Enabling Safe and Efficient Separation through Multi-Agent Reinforcement Learning

Groot, D.J.; Ellerbroek, Joost; Hoekstra, J.M.

Publication date
2022

Document Version
Final published version

Citation (APA)
Groot, D. J., Ellerbroek, J., & Hoekstra, J. M. (2022). *Enabling Safe and Efficient Separation through Multi-Agent Reinforcement Learning*. 1-4. Paper presented at 10th International Conference for Research in Air Transportation, Tampa, Florida, United States.

Enabling Safe and Efficient Separation through Multi-Agent Reinforcement Learning

Jan Groot
Control & Simulation, Faculty of Aerospace Engineering
Delft University of Technology
Delft, The Netherlands

Joost Ellerbroek, Jacco Hoesktra
Control & Simulation, Faculty of Aerospace Engineering
Delft University of Technology
Delft, The Netherlands

***Abstract*—Over the next decades, it is expected that the number of unmanned aerial vehicles (UAVs) operating in the airspace will grow rapidly. Both the FAA (Federal Aviation Administration) and the ICAO (International Civil Aviation Organisation) have already stated that aircraft operating autonomously or beyond their operators' line of sight are required to have detect and avoid capabilities. At higher traffic densities these avoidance manoeuvres can, however, lead to instabilities within the airspace, causing emergent patterns that lead to knock-on effects that can harm the safety of the operations. It might be possible to formulate a cost function that encapsulates global safety, rather than individual safety, stimulating both safety and stability. One method that lends itself for optimizing such a cost function is cooperative Multi-Agent Reinforcement Learning (MARL). It has been demonstrated that MARL can be used for optimization in both competitive and cooperative (or even mixed) environments, however, when applied in a completely decentralized manner, stability issues often arise. It is therefore proposed to investigate the application of MARL for a well known centralized domain, ATC for manned aviation. This doctoral paper breaks down the proposed research project into 4 independent phases that individually contribute to the knowledge of applying MARL for ATC.**

***Keywords*—Conflict Detection and Resolution (CD&R), Multi-Agent Reinforcement Learning (MARL), Centralized Air Traffic Control, Autonomous ATC, BlueSky ATC Simulator**

I. Introduction

With the increase in the availability of drones worldwide, combined with continuous research into the area of personal air vehicles, the overall airspace is bound to get more crowded [1], [2]. Over the course of the next decades, it is predicted that the number of Unmanned Aerial Vehicles (UAVs) flying through the air (both commercial and personal) will increase exponentially, leading to air traffic densities unprecedented by today's standards [3]. To ensure the successful deployment of such operations, the Federal Aviation Administration (FAA) and the International Civil Aviation Organisation (ICAO) have already defined that all UAVs operating in the civil airspace without a human controller need to have detect and avoid capabilities [4]. Because of this more research is being done in the areas of conflict resolution from an autonomous and decentralized perspective [5]. As conflict resolution manoeuvres deviate aircraft from their nominal path, the total occupied airspace for each aircraft increases, potentially leading to more conflicts. When this happens in an uncontrolled manner it can lead to a domino effect, destabilizing the airspace and limiting the effectiveness of the resolution manoeuvres on a global scale [6].

A potential solution for this problem can be found by defining the problem as a Multi-Agent Reinforcement Learning (MARL) problem. Recent research demonstrated that by having a common goal, it is possible to use MARL to converge to cooperative behaviour between individual agents [7]. In the case of minimizing the global intrusion rate, this implies that, when destabilizing effects from conflict resolution start to harm the safety, the model should change to a more stable policy to optimize the cost function. However, simply applying MARL to this problem does not guarantee success. Common issues with MARL include stability problems, low sample efficiency and brittle convergence. It is possible to improve upon some of these issues by applying MARL from a centralized perspective [8]. This however also entails that it might be better to apply these methods to centralized domains.

Current Air Traffic Control (ATC) is done from a centralized perspective and operates at lower traffic densities than those predicted for Unmanned aircraft system Traffic Management (UTM), which is why this doctoral paper proposes to investigate the potential of MARL when applied to a centralized ATC task.

The remainder of this doctoral paper will be structured as follows. First, the results from previous research using reinforcement learning for safe separation with drones in decentralized environments will be given. From these results, some challenges and recommendations for future work, and how they link to autonomous centralized ATC with MARL will be described in a set of research topics, subdivided into 4 separate research phases. Finally, a potential experimental scenario in which MARL can be applied for all 4 research phases will be given.

II. Initial Experiments

Previous research by the author performed experimental simulations using the BlueSky Open Air Traffic Simulator [9] to identify whether reinforcement learning can be used in a decentralized manner to improve the safety of vertical layer transitions in a high-density layered airspace [10]. In this experiment, random aircraft were selected to ascend or descend to a predetermined target layer. A positive reward was given if the target layer was reached safely. Conversely, a negative reward was given if the agent intruded the protected zone of another aircraft during these operations. Fig. 1 shows that regardless of the degrees of freedom used for the action space of the agent, improvements were made to the safety of the vertical manoeuvres when compared to taking a direct path without any detect and avoid measures (resolution manoeuvres). This performance is compared to the Modified Voltage Potential (MVP) conflict resolution algorithm [11] which illustrates that the reduction in the number of intrusions is of similar magnitude for most models.

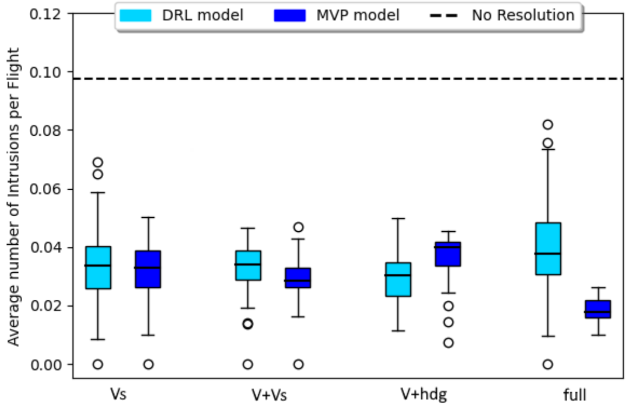

Fig. 1: Intrusion rate of the different models when compared to the baseline. The following DRL models are analyzed: 'Vs' only the vertical velocity is controllable; 'V+Vs' both the vertical and horizontal velocities are controllable; 'V+hdg' both the horizontal velocity and heading are controllable; 'Full' all aforementioned elements are controllable

If, however, instead of the number of intrusions, one compares the increase in the number of conflicts indicated in Fig. 2. It can be seen that in 2 of the 4 models, the reinforcement learning model manages to limit the domino effect, without being explicitly told so through the reward function. This does demonstrate that it is possible to find separation methods that are capable of limiting the increase in the number of conflicts. However, to explicitly stimulate this behaviour (both for decentralized and centralized applications), it will be necessary to include this in the cost function associated with the problem.

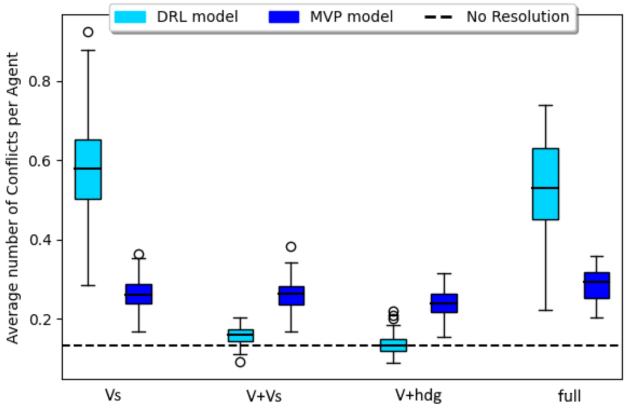

Fig. 2: Average number of conflicts encountered during vertical manoeuvres

III. Research Topics

The results from earlier research by the author [10] demonstrated some of the capabilities and potential of MARL for autonomously maintaining safe separation during vertical manoeuvres in a decentralized environment. Many of the recommendations and lessons learned following that research can however be carried over to centralized environments, such as the application of MARL in centralized ATC, proposed in this doctoral paper. This section discusses individual research topics, based on the recommendations from [10], that will aid in the successful development of an autonomous centralized ATC system with MARL.

A. *Stability of Multi-Agent Systems*

One problem with multi-agent systems is the inevitability of instabilities (for example the domino effect). However, when applying reinforcement learning in multi-agent systems, another source of instability will be introduced. As the agents are constantly changing their policy based on previous interactions with the environment, which includes other agents, the agents are essentially chasing a constantly moving target [12]. To ensure that safe and efficient operations are obtained, stability (both during training and testing) must be a main driver when designing the experiments and defining the individual components.

Multiple steps can be taken to improve the stability of multi-agent reinforcement learning methods. Many of these will be researched during Phase I of the project, such that the remainder of the project is done with as many stability enhancements as possible. The stability-enhancing methods that will be researched include using a centralized learner with a decentralized actor [8], limiting the number of concurrent actors in the environment, testing the efficacy of a variety of specialized multi-agent RL algorithms and using different models for different stages of the operations and training them in a serial instead of a parallel fashion.

B. State Representation in Partially Observable Environments

For any model to effectively and efficiently learn within an environment, consistency and completeness of the state representation is very important. Moreover, it is even a requirement for the Markov Decision Processes to ensure the theoretical convergence guarantee of Reinforcement Learning holds [13]. However, in contrast with bound environments such as Atari games, where a fixed number of pixels can be used as the state representation [14], the environment of air traffic operations is neither constant in size (changing number of concurrent aircraft) nor bound. This requires the state vector to be constructed in such a way that the majority of important information is encapsulated consistently whilst being constant in size over all time steps.

Previous research has tackled this problem by providing the number of intruding aircraft per quadrant [5], using a fixed number of concurrent agents in the environment [15] or including the N closest aircraft in the state-vector [16]. Most research looked at the state representation from a decentralized perspective. One advantage of centralized control is the availability of more high-level information, which can potentially lead to more informative and useful state representations. The conclusion of this research phase (Phase II) should include what features in the state have the most influence on the overall observed performance, if the state-representation influences the trade-off between safety and efficiency, and if there is such thing as including too much information in the state-vector.

C. Reward Function Design

The reward function is a direct way to influence the trade-off between efficiency and safety. The reward function encompasses the goal of the agent in a mathematical expression and is therefore directly related to the performance obtained in the different assessment metrics.

Designing a proper reward function is non-trivial in the case that multiple components play a role in the final learned behaviour. For example, in the case of safety versus efficiency, a shorter path should not be favoured over preventing an intrusion. Furthermore, structuring the reward function in such a manner that it completely encompasses the intended goal might still not make it a suitable reward function for reinforcement learning purposes [13]. Reward sparsity often arises in these cases, and combined with limited resources (computational power, information about the environment, training time), makes it impossible for the agent to learn the intended behaviour.

For this project, it is therefore interesting to investigate the impact of weights on the different elements of the reward function on the overall score of the defined performance metrics. During Phase III steps will be made towards defining an ideal reward structure that provides both adequate learning potential and the ability to tune the behaviour of the agents through weight selection. Defining this reward function can consist of stages of trial-and-error, online gradient ascent [17] and reward shaping, a technique first introduced in the field of animal learning [18].

D. Robustness Testing

As with any research that has real-world implications, it is important that the obtained results are, to a certain extent, still valid when carried over from the simulation environment towards a real-world domain. Researching this directly would be an expensive endeavour. However, investigating if the performance of the obtained models does not deteriorate when carried over to a different simulation environment can still provide information regarding the robustness of the models. To do this, in Phase IV of the research, the performance of a model trained in a specific environment will be compared to that of a model trained in an altered environment. These environment alterations can include, but are not limited to, change of traffic densities, change of wind intensity, and stochasticity and variations to the airspace structure. The performance of the original model in these environments will then be compared to that of the models trained for their specific environments. The observed performance differences give an indication for the sensitivity of the trained model towards environmental changes. Part of this phase of the research will include investigating which training environment conditions will eventually lead to the most robust models over all test environments.

IV. Experimental Environment

This final section of the doctoral paper proposes a simulation environment for future experiments. An overview of this simulation environment is given in Fig. 3. In this environment multiple aircraft will be flying in a section of the airspace controlled by the MARL model, the goal of the MARL model is then to guide these aircraft to their corresponding exit point of the environment. This exit point will be defined in 4D (latitude, longitude, altitude and time of arrival), enhancing the planning capabilities of potential human operators controlling the adjacent airspace sectors. The environment will be implemented in the BlueSky Open Air Traffic Simulator.

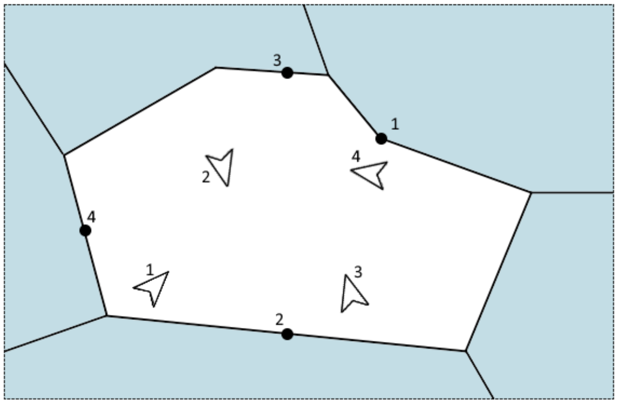

Fig. 3: Graphical representation of a potential experimental environment in which the proposed research topics can be conducted

The reward function can be composed of the following components, although many more combinations of variables for the reward function exist:

1) Number of Intrusions
2) Number of Conflicts
3) Deviation from the 4D exit constraints
4) Path efficiency

This proposed experimental environment enables all of the research phases to be conducted in the same environment, allowing direct comparison of the effect of the individual components on the performance of the MARL model.

V. Conclusion

This paper proposed a variety of steps that can be taken towards safe and efficient separation through multi-agent reinforcement learning for a centralized ATC environment, based on the results and recommendations of previous research for safe separation with UAVs in a decentralized environment. By breaking down the project into 4 separate research phases, that on their own provide valuable insights into the understanding of applied reinforcement learning, it is possible to get a glimpse of the potential of using reinforcement learning as a tool for autonomous aviation operations. At the end of the paper, an experimental environment for the to be conducted experiments is proposed that will enable direct comparison of the effects of the individual research phases on the overall performance of the MARL models.

References

[1] ”DHL Express”, “”DHL express launches its first regular fully-automated and intelligent urban drone delivery service..”

[2] D. Pierce, “Delivery drones are coming: Jeff Bezos promises half-hour shipping with Amazon Prime Air.”

[3] M. Doole, J. Ellerbroek, and J. Hoekstra, “Drone delivery: Urban airspace traffic density estimation,” *8th SESAR Innovation Days*, 2018.

[4] “Organization, i.c.a. ICAO circular 328 - unmanned aircraft systems (UAS). technical report, ICAO, 2011..”

[5] M. Ribeiro, J. Ellerbroek, and J. Hoekstra, “Determining optimal conflict avoidance manoeuvres at high densities with reinforcement learning,” *10th SESAR Innovation Days*, 2020.

[6] E. Sunil, J. Ellerbroek, and J. M. Hoekstra, “Camda: Capacity assessment method for decentralized air traffic control,” in *Proceedings of the 2018 International Conference on Air Transportation (ICRAT), Barcelona, Spain*, pp. 26–29, 2018.

[7] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.

[8] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, “Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning,” in *International Conference on Machine Learning*, pp. 4295–4304, PMLR, 2018.

[9] J. M. Hoekstra and J. Ellerbroek, “Bluesky atc simulator project: an open data and open source approach,” in *Proceedings of the 7th International Conference on Research in Air Transportation*, vol. 131, p. 132, FAA/Eurocontrol USA/Europe, 2016.

[10] J. Groot, J. Ellerbroek, and J. Hoekstra, “Improving the safety of vertical manoeuvres in a layered airspace with deep reinforcement learning,” 2022. unpublished.

[11] J. M. Hoekstra, R. N. van Gent, and R. C. Ruigrok, “Designing for safety: the ‘free flight’air traffic management concept,” *Reliability Engineering & System Safety*, vol. 75, no. 2, pp. 215–232, 2002.

[12] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, “A survey of learning in multiagent environments: Dealing with non-stationarity,” *arXiv preprint arXiv:1707.09183*, 2017.

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

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[15] J. Hu, X. Yang, W. Wang, P. Wei, L. Ying, and Y. Liu, “Uas conflict resolution in continuous action space using deep reinforcement learning,” in *AIAA AVIATION 2020 FORUM*, p. 2909, 2020.

[16] M. Brittain and P. Wei, “Autonomous air traffic controller: A deep multi-agent reinforcement learning approach,” *arXiv preprint arXiv:1905.01303*, 2019.

[17] J. Sorg, R. L. Lewis, and S. Singh, “Reward design via online gradient ascent,” *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[18] B. F. Skinner, “Reinforcement today.,” *American Psychologist*, vol. 13, no. 3, p. 94, 1958.