angular resolution, results in a better map reconstruction. Moreover, it is clearly visible that the mapping performance depends on the UAV trajectories and observation time (i.e., on the number of measurement points). Consequently, an optimization of the radar trajectory becomes essential, especially for emergency situations.

B. Target Detection Design

Figure 5 shows the detection results in terms of receiver operating characteristics (ROC) that allows the assessment of the target detection performance as a function of the intended PFA (i.e., of the threshold) and signal-to-noise ratio (SNR).

More specifically, in Fig. 5, we show results of simulations when a target is at different distances from the receiver, $d = [10, 13, 14, 15, 17, 20, 25, 30, 35]$ m. Consequently, the non-central parameter λ changes accordingly. By running the energy test in (30), we have computed the CDR and FAR as

$$\mathsf{CDR} = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbf{1}\left(\Lambda_{\mathsf{ED}} \overset{D_1}{>} \xi | \mathcal{H}_1\right), \tag{38}$$

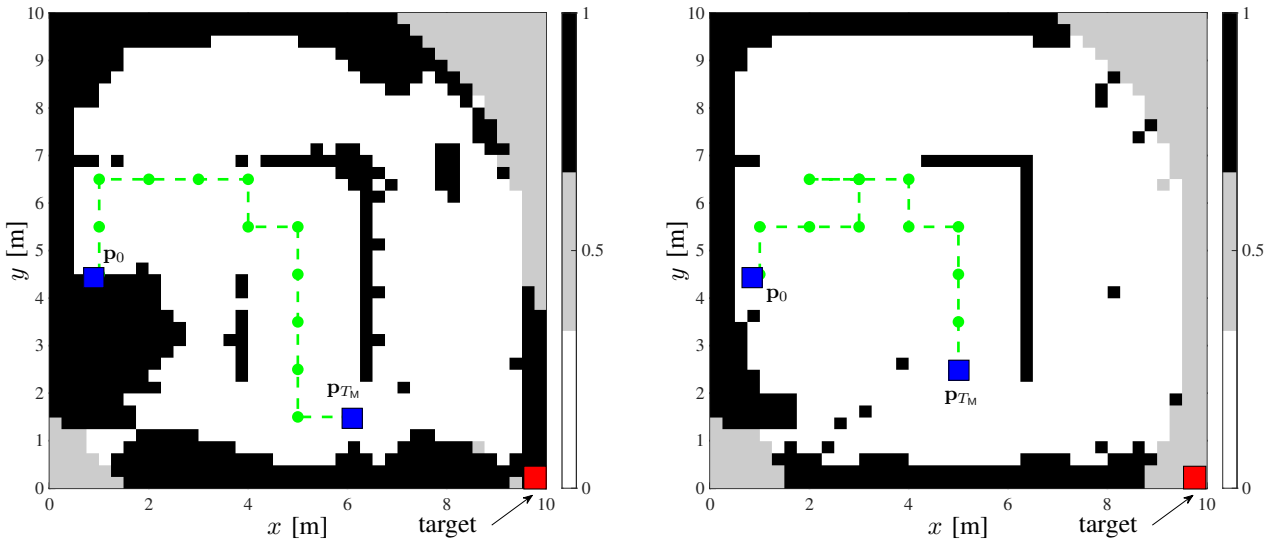

Fig. 7: UAV autonomous navigation for $N = 16$ (left) and $N = 100$ (right) and with a target placed on the bottom-right corner. The exploration rate was set according to the mission time as in (40).

$$\mathsf{FAR} = \frac{1}{N_0} \sum_{i=1}^{N_0} \mathbf{1} \left(\Lambda_{\mathsf{ED}} \overset{D_1}{>} \xi | \mathcal{H}_0 \right), \tag{39}$$

where N_1 and N_0 are the number of times that the target was present and absent in the simulations, with $N_{\mathsf{MC}} = N_1 + N_0$ being the overall number of iterations, and $\mathbf{1}(x)$ is a function that equals to one when its argument is true, and 0 otherwise. The results are compared with theoretical curves of probabilities obtained by (31)-(33).

The target detector performance is used in the sequel to model the detection reward for each action.

C. Optimized Trajectory - Mapping & Detection Results

We now describe the UAV autonomous navigation in order to optimize target detection and environment mapping. If not otherwise indicated, the same parameters of Sec. V-A were considered.

As a first approximation, we consider the agent equipped with a proximity radar for detection of the presence of close obstacles, so that collisions are avoided. The implemented "*Policy Estimator*" is outlined in Algorithm 1 and is inspired by a Q-learning approach, where we set $\gamma = 0.89$, and $T_{\mathsf{H}} = 4$ step. The state was initialized considering $\hat{m}_i = 0.5$, $\forall i$, and $\mathrm{t} = 1$. The mission time was fixed to $T_{\mathsf{M}} = 13$ step, and the radar could perform only lateral and vertical movements of 1 m. The UAV velocity was set to 1 m/step.

For each decision taken at time instant k, the agent accounted for all the possible state combinations up to time instant $k + T_{\mathsf{H}}$. The desired PFA was set to $P^{\star}_{\mathsf{FA}} = 10^{-3}$.

Figure 6 shows four examples, obtained for $N = 16$ (top) and $N = 100$ (bottom). In particular, the scenarios on the left present a target located in $x = 10$, $y = 0$, whereas the two on the right consider a target located in $x = 10$, $y = 10$. The UAV is initially located cyan at $\mathbf{s}_{\mathsf{U},0} = [1, 4.5]$. Interestingly, the agent is capable of reconstructing the environment while moving towards cells closer to the target that exhibit a higher P_{D}, and thus are more advantageous in terms of joint rewards. Even if several cells in the map are still uncertain (i.e., with final belief of 0.5), for $N = 100$, the UAV can reconstruct the environment reliably.

So far, we have investigated the following two scenarios: (i) non-optimized trajectories (see Fig. 4-top); (ii) optimized trajectories for tackling the goal of the mission (see Fig. 6). Even if in the second scenario the agent prioritizes the tasks of the mission, some parts of the map may never be explored. To mitigate such a detrimental effect, we included also a rate of exploration ϵ dependent on the mission time T_{M} as

$$\epsilon = \begin{cases} 0.8 & \text{if} \quad k < T_{\mathsf{M}}/4 \\ 0.4 & \text{if} \quad T_{\mathsf{M}}/4/ \leq k < T_{\mathsf{M}}/2 \\ 0 & \text{Otherwise}\,. \end{cases} \tag{40}$$

By this rule, initially, the agent chooses the actions randomly, whereas for $k > T_{\mathsf{M}}/2$, it evaluates the expected awards and chooses the action accordingly.

In this respect, Fig. 7 shows the obtained curves for $N = 16$ (left) and $N = 100$ (right). In both cases, the final position of the agent is not as optimized for target detection (i.e., it is more distant from the target) with respect to their counterparts in Fig. 6-bottom, but the initial randomness allows for better exploration of the environment.

Such considerations leave an open door for further research of this issue. For example, a more accurate design of ϵ can be performed according to the available T_{M}. Another solution could be based on the inclusion of an ad-hoc reward weight that accounts for the coverage of the map. This last approach is very promising but requires that different types of rewards are suitably weighted in order to form a global reward.

Finally, in the perspective of implementing collaborative approaches for multiple agents, the Q-values can be diffused

among cooperative agents to enhance the learning mechanism.

VI. Concluding Remarks

In this paper, we have investigated the possibility of empowering UAVs with RL capabilities for trajectory optimization. The applications of interest might be an emergency or safety scenario where a single UAV reconstructs a map of an unknown environment while minimizing the error of detecting the presence of a target. In contrast with non-optimized trajectory design (e.g., based on fixed waypoint paths), the results obtained using RL are promising since the agent exhibits interesting capabilities in choosing the trajectory while achieving reliable performance in terms of detection and mapping accuracy in a given mission time. Future work will consider a multiple target scenario, a performance comparison in terms of mission time, and the adoption of a POMDP for taking into consideration the state estimation uncertainty in the decision making process.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie project AirSens (grant no. 793581) and from the PRIMELOC project funded by EU H2020.

References

[1] L. Marconi *et al.*, "The SHERPA project: Smart collaboration between humans and ground-aerial robots for improving rescuing activities in alpine environments," in *Proc. IEEE Int. Symp. Safety, Security, Rescue Robot.*, 2012, pp. 1–4.

[2] A. Merwaday *et al.*, "Improved throughput coverage in natural disasters: Unmanned aerial base stations for public-safety communications," *IEEE Veh. Technol. Mag.*, vol. 11, no. 4, pp. 53–60, 2016.

[3] C. Wang *et al.*, "Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2124–2136, 2019.

[4] P. Hügler *et al.*, "Radar taking off: New capabilities for UAVs," *IEEE Microw. Mag.*, vol. 19, no. 7, pp. 43–53, 2018.

[5] I. Guvenc *et al.*, "Detection, tracking, and interdiction for amateur drones," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 75–81, 2018.

[6] P. Hügler, M. Geiger, and C. Waldschmidt, "77 GHz radar-based altimeter for unmanned aerial vehicles," in *Proc. IEEE Radio Wireless Symp.*, 2018, pp. 129–132.

[7] P. Benavidez and M. Jamshidi, "Mobile robot navigation and target tracking system," in *Proc. 6th Int. Conf. Sys. of Sys. Eng.*, 2011, pp. 299–304.

[8] F. Mohammed *et al.*, "UAVs for smart cities: Opportunities and challenges," in *Proc. Int. Conf. Unmanned Aircraft Sys.*, 2014, pp. 267–273.

[9] A. Guerra, D. Dardari, and P. M. Djuric, "Dynamic radar network of UAVs: A joint navigation and tracking approach," *arXiv preprint arXiv:2001.04560*, 2020.

[10] ——, "Dynamic radar networks of UAVs," *IEEE Veh. Tech. Mag.*, 2020.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[12] M. Kuss and C. E. Rasmussen, "Gaussian processes in reinforcement learning," in *Advances in neural information processing systems*, 2004, pp. 751–758.

[13] B. S. Ciftler, A. Tuncer, and I. Guvenc, "Indoor UAV navigation to a Rayleigh fading source using Q-learning," *arXiv preprint arXiv:1705.10375*, 2017.

[14] L. Wang *et al.*, "Reinforcement learning-based waveform optimization for mimo multi-target detection," in *Proc. IEEE 52nd Asilomar Conf. Signals, Sys., Comput.*, 2018, pp. 1329–1333.

[15] W. Jiang, A. M. Haimovich, and O. Simeone, "End-to-end learning of waveform generation and detection for radar systems," *arXiv preprint arXiv:1912.00802*, 2019.

[16] R. Malhotra, E. P. Blasch, and J. D. Johnson, "Learning sensor-detection policies," in *Proc. IEEE National Aerosp. Electron. Conf.*, vol. 2, 1997, pp. 769–776.

[17] E. Selvi *et al.*, "On the use of Markov decision processes in cognitive radar: An application to target tracking," in *Proc. IEEE Radar Conf.*, 2018, pp. 0537–0542.

[18] H. X. Pham *et al.*, "Autonomous uav navigation using reinforcement learning," *arXiv preprint arXiv:1801.05086*, 2018.

[19] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-uav networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, 2019.

[20] V. Saxena, J. Jaldén, and H. Klessig, "Optimal UAV base station trajectories using flow-level models for reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, 2019.

[21] H. Bayerlein, P. De Kerret, and D. Gesbert, "Trajectory optimization for autonomous flying base station via reinforcement learning," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun.*, 2018, pp. 1–5.

[22] M. Theile *et al.*, "UAV coverage path planning under varying power constraints using deep reinforcement learning," *arXiv preprint arXiv:2003.02609*, 2020.

[23] S. Ragi and E. K. Chong, "UAV path planning in a dynamic environment via partially observable markov decision process," *IEEE Trans. Aerosp. Electron. Sys.*, vol. 49, no. 4, pp. 2397–2412, 2013.

[24] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[25] S. Thrun, "Monte carlo POMDPs," in *Adv. Neural Info. Process. Sys.*, 2000, pp. 1064–1070.

[26] A. Guerra, F. Guidi, and D. Dardari, "Single-anchor localization and orientation performance limits using massive arrays: MIMO vs. beamforming," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5241–5255, 2018.

[27] A. Shahmansoori *et al.*, "Position and orientation estimation through millimeter-wave MIMO in 5G systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1822–1835, 2018.

[28] N. Vukmirović *et al.*, "Direct wideband coherent localization by distributed antenna arrays," *Sensors*, vol. 19, no. 20, p. 4582, 2019.

[29] F. Guidi, A. Guerra, and D. Dardari, "Personal mobile radars with millimeter-wave massive arrays for indoor mapping," *IEEE Trans. Mobile Comput.*, vol. 15, no. 6, pp. 1471–1484, Jun. 2016.

[30] C. Jouanlanne *et al.*, "Wideband linearly polarized transmitarray antenna for 60 GHz backhauling," *IEEE Trans. Antennas Propag.*, vol. 65, no. 3, pp. 1440–1445, 2017.

[31] S. Ghosh and D. Sen, "An inclusive survey on array antenna design for millimeter-wave communications," *IEEE Access*, vol. 7, pp. 83 137–83 161, 2019.

[32] A. Guerra *et al.*, "Occupancy grid mapping for personal radar applications," in *Proc. Stat. Signal Process. Workshop*, 2018, pp. 766–770.

[33] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[34] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Autonomous robots*, vol. 15, no. 2, pp. 111–127, 2003.

[35] S. Thrun, "Learning occupancy grids with forward models," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)*, vol. 3, 2001, pp. 1676–1681 vol.3.

[36] C. Robbiano *et al.*, "Bayesian learning of occupancy grids," *arXiv preprint arXiv:1911.07915*, 2019.

[37] F. Guidi *et al.*, "Joint energy detection and massive array design for localization and mapping," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1359–1371, 2016.

[38] A. Mariani, A. Giorgetti, and M. Chiani, "Effects of noise power estimation on energy detection for cognitive radio applications," *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3410–3420, Dec. 2011.

[39] N. Mafi, F. Abtahi, and I. Fasel, "Information theoretic reward shaping for curiosity driven learning in pomdps," in *Proc. IEEE Int. Conf. Develop. Learning*, vol. 2, 2011, pp. 1–7.