Multi-Agent Navigation with Reinforcement Learning Enhanced Information Seeking

Siwei Zhang*, Anna Guerra†, Francesco Guidi†, Davide Dardari‡ and Petar M. Djurić§

*Institute of Communications and Navigation, German Aerospace Center (DLR), Germany
Email: siwei.zhang@dlr.de
†WiLAB, CNR-IEIIT, Italy
Email: firstname.lastname@ieiit.cnr.it
‡WiLAB, DEI Guglielmo Marconi, CNIT, University of Bologna, Italy
Email: davide.dardari@unibo.it
§Electrical and Computer Engineering Department, Stony Brook University, USA
Email: petar.djuric@stonybrook.edu

***Abstract*—Multi-agent robotic networks allow simultaneous observations at different positions while avoiding a single point of failure, which is essential for emergency and time-critical applications. Autonomous navigation is vital to the task accomplishment of a multi-agent network in challenging global navigation satellite systems (GNSS)-denied environments. In these environments, agents can rely on inter-agent measurements for self-positioning. In addition, agents can conduct information seeking, i.e., they can proactively adapt their formation to enrich themselves with position information. Classical signal processing tools can efficiently exploit the knowledge of system and measurement models, but are not applicable for long-term objectives. On the other hand, data-driven approaches like reinforcement learning (RL) are suitable for long-term action planning but have to face the critical curse of dimensionality. In this paper, we propose a multi-agent navigation scheme with RL-enhanced information seeking, which simultaneously takes advantage of model-based and data-driven approaches to collaboratively accomplish challenging objectives while exploring a GNSS-denied environment.**

I. Introduction

Multi-agent networks have attracted an ever increasing attention in sensing and exploration applications, thanks to the increased exploration efficiency due to collaboration and the capability of observation from different points of view. The ability of ubiquitous navigation is essential for a multi-agent network. A typical navigation problem involves positioning, i.e., estimating the position of an entity like a vehicle, human or robot, and goal approaching, i.e., guiding this entity from one place to another. As specific to multi-agent navigation, collaboration among agents enhances the navigation capability by cooperative positioning and formation optimization [1], [2]. Eventually, for multi-agent information seeking [3], [4], agents can proactively adapt their formation, so that the positioning uncertainty is actively minimized while approaching a goal.

Classic signal processing methods can be employed for multi-agent information seeking [3], [5]. In [4], [6], the authors propose the use of projected steepest gradient descent (PSGD) for efficient position information seeking. In [7], PSGD is utilized for dynamic radar networks. The PSGD method utilizes the closed-form expression of gradients to efficiently exploit the knowledge of system and measurement models. However, its step-wise optimization nature makes the PSGD method not readily applicable for long-term multi-objectives missions.

When positioning is associated with control, data-driven approaches capable of learning from feedback received from the environment, such as reinforcement learning (RL), can be a viable solution. In [8], [9], RL is employed for multi-target detection, whereas [10] studied the trajectory optimization of multi-agent-radar for environment mapping and detection using an RL-based approach. Indeed, data-driven approaches are appealing as agents can learn a sequence of suitable actions to achieve multiple mission objectives, but they might suffer the curse of dimensionality in large-scale multi-agent networks.

In this paper, we propose a RL-enhanced multi-agent navigation method, combining the PSGD-based information seeking with Q-learning. This method preserves the advantages of both approaches in order to collaboratively accomplish challenging navigational objectives. To demonstrate the effectiveness of our proposed method, we consider a conceptual Mars swarm exploration mission shown in Figure 1, similar to that in [4], [6]. A swarm of rovers (i.e., agents) need to navigate themselves, with high positioning accuracy, from their landing site denoted as region A to an exploration site denoted as region B. In each region three anchor nodes with known positions, such as a lander and static rovers, are deployed with limited signal coverage area marked with magenta curves. Agents in the middle area outside the coverage of anchors cannot conduct agent-to-anchor measurement. Therefore, this area is referred to as the blind region. The mission is considered as successful if all agents reach region B, while keeping their position uncertainty below a desired limit. To this end, the swarm must autonomously adapt its formation into a dynamic rigid bridge [11] connecting regions A and B and enable all agents pass through the blind region, just like army ants assembling a "living" bridge to overcome gaps on their foraging trails in nature [12]. With this investigation, we aim at shedding light on the generic combination of model-based and data-driven

approaches, which is vital for decision making in a large-scale autonomous multi-agent network like a robotic swarm.

II. Multi-Agent Navigation

A. Network and Measurement Models

We assume a network composed of $N+M$ nodes in a set $\mathcal{V}$, including N agents in set $\mathcal{A}$ and M anchors in set $\mathcal{B}$. Agents aim at navigating themselves from region A, centered at $\mathbf{p}_A$, to region B, centered at $\mathbf{p}_B$. All positions are constrained on a horizontal two-dimensional (2D) surface. Autonomous agents are the core components of a multi-agent navigation network. At time instant k, an agent $u \in \mathcal{A}$ is located at a position $\mathbf{p}_u^{(k)} = [x_u^{(k)}, y_u^{(k)}]^T$, which needs to be estimated. The transition of an agent's position between two consecutive time instants is described by a mobility model

$$\mathbf{p}_u^{(k+1)} = f\left(\mathbf{p}_u^{(k)}, \mathbf{b}_u^{(k)}\right) + \boldsymbol{\omega}_u^{(k)}, \quad \forall u \in \mathcal{A}, \tag{1}$$

where $\mathbf{b}_u^{(k)} \in \mathcal{U}$ is the control command of agent u within a feasible control set $\mathcal{U}$ at time instant k, and $\boldsymbol{\omega}_u^{(k)}$ is additive state transition noise. We combine the control commands of all agents to $\mathbf{b}^{(k)} = \mathrm{vec}\{\mathbf{b}_u^{(k)} : \forall u \in \mathcal{A}\}$, and denote the collective feasible control set as $\mathcal{U}$. The operator $\mathrm{vec}\{\cdots\}$ arranges the variables into a column vector. An agent u emits a signal and receives the signals emitted from anchors and agents within its coverage. These nodes are included in the neighbor set of u, denoted as $\mathcal{V}_u$. A collection of all agents' positions at time instant k is denoted as $\mathbf{p}^{(k)} = \mathrm{vec}\{\mathbf{p}_u^{(k)} : \forall u \in \mathcal{A}\}$.

An agent u obtains a vector of measurements $\mathbf{z}_u^{(k)}$ from the signals emitted by neighbors. Each measurement contains the geometric relationship between u and v with v being a node in the neighboring set of u, i.e., $v \in \mathcal{V}_u$. We assume time-based ranging measurements. Hence, the measured distance $z_{uv}^{(k)}$ between agent u and its neighboring node v at time instant k is modeled as the true distance $d_{uv}^{(k)}$ distorted with zero mean additive white Gaussian noise. The noise variance is quadratically proportional to $d_{uv}^{(k)}$ until reaching the coverage limit, and then rapidly increases [4], [13]. This model captures, with simplicity, the main properties of time-based ranging under line-of-sight (LOS) condition and is sufficient for our study of multi-agent navigation. A collection of all measurements in the network at time instant k is denoted as $\mathbf{z}^{(k)} = \mathrm{vec}\{\mathbf{z}_u^{(k)} : \forall u \in \mathcal{A}\}$.

B. Objectives of Multi-Agent Navigation

Cooperative positioning aims at finding a position estimate $\hat{\mathbf{p}}^{(k)}$ at time instant k, that minimizes, for example, the overall mean-square error of the agents' positions, given the already obtained measurements $\mathbf{z}^{(k)}$. For information seeking, instead of passively utilizing the already obtained measurements, agents proactively move with a control command $\mathbf{b}^{(k)}$ into new positions $\mathbf{p}^{*(k+1)}$. At those positions, future measurements $\mathbf{z}^{*(k+1)}$ can be acquired, which minimize the (weighted) mean-square error of the agents' new position estimates $\hat{\mathbf{p}}^{(k+1)}$. We apply the Fisher information theory and utilize the Cramér-Rao bound (CRB) to predict the expected mean-square error at the new positions [1]:

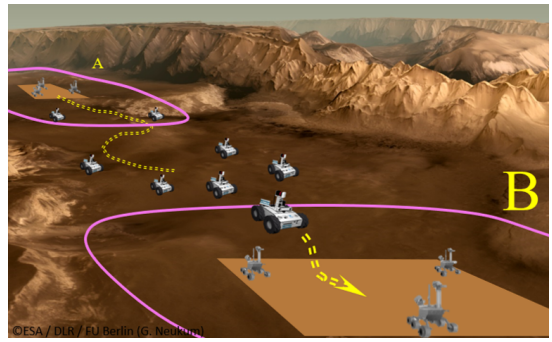


Figure 1: A conceptual Mars swarm exploration mission.

$$\mathrm{CRB}[\mathbf{p}^{*(k+1)}] \preccurlyeq \mathbb{E}\Big[(\hat{\mathbf{p}}^{(k+1)} - \mathbf{p}^{*(k+1)})(\hat{\mathbf{p}}^{(k+1)} - \mathbf{p}^{*(k+1)})^T\Big],$$

where $\preccurlyeq$ reads as 'less positive semi-definite'. Note that $\mathrm{CRB}[\mathbf{p}^{*(k+1)}]$ is expressed as a function of the control command $\mathbf{b}^{(k)}$ in closed-form in [4], which is advantageous for analytical assessment. Then, the information seeking criterion can be formulated as minimizing, with best effort, a scalar function of $\mathrm{CRB}[\mathbf{p}^{*(k+1)}]$, for example

$$f_{\mathrm{s}}(\mathbf{b}^{(k)}) = \mathrm{tr}\Big[\boldsymbol{\Lambda}^{(k)}\ \mathrm{CRB}[\mathbf{p}^{*(k+1)}]\Big], \tag{2}$$

where the diagonal weighing matrix $\boldsymbol{\Lambda}^{(k)} = \mathrm{diag}\{\lambda_u^{(k)} \otimes \mathbf{1}_{2\times 1} : \forall u \in \mathcal{A}\}$ assigns different significance to agents according to their positions in the network. The operator $\mathrm{diag}\{\cdots\}$ arranges variables into a diagonal matrix. For safety-related objectives, like collision avoidance, position uncertainty has to be limited with a higher priority. In this case, the information seeking criteria can be formulated as constraints, i.e.

$$h_{\mathrm{s},u}(\mathbf{b}^{(k)}) = \varepsilon - \mathrm{tr}\Big[\boldsymbol{\Gamma}_u^{(k)}\ \mathrm{CRB}[\mathbf{p}^{*(k+1)}]\Big] \geq 0,\ \forall u \in \mathcal{A},$$

where ε is the maximum tolerated mean square error and $\boldsymbol{\Gamma}_u^{(k)} = \mathrm{diag}\{[\mathbf{0}_{1\times u-1}, 1, \mathbf{0}_{1\times N-u}]^T \otimes \mathbf{1}_{2\times 1}\}$. Besides information seeking, other navigation criteria could be introduced including the goal approaching objective defined by

$$f_{\mathrm{g}}(\mathbf{b}^{(k)}) = \sum_{u\in\mathcal{A}} \Big\|\mathbf{p}_u^{*(k+1)}(\mathbf{b}^{(k)}) - \mathbf{p}_{\mathrm{B}}\Big\|,\ \forall u \in \mathcal{A}. \tag{3}$$

The overall navigation problem can then be formulated as a constrained multi-objective optimization problem:

$$\underset{\mathbf{b}^{(k)}\in\mathcal{U}}{\text{minimize}} \quad \Big\{f_{\mathrm{s}}(\mathbf{b}^{(k)}), f_{\mathrm{g}}(\mathbf{b}^{(k)})\Big\} \tag{4a}$$

$$\text{s.t.} \quad h_{\mathrm{s},u}(\mathbf{b}^{(k)}) \geq 0,\ \forall u \in \mathcal{A}. \tag{4b}$$

III. RL-Enhanced Information Seeking

A. PSGD-based Information Seeking

The multi-agent navigation problem formulated in (4) is a high dimensional non-convex optimization problem. Instead of finding the optimal solution in one step, a PSGD method with low complexity is proposed, which is suitable for large-scale

multi-agent navigation [4], [7]. We generate control commands with negative gradients of the cost functions of information seeking $f_{\mathrm{s}}(\mathbf{b}^{(k)})$ and goal approaching $f_{\mathrm{g}}(\mathbf{b}^{(k)})$, respectively, and linearly combine them to obtain an unconstrained control command $\mathbf{b}^{*(k)}$, i.e.,

$$\mathbf{b}^{*(k)} = -\mu \mathbf{W}^{(k)} \frac{\nabla_{\mathbf{b}^{(k)}} f_{\mathrm{s}}}{\|\nabla_{\mathbf{b}^{(k)}} f_{\mathrm{s}}\|} - \mu(\mathbf{I} - \mathbf{W}^{(k)}) \frac{\nabla_{\mathbf{b}^{(k)}} f_{\mathrm{g}}}{\|\nabla_{\mathbf{b}^{(k)}} f_{\mathrm{g}}\|}, \quad (5)$$

where μ is the step size and $\mathbf{W}^{(k)} = \mathrm{diag}\{w_u^{(k)} \otimes \mathbf{1}_{2\times 1} : \forall u \in \mathcal{A}\}$ is the trade-off weight between information seeking and goal approaching. Then we identify the activated constraint vector $\mathbf{h}(\mathbf{b}^{(k)})$, i.e., a collection of constraints that are being violated or at the boundary of violation [14, Ch. 5]. The constraint gradient matrix $\mathbf{N}^{(k)}$ is defined as

$$\mathbf{N}^{(k)} = \nabla_{\mathbf{b}^{(k)}} \mathbf{h}(\mathbf{b}^{(k)})^T|_{\mathbf{b}^{(k)}=\mathbf{0}}. \quad (6)$$

The projection matrix $\mathbf{P}^{(k)}$ defined by

$$\mathbf{P}^{(k)} = \mathbf{I} - \mathbf{N}^{(k)} \Big((\mathbf{N}^{(k)})^T \mathbf{N}^{(k)} \Big)^{-1} (\mathbf{N}^{(k)})^T \quad (7)$$

projects the unconstrained control command $\mathbf{b}^{*(k)}$ onto the tangent space of the activated constraints, i.e.,

$$\mathbf{b}^{(k)} = \mathbf{P}^{(k)} \mathbf{b}^{*(k)} - \mathbf{N}^{(k)} \Big((\mathbf{N}^{(k)})^T \mathbf{N}^{(k)} \Big)^{-1} \mathbf{h}(\mathbf{0}). \quad (8)$$

The PSGD-based multi-agent navigation exploits the domain knowledge of signal processing where control commands can be generated efficiently for large-scale agent networks. However, it is a greedy approach with step-by-step optimization, which cannot guarantee convergence to the global optimum for non-convex problems. Besides, it cannot explicitly involve multi-step or long-term objectives like "all agents reaching the goal". Last but not least, the control parameters like the weight on each agent's position in information seeking $\lambda_u^{(k)}$ and the trade-off between information seeking and goal approaching for each agent $w_u^{(k)}$ have to be chosen manually at the beginning and are often fixed during the mission. It becomes a bottleneck when the control rules have to be adjusted due to situation changes. These drawbacks of a PSGD-based method can be overcome by data-driven approaches like RL, where the ultimate mission objectives like goal approaching can be explicitly set as a long-term reward.

B. Information Seeking Enhanced by Q-Learning

Q-learning is the most popular model-free algorithm for tabular-based reinforcement learning [15]. The main goal of Q-learning is to learn a, so called, Q-function that maps a state-and-action pair to a return value, i.e., Q-value, evaluating the benefit, in terms of learning objectives, of taking such an action at the given state. One of the most successful applications of Q-learning is goal directing such as path planning and maze solving, which share great similarity with our goal approaching mission objective. In most of the applications, the actions are directly set as movements of agents, e.g. a step in a cardinal direction like {N,E,S,W}. This action choice is not exploiting any model knowledge, ergo a model-free approach.

We propose a combination of PSGD and Q-learning in order to benefit from the efficiency of model-based approaches and the long-term reward seeking ability of data-driven methods. Instead of learning directly favourable movements, we propose to employ Q-learning on formulating dynamic navigational objectives, i.e. (4). Then, the agents' movements are generated with PSGD method given the learnt objectives. With this approach, agents simultaneously benefit from the efficiency of model-based methods and the long-term reward seeking ability of data-driven methods.

a) State

It is well known that Q-learning is impractical for systems with high dimensionality like our multi-agent network. Therefore we need to choose a low-dimensional training state space which captures essential situation information of an agent. We select the agent-to-destination distance $d_{uB}^{(k)} = \|\mathbf{p}_u^{(k)} - \mathbf{p}_B\|$ and the mean distance to destination $d_{SB}^{(k)} = \sum_{u\in\mathcal{A}} \|\mathbf{p}_u^{(k)} - \mathbf{p}_B\| / N$ as the training state, i.e. $\mathbf{s}_u^{*(k)} = [d_{uB}^{(k)}, d_{SB}^{(k)}]^T$. These two dimensions describe the situation of the agent under investigation with respect to (w.r.t.) the multi-agent network and the exploration environment. The agent distance can be estimated with a positioning algorithm. The mean distance can be obtained in a decentralized fashion through a consensus algorithm. If needed, higher order statistical moments of the multi-agent formation can be added into the state space that more collective characteristics of the formation can be considered. For a tabular approach like Q-learning, the state is discretized to $\mathbf{s}_u^{(k)}$.

b) Action

Instead of choosing a cardinal direction to move, we propose a multi-agent Q-learning, where each agent learns to select appropriate control parameters, namely the agent's information seeking weight $\lambda_u^{(k)}$ and the information seeking-to-goal approaching trade-off ratio $w_u^{(k)}$. By doing so, agents are constantly adapting their implicit roles between explorers and supporters. The action space contains the discretized control parameters, i.e., $\mathbf{a}_u^{(k)} = [\lambda_u^{(k)}, w_u^{(k)}]^T$.

c) Q-table update

In the investigated scenario, we focus on a homogeneous multi-agent network, where the agents' explicit roles are identical, even though they may have different implicit roles based on their current situation. Hence, the agents are interchangeable also from a learning perspective. Therefore, every agent maintains its own Q-table $Q_u(\mathbf{s}_u^{(k)}, \mathbf{a}_u^{(k)}), \forall u \in \mathcal{A}$ and exchanges it with neighboring agents to achieve a Q-table consensus over the network. We utilize the advantage of Q-learning and directly set the instantaneous reward according to the ultimate mission objective, for example, all agents reaching region B with a continuous exponential reward function:

$$r(\mathbf{s}_u^{(k)}, \mathbf{a}_u^{(k)}) = \beta \exp\Big(-\max\Big\{ \big(d_{uB}^{(k+1)}\big)^2 / d_{\mathrm{R}}^2 : \forall u \in \mathcal{A} \Big\} \Big),$$

where β is a scaling factor and d_R is the radius defining the proximity of the destination point. The Q-table is updated, for

example at agent u, with the following procedure. First, if the state has changed, i.e., $\mathbf{s}_u^{(k)} \neq \mathbf{s}_u^{(k-1)}$, generate new actions $\mathbf{a}_u^{(k)}$ with an ϵ-greedy algorithm [15] either randomly with probability ϵ_i or from the Q-table with probability $1-\epsilon_i$. The subscript i indicates the training runs $i = 1, \cdots, E$, also known as episodes. Second, apply PSGD-based navigation with the chosen $\lambda_u^{(k)}$ and $w_u^{(k)}$ and get the new state $\mathbf{s}_u^{(k+1)}$. Third, if the state has changed, i.e., $\mathbf{s}_u^{(k+1)} \neq \mathbf{s}_u^{(k)}$, update the Q-table according to the Bellman equation [15]:

$$Q_u(\mathbf{s}_u^{(k)}, \mathbf{a}_u^{(k)}) \leftarrow \alpha_i \Big(r(\mathbf{s}_u^{(k)}, \mathbf{a}_u^{(k)}) + \gamma \max_{\mathbf{a}} Q_u(\mathbf{s}_u^{(k+1)}, \mathbf{a}) \Big) + (1-\alpha^{(k)})\, Q_u(\mathbf{s}_u^{(k)}, \mathbf{a}_u^{(k)}), \tag{9}$$

where α_i is the learning rate, decaying over episodes to guarantee convergence to an optimal solution [16], and γ is a discount factor. The convergence of the Q-table will be achieved after multiple episodes.

IV. Numerical Results

We verify the proposed RL-enhanced multi-agent navigation by simulating the Mars swarm exploration mission illustrated in Figure 1. We use a similar simulator as in [4] with a further developed Q-learning ingredient. Areas A and B are 3000 m apart from each other. A 30×30 grid is used to discretize the training state. Signals from the anchors can be detected and exploited for ranging by an agent up to 1000 m. Hence, the blind region spans over 1000 m between regions A and B. The agents can effectively range with each other up to a distance of 400 m. Eight agents depart around the origin $\mathbf{p}_A = [0, 0]^T$ and need to reach the proximity of the destination $\mathbf{p}_B = [3000, 0]^T$ with $d_{uB}^{(K)} < d_{\mathrm{R}}, \forall u \in \mathcal{A}$ within $K = 2000$ steps, where $d_{\mathrm{R}} = 100\,\mathrm{m}$. The position CRB of every agent is constrained to $\varepsilon = 100\,\mathrm{m}^2$. The agents are trained for $E = 10,000$ episodes, with a learning rate, at the ith episode, $\alpha_i = i^{-1/2}$ and an exploration factor ϵ_i which linearly decreases from 1 to 0.1 in the first 2000 episodes and remains 0.1 till the end of training for ϵ-greedy action selection. The discount factor γ is set to 0.9. The entities in Q-table initialized at the first time instant of the first episode uniformly within the interval $(0, 10^{-2}]$. The scaling factor of the instantaneous reward is $\beta = 10^4$. The agent's weight for information seeking $\lambda_u^{(k)}$ can be selected from the discrete set $\{0.01, 0.99\}$, whereas the information seeking-to-goal approaching from $\{0.3, 0.7\}$.

Figure 2 depicts the learning traces of the state, including the agent-to-destination distance $d_{uB}^{(k)}$ along the x-axis and the mean distance to destination $d_{SB}^{(k)}$ along the y-axis. In total of 500 unsuccessful episodes in black and 100 successful episodes in green are plotted. The magenta dashed lines indicate the borders of regions A and B with the blind region. In one episode, the traces evolve from the lower left corner (around the origin) aiming to reach the upper right corner (around the destination). There are two areas marked in red at the borders where the green traces diverge from the black traces. These areas reveal a decisive moment to the success of the mission

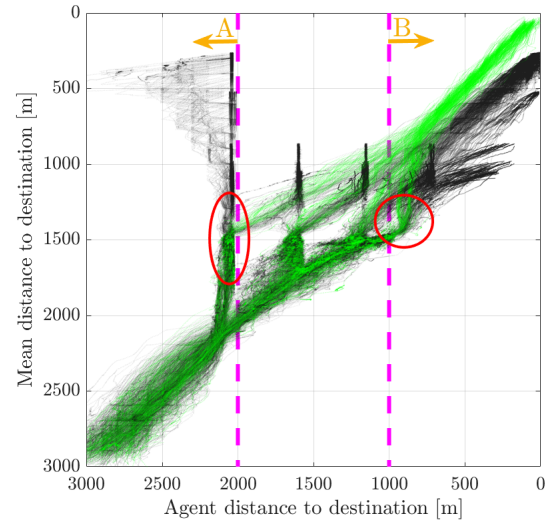


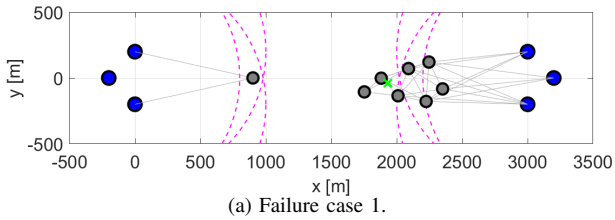
Figure 2: Traces of learning state of 500 unsuccessful episodes in black and 100 successful episodes in green.

corresponding to the establishment of the bridge connecting regions A and B. Next, the agents behind (in the left marked area) need to be ready to step outside the coverage of region A, i.e., in favor of goal approaching. In the meantime, the front agents (in the right marked area) have to wait and support the agents behind on their positioning.

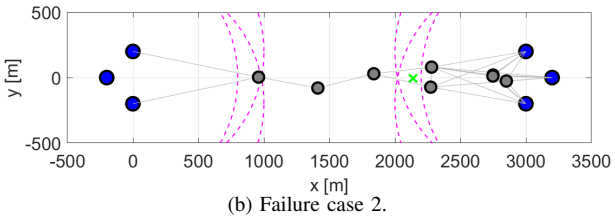
In Figure 3, two unfavorable formations from PSGD-only approach are shown. After the front agents pass the blind region, they do not wait and support the remaining ones crossing the blind region. In the first case shown in Figure 3a, all agents except the left one will reach the goal, whereas in the second case shown in Figure 3b, all agents remain in these final positions and are unable to approach the goal under constrained positioning uncertainty.

In Figure 4 we selected two crucial formation snapshots resulted from the RL-enhanced information seeking. Agents are illustrated with black and white markers. Markers' colors show the trade-off factor between information seeking and goal approaching. A white marker indicates $w_u = 0.7$ whereas a black marker indicates $w_u = 0.3$. The markers' sizes show the weight of the agent in information seeking. A small marker indicates $\lambda_u = 0.01$, whereas a large marker indicates $\lambda_u = 0.99$. At time instant 1470, the agents behind are the "pillars" for keeping the bridge rigid, but yet ready to move forward. The agents in front just reach region B. At time instant 1500, the front agents change their implicit roles from goal approaching to information seeking with a main objective of supporting the agents behind to cross the blind region.

Last, but not least, we analyse the statistical behavior of learning by plotting the number of total episodes versus the successful episodes in Figure 5. The numbers above the curves denote the slopes of the curve, i.e., the successful rates. Here we can see that the successful rate starts from 16% at the beginning of learning, i.e. applying the PSGD-only method, increases from 4500 episodes on and converges to 64%. This

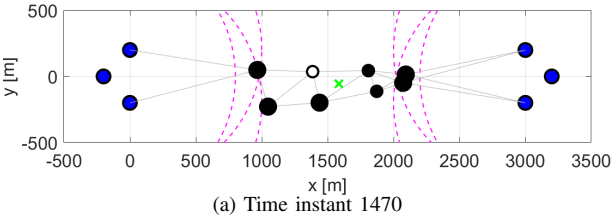


(a) Failure case 1.

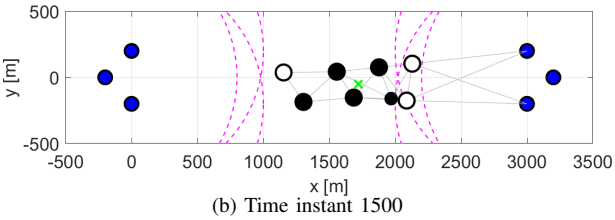


(b) Failure case 2.

Figure 3: Formations from PSGD-only information seeking with agents illustrated with gray markers.



(a) Time instant 1470



(b) Time instant 1500

Figure 4: Formations from RL-enhanced information seeking. Agents are illustrated with black and white markers whose colors and sizes indicate the taken actions, i.e. $w_u^{(k)}$ and $\lambda_u^{(k)}$, respectively.

result verifies the effectiveness of our proposed RL-enhanced multi-agent information seeking scheme. For a real-world mission, we may conduct training in simulated environment first to learn an initial Q-table to speed up online training during the operation.

V. Conclusion

In this paper, we investigate a multi-agent navigation problem. Due to highly varied scenarios and the high system dimensionality, neither classic model-based signal processing approaches nor emerging data-driven RL approaches are solely suitable to solve this navigation problem. We propose a RL-enhanced multi-agent information seeking method, which benefits from the efficiency of model-based PSGD method and the long-term objective compatibility of model-free RL. As an outcome, the agents effectively learn to change their implicit roles under different situations in order to collaboratively accomplish a challenging navigation task. In a case study of a conceptual Mars swarm exploration mission, the successful rate increases from 16% with the PSGD-only approach to 64% with the RL enhancement.

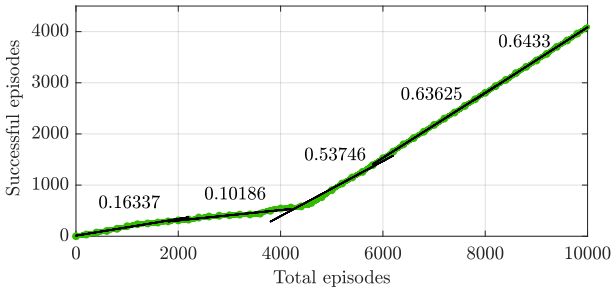


Figure 5: Number of total episodes versus number of successful episodes. Numbers above the curve indicate the successful rates.

More importantly, with this paper we shed light on the generic combination of model-based and data-driven approaches, which becomes increasingly important for decision making in an autonomous large-scale multi-agent system.

References

[1] M. Z. Win, Y. Shen, and W. Dai, "A theoretical foundation of network localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1136–1165, Jul. 2018.

[2] R. M. Buehrer, H. Wymeersch, and R. M. Vaghefi, "Collaborative sensor network localization: Algorithms and practical issues," *Proc. IEEE*, vol. 106, no. 6, pp. 1089–1114, Jun. 2018.

[3] F. Meyer, H. Wymeersch, M. Fröhle, and F. Hlawatsch, "Distributed estimation with information-seeking control in agent networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 11, pp. 2439–2456, Nov. 2015.

[4] S. Zhang, R. Pöhlmann, T. Wiedemann, A. Dammann, H. Wymeersch, and P. A. Hoeher, "Self-aware swarm navigation in autonomous exploration missions," *Proc. IEEE*, vol. 108, no. 7, pp. 1168–1195, 2020.

[5] F. Morbidi and G. L. Mariottini, "Active target tracking and cooperative localization for teams of aerial vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 5, pp. 1694–1707, 2013.

[6] S. Zhang, M. Frohle, H. Wymeersch, A. Dammann, and R. Raulefs, "Location-aware formation control in swarm navigation," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

[7] A. Guerra, D. Dardari, and P. M. Djurić, "Dynamic radar network of UAVs: A joint navigation and tracking approach," *IEEE Access*, vol. 8, pp. 116 454–116 469, 2020.

[8] A. M. Ahmed *et al.*, "A reinforcement learning based approach for multi-target detection in massive MIMO radar," *IEEE Trans. Aerosp. Electron. Syst.*, pp. 1–1, 2021.

[9] P. Liu *et al.*, "Decentralized automotive radar spectrum allocation to avoid mutual interference using reinforcement learning," *IEEE Trans. Aerosp. Electron. Syst.*, 2020.

[10] A. Guerra, F. Guidi, D. Dardari, and P. M. Djurić, "Multi-agent Q-learning in UAV networks for target detection and indoor mapping," in *Proc. Int. Balkan Conf. Commun. Netw. (BalkanCom)*, 2021, pp. 80–84.

[11] J. Aspnes, T. Eren, D. K. Goldenberg, A. S. Morse, W. Whiteley, Y. R. Yang, B. D. O. Anderson, and P. N. Belhumeur, "A theory of network localization," *IEEE Trans. Mobile Comput.*, vol. 5, no. 12, pp. 1663–1678, Dec. 2006.

[12] C. R. Reid *et al.*, "Army ants dynamically adjust living bridges in response to a cost–benefit trade-off," *Proceedings of the National Academy of Sciences*, 2015.

[13] W. M. Gifford, D. Dardari, and M. Z. Win, "The impact of multipath information on time-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 70, pp. 31–46, 2022.

[14] R. T. Haftka and Z. Gürdal, *Elements of Structural Optimization*. Springer Science+Business Media B.V., 1984.

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

[16] C. J. C. H. Watkins and P. Dayan, "Technical note: Q-learning," *Mach. Learn.*, vol. 8, no. 3–4, p. 279–292, May 1992.