Real-Time Learning for THz Radar Mapping and UAV Control

A. Guerra[⋆], F. Guidi[⋄], D. Dardari[⋆], P. M. Djurić[†]

[⋆] University of Bologna, UNIBO, Italy
[⋄] National Research Council, CNR-IEIIT, Italy
[†] Stony Brook University, SBU, USA

Email: anna.guerra3@unibo.it

ICAS 2021 – IEEE Conference on Autonomous Systems.

Outline

- Research Motivations
- Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- State Estimator: Occupancy Grid Mapping
- Policy Estimator: Off-policy Q-Learning
- Simulation Results
- Take-aways

Outline

- ◇ Research Motivations
- ◇ Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- ◇ State Estimator: Occupancy Grid Mapping
- ◇ Policy Estimator: Off-policy Q-Learning
- ◇ Simulation Results
- ◇ Take-aways

Why UAVs

◇ Advantages

- Flying sensors able to offer a privileged point-of-view for sensing;
- Autonomous, flexible, and quick to react;
- Able to access impervious or dangerous areas (e.g., mountains, oceans, etc.)

◇ Disadvantages

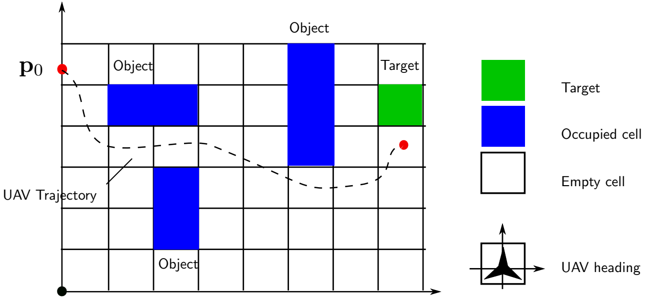
- Battery constrained;
- Lightweight and low-cost on-board sensors;
- Ethical issues when used with AI.

Possible Applications

Outline

- ◇ Research Motivations
- ◇ Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- ◇ State Estimator: Occupancy Grid Mapping
- ◇ Policy Estimator: Off-policy Q-Learning
- ◇ Simulation Results
- ◇ Take-aways

UAV for Detection and Mapping



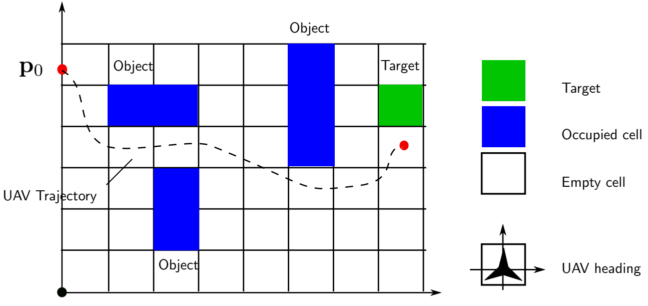
Given a fixed maximum time to complete the mission,

$\mathcal{G}_1$: *Detection*: maximize the probability of detection;

$\mathcal{G}_2$: *Mapping*: maximize the mapping coverage and accuracy.

- *State Estimator:* GLRT for Target Detection and Occupancy grid mapping;
- *Control:* Q-learning.

UAV for Detection and Mapping



Given a fixed maximum time to complete the mission,

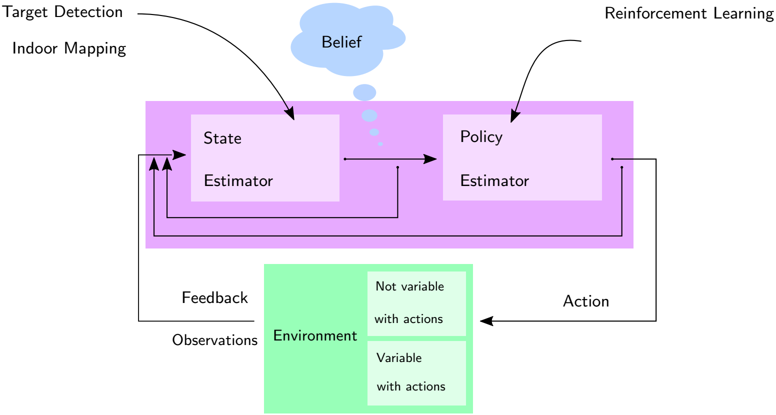
$\mathcal{G}_1$: *Detection*: maximize the probability of detection;

$\mathcal{G}_2$: *Mapping*: maximize the mapping coverage and accuracy.

- *State Estimator:* GLRT for Target Detection and Occupancy grid mapping;
- *Control:* Q-learning.

Recall on Markov Decision Processes (MDPs)

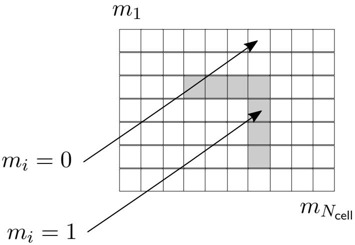
- In time-critical applications, models for navigation are often not available
- UAVs have to learn from the environment (trial and error);
- Interaction UAV-environment represented with Markov Decision Processes

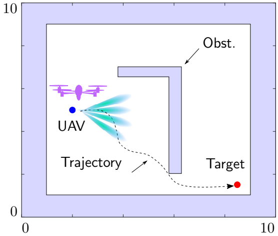


Recall on MDPs - Cont'd

◇ State Space

- ☐ The state vector is $\mathbf{s}^{(k)} = \left[\mathbf{p}^{(k)}, \mathbf{m}^{(k)}, \mathrm{t}^{(k)}\right]^T$
- ☐ $\mathbf{p}^{(k)} = \left[x^{(k)}, y^{(k)}, h\right]$ is the UAV position, that can be varied by the actions;
- ☐ $\mathrm{t}^{(k)} \in \mathbb{B}^2$ indicates the presence or absence of a target, estimated by a detection module. In the next, $\mathrm{t}^{(k)} = \mathrm{t} = 1$.
- ☐ $\mathbf{m}^{(k)} = \mathbf{m} = [m_1, \ldots, m_i, \ldots, m_{N_{\text{cell}}}]^T \in \mathbb{B}^{N_{\text{cell}}}$ is a map of the environment, estimated by the mapping module.





Recall on MDPs - Cont'd

◇ Action Space

- The UAV action is defined as $\mathbf{a}_k = \Delta\mathbf{p}_k = [\Delta x_k, \Delta y_k]^{\mathsf{T}} \in \mathbb{R}^2$ in terms of position displacement $\Delta\mathbf{p}_k$;
- $\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{a}_k$: Next UAV position;
- $\Delta\mathbf{p}_k$ is set according to $N_{\mathrm{a}} = 4$ actions;

$$\mathcal{A} = \Big\{ \underbrace{[\Delta,\, 0]}_{\text{Right}}, \underbrace{[-\Delta,\, 0]}_{\text{Left}}, \underbrace{[0,\, \Delta]}_{\text{Up}}, \underbrace{[0,\, -\Delta]}_{\text{Down}} \Big\}.$$

Recall on MDPs - Cont'd

◇ Reward Space

- r_{k+1}: Reward at $k+1$ related to the action $\mathbf{a}_k$ and the current state $\mathbf{s}_k$;
- r_{task}: Extrinsic/Task reward $\rightarrow$ *detection*;
- r_{int}: Intrinsic reward $\rightarrow$ *mapping*;
- η: Normalizing factor;

$$r_{k+1} = \underbrace{r_{\mathsf{i},\,k+1}}_{\text{Intrinsic reward}} + \eta \underbrace{r_{\mathsf{e},\,k+1}}_{\text{Extrinsic Reward}},$$

- $r_{\mathsf{i},\,k+1} = r_{\mathsf{c},k+1} + r_{\mathsf{m},k+1}$ is an intrinsic reward used for obtaining a sufficient knowledge of the surrounding environment;
- $r_{\mathsf{e},\,k+1} = r_{\mathsf{d},k+1}$ is a reward for the considered unmanned aerial vehicle (UAV) task.

Reward Shaping

- **Intrinsic Reward**: *Detection Rate*;

$$r_{\text{det},k} \triangleq f_D(\sqrt{\lambda_k},\ \sqrt{\xi}),$$

  - $f_D(\cdot)$ is a specific function depending on the particular detector statistic
  - λ_k is the measured signal-to-noise ratio (SNR) at time instant k
  - ξ is a threshold depending on the $P^{\star}_{\text{FA}}$.

- **Extrinsic Reward:** *Map entropy and coverage*;

$$r_{\text{map},k} \triangleq \frac{H_{k+1|k}(\mathbf{m})}{|\mathcal{I}_k|} \qquad r_{\text{cov},k} \triangleq \frac{1}{N_{\text{cells}}} \sum_{i \in \mathcal{I}_k} \mathbf{1}(i \in \mathcal{D}_k)$$

  - $H(\mathbf{m}) = -\sum_{i \in \mathcal{I}} b(m_i) \log_2 (b(m_i))$ is the map entropy
  - $\mathcal{I}_k$ set of illuminated cells at time instant k
  - $\mathcal{D}_k$ set of illuminated cells seen for the first timeat time instant k.

Reward Shaping

◇ **Intrinsic Reward**: *Detection Rate*;

$$r_{\mathsf{det},k} \triangleq f_D(\sqrt{\lambda_k},\ \sqrt{\xi}),$$

☐ $f_D(\cdot)$ is a specific function depending on the particular detector statistic
☐ λ_k is the measured SNR at time instant k
☐ ξ is a threshold depending on the $P_{\mathsf{FA}}^{\star}$.

◇ **Extrinsic Reward:** *Map entropy and coverage*;

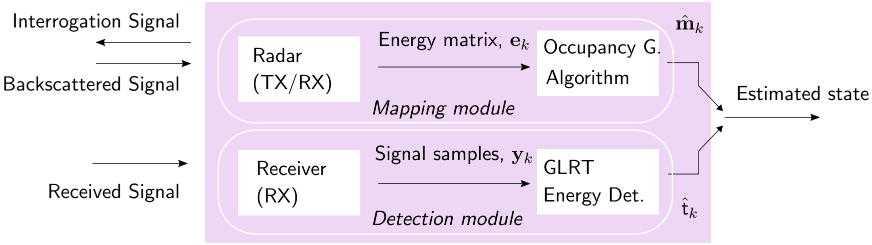
$$r_{\mathsf{map},k} \triangleq \frac{H_{k+1|k}(\mathbf{m})}{|\mathcal{I}_k|} \qquad\qquad r_{\mathsf{cov},k} \triangleq \frac{1}{N_{\mathsf{cells}}} \sum_{i \in \mathcal{I}_k} \mathbf{1}(i \in \mathcal{D}_k)$$

☐ $H(\mathbf{m}) = -\sum_{i \in \mathcal{I}} b(m_i)\ \log_2(b(m_i))$ is the map entropy
☐ $\mathcal{I}_k$ set of illuminated cells at time instant k
☐ $\mathcal{D}_k$ set of illuminated cells seen for the first timeat time instant k.

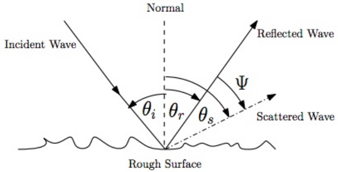
Outline

- ◇ Research Motivations
- ◇ Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- ◇ State Estimator: Occupancy Grid Mapping
- ◇ Policy Estimator: Off-policy Q-Learning
- ◇ Simulation Results
- ◇ Take-aways

State Estimator

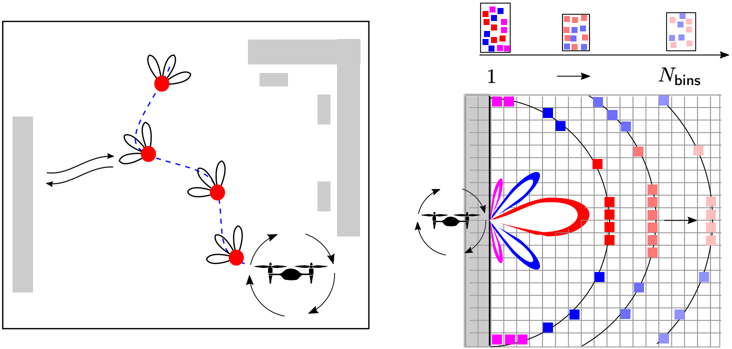


- ◇ Terahertz Radar at 140 GHz [1];
- ◇ 100 virtual antennas;
- ◇ Capability to operate in scarce visibility conditions;
- ◇ *Output:* range-angle matrix;



Ju, Shihao, et al. "Scattering mechanisms and modeling for terahertz wireless communications." ICC 2019-2019. IEEE, 2019.

UAV-Radar Mapping



- Goal: A UAV equipped with a MIMO radar explores an unknown environment and estimate a map of it;
- Interrogation Phase: For each steering direction, a train of pulses is transmitted;
- Measurement Phase: Backscattered energy measurements are accumulated in a Range-Angle matrix;
- Estimation Phase: From the Range-Angle matrix, the map is estimated using an OG algorithm.

UAV-Radar Mapping - Observations

- Scanning operation with N_{steer} beamsteering angles θ_b.
- Measurements: range-angle matrix $\mathbf{e}^{(k)}$ containing the accumulated measured energy at a certain time instant k

$$\mathbf{e}^{(k)} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{b1} & \cdots & e_{N_{\text{steer}}1} \\ e_{12} & e_{22} & \cdots & e_{b2} & \cdots & e_{N_{\text{steer}}2} \\ \vdots & & & & & \\ e_{1s} & e_{2s} & \cdots & e_{bs} & \cdots & e_{N_{\text{steer}}s} \\ \vdots & & & & & \\ e_{1N_{\text{bins}}} & e_{2N_{\text{bins}}} & \cdots & e_{bN_{\text{bins}}} & \cdots & e_{N_{\text{steer}}N_{\text{bins}}} \end{bmatrix}$$

- Statistical observation model: the measurement model at the radar is $\mathbf{z}^{(k)} = \mathbf{g}(\mathbf{m}) + \mathcal{N}\left(0, \mathbf{R}^{(k)}\right)$
  - $\mathbf{g}(\mathbf{m}) = [g_{11}(\mathbf{m}), \ldots, g_{bs}(\mathbf{m}), \ldots, g_{N_{\text{steer}}N_{\text{bins}}}(\mathbf{m})]$, $g_{bs}(\mathbf{m})$: radar range equation accounting for THz scattering model [1];
  - $\mathbf{R}^{(k)}$ is the covariance diagonal matrix whose generic element is given by σ^2_{bs}

UAV-Radar Mapping - Occupancy Grid

- Bayesian algorithm in three main steps using the following log-odd notation

$$\ell_i^{(k)}\left(m_i^{(k)}\right) \triangleq \log\left(\frac{p\left(m_i^{(k)}=1|\mathbf{z}^{(1:k)}\right)}{p\left(m_i^{(k)}=0|\mathbf{z}^{(1:k)}\right)}\right) = \log\left(\frac{p\left(m_i^{(k)}=1|\mathbf{z}^{(1:k)}\right)}{1-p\left(m_i^{(k)}=1|\mathbf{z}^{(1:k)}\right)}\right)$$

- Initialization The map is initialized with $p\left(m_i^{(k)}=1|\mathbf{z}^{(1:k)}\right)=p\left(m_i^{(k)}=0|\mathbf{z}^{(1:k)}\right)=0.5$ (complete uncertainty);

- Measurement update A new energy matrix is collected for each steering direction and time bin. The likelihood functions $p\left(\mathbf{z}^{(k)}|m_i=1\right)$ and $p\left(\mathbf{z}^{(k)}|m_i=0\right)$ are computed.

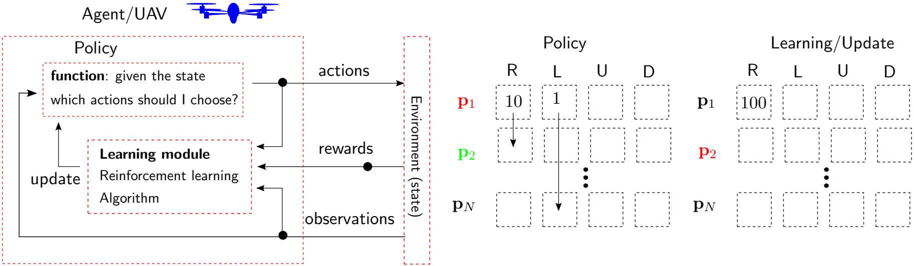
- Log-odd update the belief of the map is updated according to

$$\ell_k\left(m_i\right)=\log\left(\frac{p\left(\mathbf{z}^{(k)}|m_i\right)}{1-p\left(\mathbf{z}^{(k)}|m_i\right)}\right)+\ell_{k-1}\left(m_i\right).$$

Outline

- Research Motivations
- Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- State Estimator: Occupancy Grid Mapping
- Policy Estimator: Off-policy Q-Learning
- Simulation Results
- Take-aways

Model–free RL for UAV Control



- ◇ Tabular Policies: with discrete (few) number of actions and states, the policy can be represented as a table;
- ◇ With a Q-table, the policy is to check the value of every possible action given the current state and then choose the action with the highest value;

Q-Learning

Algorithm 1: Q-Learning Navigation for a Single Episode

Parameters: Set the learning parameters $(\gamma, \alpha, \epsilon)$ and the mission time T_{M};
Initialization: Initialize the Q-table to zeros and the initial state s_0 ;
while $k < T_M$ **do**
  Generate a random value ϵ_k;
  if $\epsilon_k < \epsilon$ **then**
    Choose a random action $\mathbf{a}_k \in \mathcal{A}$;
  else
    Choose the action $\mathbf{a}_k \in \mathcal{A}$ that corresponds to the maximum Q-value in $Q(\mathbf{s}_k, :)$;
  end
  UAV moves to the new state, collects the reward r_{k+1} and updates the Q-table according to

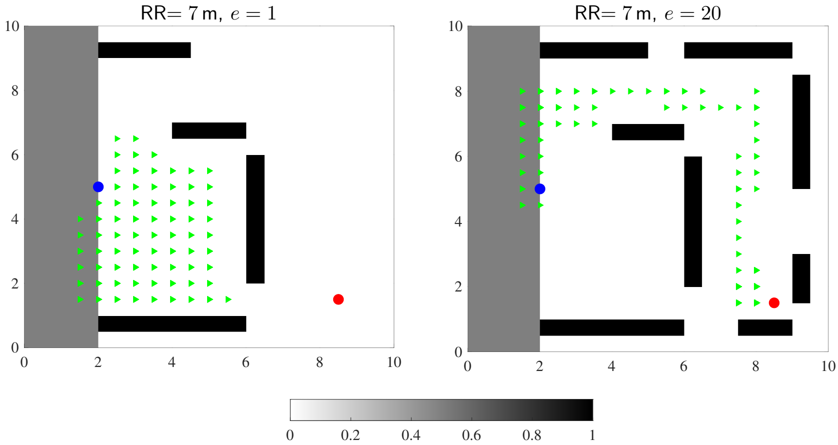
$$Q(\mathbf{s}_k, \mathbf{a}_k) \leftarrow Q(\mathbf{s}_k, \mathbf{a}_k) + \alpha \left[r_{k+1} + \gamma \max_{\mathbf{a}} Q(\mathbf{s}_{k+1}, \mathbf{a}) - Q(\mathbf{s}_k, \mathbf{a}_k) \right]$$

end

Outline

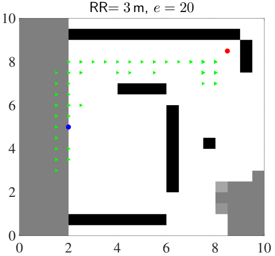
- ◇ Research Motivations
- ◇ Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- ◇ State Estimator: Occupancy Grid Mapping
- ◇ Policy Estimator: Off-policy Q-Learning
- ◇ Simulation Results
- ◇ Take-aways

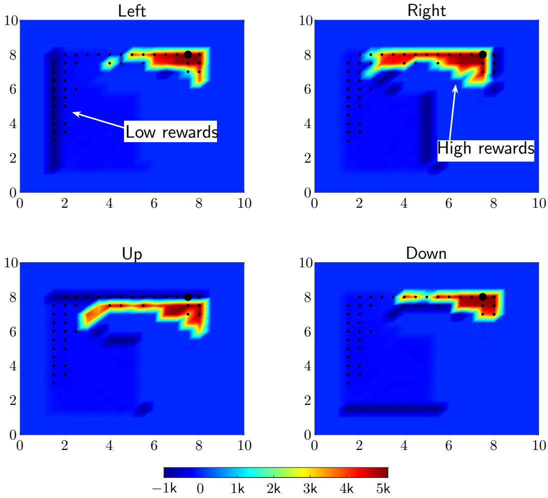
Simulation Results



Examples of estimated trajectories and maps for $e = 1$ (left) and $e = 20$ (right). Blue and red markers indicate the initial UAV and the target position, respectively.

Simulation Results - Cont.'d.





Outline

- ◇ Research Motivations
- ◇ Problem Formulation
  - Environment Mapping & Target Detection with UAVs
  - Recall on Markov Decision Processes
- ◇ State Estimator: Occupancy Grid Mapping
- ◇ Policy Estimator: Off-policy Q-Learning
- ◇ Simulation Results
- ◇ Take-aways

Conclusions

- ◇ UAVs are a promising technology to realize dynamic wireless sensor/radar networks;
- ◇ UAVs can be intelligent: optimizing their trajectory according to the assigned tasks;
- ◇ Q-learning approach with a combination of intrinsic and extrinsic rewards for target detection and environment mapping;
- ◇ Mapping aided by on-board THz radar allows for an enhanced ambient awareness;
- ◇ *Next Steps:* Distributed multi-agent learning for multi–target detection with large networks of UAVs.

Thank you

UAV-Radar Mapping - Observation Model, Cont'd

- Measurement model: $\mathbf{z}^{(k)} = \mathbf{g}(\mathbf{m}) + \mathcal{N}\left(0, \mathbf{R}^{(k)}\right)$
- The generic element $g_{bs}(\mathbf{m})$ is given by the radar range equation as

$$g_{bs}(\mathbf{m}) = \sigma^2 T_{\mathrm{ED}} N_{\mathrm{p}} + T_{\mathrm{f}} \sum_{i \in \mathcal{R}(s)} \frac{P_t\, c^2\, \rho_i^2\, G^2\, N_{\mathrm{p}}}{f^2\, (4\pi)^3 d_i^4}$$

  - $T_{\mathrm{ED}} \approx 1/W$ is the duration of a bin
  - T_{f} is the duration of a time frame
  - $\mathcal{R}(s)$ is the number of cell located at a distance d_i
  - P_t is the transmitted power
  - ρ_i^2 is the radar cross section
  - G is the radar antenna gain
  - d_i is the distance
  - N_{p} is the number of transmitted pulses
  - $\sigma^2 = N_0 W$ is the noise variance