



Saki Omi

Development and design of applications for UAV-based
satellite communication terminal antenna evaluation using
deep-reinforcement learning

School of Aerospace, Transport and Manufacturing
PhD in Aerospace

PhD

Academic Year: 2020–2023

Supervisor: Prof. Hyo-Sang Shin
Associate Supervisor: Prof. Antonios Tsourdos
August 2023



School of Aerospace, Transport and Manufacturing
PhD in Aerospace

PhD

Academic Year: 2020–2023

Saki Omi

Development and design of applications for UAV-based
satellite communication terminal antenna evaluation using
deep-reinforcement learning

Supervisor: Prof. Hyo-Sang Shin

Associate Supervisor: Prof. Antonios Tsourdos

August 2023

© Cranfield University 2023. All rights reserved. No part of
this publication may be reproduced without the written per-
mission of the copyright owner.

Academic Integrity Declaration

I declare that:

- the thesis submitted has been written by me alone.
- the thesis submitted has not been previously submitted to this university or any other.
- that all content, including primary and/or secondary data, is true to the best of my knowledge.
- that all quotations and references have been duly acknowledged according to the requirements of academic research.

I understand that to knowingly submit work in violation of the above statement will be considered by examiners as academic misconduct.

Abstract

In recent years, satellites are launched almost on a daily basis and most of them are to be operated in Non-Geostationary Orbit (NGSO). The number of user terminals communicating with satellites is rapidly increasing. The interference has become a serious issue due to the crowded communication environment and the increased popularity of NGSO. Utilization of NGSO adds more complexity for operating user terminals since it requires tracking the satellite which is not static from the terminals' points of view. Also, the risk of interference has escalated due to the greater demand for Satellite-communication-On-The-Move (SOTM), which involves the need to keep terminals constantly pointing toward the target satellite while they are installed on a moving object. To ensure a safe communication environment, the terminal antenna must be verified based on set requirements. However, the test process at conventional test facilities is inefficient and does not have a solution to test antennas in the new communication scenarios. Therefore, this thesis aims to develop in-situ Unmanned Aerial Vehicle (UAV) -based measurement applications that are autonomously guided to enhance the efficiency of the measurement and to propose novel measurement methods to verify the antennas operated in new environments. Utilizing UAVs and performing measurements onsite is challenging due to the additional error sources and uncertainties in measurements and sensor positioning. In this work, a new deep-reinforcement learning algorithm is developed which can adapt to the dynamic environment under the presence of disturbances. Using this algorithm, the applications to verify the boresight angle offset of terminal antennas and to evaluate SOTM terminal antennas are proposed. The proposed applications are tested based on the numerical simulations and the results showed that the developed applications improved the efficiency of measurements and satisfied the required measurement accuracy. The thesis investigates novel measurement approaches for a new generation of satellite communication aiming to respond to the measurement demands that currently have no solution.

Keywords:

UAV-based Antenna measurement; Non-geostationary orbit; boresight angle localization; Satellite communication; VSAT; Satellite-Communication-On-The-Move; depointing; phased array antenna; reinforcement learning; TD3; recurrent neural network; POMDP; robustness; causality

Acknowledgements

I would like to express my sincere appreciation to my esteemed superiors, Prof. Hyo-Sang Shin and Prof. Antonios Tsourdos, for their supervision and support throughout my PhD journey. They have given me countless lessons on both research and life and consistently inspired me during this period.

Secondly, I would like to thank my industrial supervisor, Joakim Espeland, for leading me to this opportunity and providing continuous support through his insightful knowledge and passion for the satellite industry and innovation.

I am also thankful to Dr. Namhoon Cho for his invaluable academic support and for sharing his research attitude.

I would like to thank Eric Chin in IT department for his support in resolving various IT challenges I encountered.

My appreciation also goes to my office mates, George, Matthew, and Rasit for encouraging me and sharing their knowledge.

I extend my gratitude to my colleagues at QuadSAT, my sponsoring company, for letting me join their journey and sharing their great aim for the world.

I would like to express my appreciation to Prof. Koji Izumi for mentoring me.

I would like to thank Maurice, Michelle, and Alex for letting me live in their beautiful home and for creating wonderful memories together.

My heartfelt appreciation goes to my friends, Aaron, Andrian, Einar, Estela, Kathrine, Marcelo, Pablo, Roberto, Ruifan, Rurika, Troels, and many more who shared their lives and supported me.

Lastly, I would like to express my deepest appreciation to my family in Japan for their unconditional and everlasting love and care.

Table of Contents

Academic Integrity Declaration	i
Abstract	ii
Acknowledgements	iii
Table of Contents	iv
List of Figures	vii
List of Tables	ix
List of Abbreviations	x
1 Introduction	1
1.1 Background and Motivation	1
1.2 Research Purpose	6
1.2.1 Aim	6
1.2.2 Hypothesis	6
1.2.3 Objectives	7
1.3 Contributions	8
1.4 Organisation	10
1.5 Publications and Patents	10
2 Dynamic deep-RL	12
2.1 Introduction	12
2.2 Preliminary	14
2.2.1 Recurrent Neural Network	14
2.2.2 Twin Delayed Deep Deterministic policy gradient algorithm	16
2.2.3 LSTM-TD3	16
2.3 Dynamic RL	17
2.3.1 Causality and Statistics	18
2.3.2 Analysis of disturbance	18
2.3.3 POMDP formulation	18
2.3.4 Information states, belief states in POMDP	19
2.3.5 Identification of transition model	20
2.3.6 LSTM to reflect the sequence on the internal representation	20
2.4 Action inclusion	20

2.4.1	Length of sequence	20
2.4.2	Experiment	21
2.4.3	Results and analysis	22
2.5	Structure	25
2.5.1	Interpretation of LSTM-TD3	25
2.5.2	Modified LSTM-TD3	25
2.5.3	H-TD3 algorithm	27
2.6	Experiment	28
2.6.1	Simulation set-up	28
2.6.2	Results	29
2.7	Related Work	32
2.7.1	Utilizing RNN in RL	32
2.7.2	Separating between estimation and control	33
2.8	Conclusion	33
2.9	Appendices	33
2.9.1	Network structure	33
2.9.2	The number of parameters	36
2.9.3	Pseudo codes	36
3	Boresight localization	39
3.1	Introduction	39
3.2	Preliminary	42
3.2.1	Antenna pattern in the scope	42
3.2.2	The airborne measurement	43
3.2.3	Guidance algorithm	44
3.3	Method	46
3.3.1	Problem Analysis	46
3.3.2	Using differential	47
3.3.3	RL setting	48
3.3.4	Scaling of sequential observation	49
3.3.5	Beam center localization	49
3.4	Experiment	50
3.4.1	Training Environment	50
3.4.2	Evaluation settings	51
3.4.3	Results	52
3.4.3.1	Climber evaluation	52
3.4.3.2	Localization evaluation	58
3.5	Conclusion	59
4	SOTM depointing for phased array	62
4.1	Introduction	62
4.2	Preliminary	65
4.2.1	FORTE method	65
4.2.2	Phased array antenna	66
4.2.3	LSTM-TD3 _{1halhc}	67
4.3	Method	68
4.3.1	Measurement setup	68

4.3.2	Objective of guidance	68
4.3.3	Problem analysis in POMDP framework	69
4.3.4	RL setup	70
4.3.5	Imaginary agent	72
4.4	Experiment	72
4.4.1	Environment	72
4.4.2	Compared Algorithms	73
4.4.3	Result	74
4.5	Conclusion	84
4.6	Appendix	85
4.6.1	Learning trajectory	85
5	Overall Discussion	87
5.1	Accomplishments of Objectives	87
5.2	Response to Hypothesis	90
5.3	Response to Aim	92
5.4	Future Work from the objectives	93
5.5	Extended scenarios of the applications	95
5.6	Implementation plan	96
6	Conclusion	99
	References	100
A	Ethical Approval Letter	110

List of Figures

1.1	Satellite communication serves and classification	1
1.2	Visualization of the connection between the chapters	9
2.1	FNN and RNN diagram	15
2.2	Simplified LSTM cell schematic	15
2.3	TD3 schematics	16
2.4	LSTM-TD3 schematics	17
2.5	Causal diagram	19
2.6	MDP	24
2.7	temporal bias	24
2.8	temporal sinusoidal wave	25
2.9	random sinusoidal wave	25
2.10	noise	26
2.11	hidden	26
2.12	Modified LSTM-TD3 structure	27
2.13	H-TD3 schematics	28
2.14	Learning trajectory	29
2.15	Final total reward	30
2.16	Iteration time comparison	31
2.17	Generalisability	31
3.1	VSAT operational θ_{3dB}	43
3.2	QuadSAT's compact UAS-APE	43
3.3	Impact of 30cm positioning error on distance	44
3.4	Meta-reinforcement learning	45
3.5	Conceptual implementation of RL system guiding a UAV to the beam peak	46
3.6	Performance comparison with various action noises	53
3.7	Performance comparison with various SNR	54
3.8	Performance comparison with various beamwidth	55
3.9	Trajectory comparison	56
3.10	Performance comparison with the distance to go	57
3.11	Fail case of Climber-sc	58
3.12	Estimation accuracy comparison	59
4.1	Simulated phased array antenna with various steering angles	67
4.2	Measurement setup	68
4.3	Structure diagram	71

4.4	Basic formation	75
4.5	Performance comparison in total UAV numbers of each algorithm	77
4.6	Trajectory example of RL^*_{rel+Pr} of 3,4 subset UAVs	78
4.7	Performance comparison in action noise	79
4.8	Estimation method accuracy comparison	80
4.9	Performance comparison in action noise and UAV numbers	81
4.10	Performance comparison in action noise	83
4.11	Learning trajectory of without imaginary agents	85
4.12	Learning trajectory of imaginary agents	85
4.13	Learning trajectory of various number agents	85
4.14	Learning trajectory of various formation	86
4.15	Learning trajectory of algorithm without past information	86

List of Tables

2.1	"Pendulum-v0" environment definition	22
2.2	Test scenario	23
2.3	Structure	34
2.4	Structure (cont.)	35
2.5	The parameter of numbers (obs dim= 3, act dim= 1)	36
3.1	GA parameters	51
3.2	Performance in empirical radiation pattern	57
3.3	Localization comparison	61
4.1	Compared algorithms	74
4.2	Performance	76
4.3	The effect of past information in SNR 40dB	84
5.1	Implementation plan	96

List of Abbreviations

ACU	Antenna Control Unit
ATE	Authorized Test Entity
AUT	Antenna Under Test
BFGS	Broyden, Fletcher, Goldfarb, and Shanno
BO	Bayesian Optimization
BPTT	Backpropagation Through Time
DDPG	Deep Deterministic Policy Gradient
DGPS	Differential Global Positioning Systems
DPG	Deterministic Policy Gradient
DRQL	Deep Recurrent Q-learning
EI	Expected Improvement
EMC	Electromagnetic Compatibility
ESA	Electronically Steerable Antenna
ETSI	European Telecommunications Standards Institute
FCC	Federal Communications Commission
FNBW	First Null Beamwidth
FNN	Feedforward Neural Network
FPA	Flat Pannel Antenna
FORTE	Facility for Over-the-air Research and Testing
GA	Gradient Ascent
GCS	Ground Control Station
GEO	Geostationary orbit
GNC	Guidance, Navigation and Control
GP	Gaussian Process
GRU	Gated Recurrent Unit
GVF	Global VSAT Forum
IID	Independent and Identically Distributed
ISL	Inter-Satellite Like
ITU	International Telecommunication Union
LCB	Lower Confidence Boundary
LEO	Low Earth Orbit
LSTM	Long Short Term Memory
MDP	Markov Decision Process
MEO	Midium Earth Orbit
ML	Machine Learning
MPI	Maximum Probability of Improvement
NFFFT	Near-Field Far-Field Transformation

NGSO	Non-Geostationary Satellite Orbit
POMDP	Partially Obaservable Markov Decision Process
RF	Radio Frequency
RFI	Radio Frequency Interference
RNN	Recurrent Neural Network
RTK	Real Time Kinematics
SNR	Signal-to-Noise Ratio
SOMAP	Satellite Operators Minimum Antenna Performance
SOTM	Satellite-communication On the Move
TD3	Twin Delayed Deep Deterministic Policy Gradient
UAV	Unmanned Aerial Vehicle
VSAT	Very Small Aperture Antenna

Chapter 1

Introduction

1.1 Background and Motivation

In recent years, satellites are launched on a daily basis all over the world. According to a survey [1], the number of launched satellites in 2020 was around 1000 whereas it was less than 100 in 2015. It is also estimated that this trend would continue for the next decade in [1]. Ka-band communication using Low-Earth-Orbit (LEO) will be the dominating frequency for satellite communication and orbit for the coming years with the key phrase of "anywhere, anytime" available broadband connectivity.

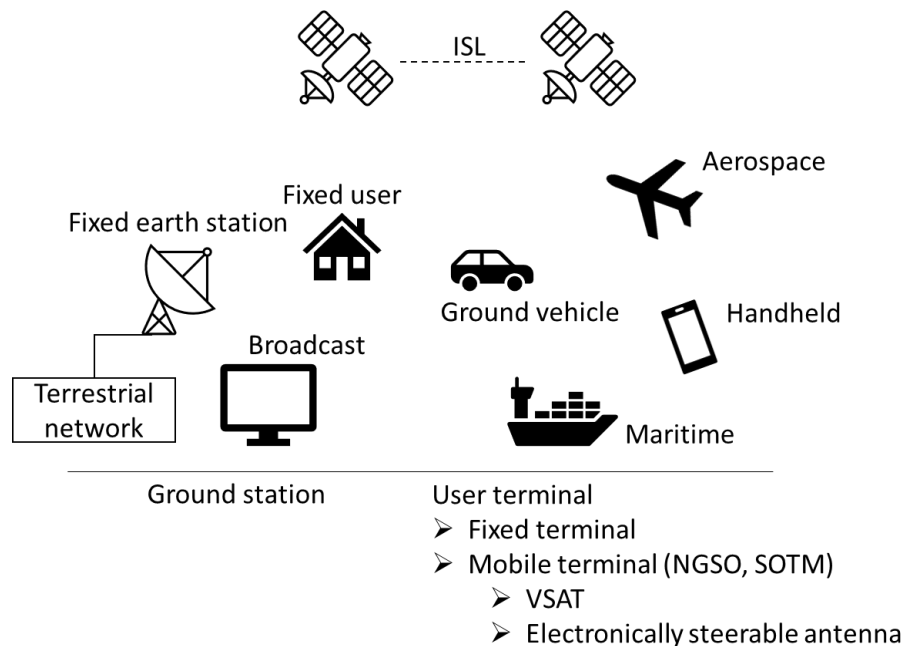


Figure 1.1: Satellite communication serves and classification

Fig. 1.1 shows the overview of satellite communication applications and the user terminal's description. The satellites are communicated via Inter-Satellite Link (ISL) and relay the data. The earth stations are usually equipped with large 15 – 30 m diameter antennas and connected to a terrestrial network. It is normally owned by the terrestrial network operator. In the broadcasting network, typically TV channel contents

are distributed via one-directional communication to the end-users having receiving systems referred to as television receive-only. The end-users can also be categorized as mobile users and fixed users. Normally these communications are bi-directional where the terminals receive and transmit. "User terminal" is a terminology to describe the communication terminals owned by end-users while "ground station" represents the system owned and operated by satellite operators or service providers. In other words, ground stations belong to the infrastructure of the communication network, like cell towers. There are two types of user terminals: fixed terminals and mobile terminals. The antenna beam of the fixed terminal does not move actively and is intended to be used for communication with Geostationary orbit (GEO) satellites. The mobile terminals actively steer the beam either mechanically or electronically to keep the communication. Communication with Non-Geostationary Satellite Orbit (NGSO) is also included in this category. With the transition from GEO to NGSO, all antennas will be required to track moving target satellites and perform handover. As such, the chances for the antennas to mispoint and the risk of generating interference will be significantly increased. Satellite-communication-On-The Move (SOTM) refers to the application when the terminal is mounted on a moving vehicle, which is a ship, a train, a vehicle, or an airplane [2], and is required to compensate for the motion generated from the vehicle by steering the beam. The source detection of interference generated from SOTM is quite difficult since it randomly and intermittently appears and hence, a thorough examination of the antenna is important to mitigate the interference issue [2]. The antennas for mobile terminals may be mainly classified as either Very-Small-Aperture Terminal (VSAT) or Electronically Steerable Antennas (ESAs). VSAT antennas are parabolic antennas with a dish-shaped reflector, and they mechanically steer the beam by controlling the orientation of the dish using a servomotor. In contrast, ESAs or Flat Panel Antennas (FPAs) such as phased array antennas do not require reflectors. There were 5 million of commercial user terminals in 2020 and they are estimated to reach more than 10 million in 2030 supported by an accelerated launch in NGSO and FPA technology developments [1]. According to [1], the market of FPA terminals is growing rapidly and will represent the bulk of mobile terminals over time after their initial implementation period. The acceleration of technology has been driven by the enrollment of private companies, e.g., SpaceX, OneWeb, and Amazon, in the satellite communication business. As the number of businesses serving end-users has massively increased at a reduced cost per bit, businesses require cost-effective communications, which are enhanced by the effective management of the quality and quantity of data communicated within budgetary constraints to be sustainable.

As understood from the general link analysis [3], antenna performance and signal noise determine the quality and quantity of communicated data in satellite communication. Hence it is critical to evaluate the antenna performance numerically and ensure that it does not create unwanted interference which increases the noise present within communications. Therefore, each satellite operators and authorities have requirements for antennas and it is required that all of the antennas should satisfy the requirements before it is operated. Global VSAT Forum (GVF) and a group of leading satellite operators such as Eutelsat, Intelsat, and SES produced guidelines regarding the satellite operators' expectations which is referred to as Satellite Operator Minimum Performance Specifications (SOMAP), and how to demonstrate compliance [4]. For example, a list

of the requirements for Ka-band in SOMAP type approval can be found in [5]. Let us review some of the listed requirements.

One of the common requirements, regardless of whether the Antenna Under Test (AUT) is SOTM, is "Antenna sidelobe characteristics". The sidelobe mask mostly defined by $29 - 25 \log \theta$ is applied and the sidelobes of the AUT should be under the mask to ensure that the AUT does not interfere with the adjacent communication.

Another example from the list is "X-pol isolation" which is also known as cross-polar discrimination (XPD). XPD is the magnitude of the relative power of the cross-polarized pattern with respect to the co-polarized pattern at the boresight angle. XPD measures how well the AUT can reject the orthogonally polarized signal. This is an important characteristic when the terminal expects a duplex link and needs to isolate the signals between the transmit and receive channels as both of the channels are operated on the same antenna. Since the accuracy of XPD measurement is quite sensitive to the position and polarization angle of the measurement system, it is important to localize the boresight angle accurately.

There are also special requirements for particularly SOTM terminals, these requirements are summarised as follows:

- Pointing error (depoining) is kept below $\pm 0.5^\circ$ in any azimuth and elevation angle.
- Cessation of emission occurs within 100 milliseconds if depointing exceeds $\pm 0.5^\circ$.
- Transmission resumes when depointing is below $\pm 0.2^\circ$.

Note that the AUT is evaluated with respect to the above requirement when it is in "auto-tracking mode" on the move. By using its algorithm, the AUT generates its motion to track the satellite. The measurement system developed in this thesis provides evaluations with respect to these criteria and an indication of the failure of the tracking accuracy of the AUT rather than providing the tracking or estimation function of the AUT since these methods vary depending on the AUT. In the case of the evaluation of pointing accuracy, the measurement system is required to estimate the boresight directions. There are two approaches to measuring depointing: (a) Mechanical measurement with respect to a mechanical coordinate system of AUT using the angle measurement provided by the Antenna Control Unit (ACU) [6], (b) Electrical measurement with respect to a boresight coordinate system using the correlation between radiation pattern [7]. The electrical measurement might be more trustworthy since there are no concerns of misalignment between the mechanical and electrical coordinate systems, although the mechanical measurement can be verified once the radiation pattern is obtained. However, these methods might not be feasible for ESAs (e.g., phased array antenna) since their radiation patterns vary depending on the steering angle and collecting all possible radiation patterns is not practically possible.

The test and measurement industry provides solutions to evaluate communication systems. To perform the aforementioned measurements, the measurement equipment needs to collect the information around the antenna by either placing the sensors around the AUT or mounting the AUT on the turn table. Antenna measurement is typically performed in the range called "far-field" which is approximated by $2D^2/\lambda$ where D is the aperture diameter and λ is the wavelength of the operational frequency.

For example, the far-field starts from 134 m for a VSAT antenna with operation frequency 14 GHz and aperture size 1.2 m. In this range, the electromagnetic wave is well constructed. Another solution is called near-field measurement which requires higher accuracy in sensor positioning and a mathematical calculation called Near-Field Far-Field Transformation (NFFFT).

Traditionally, antenna measurements used to be performed in a special facility. The facility used for antenna measurements can be put into three categories:

- Outdoor far-field range: Typically it offers 100 m to 2 kilometers range
- Compact range: Using single or dual reflectors, the facility creates far-field conditions
- Near-field scanners: Using NFFFT method, the facility is compact.

However, there are considerable drawbacks to this method. Since it requires assembling/reassembling and transporting the AUT, the process is logistically heavy and slow. The AUT and the bulky facility have to be set up one by one, the evaluation process is also time-consuming and the measurement scenario cannot be easily extended to the one considering NGSO or FPA communication. Also, using conventional facilities could be a costly solution as it requires human resources and their time, and a large facility with a spacious area to perform far-field measurements. In addition, the number of facilities is limited. For instance, currently, seven facilities are used for VSAT type approval in Europe and the US under the approval of Eutelsat, which is one of the world's leading satellite operators [6]. Unmanned Aerial Vehicle (UAV)-based measurement has several benefits which might solve the limitation of the measurement at conventional facilities. The UAV-based measurement system is portable and hence, it makes in-situ measurements possible in the operational environment of the AUT. Also, the test scenario and configuration can be flexibly and easily designed. The satellite industry and technology are undergoing a major transformation, however, current test facilities are not designed to meet the needs of changing satellite industry while UAV-based measurement might be able to adapt easily and promptly to the requirements that emerged from this transition. Given these benefits, airborne measurements were initially investigated using a helicopter [8, 9, 10], a balloon [11], and a UAV with fixed-wing [12], yet they did not spread as a common solution due to the operational cost and inaccuracy of positioning. In recent years, the improvement and accessibility of UAVs have led to renewed interest in airborne measurement specifically by utilizing UAVs, more specifically drones.

UAV-based measurement using drones, not limited to satellite communication antennas, was developed with a focus on demonstrating feasibility. This was done by presenting experimental results and analyses on radiation pattern measurement in the early stages [13, 14, 15, 16]. The development showed promising performance and nowadays, antenna measurement conferences (e.g., European Conference on Antennas and Propagation and Antenna Measurement Techniques Association) host UAV-based measurement sessions. From the literature, the current main research directions could be put into three categories as follows:

- Solving / Investigating drone and in-situ measurement issues

- Development of NFFFT method for UAV measurement
- Utilizing UAV measurement for development of new technology

The first research direction is to evaluate and solve the newly emerging issues. The in-situ UAV measurement has introduced new error sources and uncertainties in contrast to the measurement in conventional methods using anechoic chambers where reflections disappear or using facilities fixed to the ground in a known test environment. In [17], it was identified that the largest uncertainty was generated from the environmental reflections while reflections disappear if measurement is performed in anechoic chambers, and a method to obtain the reflective characteristic of the environment on a pre-planned simple flight path was developed [18]. Also, a different measurement equipment (network analyzer) was tested to eliminate the effect of unwanted reflections [19]. As pointed out in [20], electromagnetic compatibility and interference from UAV components are another concern. The majority of propellers of the drones are made of carbon fiber which is conductive. The impact of interference was investigated by comparing it with another material in [21]. The localization and positioning error of UAVs and the ripple of the platform contribute to the measurement accuracy. The accuracy RTK positioning system was verified in [22] and a new design of the Radio Frequency (RF) payload was proposed [23]. To cope with the positioning uncertainty of the UAVs, the method to calculate the time duration for the UAV to hover in one spot in order to perform accurate gain measurement was developed in [24].

Another key topic is the development of NFFFT for UAV measurement. Conventionally, NFFFT requires phase data of the signal and theoretically 0.5λ sampling density [25], which is approximately 10 cm for the frequency of 14 GHz. Near-field measurement using UAV was introduced in [26] and it has gathered a lot of research interest since it does not require the UAV to fly at far distances and the required setup in the conventional method could be quite heavy. However, reaching the required localization accuracy and reconstructing accurate radiation patterns via NFFFT without having a complicated measurement setup like in [27] is challenging. Many approaches by modifying the mathematical operation and designing the sample points for NFFFT have been proposed [28, 29, 30, 31, 32].

Another topic is utilizing UAV measurement for new developments which shows the significance and potential of UAV measurements. This topic includes verification as array radio telescopes [33, 34, 35] which used to be done by astronomical calibration. A new design for beamforming was verified by using the UAV measurement system [36, 37]. The quality of service of the 5G network was analyzed in an urban environment at different heights in [38] and three-dimensional latency measurement for the 5G network was investigated in [39].

The developments are promising and improving UAV-based measurement technology. However, the current direction of the development is not focused on making the system autonomous and intelligent.

On the other hand, the popular UAV applications in the literature outside of the antenna measurement community are Search and Rescue (SAR), geographical Simultaneous Localization And Mapping (SLAM), remote sensing, construction and inspection, precision agriculture, delivery of goods, real-time monitoring of road traffic, surveillance applications of UAVs, and providing Wireless Coverage [40, 41]. Typical sensor

payloads are video camera, radar, LiDAR, and sonar. The applications with RF payloads are either extending network coverage [42] or Radio Frequency Interference (RFI) [43, 44, 45]. Approximately 45% of the UAV market is occupied by construction and inspection which mainly monitors construction projects and inspects power lines and gas pipelines. The predicted value of UAV solutions in telecommunication was \$6.3 billion which was the second lowest in 8 categories [46]. The UAV-based antenna measurement has very little recognition currently although there are huge potential benefits and demands in the satellite industry and there might be a lot to contribute from outside of the electrical engineering field.

Therefore, this thesis is motivated to develop autonomous and hence efficient measurement applications for UAV-based satellite communication terminal antenna evaluation. This is approached from the perspective of Guidance, Navigation, and Control (GNC) in the aerospace field, and it proposes new measurement methods that reflect the undergoing transition in the satellite industry.

1.2 Research Purpose

1.2.1 Aim

Although the UAV-based measurement might solve the issues of conventional methods as explained in Section 1.1, it introduces new challenges to the antenna evaluation due to the nature and operational and legal limitations of UAVs e.g., flight time restrictions, a limited number of sensor numbers, and positioning errors and uncertainties of the sensors. The antenna measurement used to be led by electrical engineers and researchers who might not be familiar with these problems, whereas these are common considerations in UAV applications. However, antenna evaluation is still a minor topic in the field of aerospace engineering. Given the promising benefits of UAV-based measurements, this existing gap should be bridged. Doing so would allow for further development, drawing from knowledge in both domains to realize more efficient and trustworthy antenna evaluations. Also, test and measurement industry should respond to the technical and industrial transition of the satellite industry. The possibility of UAV-based measurement should be explored from this point of view. Hence, the aim of the research is:

To extend knowledge and develop applications of autonomous UAV-based measurement system for satellite terminal antenna evaluation.

1.2.2 Hypothesis

It is critical that the antenna evaluation system can deliver accurate and robust results. The data to generate the test result is based on the measured data on the UAV's trajectory. Therefore, the development is focused on guiding the UAV to accomplish the desired evaluations. The measurement and the UAV positioning are always affected by noise. The measurement model is often not available when the radiation pattern of the AUT is unknown. Also, the data which can be sampled simultaneously at different points are sparse due to the limited number of sensors (UAVs). In this circum-

stance, it is not practical or sometimes impossible to utilize an external estimator to reveal the state of the AUT. For example, the radiation pattern of phased array antennas varies depending on direction of the steered beam. There is no existing solution to verify the sidelobe characteristics of phased array antennas based on measurements and the evaluation currently relies on antenna simulators. The decision-making problem where we have no access to the true state of the system can be classified as Partially Observable Markov Decision Process (POMDP).

Responding to these challenges, we define the following research hypothesis to be tested:

Accurate and robust UAV antenna evaluation applications can be designed by utilizing an algorithm which performs in a POMDP environment under the presence of disturbances without the specific knowledge of models of AUT.

1.2.3 Objectives

Based on the hypothesis, the following three objectives are specified as follows:

1. Develop a deep-RL algorithm with improved robustness of optimality which can dynamically adapt to the POMDP environment under the presence of non-static disturbance and which provides an end-to-end solution.
2. Develop a boresight angle localization application using the developed RL algorithm.
3. Develop a depointing measurement application for SOTM phased array antenna using the developed RL algorithm.

Objective 1 The algorithm in the hypothesis is identified as deep-RL algorithm since it does not require external models of the environment, instead it requires data from simulation or real-world to train the internal model and the trained agent provides instantaneous decision making which is beneficial for antenna evaluation in real-time. Signal propagation models have been extensively studied and high-fidelity models are available although the development of the thesis considered the basic model and utilizing the high-fidelity models is out of the scope of this thesis. The disturbance in observation would be non-static and the antenna pattern varies during the measurement in the case of phased array antenna. Thus, the agent is required to adapt to this dynamic environment. The robustness of optimality is measured as the achieved total reward and the developed algorithm is evaluated based on this value in non-static environments. In the standard POMDP framework, the algorithm comprises an estimator and policy. However, this standard framework is not available since the estimator is not available or unreliable in the targeted applications. Therefore, the algorithm is developed to provide an end-to-end solution without using an external estimator.

Objective 2 Boresight angle localization contains the scenario where the developed RL algorithm is effectively applied and could make a great impact on the antenna measurement procedure. A single UAV with RF payload is guided to the static boresight

angle without specific prior knowledge of the radiation pattern of AUTs. The application must be robust and efficient and it should be applicable for the entire range of terminal antennas without being limited by the configuration of the evaluation. The implementation of RL algorithm should be devised to satisfy this requirement. The primary requirement regarding localization accuracy is aimed to be 0.05° which is an order of magnitude more than the performance requirement of AUT at 0.5° in [47].

Objective 3 Depointing measurement is a specific requirement for SOTM application to evaluate if AUT can keep tracking the target directions when it is on the move. The radiation patterns of phased array antennas vary depending on the beam steering angle and it is infeasible to collect all possible radiation patterns to verify pointing accuracy. Thus, there is no currently existing solution for phased array antenna de-pointing measurement. Application of de-pointing measurement for phased array antenna is developed by utilizing multiple UAVs guided by the developed RL algorithm in the environment where the radiation pattern and the boresight angle are dynamically changing during the measurement. The targeted measurement accuracy is 0.05° and the minimum number of the UAVs to satisfy the requirement should be addressed.

1.3 Contributions

To accomplish the stated aim and objectives, each chapter is structured as follows. Fig. 1.2 illustrates the connection between each chapter and objective and contrasts the application with respect to objective 2 and objective 3. The objective 1 is developed in Chapter 2 and the hypothesis is proven in Chapter 3 and Chapter 4. The applications of boresight angle localization (Objective 2) and de-pointing measurement for SOTM phased array (Objective 3) are developed in Chapter 3 and Chapter 4 respectively. The application of Objective 2 and Objective 3 could be seen as alike since both measure the boresight angle. However, there are conditions that are fundamentally different, and hence they require different development. Boresight angle localization application considers the scenario where the boresight and antenna pattern are fixed during the measurement. However, there is no available knowledge about the antenna pattern specific to the AUT, and without having this knowledge, the expected range of signal strength to be received during the measurement cannot be defined. Hence, the application should be flexible for any test configuration. In contrast, the de-pointing measurement application is performed as a time series estimation. The initial boresight angle is assumed known, yet the direction of the boresight and the radiation pattern dynamically changes during the measurement. The detailed contribution of each chapter is summarized as follows:

Contribution of Chapter 2

One of the Deep RL algorithms, Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm, is extended to solve POMDP under the presence of various non-static disturbances. Implementing an enhanced Recurrent Neural Network (RNN) is a recent popular approach to make the agent perform in POMDP environment. However, most of the implementation utilizes only past sequential observations but not action

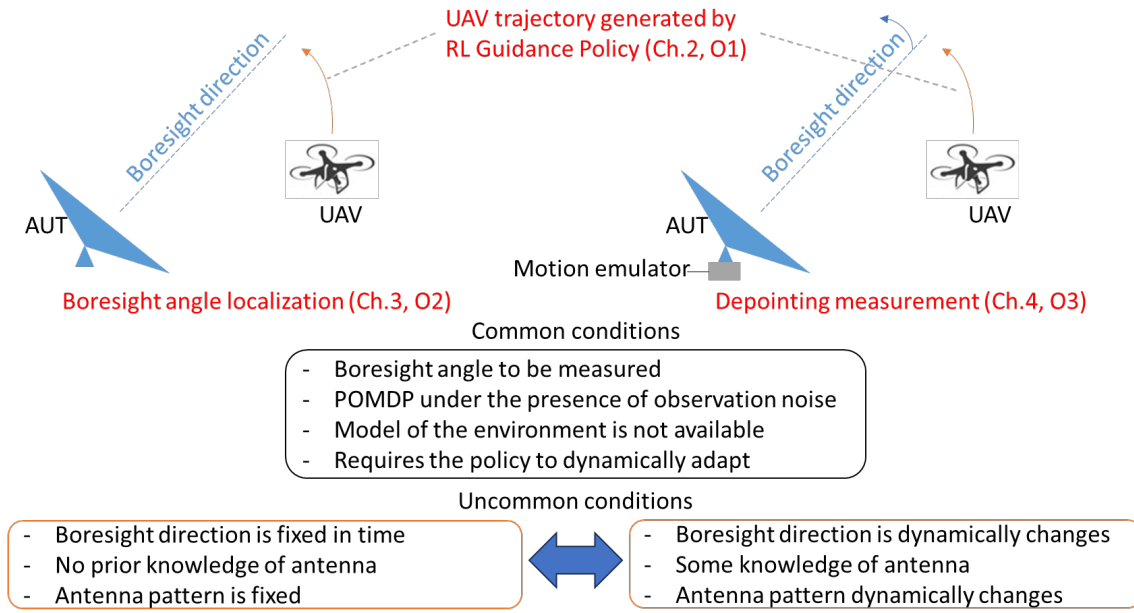


Figure 1.2: Visualization of the connection between the chapters

sequences. It is hypothesized that incorporating the past action sequence will enhance the robustness of optimality. This hypothesis is analyzed based on the causality of observation and action, by breaking down the disturbance characteristics into two categories. It is further analyzed based on the update of the belief state, as found in the standard POMDP framework. The hypothesis is tested by using an existing algorithm that has the RNN in TD3. Secondly, a structure that processes sequential information in a logical flow from the perspective of Bayesian inference is proposed, modifying the existing algorithm based on the initial analysis. Numerical simulations showed an improvement in the robustness of optimality when introducing this new structure. However, when sequential data is repeatedly processed in each network, the training can take a considerable amount of time. To mitigate this issue, another structure, which shares the hidden states of the enhanced RNN, is proposed. The results showed that the proposed method could significantly decrease the training time, although the robustness of optimality was not the best compared to other algorithms.

Contribution of Chapter 3

The application for localizing boresight angle is proposed by decomposing the strategy into two parts. The first part is to guide the UAV to the boresight angle by using the developed RL algorithm in Chapter 2 and the second part is estimating the boresight angle by utilizing Gaussian Process regression from the sampled data on the trajectory. The application is trained by targeting VSAT terminals. The range of beamwidth for VSAT terminal antennas and realistic positioning noise are analyzed based on literature and data provided by the sponsoring company, QuadSAT. A large variety of beamwidths exist in the VSAT category and the range of the measured signal strength and signal-to-noise ratio are not fixed. The RL elements are carefully defined to prevent the creation of a massive state space, thereby enabling the agent to learn effectively.

The application aided by the RL algorithm is trained by using the concept of meta-RL. The proposed method is evaluated through numerical simulations based on both theoretically and empirically obtained radiation patterns. These evaluations demonstrate that the method can perform robustly without being constrained by unknown environmental factors such as beamwidth and signal strength range. By comparing with the other algorithms, the developed algorithm showed several benefits such as no requirement to adjust parameters, the robustness of the performance in accuracy, and the efficiency of the generated trajectory. With this application, the efficiency of the measurement process should be improved and the new framework to evaluate side-lobe characteristics for phased array antenna is proposed in Chapter 5.

Contribution of Chapter 4

The application for depointing measurement for SOTM phased array antenna is proposed. The solution is novel since the solution for this context currently does not exist due to the dynamically changing radiation patterns depending on the steering angle. The depointing is measured by dynamically locating the multiple UAVs to track the boresight angle and the UAVs are guided based on the received signal strength at each UAV by using the developed RL algorithm in Chapter 2. The proposed application does not utilize an external estimator since the available estimation can be unreliable due to the limited number of sensors and the state including the angular velocity of the boresight angle of AUT is not fully observable. The developed application is evaluated on the simulated phased array antenna in terms of the estimation accuracy, trajectory stability, the required number of UAVs, their formation, and the robustness of the accuracy under the involvement of positioning noises. The result demonstrated that it is feasible to measure depointing by the proposed method with an estimation accuracy that satisfies the requirement. Furthermore, the number of required UAVs is effectively reduced to three by introducing an imaginary agent. Finally, extending the measurement scenario to SOTM depointing for NSGO satellite communication is discussed in Chapter 5.

1.4 Organisation

The organization of the rest of the thesis is as follows. A deep-RL is developed which performs in POMDP under the presence of dynamic disturbances in Chapter 2. By implementing the developed RL an applications are designed for boresight angle localization in Chapter 3 and depointing for SOTM phased array antenna in Chapter 4. Overall discussion as a synthesis of Chapter 2 to Chapter 4 is included in Chapter 5. Finally, the thesis is concluded in Chapter 6.

1.5 Publications and Patents

- Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "Introduction to UAV swarm utilization for communication on the move terminals tracking evaluation with reinforcement learning technique." In 2021 15th

- European Conference on Antennas and Propagation (EuCAP), pp. 1-5. IEEE, 2021.
- Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "Stochastic Filtering Technique for UAV-Based Communications On The Move Terminal Tracking Accuracy Evaluation." In 2021 Antenna Measurement Techniques Association Symposium (AMTA), pp. 1-6. IEEE, 2021.
 - Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "Reinforcement Learning System of UAV for Antenna Beam Localization." In 2021 IEEE Conference on Antenna Measurements & Applications (CAMA), pp. 61-65. IEEE, 2021.
 - Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "Multi-beam tracking for phased array antenna measurement by multiple UAVs." In 2022 16th European Conference on Antennas and Propagation (EuCAP), pp. 1-5. IEEE, 2022.
 - Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "Reinforcement learning aided path planning algorithm for multi-UAV based in-situ satellite terminal antenna evaluation system." 73rd International Astronautical Congress (IAC), Paris, France, 2022
 - Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos. "Learning from hidden state to solve POMDP." 10th International Conference on Robot Intelligence Technology and Applications (RiTA), 2022 (Not published but accepted and presented)
 - Omi, Saki, Hyo-Sang Shin, Namhoon Cho, Antonios Tsourdos. "Dynamic deep-reinforcement-learning algorithm in Partially Observed Markov Decision Processes" (Submitted to IEEE Transactions on Neural Networks and Learning Systems (T-NNLS))
 - Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "Autonomous boresight localization for UAV-based satellite terminal antenna measurement" (To be submitted to IEEE Transactions on Instrumentation and Measurement)
 - Omi, Saki, Hyo-Sang Shin, Antonios Tsourdos, Joakim Espeland, Andrian Buchi. "UAV-based satellite communication on the move depointing measurement for a phased array antenna" (To be submitted to Journal of Aerospace Information Systems)
 - QuadSAT ApS. 4229780. ANTENNA EVALUATION TEST SYSTEM. European Patent Office.; 2023. (To be published on 23 August 2023).
 - QuadSAT ApS. AIRBORNE SATELLITE ORBITAL SIMULATOR. (Patent application has been filed as United Kingdom Patent Application No. 2303672.6 in 2023).

Chapter 2

Dynamic deep-reinforcement-learning algorithm in Partially Observed Markov Decision Processes

Abstract

Reinforcement learning has been greatly improved in recent studies and an increased interest in real-world implementation has emerged in recent years. In many cases, due to the non-static disturbances, it becomes challenging for the agent to maintain its performance. The disturbance results in the environment known as Partially Observable Markov Decision Process (POMDP). Typically, POMDP is handled by introducing an additional estimator, or Recurrent Neural Network is utilized in the context of reinforcement learning. Both of the cases require to process sequential information on the trajectory. However, there are only a few studies investigating the effect of information to consider and the network structure to handle them. This study shows the benefit of action sequence inclusion in order to solve POMDP. Several structures and approaches are proposed to extend one of the latest deep reinforcement learning algorithms with LSTM networks. The developed algorithms showed enhanced robustness of controller performance against different types of external disturbances that are added to observation.

2.1 Introduction

Machine Learning (ML) techniques are widely applied for control tasks. Data-driven approach aided by ML algorithms could be beneficial since it does not necessarily require a system model. More specifically, recent developments in deep Reinforcement Learning (RL) techniques have been adapted to generate near-optimal actions in the given state of the agent in the continuous action and state space. The latest development of deep RL has been successfully tested in many complex simulation environments from the standardized framework in Open AI gym [48].

However, it is not guaranteed that the RL algorithm showing good performance in the simulation environment will perform as intended in the real world. Although most of

the RL algorithms have been built upon the assumption of the Markov Decision Process (MDP), this assumption is not valid in many real-world applications. The information collected from the sensors does not fully describe the true and full state of the agent due to the limited measurement, noise, and disturbances. This partially observable and non-deterministic scenario is categorized as Partially Observable Markov Decision Process (POMDP) [49]. In the MDP condition, it is possible to map the action based on the observation which uniquely represents the states of the system. However, in POMDP, the agent has no access to the true system state and the agent only receives the observation which is generated from the latent state of the system [50]. In this case, the agent is required to construct its state representation from the available information to restore the MDP condition. This state representation can be described in the same state space as the systems but also it can be in the encoded internal state space.

One of the approaches to address POMDP is using sequential information which the agent has experienced on its trajectory. By processing the sequential information, the agent could generate the internal state representation and restore the MDP condition. Recurrent Neural Networks (RNN) or other advanced gated RNN such as Long-Short Time Memory (LSTM) [51] and Gated Recurrent Unit (GRU) [52] can process the sequential data by carrying hidden states. These networks are implemented RL and they have shown improved performances in various POMDP environments [53, 54, 55, 56]. However, most of the development only utilizes sequential observation and not uses all available information. There are also other recent approaches to work on time series data in this context e.g., transformer [57, 58] and Liquid Time-constant Network [59].

Although many algorithms have been developed to deal with POMDP, further investigation and development are required regarding the robustness under disturbances. Very little research has studied the effects of selection of information, length of the utilized action/observation sequence, and network structure on the robustness. Also, the analysis of the disturbance characteristics is lacking, even though algorithms are tested in POMDP environments. Robustness is an important element in practical applications. Although the policy is optimized in certain conditions, the solution does not necessarily remain optimal once the configuration of the environment varies. The recent discussion of Sim2Real transfer is more focused on increasing the fidelity of the simulator [60, 61] but it is not possible to model the real world perfectly [62]. Therefore it is important to increase the robustness of the trained system.

The robust system with generalizability can keep the performance well in any condition once it is trained. In an effort to build such generalizability into the system, we consider the robustness of optimality under the specified non-static environment in this work. Namely, the networks are evaluated based on the achieved total reward per episode in POMDP environment. The robust agent, in this study, should dynamically adjust the policy in the environment with disturbances.

Firstly, the related background knowledge is briefly explained in Section 2.2. Secondly, we discuss the importance of including action sequences in the POMDP framework in Section 2.3. Then, we evaluate one of the existing algorithms which has the LSTM layer and discuss the effect of the sequential information component and the length of time window to look back on different types of disturbances in Section 2.4. Section 2.5

proposes a new structure of the network to improve the existing algorithm. With this structure, the sequential information is processed individually in the actor and critic network. Also, in Section 2.5, we discuss a different methodology to share the internal state representation between both networks. Finally, all developed algorithms are compared and evaluated in Section 2.6. The contribution of this work can be summarised as follows:

1. We explain how the selection of information matters in solving POMDP. Past action sequences can help to increase the robustness of optimality due to their causality in the environment (Section 2.3). The effect of information on the robustness is identified and discussed (Section 2.4).
2. The network structure to attend to the sequential data is discussed carefully in Section 2.5. By following the evolution of internal state construction, we show that it is more reasonable to treat the past and current information as one sequence instead of processing them independently.
3. When LSTM cells are applied in the network, the sequential information is carried by hidden states h_t and cell states c_t . We propose a new algorithm to train the critic network based on the generated hidden states from the actor network (Section 2.5). The evaluation of the proposed methodology shows the possibility to utilize them as the initial state of LSTM cells in training. This approach makes it possible to avoid reprocessing the sequence at each network during the training.

2.2 Preliminary

The neural network and algorithm related to this work are introduced in this section. Further basics of RL can be found in [63].

2.2.1 Recurrent Neural Network

RNN is nowadays a widely used technique, yet it is worth briefly revisiting the concept since later discussion and algorithm development require the detail of RNN. RNN is a suitable learning network to deal with sequential data as input. Fig. 2.1 contrasts the difference between Feedforward Neural Networks (FNN) and RNN. Unlike FNN, RNN carries the context of sequential observation and make predictions. On the other hand, FNN treats every input independently from other sequentially provided inputs and it returns outputs only depending on the current neural state [64]. For example, the text generation application requires the prediction of the next word coming based on not only the last word but also the previous words. RNN can predict output as a function of the current input and the inherited hidden states that capture contextual information from all of the previous inputs based on its trajectory. As a deep neural network application, RNN is widely applied for cases where sequential data is involved in the prediction (e.g. audio speech, time series prediction). In Fig. 2.1, the hidden state h_{t-1} from the hidden layer is carried to generate output data y_t at the output layer when new input data x_t enters in RNN. In case of RL, input data x_t can be direct observation

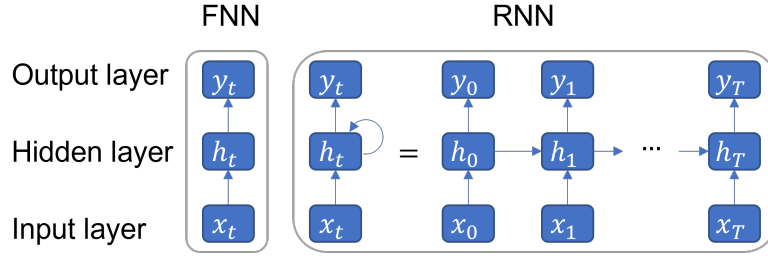


Figure 2.1: FNN and RNN diagram

o_t or output from the previous layer originated from o_t . The original RNN suffered from a vanishing gradient problem where the gradient becomes extremely small and has no effect on backpropagation through time (BPTT). To address this issue, advanced recurrent networks, such as LSTM network and GRU, with the gating network were developed. Fig. 2.2 shows the simplified schematic of LSTM cell. LSTM consists of three gates named input gate, output gate and forget gate and inherits hidden states h_t and cell states c_t to make an estimation for the next timestep. More specifically, the input gate decides what information should be stored in the cell state c_t based on the current input and inherited hidden states. If the value is closer to 1, the particular information generated from input data x_t and previous hidden states h_{t-1} has more weight to update the cell state. Also, forget gate selects how much information from the previous cell states should be forgotten or kept. If the values are closer to 0, less information from the previous cell states c_{t-1} is kept in the updated cell states c_t . By using these two gates, the network can selectively process the most informative (related) data in the input sequence. The output gate finally weights the information to be taken from the cell state as an output. GRU also has a similar gating mechanism. Due to these gates, LSTM and GRU can effectively carry the information and have a lower chance of facing the vanishing gradient problem. In this work, LSTM is utilized in one of the layers of the neural networks to approach POMDP. The network is trained via BPTT, which goes back through the entire sequence used to generate the output.

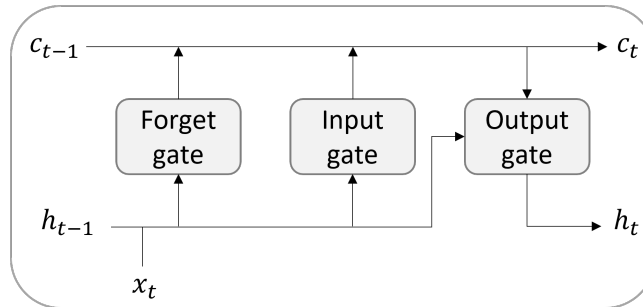


Figure 2.2: Simplified LSTM cell schematic

2.2.2 Twin Delayed Deep Deterministic policy gradient algorithm

The developments in this work are based on Twin Delayed Deep Deterministic policy gradient algorithm (TD3) [65]. The schematics of TD3 are illustrated in Fig. 2.3. o_t , a_t , r_t and a'_t in the figure represent observation, action, reward and action generated by target actor network at timestep t . Deep Deterministic Policy Gradient (DDPG) algorithm [66] showed great performance with an efficient number of iterations. However, the issue of Q-learning overestimation remained. In [65], this issue was investigated and the extended algorithm was presented as TD3. Like in the Double Q-learning algorithm [67], TD3 has a double critic network. Each critic network has a target critic network and actor network also consists of the target network and behavior network. Thanks to several modifications, TD3 showed better performance in achievable total reward compared to the other latest deep RL algorithms. Although the original TD3 might suffer from performance degradation, there are only few developments for TD3 to solve POMDP.

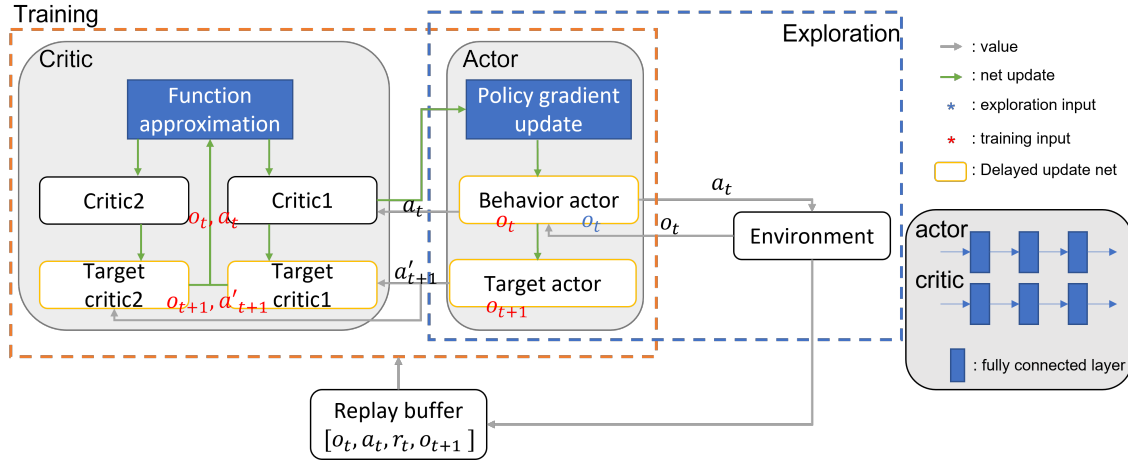


Figure 2.3: TD3 schematics

2.2.3 LSTM-TD3

In [68], the LSTM layer was implemented in the TD3 algorithm which is named LSTM-TD3. LSTM-TD3 has LSTM layers in actor and critic networks. The past sequential information is processed through LSTM layer. Processing sequence in an entire episode is time-consuming to train the network. Hence, in LSTM-TD3, the networks process partial trajectory instead of entire trajectory up to the current timestep by setting a specific sequence length l to generate the action from the policy and to roll out the trajectory during the update as a hyper-parameter. This method to train a network with the partial history of a specific length is also applied in [55] as described in Section 2.7. Another aspect to be mentioned is that LSTM-TD3 has a double input channel as network architecture as shown in Fig. 2.4. The input data for the first input channel is sequential past information $[o_{t-l:t-1}, a_{t-l:t-1}]$ from the memory and the second input channel takes the current information $[o_t, a_t]$. The expression $o_{t-l:t-1}$ and $a_{t-l:t-1}$

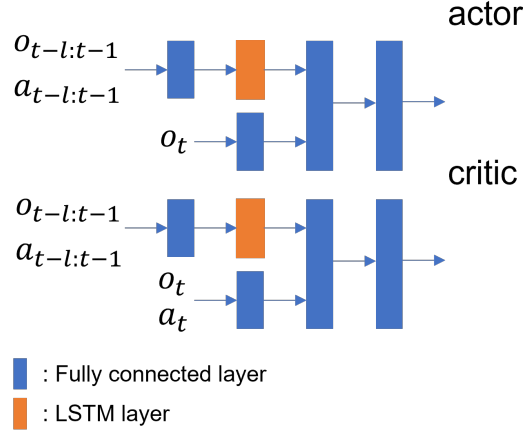


Figure 2.4: LSTM-TD3 schematics

describes the sequential observation and action between timestep $t-l$ and $t-1$ respectively. This structure was selected in order to prioritize the most recent observation since it is considered that the recent observation is more valuable in decision-making than earlier observations. Note that this hypothesis is questionable because it might not be true that the latest observation is more informative to reveal the true state of the system. This will be discussed in detail in Section 2.5. LSTM-TD3 [68] was tested for robot locomotion tasks in open AI gym simulation environment [48] with several types of POMDP such as removing one of the state elements from the observation and adding random noise to observation. Especially in the scenario where one or all of the observation is lost with probability p , LSTM-TD3 showed greater performance compared to original TD3 and other Deep RL algorithms.

2.3 Dynamic RL

The challenge of POMDP in this study is that we have no access to the true state of the system s_t due to the disturbances in observations. In our scenarios, disturbances are realized by stochastic noise or dynamic noise, or by hiding one of the elements of the observation. The details of disturbances are defined in Section 2.4. If s_t can be estimated or represented in an internally encoded format, it is possible to restore the MDP condition. Then, an already established RL algorithm can be utilized to identify the (sub-)optimal policy. Therefore, recovering the MDP condition through certain architecture design choices such as the definition of input/output variables should be important for the policy robustness. Hence our focus is to develop an RL agent which can dynamically adapt to the observation containing the disturbance with their own dynamics. To make a dynamic RL agent, the system should be able to generate the internal state representation by assessing causal information from the trajectory.

2.3.1 Causality and Statistics

The dichotomy of statistics and causality are emphasized in [69]. A statistical model works on the joint probability of observed variables. A causal model, on the other hand, assumes that certain variables are unobserved. The causal analysis enables inference of dynamics of events under changing conditions, not only the likelihood of events under static conditions.

In MDP, the static transition model is available (e.g., $P(o_{t+1}(=s_{t+1})|o_t(=s_t), a_t)$), where o_t , s_t , a_t are observation, state, and action at time step t . However, in POMDP with non-static disturbances, the model of the environment dynamically changes. Therefore, it requires an additional mechanism to identify the causal effect and recognize patterns based on available information from the past and current.

2.3.2 Analysis of disturbance

When the decision-making is performed based on information on the trajectory, the causal system is illustrated in Fig.2.5. In this work, the considered disturbances can be categorized into two. The first case is where the disturbance can be described by a dynamic model (Fig. 2.5a). Therefore, the trained agent may be able to learn a model comprising the system and the disturbance. The second case is where the disturbance does not have a temporal correlation between them and cannot be modeled (Fig. 2.5b). In this case, the agent should learn to eliminate the effect of the disturbance to reveal the true state. The trained agent in the environment of which category described in Fig. 2.5a might show generalizability in the environment where the existing disturbance has temporal correlation but not the other one. This will be demonstrated in Section 2.6.

2.3.3 POMDP formulation

In general, the POMDP model is defined as 6-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, \mathcal{R} \rangle$ [49], where

- $\mathcal{S}$ is the state space,
- $\mathcal{A}$ is the action space,
- $\mathcal{O}$ is the observation space,
- $\mathcal{T}(s_t, a_t, s_{t+1})$ is the state transition which is a conditional probability distribution $P(s_{t+1}|s_t, a_t)$,
- $\mathcal{Z}(s_{t+1}, a_t, o_{t+1})$ is the observation function which is a conditional probability distribution $P(o_{t+1}|s_{t+1}, a_t)$,
- $\mathcal{R}$ is an immediate reward function which represents $\mathcal{R} = \mathbb{E}[r_{t+1}|s_t, a_t]$.

When observations are corrupted by additive dynamic disturbances, the observation function $\mathcal{Z}$ is not static. Namely, the observation function can be described as $P(o_{t+1}|s_{t+1}, a_t, d_t)$. In addition, if the disturbance is modeled and it is independent of the states and action, a new disturbance transition function, which represents a conditional probability distribution could be introduced.

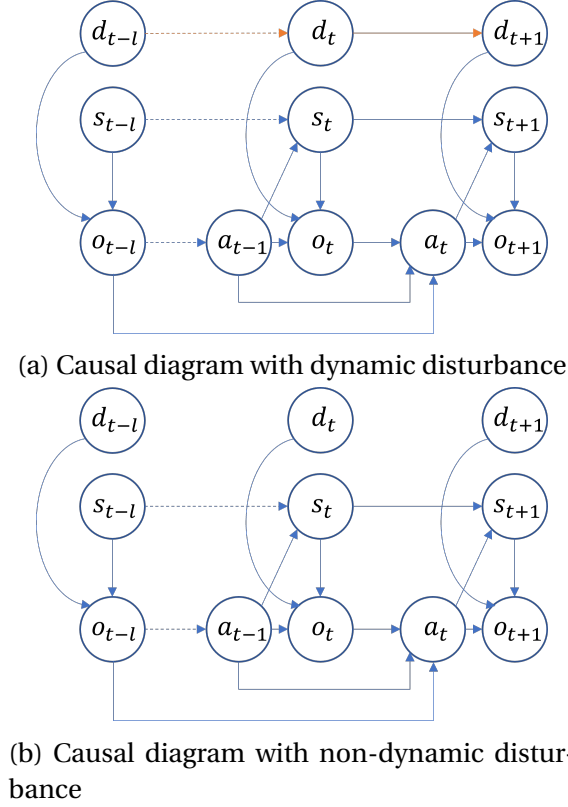


Figure 2.5: Causal diagram

2.3.4 Information states, belief states in POMDP

In the context of dynamic programming, the optimal controller in POMDP can be separated into two parts: (a) an estimator and (b) an actuator [70, 49]. The information available at time t is called complete information state I_t^C [71] which consists of all historical observation and action i.e., $I_t^C = (o_{0:t}, a_{0:t-1})$, where $o_{0:t} = o_0, \dots, o_t$ and $a_{0:t-1} = a_0, \dots, a_{t-1}$. In the case of POMDP, the estimator could make the conditional expectation $\hat{s}_t = \mathbb{E}(s_t | I_t^C)$. The challenge of utilizing the complete information states is that the size of the data expands as time elapses. To resolve this issue, quantities known as sufficient statistics are normally considered. The sufficient statistics has a smaller dimension and preserves the essential content of I_t^C . The sufficient statistic is generated at the estimator and the actuator generates control inputs to the system based on the sufficient statistic. In the standard POMDP model, a belief state b_t is commonly used as sufficient statistics. The belief state assigns conditional probabilities of every possible state [72]. As illustrated in [49], the policy of the agent receives the belief states to make optimal decisions. The belief states could be recursively updated via Bayesian inference. The update of belief states can be described as (2.1).

$$\begin{aligned}
 b_{t+1}(s_{t+1}) &= P(s_{t+1} | o_t, a_t, b_t) = \frac{P(o_t | s_{t+1}, a_t, b_t) P(s_{t+1} | a_t, b_t)}{P(o_t | a_t, b_t)} \\
 &= \frac{\mathcal{Z}(s_{t+1}, a_t, o_t) \sum_{s_t \in \mathcal{S}} \mathcal{T}(s_t, a_t, s_{t+1}) b_t(s_t)}{\sum_{s_{t+1} \in \mathcal{S}} \mathcal{Z}(s_{t+1}, a_t, o_t) \sum_{s_t \in \mathcal{S}} \mathcal{T}(s_t, a_t, s_{t+1}) b_t(s_t)}
 \end{aligned} \tag{2.1}$$

2.3.5 Identification of transition model

The objective of RL is to maximize the future cumulative reward G_t by learning policy π . In model-based RL, it is solved by decomposing the problem into learning and planning. From learning, the transition model is explicitly obtained from experience, and the value function is established based on the learned model in planning. On the other hand, the agent is directly trained from the experience in model-free RL. Thus, the explicit transition model and its update operation formulated in (2.1) are not required. However, the concept itself is still applicable since it is established based on it. In model-free RL, the model is internally generated based directly on the collected data. Actions are part of the transition model and, thus it is part of a recursive update. Therefore, it is clear that taking action sequences in model-free RL will help to improve the robustness of optimality in POMDP.

Regardless of whether the complete information state I_t^C or belief state b_t is used, the identification of the dynamic transition model depends not only on the historical information of the observation but also that of the action. This is because both have causal connections with the current observation, as illustrated in Fig. 2.5. As we are considering the environment where the observation is disturbed, our discussion is focused on generating internal representations of s_t which is symbolized by s_t^* in the rest of this paper. The key idea is to formulate a structure of an RL agent which can dynamically adapt the state representation s_t^* to appearing disturbances.

2.3.6 LSTM to reflect the sequence on the internal representation

When the model of the system is not available, s_t^* could be generated via RNN or enhanced RNN by processing the sequential input. s_t^* could contain the identified information regarding the pattern of the dynamic model of the system and the disturbance or distilled information which is less affected by the disturbance. Most of the existing RL algorithms with RNNs only utilized the sequential information of observation and there are only very few cases taking action sequence as input. We suggest utilizing the sequence of action based on the above discussion in this study. Therefore, the effect of sequential action inclusion with the LSTM-TD3 algorithm [68] is investigated in the next section.

2.4 Action inclusion

In the previous section, we have discussed that the belief state evolves based on the observation and action on the trajectory. In this section, the impact of the inclusion of the action sequence is demonstrated using LSTM-TD3. Also, the effect of the length of the sequence for the individual disturbance condition is investigated.

2.4.1 Length of sequence

When the partial trajectory is taken as the input sequence, the selection of its length l can be a design parameter as in LSTM-TD3. In general, RL algorithms with a RNN layer require repeating either entire or partial episodes during the update in order to

train the network via BPTT. When entire episodes are repeated in training, it can be considered that the agent is trained using complete information I^C . As stated in Section 2.3, the size of I^C increases as time elapses. Thus, repeating entire trajectories to update the network is a time-consuming process. In case of the on-policy algorithm with RNN, utilizing I^C might be a more natural approach since it follows the same trajectory during the training. Several works have explored the methods combining RNN with on-policy algorithms (e.g. with PPO[73] and A3C [74]). However on-policy algorithms are much less sample-efficient compared to off-policy algorithms.

For state value function based algorithms, parameter sharing is a common technique where the actor and critic share the same network parameters and only the last layer of each network is different. However, this technique is not as common for action value function based algorithms like TD3 since the parameter sets are different between actor and critic networks. When parameter sharing is not used, the same trajectory needs to be repeated multiple times to update multiple networks.

If the whole episode is repeated, the benefit provided by random sampling can be lost in the experience replay which is the mechanism for improving data efficiency and stability. Experience replay effectively minimizes the correlations between samples [66] so that the data distribution becomes closer to Independent and Identically Distributed (IID) condition which is preferable in RL. In [75], it was pointed out that if the whole trajectory is sampled together, the bias in the learned policy might become more severe in the update because the state distribution under the current policy no longer corresponds to the one from the replay buffer.

If a few steps of sequential data are randomly sampled as a unit, the sample correlation issue could be mitigated. This approach captures a shorter history of fixed length in the environment. LSTM-TD3 offers the opportunity to define the length of the sequence. The agent should be provided with enough length of sequential data so that it can capture the characteristics of the system and the disturbance. In the experiment in this section, the relationship between several characteristics of the disturbance and the length of sequence that is provided to the network are investigated.

2.4.2 Experiment

To validate the discussion in Section 2.3 and investigate the effect of including action sequence on the different types of disturbances, we compare learning trajectories of the LSTM-TD3 algorithm with or without the past action sequence in the first input channel (see Fig. 2.4) in classic control "Pendulum" environment from OpenAI gym. In this environment, the observation consists of three elements $[x, y, \dot{\theta}]$, where x and y represent the position of the free end of the pendulum and $\dot{\theta}$ is its angular velocity. The action is defined as the torque to be added and the agent learns to keep the pendulum upright from a random initial condition. The definition of the "Pendulum-v0" environment is summarized in Table 2.1.

The motivation for this work is to develop an algorithm that robustly provides optimality in the presence of disturbance. As summarized in Table 2.2, we classified disturbance under condition into five types: "temporal bias", "temporal sinusoidal wave", "random sinusoidal wave", "noise" and "hidden" conditions. In the "temporal bias" and the "temporal sinusoidal wave" cases, the observation of x and y are disturbed

Table 2.1: "Pendulum-v0" environment definition

action	torque	min -2.0	max 2.0
observation	$x = \cos \theta$	min -1.0	max 1.0
	$y = \sin \theta$	min -1.0	max 1.0
	angular velocity	min -8.0	max 8.0
reward	$r = -(\theta^2 + 0.1\dot{\theta}^2 + 0.001(\text{torque})^2)$		
termination	at 200 timesteps		

by additive constant biases and sinusoidal waves for specific lengths and times at random timesteps. In the "random sinusoidal wave" case, the sinusoidal wave is randomly configured at each episode and all three elements are disturbed. In the "noise" case, zero-mean Gaussian noise with standard deviation σ , is added to each element of the observation during the entire episode. In the "hidden" condition, $\dot{\theta}$ is removed from the observation during the whole episode reducing the dimension of the observation to two. Scenarios 1,2, and 3 in Table 2.2 are the cases where the disturbance has its dynamics and patterns could be identified from the observation. The other remaining two cases have no dynamics in the disturbance. The networks should rather work on recovering the dynamics of the system by eliminating the effect of the disturbance from the observation than establishing a model that includes both the system and the disturbance.

2.4.3 Results and analysis

Fig. 2.6 to Fig. 2.11 show averaged learning trajectories after 5 trials of the LSTM-TD3 algorithm in the discussed scenarios. The orange lines, labeled "w/o action", show the case in which sequential action information is not considered. Specifically, the first input channel of actor and critic networks process only sequential $l - 1$ length of historical observation $o_{t-l:t-1} (= o_{t-l}, \dots, o_{t-1})$. The blue lines, labeled "w/ action", show the case where the historical action $a_{t-l:t-1}$ is used as additional information. The columns represent the length of the history to be processed. Fig. 2.6 and Fig. 2.7 illustrate the results of the MDP case and the "temporal bias" case respectively. In the MDP case, LSTM-TD3 with or without action performs the same. Also, the final performance at the end of training in the "temporal bias" case did not show obvious differences. In this condition, a constant value is added for three timesteps that randomly start appearing 10 times. Since the observation is biased by a constant value, the observation is still almost unique. Therefore, it can be considered to be similar to the MDP condition.

Fig. 2.8 shows the results of the "temporal sinusoidal wave" scenario. As l gets longer, the robustness improves. The wavelength of the defined sinusoidal wave is 70. Longer l helps to identify the system dynamics and the disturbance dynamics. Also, the case with action sequence included as the input to the agent achieved higher total rewards. The result indicates the impact of information selection on robustness.

Fig. 2.9 represents the learning trajectory of the "random sinusoidal wave" scenario. Establishing the observation function in this scenario is much harder than in the "tem-

Table 2.2: Test scenario

Scenario	Name	Description
1	temporal bias	Disturbance defined by $A \in [0.5, 1.0]$ appearing randomly is added to observation in x and y lasting 3 timesteps for 10 times per episode.
2	temporal sinusoidal wave	Disturbance defined by $A \sin 2\pi t / T$ ($A = 1$ and $T = 70$) which starts appearing at randomly chosen timesteps for the period of T , is added to observation x and y .
3	random sinusoidal wave	Disturbance randomly defined in every episode by $A \sin 2\pi t / T$ ($0.5 \leq A \leq 2$ and $10 \leq T \leq 100$) which starts appearing at random timestep for the period of T , is added to all 3 observation elements.
4	noise	Noise defined by zero-mean normal distribution with standard deviation $\sigma \in [0.5, 1.0]$ is constantly added to all 3 observation elements.
5	hidden	Angular velocity $\dot{\theta}$ is removed from observation.

poral sinusoidal wave" due to the infinite variety of sinusoidal waves. It requires the agent to dynamically adapt to the environment based on causality of action and observation and hence, the network with action sequence achieves greater optimality. Fig. 2.10 illustrates the learning trajectory under the "noise" condition. This scenario is more challenging than the "temporal sinusoidal wave" since there are no specific dynamics in the disturbance and all of the elements of the observations are constantly disturbed. Although the network considering sequential action information performs better, the performance degrades with the larger variance of the noise. The performance improves with longer l . Longer sequential data contributes to extracting the system state transition characteristic and eliminating the effect of the noise in LSTM. In the white noise case, some data randomly contain less noise. By focusing on these data and putting less focus on more noisy data, the underlying pattern can be restored. When the length of the input sequence is longer, it has more chance to have informative data that is less affected by noise. In addition, as more episodes are experienced, degradation of achieved total reward was observed in the "noise" scenario. This would be due to the occurrence of overfitting. It might be prevented by changing the network configuration, e.g., reducing the number of parameters, and adding dropout layers. Fig. 2.11 presents the result of the "hidden" scenario. The final performance is the same regardless of the action inclusion and length of the sequence. In this environment, only current observation o_t and one step before o_{t-1} should be sufficient to re-

store $\dot{\theta}$ since θ is updated as $\theta(k+1) \leftarrow \theta(k) + \dot{\theta}(k+1)dt$. The formula only requires the previous observation to restore $\dot{\theta}$ from the equation and therefore the performance is not affected by l . However, the past action information is making the learning process more efficient. The network does not have this mentioned interpretation and the reward signal is mixed with θ , $\dot{\theta}$, and torque. The inclusion of action information could make learning efficient because the network learns to process the dynamics under more specific conditions. The most reasonable network structure to capture this interpretation might instead just consist of fully connected layers by having the input of $[x_t, y_t, x_{t-1}, y_{t-1}]$ because the sequential process is not required.

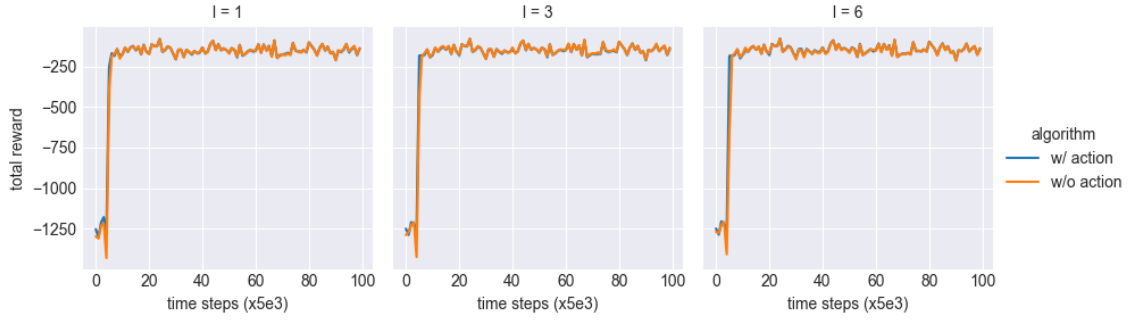


Figure 2.6: MDP

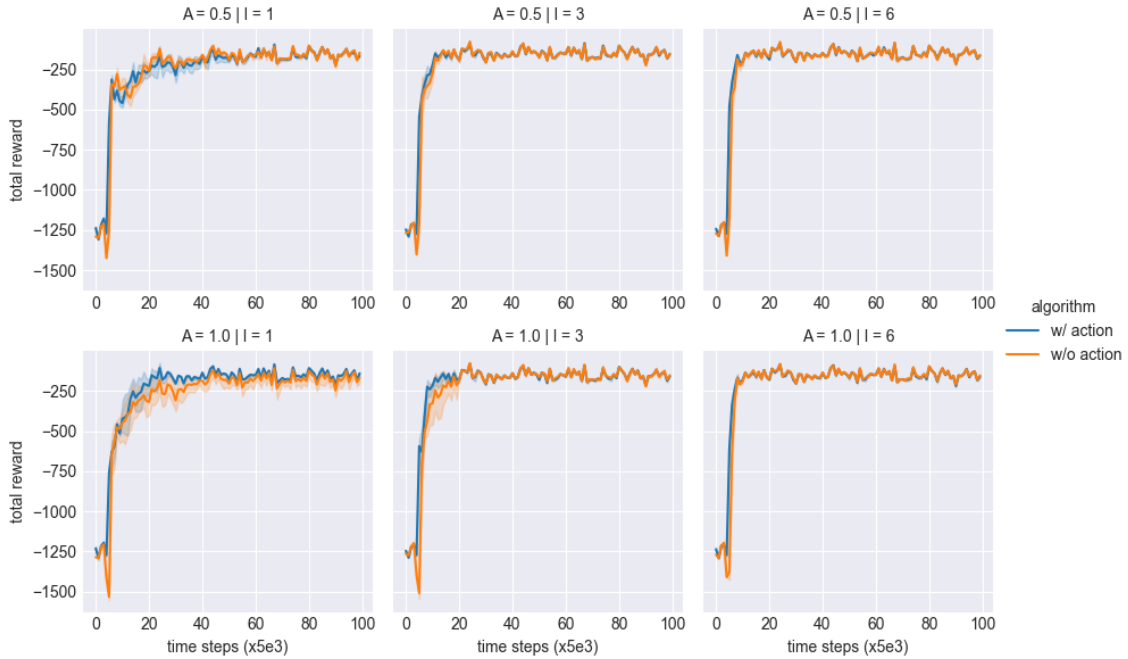


Figure 2.7: temporal bias

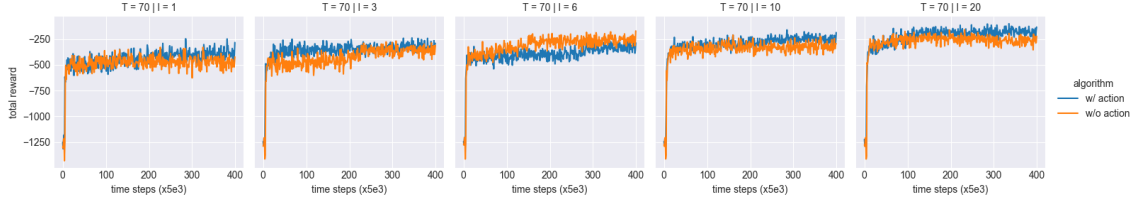


Figure 2.8: temporal sinusoidal wave

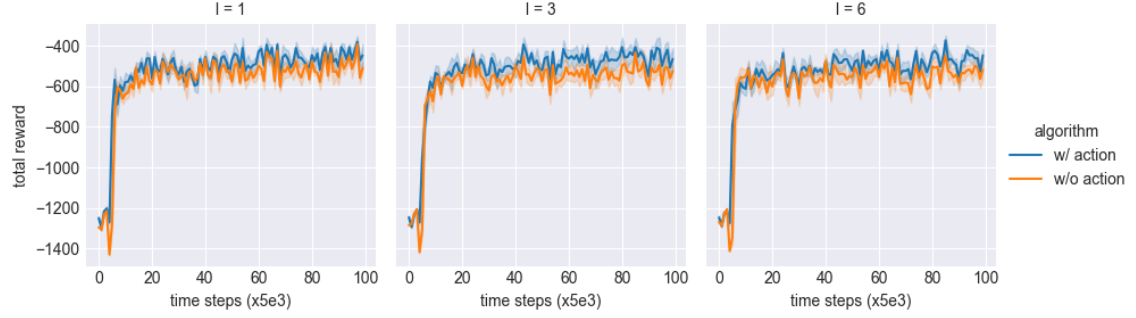


Figure 2.9: random sinusoidal wave

2.5 Structure

In Section 2.4, we demonstrated that including the action sequence can improve the robustness of learning. In this section, let us discuss the structure.

2.5.1 Interpretation of LSTM-TD3

The LSTM-TD3 has two input channels in both actor and critic networks. They could be interpreted as the first input channel to generate s_{t-1}^* from complete information $I_{t-l:t}^C$ taken from the memory. Then, the s_{t-1}^* is updated to s_t^* by combining the current observation information from the second channel at the actor. Our hypothesis is that it might not be necessary to have a double-headed structure as presented in Fig. 2.4 because the information up to the current timestep can be treated as one sequence according to the generation of the belief states.

2.5.2 Modified LSTM-TD3

Based on this hypothesis, this paper proposes an actor network with a single input channel as illustrated in Fig. 2.12. This input channel processes $I_{t-l:t}^C = (o_{t-l}, \dots, o_t, a_{t-l-1}, \dots, a_{t-1})$. Internal state representation s_t^* is generated in the process of the LSTM layer and action is decided based on s_t^* by the succeeding layers. For the critic network, two types of structure are considered. LSTM-TD3_{1ha2hc} (1 headed actor and 2 headed critic) in Fig. 2.12a has a clearer interpretation for s_t^* to the information states. s_t^* is generated in the same way as actor network at the first input channel and it is concatenated with the processed a_t at another input channel. The last two layers can be

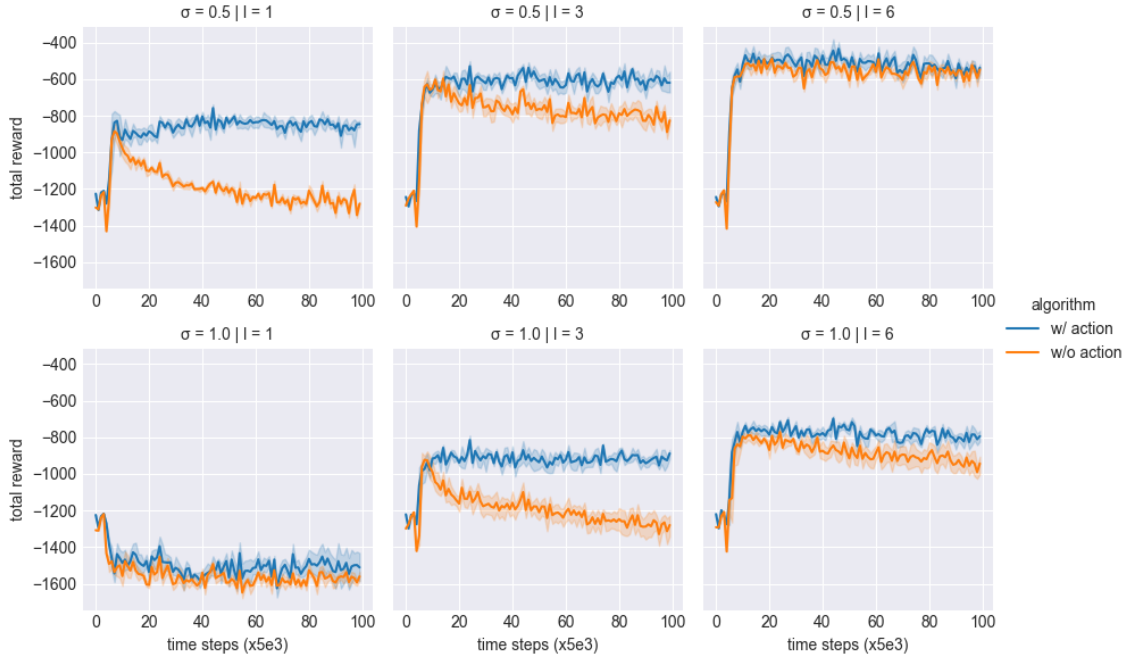


Figure 2.10: noise

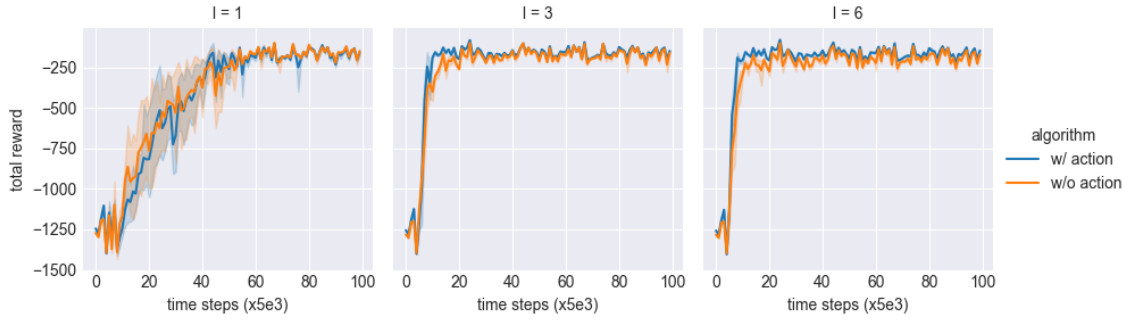


Figure 2.11: hidden

understood as $Q(s_t^*, a_t')$. On the other hand, $\text{LSTM-TD3}_{\text{1ha1hc}}$ has single input channel and modified complete information $I_{t-l:t}^{C'} = (o_{t-l}, \dots, o_t, a_{t-l}, \dots, a_t)$ is taken as input. $I_{t-l:t}^{C'}$ differs from the generation of s_t^* based on the Bayesian inference and hence, the output from the LSTM layer is described as sa_t in the figure. However, the definition of action value function Q is the expected return at s_t , taking action a_t when following policy π . When the critic network is given with the trajectory, $I_{t-l:t}^{C'}$, which is the combination of o_t and a_t from π , it should be able to still generate the expected return for the policy π .

Unlike the original LSTM-TD3, $\text{LSTM-TD3}_{\text{1ha2hc}}$ and $\text{LSTM-TD3}_{\text{1ha1hc}}$ do not have an explicit intention to prioritize the current information. These proposed algorithms follow the update of the belief state and let LSTM to evaluate the importance of the data. Moreover, since the system model does not change in any timestep, all data in a sequence should be treated in the same path. Training the agent with separate sequences

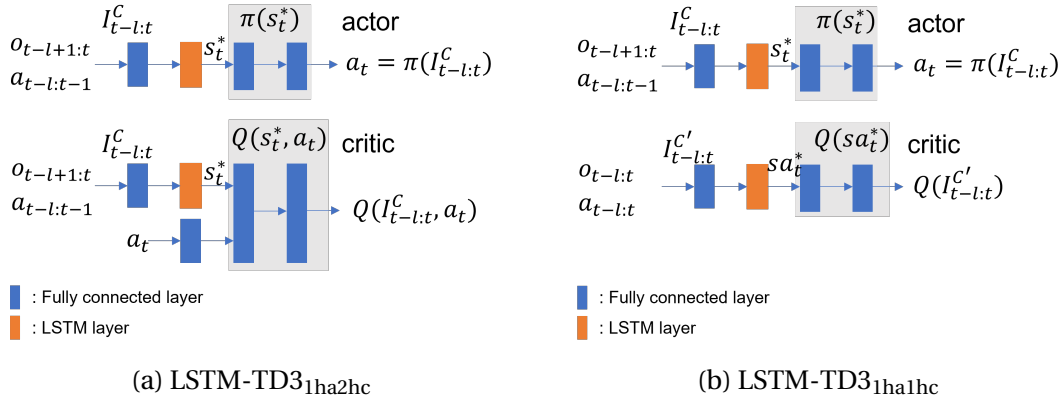


Figure 2.12: Modified LSTM-TD3 structure

might be more complicated because the data is processed differently and combined which might be unnecessary according to the belief state update. Therefore, LSTM-TD3_{1ha2hc} and LSTM-TD3_{1ha1hc} are expected to show better robustness.

2.5.3 H-TD3 algorithm

The previously discussed algorithms generate s_t^* in actor and critic networks independently by taking complete information at each network. Particularly, we have introduced the interpretation where s_t^* is produced in the LSTM layer at actor network. The research question to be also investigated is whether or not it is necessary to have a different representation of s_t^* between the two networks. If s_t^* generated from the actor network can be shared with the critic network, it will drastically reduce the computational costs. Because TD3 has double critics as seen in Fig. 2.3, $I_{t-l:t}^C$ is processed four times at each critic network during the update. Hence we propose, a new algorithm, named H-TD3 (Hidden-state-based TD3) in which the critic networks are trained based on the shared s_t^* from the behavior actor network during the exploration. Here, we consider the same network structure with LSTM-TD3_{1ha1hc} which has a LSTM layer with a single input channel for the critic network as it is illustrated on the right side of Fig. 2.13. The main diagram of Fig. 2.13 describes the data flow of the H-TD3 algorithm. During exploration, the behavior actor and environment interact and the collected data is stored in the replay buffer. In addition to the normal data set, the hidden state and the cell state of the LSTM cell are stored. These states are generated at the behavior actor network when LSTM processed the sequence of $a_{t-l-1:t-2}$ and $o_{t-l:t-1}$. During the training, the critic network is trained based on this tuple. Critic networks take a_t and o_t as input data as the original TD3 algorithm. However, the LSTM cell is initialized with the stored $(h_{t-l:t-1}, c_{t-l:t-1})$. In this way, input data a_t and o_t is treated as if a continuous sequence of $a_{t-l-1:t-2}$ and $o_{t-l:t-1}$ without repeating the entire trajectory between $t-l$ to t . In the same way, at the target critics, the target Q-values are calculated based on input data a'_{t+1} and o_{t+1} with initialized LSTM state with $(h_{t-l+1:t}, c_{t-l+1:t})$. The critic network is required to learn parameters based on $(a_t, o_t, h_{t-l:t-1}, c_{t-l:t-1})$. The criticism of this approach may be the dismissed action a_{t-1} and a_t at critic and target critic networks respectively. However, if the hidden states $(h_{t-l:t}, c_{t-l:t})$ of LSTM after processing all sequence $a_{t-l-1:t-1}$ and $o_{t-l:t}$ are used

to initialize critic networks, o_t is counted twice and it would break the sequence at the critic. Another possibility could be to limit the input just to a_t , but changing the character of input data to LSTM for the given $(h_{t-l:t}, c_{t-l:t})$ in critics would add difficulty in learning. During the development, we observed the degradation of performance in both cases compared to using $(h_{t-l:t-1}, c_{t-l:t-1})$.

Considering the developed LSTM-TD3_{1ha2hc} and LSTM-TD3_{1ha1hc}, the output data from LSTM in actor network is the last hidden states h_t after processing l length of the sequential input. After the LSTM layer, h_t is carried to the following last two layers. Hence, the simplest method for the critic networks to learn based on the shared s^* would be taking h_t as input data at the critic network in addition to the current observation. However, this operation would extend the input dimension excessively as the dimension of the hidden state is usually much larger than that of the input. It increases the difficulty to train the network and the risk of failure of learning as known as the "curse of dimensionality". Training network with larger input dimensions increases the number of parameters drastically, leading to the much greater computational complexity of training. Also, ending up with over-fitting is a common issue. Therefore, it is more scalable to share the hidden and cell states produced in the actor network to initialize the hidden state of the critic network as in the proposed H-TD3 algorithm.

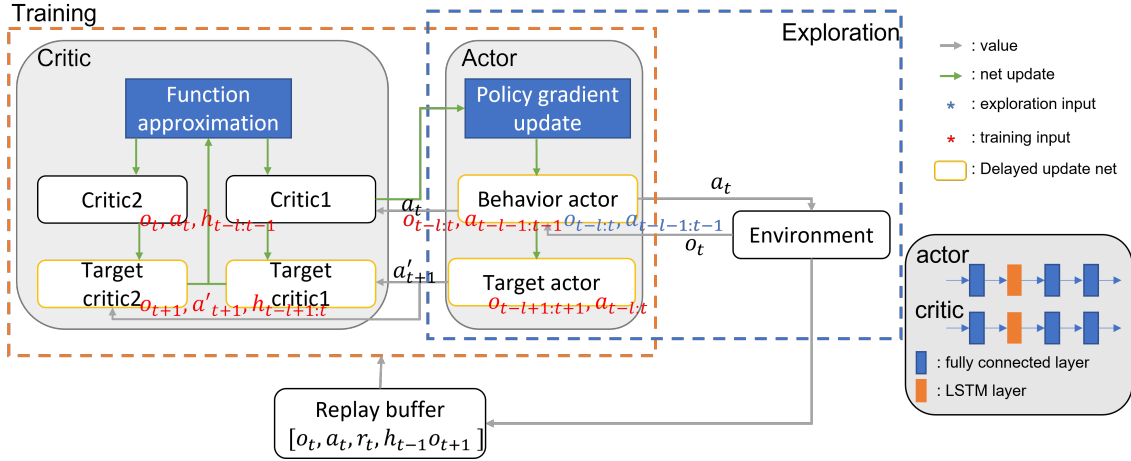


Figure 2.13: H-TD3 schematics

2.6 Experiment

2.6.1 Simulation set-up

In this section, the algorithms developed in Section 2.5 are tested in the "Pendulum" environment with disturbance conditions discussed in Section 2.4. The applied disturbance conditions are as follows:

- "temporal sinusoidal wave" with $\sin \frac{2\pi}{70} t$
- "random sinusoidal wave"

- a zero-mean Gaussian "noise" with $\sigma = 0.5$
- "hidden"

2.6.2 Results

In Fig. 2.14, the learning trajectories of the proposed algorithms are depicted with trajectories being averaged over 5 trials. We experiment with various time windows l specifically 1, 3, 6, 10, 20. In the array of figures, each row corresponds to a specific disturbance condition, while each column represents an algorithm. The algorithms compared are original LSTM-TD3 without past action sequence, LSTM-TD3 with past action sequence, LSTM-TD3_{1ha1hc}, LSTM-TD3_{1ha2hc}, H-TD3 and original TD3 from the left column. The disturbances are "temporal sinusoidal wave", "random sinusoidal wave", "noise" and "hidden" from the top. The number of the iterations varies depending on the disturbance conditions.

Fig. 2.15 shows the averaged total reward of the last ten records by taking algorithm types on the x-axis. Each figure describes the result of "temporal sinusoidal wave", "random sinusoidal wave", "noise" and "hidden" conditions respectively. The lines are drawn to compare the effect of the time window l .

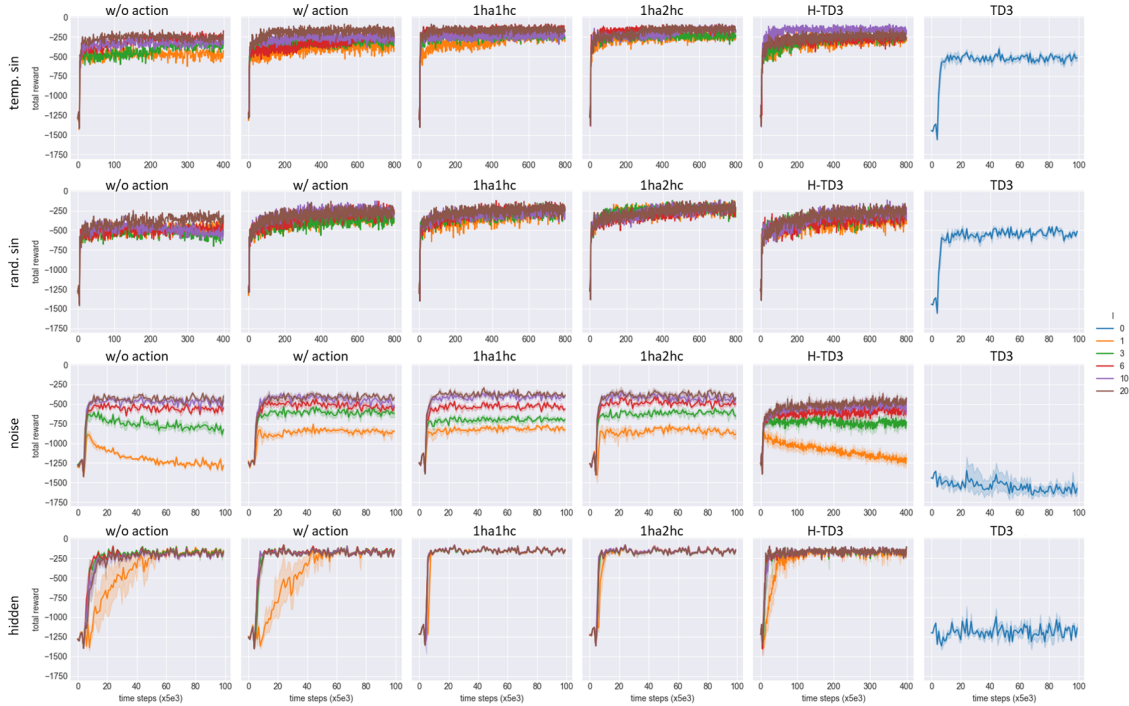


Figure 2.14: Learning trajectory

Overall analysis Compared to TD3, all algorithms containing LSTM layer demonstrated superior performance in the tested POMDP conditions. Particularly, in the "noise" and the "hidden" cases, TD3 struggled to learn. The "noise" environment generates noisy observations. Without access to a sequence, it is difficult to learn since the

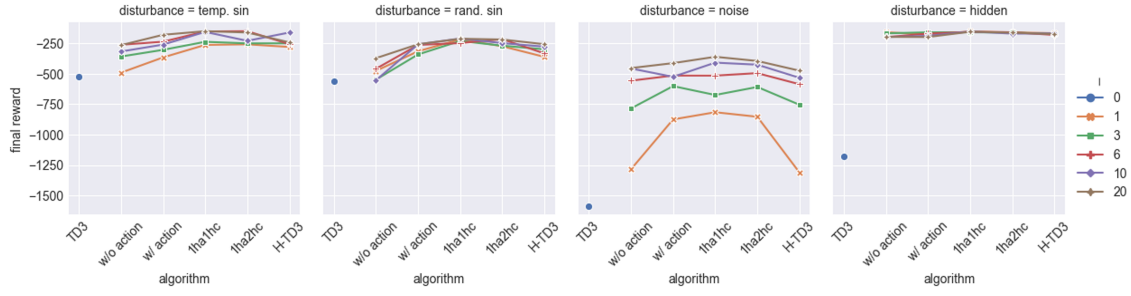


Figure 2.15: Final total reward

observation values are not unique to the reward signal. Also, the "hidden" scenario, a network without an LSTM layer struggles to recover information related to angular velocity which is part of the reward function.

As observed in Section 2.4, all algorithms incorporating action sequences performed better than LSTM-TD3 without an action sequence. The variants LSTM-TD3_{1ha1hc}, LSTM-TD3_{1ha2hc} generally performed better performance than with the original LSTM-TD3 with action sequence. From the results, it can be understood that it is better to treat the past sequence and current data as one sequence, unlike LSTM-TD3. LSTM-TD3_{1ha1hc} exhibited the best robustness of optimality of all algorithms. The simplicity of the structure could have helped the network to learn. H-TD3 algorithm delivered a comparable result with LSTM-TD3 with action case except for the "noise" case. It confirms that it is possible to learn from shared hidden states as demonstrated by the H-TD3 results. However, H-TD3 exhibited slower learning trajectories for a given number of iterations than other algorithms in Fig. 2.14 due to the critic's dependence on the actor network to generate the hidden states and cell states. Interestingly, the results of the "noise" case in H-TD3 were significantly degraded and mostly worse than LSTM-TD3 without action sequence. In the H-TD3 algorithm, a_{t-1} is not included in the information. The effect of omitting this information could become prominent in some conditions as in the "noise" case.

Computational time and size Let us examine the computational time for training the network. Fig. 2.16 shows the comparison of the training time with LSTM-TD3 with $l = 1$ serving as a benchmark. Due to the added operations in the code, LSTM-TD3_{1ha1hc} and LSTM-TD3_{1ha2hc} exhibit a larger iteration time. The code is not optimized and there remains room for improvements in computation. As shown in the figure, the H-TD3 provides a much shorter time per iteration and is less affected by the length l . It is because it avoids repeating the trajectory in critic networks.

Generalisability So far, the networks have been evaluated in the same environment in which they were trained. However, in practice, the agent may experience different types of disturbances from the trained environment. For this reason, we tested how the networks trained in the "random sinusoidal wave" environment perform in "bias" and "noise" cases and also newly added environments: "combinational sinusoidal wave" and "damped sinusoidal wave". In the "combinational sinusoidal wave",

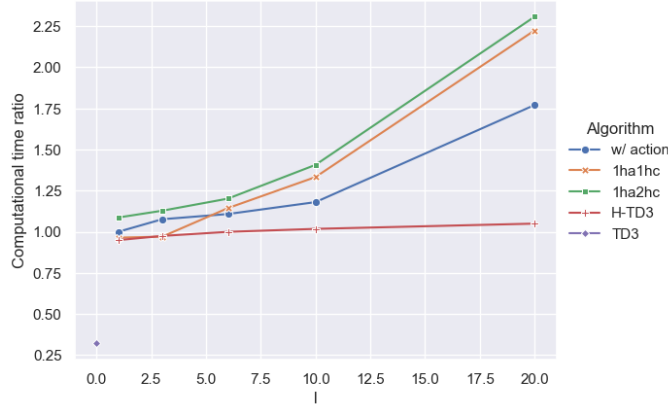


Figure 2.16: Iteration time comparison

two randomly configured sinusoidal waves with the same definition of "random sinusoidal wave" simultaneously appear, i.e., disturbance = $A_1 \sin 2\pi t / T_1 + A_2 \sin 2\pi t / T_2$ where $A_1, A_2 = [0.5, 2]$ and $T_1, T_2 = [10, 100]$. On the other hand, the "combinational sinusoidal wave" adds damped sinusoidal waves on all 3 observation elements independently defined by $e^{-(t-t_0)/T} A \sin 2\pi(t-t_0)/T$ ($A = [0.5, 2]$ and $T = [10, 100]$). The timesteps to initiate damped sinusoidal wave t_0 are randomly selected and this disturbance continues till the end of the episode. The "bias" cases with the amplitude 1.0 and the "noise" cases with $\sigma = 1.0$ are also selected in this experiment. The normalized total reward of the episode after 1000 time of trials is illustrated in Fig. 2.17. The "comb. sin" and "damped sin" represent "combinational sinusoidal wave" and "damped sinusoidal wave" respectively. In the first column, the result of the "random sinusoidal wave" is shown as a reference. The trained network performed well except in "noise" disturbance. The dynamic model of disturbance exists in "combinational sinusoidal wave", "damped sinusoidal wave" and "bias" cases. As discussed in Section 2.3, the required behavior is different between the "noise" case and the other three cases. This result reinforces the hypothesis that the network trained in "random sinusoidal wave" utilizes the dynamic model of the disturbance and the system dynamic model and is compatible with environments that have temporally correlated disturbances.

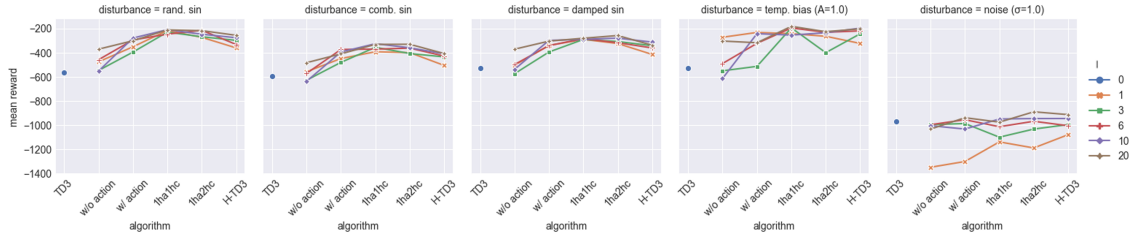


Figure 2.17: Generalisability

2.7 Related Work

2.7.1 Utilizing RNN in RL

There are many existing developments that utilize RNN for RL algorithms to solve POMDP. A few examples are reviewed in this section.

RL-LSTM The utilization of LSTM in RL has been discussed for a long time. In the earlier development, RL-LSTM [76] was presented with a simple grid T-maze experiment using Advantage(λ) learning [77]. In their T-maze experiment, the agent is required to remember the first observation in the entire episode to reach the goal with optimal action. It was shown that RL-LSTM could effectively perform in this non-Markovian scenario. In the recent context of RL, the deep neural network is coupled with the gated recurrent neural network to address the challenge of solving POMDP.

Deep Recurrent Q-learning (DRQL) In Deep Recurrent Q-learning (DRQL) [55], LSTM cells were added to the network by replacing the first dense layer after a convolutional network of Deep Q-learning (DQN) [78] in discrete action space. In their work, two types of updates were considered; Bootstrapped Sequential Updates and Bootstrapped Random Updates. The former randomly selects episodes and selected full episodes are replayed from the first step to the end. Hence, the hidden state of LSTM is carried out without initialization throughout one episode. The latter case randomly selects the point of the episode and unrolls the experience for certain timesteps. Bootstrapped Sequential Updates can capture longer sequential information but it violates the random sampling policy of DQN. Bootstrapped Random Updates can keep the random sampling but captures the information in a shorter time scale since it requires hidden states of LSTM to be initialized at the beginning of each step. Both types showed similar performance in their investigation, and Bootstrapped Random Updates was finally applied for further discussion because it takes a shorter time to train. In their work, the past action information was not considered, instead, integrated visual information from the console screen is received as input.

Recurrent DPG (RDPG) Heess et al. [75] developed Recurrent DPG (RDPG) by extending Deterministic Policy Gradient (DPG) [79] with LSTM in actor-critic style by defining actor and critic as $\mu^\theta(h_t)$ and $Q^\mu(h_t, a_t)$ respectively, in terms of observation-action history $h_t = (o_1, a_1, o_2, a_2, \dots, o_{t-1}, a_{t-1}, o_t)$. $\mu^\theta(h_t)$ represents a deterministic policy μ with parameters θ . LSTM is used to summarize the history to make it scalable in POMDP. Similar to the Bootstrapped Sequential Updates in DRQL, during the update, a minibatch of N episodes is sampled to construct histories h_t^i ($i = 1 \dots N$), and the target values are computed for each sampled episode through BPTT. It was tested in several environments, with the most related ones being "the under-actuated pendulum" and "cartpole swing up" examples. In the experiments, the velocity-related information was removed from the observation. RDPG could learn the tasks that DDPG could not perform.

2.7.2 Separating between estimation and control

It is also a common approach to separate the estimation and control strategy using information states in reinforcement learning for POMDP conditions [80, 81]. The information states are updated based on Bayesian inference. Then, the control strategies are learned by taking the information state into account. Unlike the methods discussed previously, this approach explicitly generates the information state and the policy considers estimated states. It supports the idea to consider past action information and also processes the sequence up to the current information in the same stream as discussed in Section 2.5.

2.8 Conclusion

In this paper, the TD3 algorithm has been extended using the LSTM network to solve POMDP. When the disturbance is dynamically changing, it is beneficial to have a policy that dynamically adapts to the changing environments. In this context, the inclusion of only observation sequences is not sufficient to construct an internal state representation. When the policy is separated into estimation and control parts, belief states are generated from the sequence of state and action in POMDP. This framework is implicitly applicable to the end-to-end model-free RL. We have presented that adding an action sequence can improve the robustness of the achieved performance without the separation of the policy into these two components. The structure to treat the sequence was also discussed. The results showed that it is better to process the past sequence and the current information in the same channel. Also, we proposed a new method called H-TD3 that can significantly reduce the computational time by using the representation of the trajectory information encoded in the actor network for initialization of the critic network.

For our future work, the generalizability should be further investigated. It was observed that the trained role of the network would differ depending on the characteristics of the disturbances. Developing an algorithm that covers both types of disturbances would bring the RL application closer to real-world implementation.

2.9 Appendices

2.9.1 Network structure

The used network structures are summarized in Table 2.3 and Table 2.4.

Table 2.3: Structure

Algorithm	Actor network	Critic network
LSTM-TD3	Lin(obs dim+act dim, 128), Lin(obs dim, 128) ReLU, ReLU	Lin(obs dim+act dim, 128), Lin(obs dim+act dim, 128) ReLU, ReLU
	LSTM(128,128), N/A	LSTM(128,128), N/A
	Lin(256,128) ReLU	Lin(256,128) ReLU
	Lin(128,1) tanh	Lin(128,1)
LSTM-TD3 _{1ha1hc}	Lin(obs dim+act dim, 128) ReLU	Lin(obs dim+act dim, 128) ReLU
	LSTM(128,128)	LSTM(128,128)
	Lin(128,128) ReLU	Lin(128,128) ReLU
	Lin(128,1) tanh	Lin(128,1)
LSTM-TD3 _{1ha2hc}	Lin(obs dim+act dim, 128) ReLU	Lin(obs dim+act dim, 128), Lin(act dim, 128) ReLU, ReLU
	LSTM(128,128)	LSTM(128,128), N/A
	Lin(128,128) ReLU	Lin(256,128) ReLU
	Lin(128,1) tanh	Lin(128,1)
Lin: linear layer		

Table 2.4: Structure (cont.)

Algorithm	Actor network	Critic network
H-TD3	Lin(obs dim+act dim, 128)	Lin(obs dim+act dim, 128)
	ReLU	ReLU
	LSTM(128,128)	LSTM(128,128)
	Lin(128,128)	Lin(128,128)
	ReLU	ReLU
TD3	Lin(128,1)	Lin(128,1)
	tanh	
	Lin(obs dim, 128)	Lin(obs dim+act dim, 128)
	ReLU	ReLU
	Lin(128,128)	Lin(128,128)
Lin: linear layer	ReLU	ReLU
	Lin(128,1)	Lin(128,1)
	tanh	

2.9.2 The number of parameters

The number of the parameters are summarized in Table 2.5.

Table 2.5: The parameter of numbers (obs dim= 3, act dim= 1)

Algorithm	The number of parameters (actor, critic)
LSTM-TD3	166273, 166401
LSTM-TD3 _{1ha1hc}	149377, 149377
LSTM-TD3 _{1ha2hc}	149377, 166017
H-TD3	149377, 149377
TD3	17153, 17281

2.9.3 Pseudo codes

Pseudo codes of LSTM-TD3_{1ha2hc}, LSTM-TD3_{1ha1hc} and H-TD3 are described below based on the original TD3 algorithm[65].

Algorithm 1 LSTM-TD3_{1ha2hc}

Initialization of critic networks Q_{θ_1} , Q_{θ_2} and actor network π_ϕ with random parameters θ_1 , θ_2 , and ϕ
Initialization of target networks $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$, and $\phi' \leftarrow \phi$
Initialization of replay buffer $\mathcal{B}$
for $episode = 1 \dots T$ **do**
 Select action with exploration noise ϵ as $a \sim \pi_\phi(o_{t-l+1:t}, a_{t-l:t-1}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma)$, and receive reward r and new observation o_{t+1}
 Store transition tuple (o_t, a_t, r_t, o_{t+1}) in $\mathcal{B}$

 Sample mini-batch of N transition blocks $([o_{t-l+1:t}, a_{t-l:t-1}] \ a_t, r_t, [o_{t-l+2:t+1}, a_{t-l+1:t}])$ from $\mathcal{B}$
 $\bar{a} \leftarrow \pi_{\theta'}(o_{t-l+2:t+1}, a_{t-l+1:t}) + \epsilon$, $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c)$
 $y \leftarrow r_t + \gamma \min_{i=1,2} Q_{\theta'_i}([o_{t-l+2:t+1}, a_{t-l+1:t}], \bar{a})$
 Updating critics $\theta_i \leftarrow \arg \min_{\theta_i} N^{-1} \Sigma(y - Q_{\theta_i}([o_{t-l+1:t}, a_{t-l:t-1}], a_t))^2$
 if $t \bmod d$ **then**
 Update ϕ deterministic policy gradient:
 $\nabla_\phi(J(\phi)) = N^{-1} \Sigma \nabla_a Q_{\theta_1}([o_{t-l+1:t}, a_{t-l:t-1}], a_t)|_{a=\pi_\phi(o_{t-l+1:t}, a_{t-l:t-1})}$
 $\nabla_\phi \pi_\phi(o_{t-l+1:t}, a_{t-l:t-1})$

 Update target networks:
 $\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i$
 $\phi' \leftarrow \tau \phi + (1 - \tau) \phi'$
 end if
end for

Algorithm 2 LSTM-TD3_{1ha1hc}

Initialization of critic networks Q_{θ_1} , Q_{θ_2} and actor network π_ϕ with random parameters θ_1 , θ_2 , and ϕ
 Initialization of target networks $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$, and $\phi' \leftarrow \phi$
 Initialization of replay buffer $\mathcal{B}$
for $episode = 1 \dots T$ **do**
 Select action with exploration noise ϵ as $a \sim \pi_\phi(o_{t-l+1:t}, a_{t-l:t-1}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma)$, and receive reward r and new observation o_{t+1}
 Store transition tuple (o_t, a_t, r_t, o_{t+1}) in $\mathcal{B}$

 Sample mini-batch of N transition blocks $([o_{t-l+1:t}, a_{t-l:t-1}], r_t, [o_{t-l+2:t+1}, a_{t-l+1:t}])$ from $\mathcal{B}$
 $\bar{a} \leftarrow \pi_{\theta'}(o_{t-l+2:t+1}, a_{t-l+1:t}) + \epsilon$, $\epsilon \sim clip(\mathcal{N}(0, \sigma), -c, c)$
 $y \leftarrow r_t + \gamma \min_{i=1,2} Q_{\theta'_i}([o_{t-l+1:t+1}, a_{t-l+1:t+1}])$, where $a_{t+1} = \bar{a}$
 Updating critics $\theta_i \leftarrow \arg \min_{\theta_i} N^{-1} \Sigma(y - Q_{\theta_i}([o_{t-l:t}, a_{t-l:t}]))^2$
 if $t \bmod d$ **then**
 Update ϕ deterministic policy gradient:
 $\nabla_\phi(J(\phi)) = N^{-1} \Sigma \nabla_a Q_{\theta_1}([o_{t-l:t}, a_{t-l:t}])|_{a=\pi_\phi(o_{t-l+1:t}, a_{t-l:t-1})} \nabla_\phi \pi_\phi(o_{t-l+1:t}, a_{t-l:t-1})$

 Update target networks:
 $\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i$
 $\phi'_i \leftarrow \tau \phi + (1 - \tau) \phi'$
 end if
end for

Algorithm 3 H-TD3

Initialization of critic networks $Q_{\theta_1}, Q_{\theta_2}$ and actor network π_ϕ with random parameters θ_1, θ_2 , and ϕ

Initialization of target networks $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2$, and $\phi' \leftarrow \phi$

Initialization of replay buffer $\mathcal{B}$

for $episode = 1 \dots T$ **do**

Select action with exploration noise ϵ as $a \sim \pi_\phi(o_{t-l+1:t}, a_{t-l:t-1}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma)$, and collect reward r , hidden states of actor h_{t-1} , and new observation o_{t+1}

Store transition tuple $(o_t, a_t, r_t, h_{t-1}, o_{t+1})$ in $\mathcal{B}$

Sample mini-batch of N transition blocks $([o_{t-l+1:t}, a_{t-l:t-1}], o_t, a_t, h_{t-1}, r_t, [o_{t-l+2:t+1}, a_{t-l+1:t}], o_t, h_t)$ from $\mathcal{B}$

$\bar{a} \leftarrow \pi_{\phi'}(o_{t-l+2:t+1}, a_{t-l+1:t}) + \epsilon, \epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c)$

$y \leftarrow r_t + \gamma \min_{i=1,2} Q_{\theta'_i}(o_{t+1}, \bar{a})$, where the initial state of LSTM in target critic is by h_t

Updating critics $\theta_i \leftarrow \text{argmin}_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(o_t, a_t))^2$, where the initial state of LSTM in critic is by h_{t-1}

if $t \bmod d$ **then**

Update ϕ deterministic policy gradient:

$$\nabla_\phi(J(\phi)) = N^{-1} \sum \nabla_a Q_{\theta_1}(o_t, a_t)|_{a=\pi_\phi(o_{t-l+1:t}, a_{t-l:t-1})} \nabla_\phi \pi_\phi(o_{t-l+1:t}, a_{t-l:t-1}),$$

where the initial state of LSTM in target critic is by h_{t-1}

Update target networks:

$$\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i$$

$$\phi' \leftarrow \tau \phi + (1 - \tau) \phi'$$

end if

end for

Chapter 3

Autonomous boresight localization for UAV-based satellite terminal antenna measurement

Abstract

Localizing the boresight angle of the satellite terminal antenna is a necessary step for antenna verification in UAV-based antenna measurement. The procedure for this step should be efficient and generalizable to successfully localize the beam center within the limited flight time without knowing the radiation pattern of the antenna under test. We propose a method to autonomously guide the vehicle to the boresight angle by using a reinforcement learning technique. The proposed method does not require parameter tuning and specific prior knowledge of the antenna pattern. The developed method is tested via numerical simulation with theoretical and empirically obtained radiation patterns. The results showed that the proposed method could localize the boresight angle accurately and robustly without being limited by conditions such as Signal-to-Noise Ratio, beam shape, and distance to go. The method was compared with other approaches and showed an advantage in trajectory efficiency.

3.1 Introduction

Satellite communication currently plays an important role. Keeping connectivity is an essential requirement for safety, commercial services, and entertainment. Approximately 8,000 satellites were orbiting the earth in 2021 and OneWeb and SpaceX are planning to launch more than 40,000 satellites in Low-Earth-Orbit in the upcoming years [82]. The number of satellite user terminals is forecasted to exceed 10 million by 2030 [1]. As communication usage expands, it is critical to keep reliable connectivity without interfering with other communications. One of the key components of satellite communication on the ground segment is terminal antennas, which are devices that convert electrical energy into electromagnetic waves. The communication could fail or become inefficient if the antenna is not designed or not calibrated well. Furthermore, it could interfere with other communications. Satellite operators have certain re-

sponsibilities to protect other communications under national laws and international treaties [83]. As such, satellite operators and authorities have posed requirements for terminal antennas with respect to several aspects such as radiation pattern and pointing accuracy to mitigate these risks [84, 85].

There are many occasions where the performance of the terminal antennas needs to be evaluated at any stage from development to after installation. The antenna measurements are commonly performed at the distance called "far-field" which can be a few hundred meters to kilometers depending on the antenna characteristics and operation frequency. Conventionally, the satellite Antenna Under Test (AUT) is transported to special facilities with a large space which can provide a far-field for the evaluation. The disadvantage of using these facilities is that this approach might be logistically heavy since the facility is locally fixed and a time-consuming process due to the lack of flexibility. This approach would not be able to provide an adequate service in the current trend since the number of antennas to be tested is rapidly increasing.

In this context, in-situ airborne measurement using Unmanned Aerial Vehicles (UAVs) is an emerging technology that would resolve the issues of the conventional method and it has received a lot of research attention [86, 87, 88, 89]. The UAV-based measurement can be performed on-site in a realistic operation environment and provide the desired evaluations for AUTs. The system enables flexible test configuration and new test scenarios, such as Non-Geostationary Orbit (NGSO) communication [90], thanks to the UAVs which can dynamically change their positions during the measurement. However, it comes with the restriction of battery time and additional consideration of uncertainties and errors in the measurements since the UAV-based measurement system is not attached to the ground and the measurement environments are different in each test site. The effect of the positioning error becomes less when the distance between the AUT and the UAV is large. However, it will increase the flight time and distance. Therefore, the test should be carefully designed by considering the effect of the positioning error on the accuracy of test results and flight time. It is beneficial to have a method to generate efficient flight trajectories that provide shorter flight times while keeping the accuracy within the desired range.

In this work, we will focus on the localization of a beam center (boresight angle) of AUT for UAV-based measurement. Finding a beam center is an essential step to initiate the measurement. By finding the boresight angle, the electrical coordinate system of the AUT can be defined based on Radio Frequency (RF) characteristics. The boresight offset from the target angle and actual boresight angle can be directly identified once the boresight angle is known. From this point, the required antenna evaluation can be executed. A typical list of requirements for antennas operated in the Ka-band can be found in [5]. For example, taking the found electrical coordinate as a reference frame, the radiation patterns of the AUT can be measured and a corresponding regulatory mask can be applied to verify sidelobe characteristics. Another common requirement is the verification of cross-polar discrimination of which measurement is very sensitive to accurate boresight angle localization.

There are existing techniques used for the beam localization such as conical scanning [91] and monopulse tracking [92] in satellite communication. These are common techniques for the antennas in the operation to align their boresight to the intended satellite. However, these techniques rely on the symmetry of the radiation pattern and

they are not suitable for antenna evaluation where high accuracy is required.

If there is no information on the beam shape, localizing the beam center could require exhaustive data sampling in the search domain. For instance, the beam center could be localized by multiple raster scans [24]. The raster scan refers to a pattern of flight trajectory on a rectangular area defined by azimuth and elevation angle of the AUT. The first raster scan is performed with low resolution to find the beam in a large search domain. Then, the following raster scans are continued with higher resolutions in smaller areas. However, if the resolution is not chosen adequately, it causes under/over-sampling. The under-sampling would fall in the failure to find the beam in the first scan or to achieve the desired accuracy. The over-sampling increases the flight time and the number of data to process.

Bayesian optimization (BO) technique has been studied in [93] to find the boresight and characterize radiation pattern. The proposed method showed the initial results which reduced the total travel distance compared to the method using raster scans. Regardless of the fact that the main beam generally has a convex shape, the guidance policy in BO assumes no available model. Both a raster scan and BO may not generate an efficient trajectory in the considered scenario because it does not take the general model of the antenna pattern into account.

Sampling-based Gradient Ascent (GA) or GA momentum may provide an efficient trajectory compared to raster scan and BO. Nonetheless, the suitable parameters of GA may depend on the beam shape and this approach may not be robust when the sensor positioning and observed signals contain errors and noises.

As found in the literature, beam localization application comprises guidance and estimation algorithms except for GA. The RF and positional data are collected on the generated trajectory by the guidance algorithm and the boresight angle is finally estimated by introducing a model. The issues of the aforementioned methods are: (1) the accuracy would not be enough as the trajectory is pre-planned and the estimation relies on the assumption of the symmetry (conical scanning and monopulse tracking), (2) the trajectory is not efficient because the guidance algorithm does not utilize the available assumption, while the general shape of the beam is convex (raster scan, BO), (3) it lacks the robustness of the performance and requires manual parameter tuning for each shape of the beam (GA) (4) the environment is different for each measurement and the measurement experience the positioning and RF noise. To overcome these issues, we propose a guidance algorithm based on Reinforcement Learning (RL) for boresight localization. TD3_{1ha1hc}[94] is selected as RL algorithm. TD3_{1ha1hc} is developed to perform in the environment where noise exists by processing sequential action and observation through Long Short-Term Memory (LSTM) network [51]. Although the beam shape is not known in advance, the general shape of the beam is convex. By training the agent in various environments, the RL algorithm would be able to adapt its policy to any beams with general shapes and autonomously generate flight trajectories without manual parameter tuning. Also, the RL algorithm would effectively mitigate the issue caused by the positioning and RF noise.

The scope of AUTs in this work is focused on Very-Small-Aperture Terminal (VSAT) antennas in Ka and Ku bands due to their popularity as terminal antennas [1] although it can be extended to other types of antenna evaluations. The contribution of this work can be summarised as follows:

- We propose a method to guide a UAV to the boresight angle by using an RL algorithm in Section 3.3. The RL elements are defined based on differentials of the available information to increase the generalizability of the application.
- We further introduce scaled sequential observation (Section 3.3). The range of beamwidth in Ka and Ku bands is quite large. Thanks to the scaled observation, the performance of the developed guidance algorithm is not affected by beamwidths.
- The performance of the guidance algorithm and the final estimation results are evaluated by comparing them with other methods through numerical simulations using both theoretical and empirical radiation patterns, as detailed in Section 3.4.

3.2 Preliminary

In this section, we review the VSAT antenna characteristics and introduce the UAV system under consideration. Then, the challenge of the boresight angle localization application is discussed. Based on this analysis, the application is developed in the next section. Basic knowledge related to the developed algorithm is also introduced.

3.2.1 Antenna pattern in the scope

In this work, the AUT is focused on VSAT terminal antenna in operation frequency of Ku and Ka bands. The Ku and Ka bands are defined by the frequency range between 12–18 GHz and 26.5–40 GHz respectively. The diameter of the antenna would be categorized below 3.8 m for Ku-band and 1.8 m for Ka-band according to SOMAP [47, 5]. The 3 dB bandwidth θ_{3dB} is commonly approximated as (3.1):

$$\theta_{3dB} = \frac{70\lambda}{D} \quad (3.1)$$

where λ is the wavelength and D is the diameter [3]. The first null bandwidth θ_{FNBW} (FNBW) is also generally approximated as (3.2):

$$\theta_{FNBW} = 2\theta_{3dB}. \quad (3.2)$$

Fig. 3.1 illustrates the operational θ_{3dB} when the diameter is limited to 0.5–1.8 meters, with D/λ from equation (3.1) as the x-axis. λ is calculated by $\lambda = c/f$ (where c is the speed of light and f is frequency) based on the stated frequency range. The crosses in the figure represent the start and stop θ_{3dB} of the range for each band. We observe that the beamwidth to be considered is approximately $0.3 \leq \theta_{3dB} \leq 3.5$ degrees. Thus, the ranges of θ_{FNBW} in Ka and Ku bands are $0.6 \leq \theta_{FNBW} \leq 7.0$ degrees.

Let us consider a search domain defined by azimuth and elevation angles on a spherical surface centered at the AUT. If there is an expected maximum boresight offset angle in any direction, the search domain is defined to cover the expected offset range. If half of FNBW θ_{FNBW} is larger than the maximum possible offset, the measurement in the search domain is filled by the main beam. Otherwise, there is a possibility that

the main lobe and side lobes exist in the search domain. The strategy for boresight localization might differ in these two cases since the latter case could be considered as a global optimization problem. An initial survey of the AUT would help to select the strategy in real-world evaluation, although this work assumes that the initial position of the UAV is located within half of the FNBW from the boresight angle.

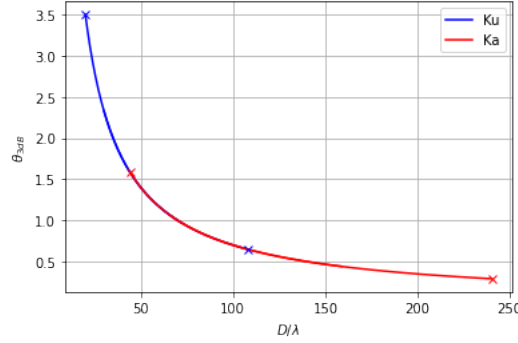


Figure 3.1: VSAT operational θ_{3dB}

3.2.2 The airborne measurement

Our measurement system consists of two main parts: the RF payload and the Ground Control Station (GCS). Fig. 3.2 shows an example of a UAV with a RF payload. This system was developed by QuadSAT and is named UAS-APE. The RF payload primarily consists of a receiver or transmitter, a probe antenna, a microcomputer, and a 3-axis (yaw-pitch-roll) gimbal system for the stabilization and orientation of the probe antenna. The gimbal's control is based on the orientation estimation acquired from onboard Inertial Measurement Units (IMUs), encoders placed at the payload's joints, and the UAV itself. The system precisely maintains the antenna's orientation towards the AUT during flight and ensures polarization alignment. Additionally, the antenna at the payload has a large beamwidth, making the measurement less affected by the payload's ripples during flight. The UAS-APE's safe maximum flight time is around 30 minutes.



Figure 3.2: QuadSAT's compact UAS-APE

The GCS is composed of a receiver or transmitter (to pair with the payload) which is connected to the AUT, a Real Time Kinematics (RTK) station, and a main computer.

The RTK can provide positioning error corrections to the UAV. Along with the GPS receiver on the UAV, Differential Global Positioning Systems (DGPS) are employed for the centimeters-level localization accuracy of the UAV. Specifically, the reported localization accuracy of the UAV with DGPS is 2 cm horizontally and 5 cm vertically. The UAS-APE and DGPS help stabilize the UAV platform within a square with sides of 30 cm even in wind speeds up to 15 m/s [17]. The main computer collects RF and positional data from both the UAS-APE and the measurement equipment. The data is post-processed into a specific format, generally consisting of signal levels at the azimuth and elevation angle with respect to the AUT, and present it in the desired form. More detail of QuadSAT system can be found in [17]. When the UAV has a positioning error of around 30 cm, the error in azimuth and elevation angle with respect to the AUT coordinate system varies depending on the distance from the AUT. Fig.3.3 is visualization of the effect of the positioning error 30 cm on a principal plane. As the distance increases, its effect is less. In this work, 30 cm is considered to be the maximum positioning error, and the standard deviation of positioning error in the simulation is defined based on the analysis in Fig.3.3.

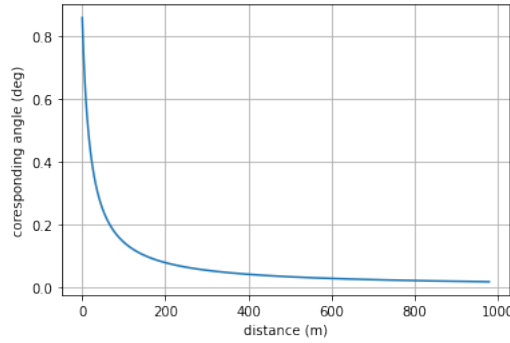


Figure 3.3: Impact of 30cm positioning error on distance

3.2.3 Guidance algorithm

The source of measurement error is not just limited to the positional error of the UAV. The received signal can be influenced by many sources such as losses at the components, depointing of the probe antenna, polarization mismatch, and atmospheric loss. Moreover, RF additive noise always exists in observed signals. Typically the relationship between the signal and RF noise is described by the Signal-to-Noise Ratio (SNR). The observed signal strength in decibels is more affected by RF noise when the power level is lower. In this work, SNR is defined with respect to the received signal at the beam peak. It is necessary for the measurement system to be robust against these noise and uncertainties. Additionally, the shape of the beam widely varies in VSAT antennas as seen in Fig.3.1. To make the solution universal, it should be able to autonomously generate its trajectory to collect data and should not require manual parameter tuning.

Meta-RL The UAV is guided by an RL-based policy, as developed in Section 3.3. RL is a branch of machine learning that learns optimal policy by interacting with the environ-

ment as demonstrated in the inner loop of Fig. 3.4. The agent receives an observation o_t and the reward r_t from the environment at timestep t . The agent generates an action a_t to take based on o_t . The reward r_t is interpreted as an evaluation signal of the action a_t when the agent is in the state s_t from the environment. The objective of RL is to maximize the future cumulative reward at the end of an episode.

Typically, RL is built under the assumption of a Markov Decision Process (MDP), where

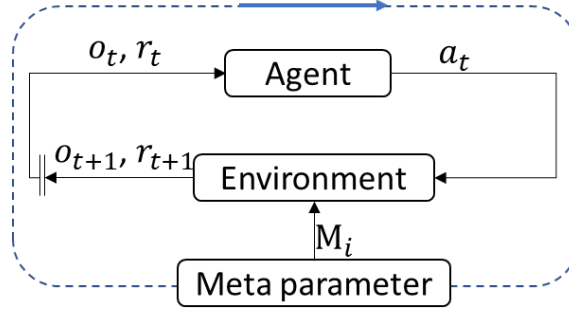


Figure 3.4: Meta-reinforcement learning

observation o_t can be uniquely mapped to the true state s_t . However, in our application, we need to consider multiple MDPs as the solution should be applicable to any antenna pattern shape. Therefore, we introduce the concept of meta-RL. Generally, an MDP is defined by four components in a tuple, $\mathcal{M}_i = [\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}]$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathcal{T}$ represents the state transition $P(s_{t+1}|s_t, a_t)$, and $\mathcal{R}$ is an immediate reward function $\mathcal{R} = \mathbb{E}[r_{t+1}|s_t, a_t]$. The agent learns its policy π for the environment governed by $\mathcal{M}_i$. In meta-reinforcement learning, the agent interacts with multiple MDPs to perform towards a more general purpose. As suggested in [95, 96], during the training procedure, a new $\mathcal{M}_i \sim \mathcal{M}$ is sampled as depicted in the outer loop of Fig. 3.4. The policy of meta-RL incorporates past action in addition to observation. Enhanced recurrent networks, such as the LSTM network [51], are often implemented in the network and can help identify the characteristics of the MDP from the information sequence on the trajectory.

POMDP and LSTM-TD3_{1halhc} In this application, the observation does not fully represent the true state due to disturbances, resulting in a Partially Observable Markov Decision Process (POMDP). In POMDP, two additional elements are considered in the tuple: $\mathcal{O}$ and $\mathcal{Z}$. $\mathcal{O}$ is the observation space and $\mathcal{Z}$ is the observation function $P(o_{t+1}|s_{t+1}, a_t)$. LSTM-TD3_{1halhc}, together with the other two algorithms, has been developed to enhance the robustness of optimality in POMDP environments [94]. The l length of sequential action and observation on the trajectory up to the current timestep is processed via a neural network, which includes an LSTM layer, to generate the next action. The developed guidance algorithm employs LSTM-TD3_{1halhc}, which showed the best overall performance in comparison [94], and it is trained using Meta-RL concepts.

3.3 Method

The proposed method comprises a guidance algorithm, referred to as "Climber", and a beam center "localization" process. The Climber incorporates the LSTM-TD3_{1h1hc} algorithm as a policy to guide the UAV to the boresight angle in an environment with existing measurement noise and positioning errors. As depicted in Fig. 3.5, Climber takes measured data from the UAV, including the UAV's position x_t and the measured signal strength Pr_t , and generates the next waypoint x_{t+1} for the UAV. The input data to Climber is transformed into the observation o_t for LSTM-TD3_{1h1hc}. Together with the sequence of past observations $o_{t-l:t-1} = \{o_{t-l}, \dots, o_{t-1}\}$, the sequence of observations up to the current timestep $o_{t-l:t}$ is formed. The input to LSTM-TD3_{1h1hc} is the combined sequential observation $o_{t-l:t}$ and the sequence of past actions $a_{t-l+1:t-1} = \{a_{t-l+1}, \dots, a_{t-1}\}$, and the output is the action a_t . Here, a_t is defined as the angular displacement of the UAV from the current step. Lastly, the action is converted into the next waypoint x_{t+1} for the UAV.

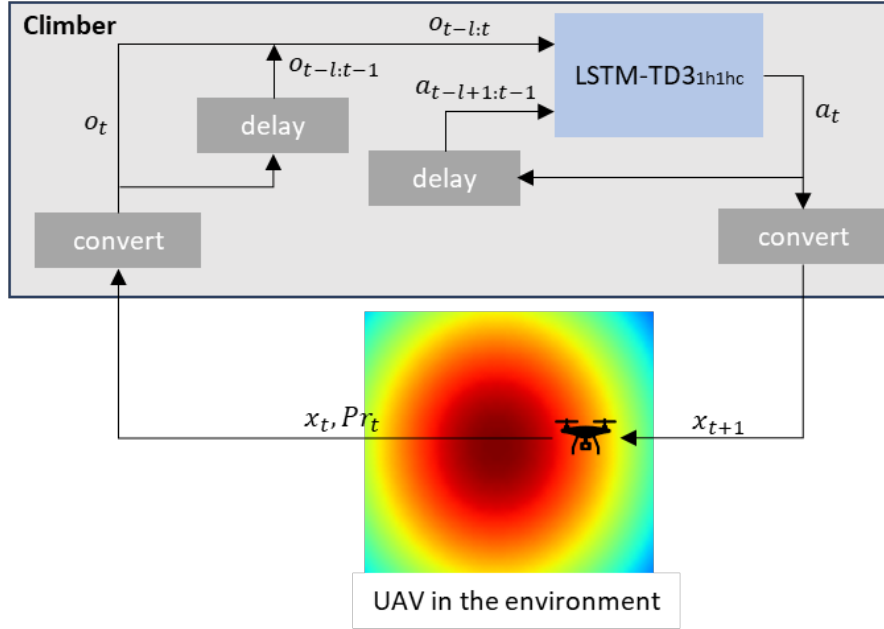


Figure 3.5: Conceptual implementation of RL system guiding a UAV to the beam peak

3.3.1 Problem Analysis

The state needs to represent the UAV position with respect to the boresight angle. The obtainable information from the sensors in this application is the position of the UAV in GPS coordinates and received signal strength. However, the position in GPS coordinates does not provide informative information for decision-making, because the positional information on GPS coordinates does not correlate to the boresight angle. Namely, states cannot be constructed based on positional information on GPS coordinates. The objective of Climber is to guide the UAV to the boresight angle. Therefore, it requires converting the measured data to some value that has an interpretation of the

position of the UAV with respect to the boresight angle.

The received signal strength from the sensor is not suitable as the input either. According to the link analysis calculation [3], the received signal strength is calculated by (3.3) in dB.

$$P_{RX} = P_{TX} + G_{TX}(\theta_{TX}) - L_{FS} + G_{RX}(\theta_{RX}) \quad (3.3)$$

where P_{RX} and P_{TX} are received and transmitted power, G_{RX} and G_{TX} are gains of the receiving and transmitting antennas, θ_{TX} and θ_{RX} are the angles of transmission and reception at each antenna coordinate system centered at boresights, and L_{FS} is the free space loss modeled as $L_{FS} = (4\pi R/\lambda)^2$ where R is the distance. Note that P_{RX} in (3.3) at timestep t is described as Pr_t for the remainder of this work. Assuming that P_{TX} , L_{FS} , and G_{RX} remain constant during the measurement since the probe antenna on the payload has a wide beam and the distance between the UAV and the AUT is kept constant in our scenario, the received signal strength is

$$P_{RX} = G_{TX}(\theta_{TX}) + C \quad (3.4)$$

where C is constant. The constant value in (3.4) would change depending on the test configuration, and $G_{TX}(\theta_{TX})$ is assumed to be unknown in this scenario since the boresight angle should be localized before a radiation pattern measurement.

3.3.2 Using differential

Based on the above analysis regarding state representation, the input to the RL policy is defined as the differentials of the sensor measurements from the previous timestep. Namely, the input to the policy is $\Delta Pr_t = Pr_t - Pr_{t-1}$ and $\Delta x_t = x_t - x_{t-1} = a_{t-1}$. a_{t-1} is the action which is already a part of the input in LSTM-TD3_{1halhc}[94]. LSTM-TD3_{1halhc} takes an l length sequence of observations and actions. Hence, the final input to the policy is $\Delta Pr_{t-l:t} = \{\Delta Pr_{t-l}, \dots, \Delta Pr_t\}$ and $a_{t-l-1:t-1} = \{a_{t-l-1}, \dots, a_{t-1}\}$. In combination with the meta-RL method, LSTM-TD3_{1halhc} helps the agent to autonomously adjust its policy in dynamic environments.

This approach can also be related to sampling-based GA with momentum[97], which is a first-order optimization algorithm for finding local maxima. The policy of GA with momentum is defined by (3.5).

$$x_{t+1} = x_t + \alpha \nabla f(x_t) + \gamma(x_t - x_{t-1}) \quad (3.5)$$

where f is the objective function, x_t is the last sampling point, and α and γ are tuning parameters for the learning rate and momentum. When the GA is applied in boresight angle localization, (3.5) is rewritten as (3.6)

$$x_{t+1} = x_t + \alpha \frac{Pr_t^{sample} - Pr_t}{h} + \gamma(x_t - x_{t-1}) \quad (3.6)$$

where Pr_t^{sample} is the sample used to generate the angular distance h apart from the point where Pr_t is sampled. Instead of taking $Pr_t^{sample} - Pr_t$ and $x_t - x_{t-1}$, the developed Climber algorithm determines the next waypoint x_{t+1} based on historical values $\Delta Pr_{t-l:t}$ and $\Delta x_{t-l:t} = a_{t-l-1:t-1}$, which are longer correlations of the differentials of

measured data on the trajectory. The proposed algorithm does not require manual tuning of parameters and is expected to generate the trajectory more robustly than GA by taking l length of historical information, even in the presence of observation disturbances by RF noise, as will be confirmed in Section 3.4.

3.3.3 RL setting

From the above discussion, the observation is defined as ΔPr_t . Let us discuss the other RL elements in this section. The state would be defined as $\Delta \hat{P}r_t$ which is not affected by the RF noise whereas ΔPr_t is disturbed by the noise. The relationship between the state and observation is described as (3.7) and the observation is not the true state i.e., POMDP.

$$\Delta Pr_t = \Delta \hat{P}r_t + \text{noise} \quad (3.7)$$

POMDP tuple contains the observation function in addition to MDP tuple as described in Section 3.2. The RL agent should internally learn the model of the observation function during training. The observation function is described as (3.8) when the state is defined as ΔPr_t .

$$P(\Delta Pr_{t+1} | \Delta \hat{P}r_{t+1}, a_t) \quad (3.8)$$

The state could also be defined as the position of the UAV in the boresight coordinates. However, if the observation function is defined as (3.9), it might cause extra complexity to map observations to states besides the disturbance and the mapping between positional state from ΔPr_{t+1} and a_t may not be unique.

$$P(\Delta Pr_{t+1} | \text{positional state}, a_t) \quad (3.9)$$

Therefore the state is defined as $\Delta \hat{P}r_t$.

Next, the reward is defined. The reward function is represented as $\mathcal{R} = \mathbb{E}[r_{t+1} | s_t, a_t]$. For example, the reward could be defined as the negative angular distance between the agent position and the boresight angle. With this definition of the reward, the agent might naturally learn the efficient trajectory to reach the boresight angle with the minimum timesteps. However, since the state is defined as $\Delta \hat{P}r$, the reward needs to be as a function of $\Delta \hat{P}r$ rather than positional relation to the boresight angle for the same reason used in state definition. Thus, one element of the reward is defined as $\Delta \hat{P}r$ so that the objective of the agent is to maximize the difference of the received signal strength between timestep t and the end of the episode. To encourage the agent to learn an efficient trajectory to reach the boresight angle with minimum timesteps, a negative constant c is added to the reward at each step, and the episode is terminated if the agent reaches the boresight angular position with threshold i or if the total steps exceed maximum steps.

In summary, the RL setting is configured as below in Climber.

- State: differential of the signal strength without noise $\Delta \hat{P}r_t$
- Observation: differential of the received signal strength ΔPr_t
- Action: displacement Δx_t
- Reward: $\Delta \hat{P}r_t - c$

- Termination condition: the angular distance reaches below threshold i or total steps reach the maximum.

3.3.4 Scaling of sequential observation

The obtainable ΔPr_t from the same action becomes smaller as the sampling point gets closer to the beam peak. Also, the range of differential in observation gets smaller with flatter beams. Since the shape of the beam has a large variety, the proposed algorithm might suffer when the AUT has a wide beam. To mitigate this issue, the sequential observation could be transformed into a certain range by scaling them. The scaling function is described as (3.10) when the range of observation sequence is fixed between $[0, 1]$.

$$\Delta Pr_i^{sc} = \frac{\Delta Pr_i - \min(\Delta Pr_{t-l:t})}{\max(\Delta Pr_{t-l:t}) - \min(\Delta Pr_{t-l:t})} \quad (3.10)$$

where ΔPr_i^{sc} is the i th element of the scaled value in the sequential observation between timestep $t-l$ and t . The sequential scaled observation $\Delta Pr_{t-l:t}^{sc}$ still holds the information related to the beam shape unless the derivative of the objective function, i.e., the antenna pattern, is described in the second order and all actions in the sequences are the same. On the other hand, the definition of the reward should not be affected by this transformation. The reason is explained in the following example. As long as the maximum signal strength on the trajectory between $t-l+1$ and $t+1$ is obtained by taking action a_t , then the ΔPr_{t+1}^{sc} becomes 1 for any action in a certain range. Namely, $\mathbb{E}[r_{t+1}|s_t, a_t]$ is not affected by any action a_t in the range. Therefore, this scaling function is applied solely to the sequential observation, and it is executed before the input is generated for each neural network in LSTM-TD3_{1halhc}. The pseudo-code to implement Climber guidance algorithm with scaling is described in Algorithm4.

Algorithm 4 Beam localization with Climber-sc

- 1: Initialize length l of observation, action array $\mathbf{o}$ and $\mathbf{a}$
 - 2: Get initial observation o_0
 - 3: **while** the angular distance $\geq$ threshold **do**
 - 4: Calculate scale ΔPr_i^{sc} in array $\mathbf{o}$
 - 5: Get action $\mu(\Delta Pr_i^{sc}(= \Delta Pr_{t-l:t}^{sc}), a_{t-l-1:t-1})$
 - 6: Move drone and get measurement Pr_{t+1} and calculate observation o_{t+1}
 - 7: Store action and observation in array $\mathbf{o}$ and $\mathbf{a}$
 - 8: **end while**
-

3.3.5 Beam center localization

The guidance algorithm itself does not have the function to determine the beam center position. The simplest method to produce the boresight angle estimation is to take the position where the maximum power level was observed on the trajectory. However, it only relies on the data on the trajectory which may not hit the exact boresight angle and would lack repeatability due to the disturbances. The data can be approximated by introducing a mathematical model to resolve the above-mentioned issue. A

suitable model may vary depending on the trajectory. If enough amount of data is collected uniformly around the peak of the beam and the shape is assumed to be smooth, the antenna pattern around the area could be approximated by introducing a polynomial model regression. Nonetheless, the trajectory to sample a number of data which meets Independent and Identically Distributed (IID) condition may not be produced by Climber. Since we aim to generate an efficient trajectory to reach the boresight angle with this application, Gaussian Process (GP) Regression is considered. As found in Bayesian Optimization[98], GP regression provides the conditional distribution of the objective function via Bayes' theorem based on the finite samples and does not assume IID conditions explicitly. GP regression would be a suitable method to estimate boresight angles based on the collected data by Climber.

3.4 Experiment

In this section, the proposed algorithm is evaluated based on a numerical experiment which consists of two parts. Firstly we evaluate the Climber in comparison with GA[97]. Finally, the evaluation of methods for the localization algorithm and overall evaluation as an application are conducted in the second part.

3.4.1 Training Environment

Meta parameters The radiation pattern is simulated by taking D/λ as variables using the circular aperture model [99]. The constant in (3.4) is neglected, thus normalized signal strength in the training environment is measured and the signal strength at the boresight is 0 dB. D/λ and SNR are randomly configured in every episode as meta-parameters to create a new MDP. The range of D/λ is set to cover the entire VSAT antenna beamwidth θ_{3dB} (Fig.3.1) and the SNR range is set between 60 and 100 dB.

RL training environment The RL training environment is configured as follows:

- **Reward:** the constant c in reward is defined as 0.1.
- **Termination:** The threshold i is set to be 0.05° and the maximum episode length is 100 steps.
- **Initial condition:** A reference coordinate system with azimuth and elevation axis to describe the position of the UAV and boresight angle is introduced by taking the initial position of the UAV as its origin. This point can be understood as the initially estimated boresight angle in a real-world measurement scenario. The boresight offset angle is randomly allocated within the angular distance of $\frac{\theta_{FNBW}}{2}$ from the origin so that the UAV is placed on the main beam.
- **Search range:** The movable range of the agent is set to the range of $[-1.5 \times \frac{\theta_{FNBW}}{2}, 1.5 \times \frac{\theta_{FNBW}}{2}]$ to prevent the UAV from being guided outside of the main beam.

3.4.2 Evaluation settings

Climber evaluation The same environmental configuration is applied in the evaluation environment. The algorithms are compared in the scenario where the generated action from each policy is disturbed by zero-mean Gaussian noise. Based on the positioning accuracy of 30 cm cube (Section 3.2), the standard deviation of the action noise is defined as $\sigma \in [0.01, 0.02, 0.04]$ by roughly indicating the distance between the AUT and the UAV as 600, 300, 150m respectively. The algorithms selected for the performance comparison are listed below.

- GA[97]: Sampling-based GA method with various parameter settings. The next sampling point is decided based on two samples ideally on horizontal and direction though it can be affected by action noise. The parameters used in (3.6) to configure GA are provided in Table 3.1.
- Climber: The algorithm developed in Section 3.3 without scaling of the observation sequence.
- Climber-sc: The algorithm developed in Section 3.3 with the scaled observation sequence.
- Climber-d: The algorithm developed in Section 3.3 without scaling of the observation sequence. The reward is replaced as $r_t = -\Delta d_t$, where Δd_t is the angular distance between the UAV and the boresight angle, and the agent is trained based on this reward to confirm the discussed hypothesis regarding the reward definition.

Table 3.1: GA parameters

Name	α	h	γ
GA-0	0.01	0.01	0
GA-1	0.01	0.05	0
GA-2	0.05	0.01	0
GA-3	0.01	0.01	0.2
GA-4	0.01	0.01	0.8

Localization evaluation The evaluation environment of localization evaluation slightly differs from the training environment in order to compare with BO. The changes in the environment are listed below:

- Termination: The maximum episode length is 50 steps without the threshold i condition.
- AUT: AUT is configured to have $\theta_{3dB} = 3.0^\circ$ (i.e., $\theta_{FNBW} \approx 6.0^\circ$).
- Boresight offset angle: The angle is randomly allocated within a circle with a radius of 3° from the origin of the reference coordinate.

- Search range: The movable range of the agent is set to the range $[-3.0^\circ, 3.0^\circ]$ along the azimuth and the elevation axes.

The termination condition is only defined by the number of steps so that the comparison focuses on which algorithm can collect informative data to estimate the boresight offset within the budget of steps with efficient trajectories. Climber-sc is selected as a guidance policy in this evaluation. The three estimation methods are compared using the collected data by Climber-sc policy to estimate the boresight angle.

- RL: It takes the position where the strongest signal is observed on the trajectory.
- GP: The collected data on the trajectory is modeled in GP regression and the position where the highest value is found is taken. Particularly with this method, the estimation includes the estimated mean value and its variance.
- Polynomial estimating (Poly): The estimation is made by polynomial model optimization using Broyden, Fletcher, Goldfarb, and Shanno (BFGS) method [100] based on the collected data. This estimation algorithm was developed by Quad-SAT.

The final application performance in this scenario is compared with BO [98]. The "Lower Confidence Boundary (LCB)" [101] is selected as the acquisition function for BO since it showed better accuracy and total travel distance compared with "Expected Improvement (EI)" [98] and "Maximum Probability of Improvement (MPI)" [102] during the preliminary experiments. The implementation of the BO is based on GPyOpt [103].

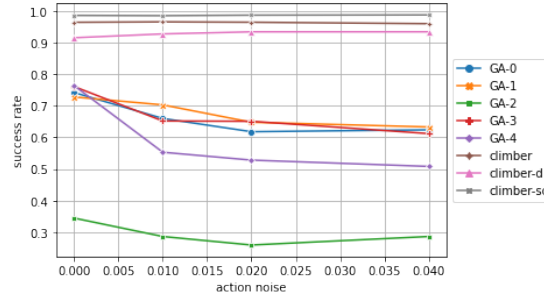
3.4.3 Results

3.4.3.1 Climber evaluation

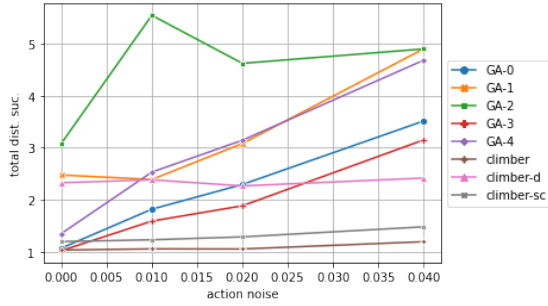
The results of the Climber evaluation were averaged from 5000 trials. In this evaluation, the performance of algorithms is evaluated in terms of success rate, total distance in success, and total distance standard deviation in success. If the guided waypoint reaches threshold 0.05° within 100 steps, the trial was considered to be successful.

Effect of action noise Firstly, the performance is compared between the aforementioned Climber-based algorithms and GA-based algorithms with various action noises as illustrated in Fig. 3.6. The horizontal axis shows the standard deviation of the action noise σ . The Climber-based algorithms (Climber, Climber-sc, Climber-d) achieved a success rate of more than 90% regardless of the existence of action noise while GA-based algorithms showed degradation of success rate with increasing action noise (Fig. 3.6a). Climber-sc achieved the highest success rate in all action noise cases, followed by Climber. In the total traveling distance in successful cases (Fig. 3.6b), GA-0, GA-3 and Climber showed the shortest distance when there was no action noise. However, the performances in the total distance of GA-0 and GA-3 together with GA-1 GA-4 were greatly affected by action noises. These four GA-based algorithms have a common parameter $\alpha = 0.01$ which decides the learning rate. Since the step size is

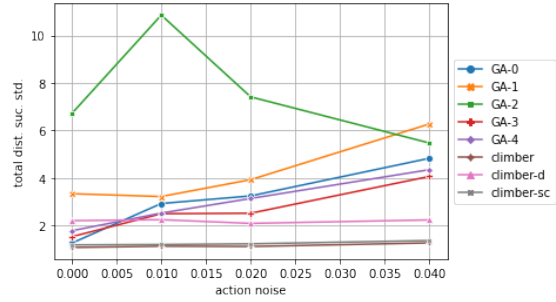
smaller than GA-2 with $\alpha = 0.05$, their trajectories were more sensitive to action noise. Meanwhile, the total distance traveled in successful cases by Climber-based algorithms was less influenced by action noise and could generate stable trajectories as it showed smaller standard deviations of the trajectory compared with GA-based algorithms. The results demonstrate that Climber-based algorithms could robustly reach the boresight angle without the need for parameter tuning regardless of the existence of action noise. Among the Climber-based algorithms, Climber-d performed the worst across all evaluation parameters. As we hypothesized, the mismatch between the reward definition (negative angular distance to the boresight angle) and the observation definition (ΔPr) caused the difficulty in learning. The success rate of Climber-sc is higher than that of Climber, and the total distance traveled by Climber-sc is slightly longer than that traveled by Climber. This is because of the trials where Climber-sc successfully finished the mission with long travel distances whereas Climber did not as it will be confirmed in later sections.



(a) Success rate



(b) Total distance in success

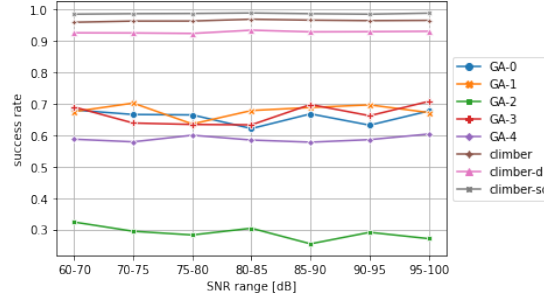


(c) Total distance std. dev. in success

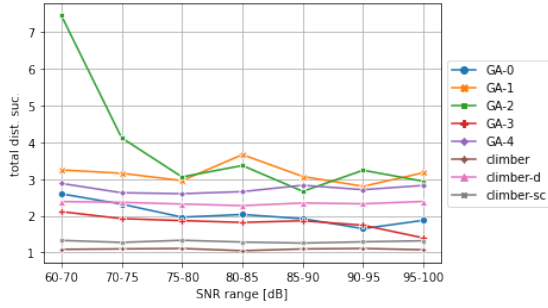
Figure 3.6: Performance comparison with various action noises

Effect of SNR Secondly, the effect of SNR is examined. SNR is divided into seven categories shown on the horizontal axis in Fig. 3.7 and data within each category were averaged. Although the success rates of any algorithms did not change, the total traveling distances and their standard deviations of GA-based algorithms except GA-1 and GA-4 were affected by SNR range. The trajectory became less efficient and got more variance as SNR got smaller. Meanwhile, the Climber-based algorithms maintained consistent performance across all SNR ranges. Note that we propose to use the sequential information through LSTM network to overcome the POMDP issue. The simulation results

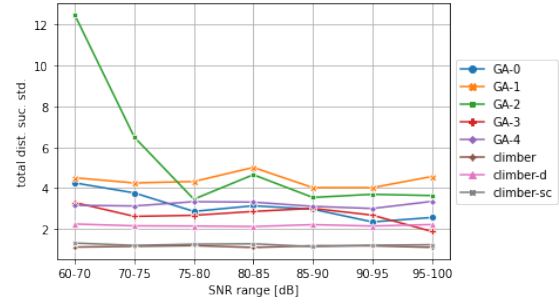
confirm the issue caused by the noise is effectively mitigated. GA-4 has a high value of momentum parameter γ and hence its trajectory was generated by the highly weighted previous direction of move. Therefore, like the Climber-based algorithms, its trajectory was not influenced by SNR. The characterizing parameter of GA-1, which is $h = 0.05$, defines the distance to the sampling points. The trajectory of GA-1 did not show the effects of changes in SNR since the measured signal strengths at sampling points were distinctive regardless of the RF noise.



(a) Success rate



(b) Total distance in success



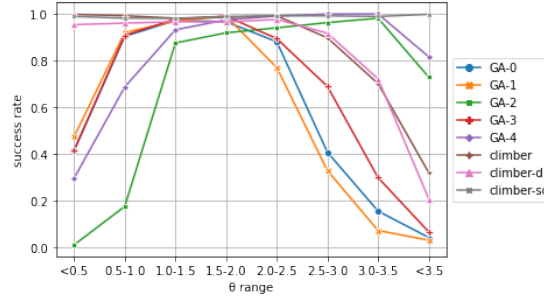
(c) Total distance std. dev. in success

Figure 3.7: Performance comparison with various SNR

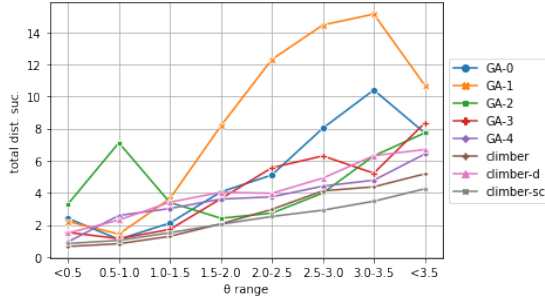
Effect of beamwidth Next, the influence of the beam shape is assessed. The shape of the beam is categorized by θ_{3dB} and the results are shown in Fig. 3.8 of which the horizontal axis describes the range of θ_{3dB} . As Fig. 3.8a illustrates, the success rate of algorithms varies depending on the shape of the beam. The success rates of GA-0, GA-1 and GA-3 improved as θ_{3dB} increased up to the range of 1.5° - 2.0° , after which the performance started to degrade. When the beam is narrow, they did not converge because the step size was too large for the beam shape. This also explains the low success rate found in GA-2 and GA-4. In the given environment setting, a wider beam results in a larger mean initial angular distance between the UAV and the boresight angle. The performance degradation beyond this range is likely due to the steps being too small to reach the boresight angle within 100 steps. This will be confirmed in the following evaluation. Climber and Climber-d showed significantly lower success rates when θ_{3dB} was larger than 2.5 , while Climber-sc kept high success rates in the entire range. As discussed in Section 3.3, the observation without scaling was less distinctive in large areas when the beam was wide. Fig. 3.9 illustrates the trajectory of Climber and Climber-sc

when the same environmental parameters were set ($\text{SNR} = 80\text{dB}$, $\theta_{3dB} = 3.5$, boresight angle was $[-2.0, 2.5]$). The position of the UAV, the received signal strength, and the angular distance to the beam center on the trajectory are represented in the figure. The UAV guided by both algorithms arrived near the beam center at around step 30. However, the Climber algorithm could not guide the UAV precisely after this step and ended up loitering around the peak till the end of the episode. Climber-sc reached the threshold distance at step 38, as indicated by the "done" label in the figures. Although the episode should be terminated in this evaluation environment, the episode was extended to 100 steps to show the behavior for the next localization evaluation. The success rate of Climber without scaling could be improved by training the network on a smaller dedicated range of the beamwidth. However, the solution would no longer cover the entire targeted range of beamwidth. Therefore, the Climber-sc approach was effective.

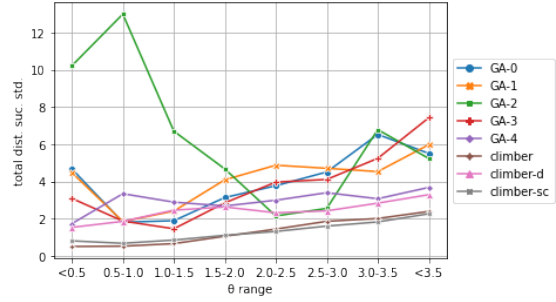
The average total distance in success and its standard deviation became larger as the beam width increased. The total distances in the success of Climber and Climber-sc were similar until θ_{3dB} is 2.5 (Fig. 3.8b). Beyond this range, the trajectories became longer due to the loitering behavior.



(a) Success rate



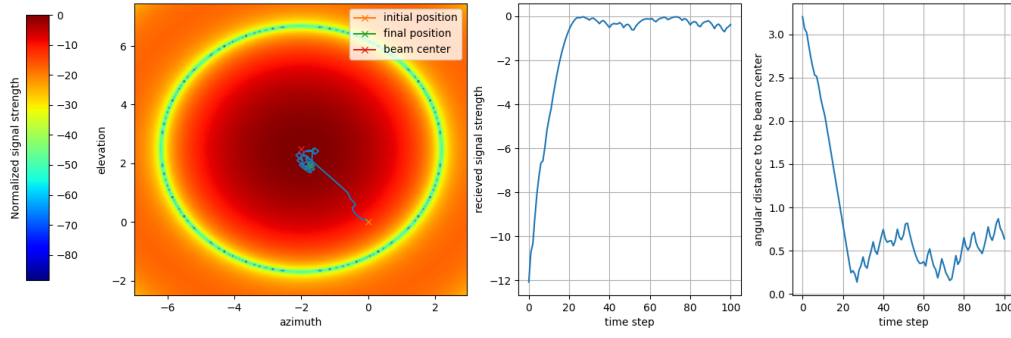
(b) Total distance in success



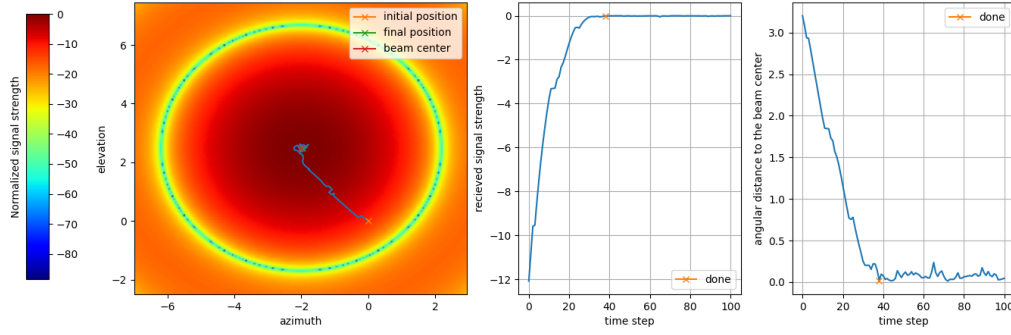
(c) Total distance std. dev. in success

Figure 3.8: Performance comparison with various beamwidth

Effect of distance The performances are compared by taking the initial angular distance between the UAV and the boresight angle as shown on the horizontal axis of Fig. 3.10. The initial angular distance is categorized into 10 by 0.5° intervals, ranging from 0° to 5° . As discussed in the previous evaluation with the beam shape, GA-0, GA-1, and GA-3 could not reach the beam center due to the small step size when the



(a) Trajectory of Climber



(b) Trajectory of Climber-sc

Figure 3.9: Trajectory comparison

distance to cover was large, while GA-2 and GA-4 could thanks to the increased step size produced by the high learning rate α or momentum weight γ .

The success rate of Climber and Climber-d showed degradation as the initial distance increased. Owing to the UAV's initial position being set within the range of $\frac{\theta_{FNBW}}{2}$ to position the UAV on the beam, the larger initial distance corresponds to larger beam width. Therefore, the success rate deteriorated because of the flatness of the beam. Climber-sc maintained high success rates, and the total distance increased almost linearly as the initial distance to the boresight angle increased without increasing the standard deviation. This confirms the stable performance of Climber-sc across all distances and beam shapes.

Fail case of Climber-sc The trials in which Climber-sc failed are examined. Fig. 3.11 plots all trials with respect to the relationship between the initial distance to the boresight angle, the beam center, and FNBW. The blue dots and orange dots represent successes and failures respectively. As it illustrates, Climber-sc failed only when the initial position was at the edge of the main beam.

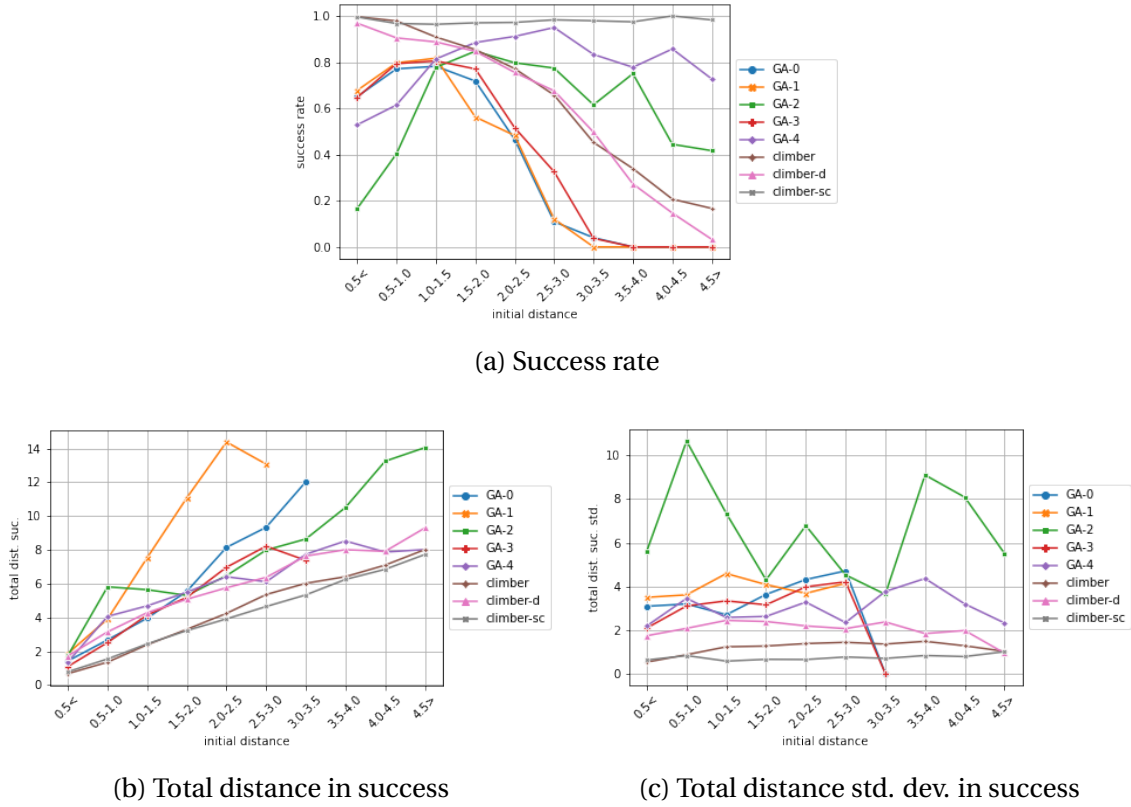


Figure 3.10: Performance comparison with the distance to go

Empirical radiation pattern To evaluate whether the developed algorithms can be applied to more realistic environments, they were tested on the antenna pattern constructed based on empirical measurements provided by QuadSAT. The patterns were asymmetrical, and the signal strength was based on actual measured values, which do not neglect the constant in (3.4), unlike the training environment. Table 3.2 shows the performance results of the Climber-based algorithms. The θ_{3dB} values of "Pattern1" and "Pattern2" range from 2.99° - 3.28° and 2.52° - 2.6° respectively, depending on the cut direction of the beam. The maximum initial distance to the boresight angle was set at 3° , and the UAV's angular position was randomly initialized within this range. The results demonstrated that the developed algorithm is effective in realistic environments. Notably, Climber-sc achieved a 100% success rate after 1000 trial runs.

Table 3.2: Performance in empirical radiation pattern

pattern	method	success rate	total dist. in suc.	total dist. std. in suc.
Pattern1	Climber	0.798	4.068	1.607
	Climber-d	0.814	5.838	2.555
	Climber-sc	1.0	2.848	1.445
Pattern2	Climber	0.969	3.28	1.638
	Climber-d	0.994	4.23	2.147
	Climber-sc	1.0	2.632	1.38

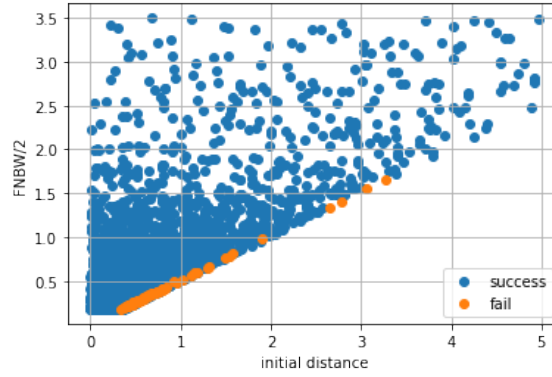


Figure 3.11: Fail case of Climber-sc

3.4.3.2 Localization evaluation

In this section, the boresight angle is numerically estimated by applying the data collected by the Climber-sc policy to the described estimation methods. Then, the overall application performance is compared with the BO beam localization method.

Estimation method Fig. 3.12 shows the estimation error and its standard deviation comparisons using "RL", "polynomial estimation" (labeled as "poly" in the figure), and "GP", as discussed previously. The positions of the data sampling before reaching the area of the boresight angle are biased, while the polynomial regression expects IID conditions from the provided data. Hence, the collected data were omitted if the power level was below a specific criterion in dB relative to the maximum signal strength on the trajectory. The numbers in brackets on the label represent this criterion. The accuracy of the polynomial estimation varied depending on the criterion (Fig.3.12a). Additionally, the performance improved or unchanged when the action noise was larger because the data was more randomly distributed as the data set became closer to the IID condition. Although this method could achieve the highest accuracy with the smallest standard deviation of error (Fig.3.12b) when the criterion was set at 1dB, "poly(1.0)", and the best criterion would be different depending on the beam shape. The Climber-based policy is not suitable unless its trajectory is designed to satisfy the IID condition. The polynomial estimation method would be more appropriate when the sampling points are uniformly distributed, like in a raster scan. The estimation from RL exhibited the worst estimation performance because the estimation relies solely on the corrected data. This method also lacks repeatability, as shown by the high estimation error standard deviation. This result underscores the importance of constructing a model based on the samples. GP regression is the most suitable estimation method to apply to the data from Climber-sc since GP regression does not explicitly assume IID data and it showed comparable and robust accuracy in the results, regardless of the action noises.

Comparison with BO Climber and BO [98] are evaluated in this scenario where the initial position of the UAV is on the main beam. Table 3.3 describes the numerical simulation results when the step size is limited to 50. BO does not have prior knowl-

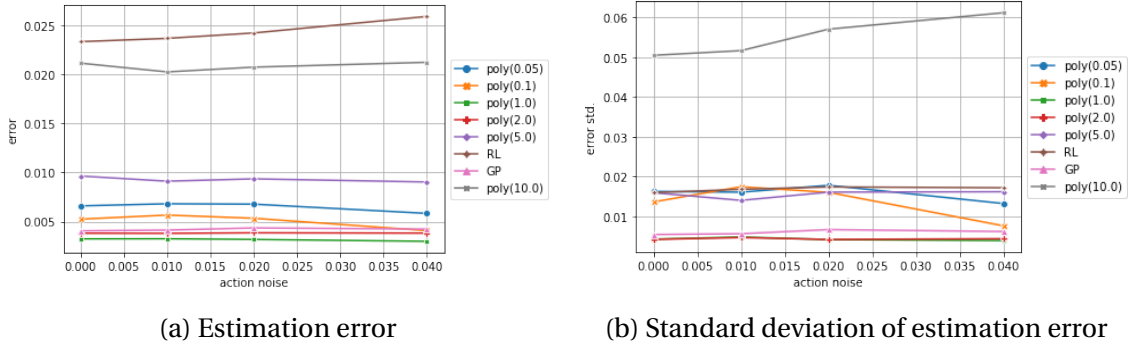


Figure 3.12: Estimation accuracy comparison

edge of the general beam pattern. Therefore, its trajectory is initially highly focused on exploring the entire search domain, resulting in a much higher total travel distance compared to Climber-cs. The acquisition function LCB in BO seeks to find points in the search domain where the antenna gain is expected to be high and uncertain, and the trade-off between exploration and exploitation is decided by a tuning parameter. The estimation accuracy of Climber-sc was higher than that of BO. The accuracy of BO improved as the action noise increased. This would indicate that the parameter to balance exploitation and exploration in BO was not optimally tuned because the accuracy was improved by consequently encouraging exploration by the action noise. The estimation from Climber-sc was superior because the UAV, guided by Climber-sc, moved slightly but more consistently than BO after reaching the center, as shown in Fig. 3.9b. However, the estimated amplitude and its variance of Climber-sc were inferior to those from BO, as it does not interact with the GP regression model.

3.5 Conclusion

In this work, an application to localize the boresight angle of the AUT for UAV-based measurements was developed. We devised a guidance algorithm, Climber-sc, based on an RL algorithm. The RL algorithm takes historical actions and observations on the trajectory as its inputs to overcome the challenge of POMDP. By defining RF elements appropriately, Climber-sc could robustly guide the UAV to the boresight angle, irrespective of conditions such as SNR, beamwidth, and initial angular distance to the target angle under positioning disturbances, if the UAV was initially located on the beam. Together with the GP regression model, the boresight angle was localized with an accuracy of around 0.004° .

The trajectory of Climber-sc after reaching the beam center was not developed in this study. The guidance policy could be a hybrid of Climber-sc and other algorithms such as the combination of raster scan and polynomial estimation or BO in a small search space. The mechanism of handing over the guidance policy should be explored further. More importantly, this work assumes that the UAV was initially located on the beam. The likelihood to have this condition depends on the beamwidth of the AUT and the expected boresight offset which might vary in each AUT evaluation. The development of an algorithm to roughly localize the beam should be a focus of our future work.

Lastly, the beam was localized considering a two-dimensional surface. This methodology could be extended to a three-dimensional strategy. If the beam center can be localized on the way to reach the evaluation range, the continued AUT evaluation can be instantly started, thus potentially saving significant operational time.

Table 3.3: Localization comparison

method	action noise	error	error std.	amp. est.	amp. est. var	total dist.	total dist.std
Climber-cs	0.00	0.0041	0.0054	-0.5492	0.0041	4.8869	0.5559
	0.01	0.0041	0.0056	-0.5525	0.0040	4.9469	0.5472
	0.02	0.0044	0.0067	-0.5571	0.0035	5.1071	0.5215
	0.04	0.0042	0.0062	-0.5829	0.0028	5.6980	0.4484
BO	0.00	0.0188	0.1677	-0.3802	0.0001	43.2261	18.3536
	0.01	0.0125	0.1520	-0.3775	0.0000	45.0424	19.6226
	0.02	0.0164	0.2146	-0.3812	0.0000	45.2009	19.7485
	0.04	0.0059	0.0098	-0.3848	0.0000	46.9875	19.8396

Chapter 4

UAV-based satellite communication on the move depointing measurement for a phased array antenna

Abstract

A method for numerically measuring depointing angles of Satellite-Communication-On-The-Move terminal antenna is developed. The developed application is dedicated to the evaluation of phased array antenna where the radiation pattern varies depending on its steering angle. Multiple unmanned aerial vehicles with radio frequency payloads are guided by a reinforcement learning-based algorithm to track the boresight of the antenna under test. The problem is analyzed as a Partially Observable Markov Decision Process to implement a reinforcement learning algorithm. Due to the limitation of the practically available number of sensors and the infeasibility of collecting all possible radiation patterns, a standard framework for Partially Observable Markov Process, which generally contains an external estimator, is not applicable. The proposed method does not utilize an estimator and is not restricted by the range of depointing. The results showed accuracy that meets the industrial requirements and demonstrated the feasibility of a new way of depointing measurement. Additionally, by introducing an imaginary agent, the required number of UAVs is effectively reduced without compromising the performance quality. The proposed method would enable on-site phased array terminal antenna depointing measurement without exhaustive data collection.

4.1 Introduction

Satellite-Communication-On-The-Move (SOTM) is nowadays a commonly used application that maintains communication between mobile terminals and satellites. A SOTM system comprises an antenna, which is mounted on the vehicle, a satellite modem, and an Antenna Control Unit (ACU). The angle of offset between the boresight and its target direction is numerically measured as "depointing". The ACU generally has an automatic pointing algorithm to compensate for the motions generated by the

vehicle so that the antenna stays pointed toward a target satellite, minimizing the risk of undesirable adjacent satellite interference. SOTM systems are widely applied in land mobile, maritime, and aeronautical applications [104, 2]. The types of terminal antennas used for SOTM systems are Very Small Aperture Terminals (VSATs) or Electronically Steerable Antennas (ESAs). For installing the SOTM terminal antenna on the roof of vehicles, ESAs are more beneficial solutions compared to conventional reflector antennas, such as VSAT, since they are more compact and offer instantaneous beam steering [105, 106]. While conventional mobile communications are established between satellites in geostationary orbit (GEO), a number of satellites were recently launched in Low-Earth-orbit (LEO) and Medium-Earth-orbit (MEO). Establishing seamless communication with satellites on Non-Geostationary Satellite Orbit (NGSO) is more challenging for SOTM applications since it requires further dynamic tracking of the intended satellite and handover.

Harmful interference is a critical issue. According to a survey, the satellite industry frequently suffers from satellite interference [107]. The characteristics of interference caused by SOTM systems are fundamentally different from the ones caused by fixed terminals since they appear randomly and intermittently. Hence, it is difficult to detect the source of interference and rigorous verification of SOTM systems is essential to prevent interference [2]. Satellite operators are responsible for ensuring that their users' communications do not interfere with other satellite communications and every satellite operator has their own requirement for type approval [6]. International organizations, such as Federal Communications Commission (FCC), International Telecommunication Union (ITU), and European Telecommunications Standards Institute (ETSI), also provide their recommendations [108, 109, 110]. Particularly, Satellite Operator's Minimum Antenna Performance (SOMAP) was generated by Global VSAT Forum (GVF) and industry-leading satellite operators to set minimum performance criteria [4]. Specifically, SOTM systems are subject to special requirements in terms of depointing and cessation of emissions. SOMAP stipulates that the allowable depointing is 0.5° and transmission cessation should occur within 100 milliseconds if depointing exceeds 0.5° [5, 47]. To provide a comprehensive evaluation of SOTM terminal antennas in the far-field, Fraunhofer IIS and Ilmenau University of Technology have developed a Facility for Over-the-air Research and Testing (FORTE), accompanied by extensive research [111]. FORTE is currently operated as an Authorized Test Entity (ATE) of the GVF. In depointing measurements conducted at FORTE, a signal is transmitted from an Antenna Under Test (AUT), with the signal being received at sensors placed in the far-field. The AUT is mounted on a motion emulator which simulates the angular motion of the intended vehicle based on motion profiles [7]. Five sensors are attached to a tower on a horizontal and vertical axis and their intersection point. Depointing is derived by calculating the correlation between the measured signal strength and the reference table. This method may have several drawbacks, despite the test result satisfying the required measurement accuracy of 0.05° . First, it requires a vast, stationary facility. The measurement process might be too time-consuming to handle the demand for evaluations, given the rapidly increased number of user terminals due to the growth of the satellite industry. Second, sensors are fixed at specific positions. The accuracy of the method is sensitive to sensor positioning with respect to boresight angle. However, it does not allow dynamic sensor positioning during the measurement. Therefore the measure-

ment accuracy may vary depending on the depointing and the measurable range of depointing is limited. Also, due to the static sensors, the facility is incompatible with the scenario where AUT communicates with Non-Geostationary Orbit (NGSO) satellites. Finally, the method requires reference tables. The radiation patterns of phased array antennas vary depending on the steering angle, and collecting all possible reference data is not feasible, as pointed out in [111].

In recent years, Unmanned Aerial Vehicle (UAV)-based antenna measurement has received a lot of research and industrial attention. Airborne RF measurements leverage UAVs to gather Radio Frequency (RF) data, facilitating antenna measurements [112, 28]. The UAV carries an RF payload to transmit or receive a signal from the AUT. Based on the collected RF data, such as power level and phase, and positional data of the UAV, antenna diagnoses are performed. The UAV-based measurement has practical benefits for the industry since it does not require bulky test facilities and the AUTs are evaluated on-site in a real operating environment. Furthermore, the UAV-based measurement system can adjust the sensor positions during the test and offer flexible testing scenarios, e.g., [113]. However, to the best of our knowledge, there is no existing work regarding UAV-based depointing measurement despite its importance.

In this study, we propose a method to guide UAVs with RF sensors to accurately evaluate SOTM terminal antenna in terms of depointing, especially for phased array antennas. To evaluate a phased array antenna, the previously mentioned method used in FORTE is not applicable, since available reference data are limited. Therefore, we propose to estimate the direction of boresight using multiple UAVs, under the assumption that no specific reference data are known in advance. The multi-agent system could be categorized as "team", "formation", and "swarm" based on the number of agents [114]. The "team" consists of fewer than 10 agents and "formation" and "swarm" assign categories include more than 10s of agents. If a "formation" or "swarm" of UAVs is available, the beam direction could be estimated by simultaneously reconstructing the shape of the beam based on the measured data at each UAV. There might be even may even be no need to reposition the UAVs during the measurement. Nonetheless, operating more UAVs increases logistical and operational complexity, as well as costs, and time for evaluations. It would be more practical to utilize UAVs in the range of "team" for real-world implementation as a service. When the number of available sensors is limited, UAVs should be dynamically positioned to maintain the estimation accuracy. Consequently, the decision-making process to guide UAVs must be instantaneous without exhaustive calculation. We propose the use of a Reinforcement Learning (RL) algorithm for this problem which maps observation to action spontaneously via a pre-trained neural network. The contributions of this work are summarized below:

- The problem is analyzed in Section 4.3. The standard framework of Partially Observable Markov Decision Problem requires an external estimator to maintain recursive updates. However, this approach is not feasible in this case, as the explicit measurement model is unavailable and the full state of the AUT is not directly observable. Hence, new frameworks are proposed to cope with this issue.
- The RL-aided guidance algorithm for UAVs without using an estimator is proposed in Section 4.3. The proposed algorithm identifies the transition of the motion of the boresight angle and the pattern transition. The algorithm is inde-

pendent of the reference coordinate system and hence, the developed method is applicable to the NGSO scenario.

- Keeping the required number of UAVs as small as possible is beneficial in practice. We discuss how the measurements taken at the boresight provide less useful information for decision-making. From this analysis, we introduce the concept of an imaginary agent (Section 4.3).
- The proposed algorithms are evaluated by numerical simulation results (Section 4.4). The performance is compared with other methods from the perspective of accuracy, trajectory stability, the required number of UAVs, their formation, and the robustness of the accuracy when action noise is considered. It was identified that the accuracy of the depointing measurement could reach 0.05° with a minimum of 3 UAVs.

4.2 Preliminary

In order to describe the problem and characteristics of the proposed RL algorithm, we provide some preliminary information in this section.

4.2.1 FORTE method

The method for depointing measurement at FORTE [111] can be described in three steps.

1. Create a reference table based on radiation patterns for five sensors placed at specific positions. The reference table may consist of the following items: depointing angle and expected signal strength at each sensor.
2. Record the received signal strength received at each sensor when the AUT is on the move.
3. Calculate the correlation between the measured signals and the reference table. Depointing is indicated by the angle which shows the maximum agreement.

The targeted measurement accuracy was set to 0.05° in the development [111] which is 10% of the allowable depointing stated in [5, 47]. Hence, this work also targets to reach 0.05° measurement accuracy. The measurement accuracy of FORTE to evaluate reflector antennas, which have static radiation patterns, satisfied this requirement when the sensors were placed at the optimal positions based on the function developed in [111]. The function for optimal sensor placement is a function of Signal-to-Noise Ratio (SNR) and 3 dB beamwidth. The physical meaning of this function is that the received power is increasingly affected by RF noise as the power level decreases when it is measured in dB units. Also, as the beam becomes flatter, a smaller deviation between measurements is obtained, making it more difficult to detect depointing based on the reference table. The parameters of the function were tuned to find an optimal balance between the noise effect and the distinctiveness of the observed signal, based on numerical simulation and statistical analysis. In [111], the method was also applied

in phased array antenna measurements. The reference table was prepared for a specific steering angle. However, due to the mismatch between the reference data and the actual steering angles operated during the measurement, the accuracy did not meet the expected requirement. An alternative method proposed to increase the number of sensors to reconstruct the beam pattern simultaneously instead of using a reference table. However, increasing the number of sensors is not preferable, especially for UAV-based measurements, as it may increase operational complexity, cost, and accident risk.

4.2.2 Phased array antenna

To illustrate the challenges of the method in FORTE that utilizes a reference table for phased array measurements, and to analyze the problem, we describe the behavior of a phased array antenna as follows.

A phased array antenna comprises multiple stationary antenna elements. The beam is concentrated toward a target direction by controlling the phase of the signals. With this technology, a higher antenna gain can generally be achieved compared to reflector antennas. The elements need to be fed coherently to steer the main beam toward target angles in space using variable phase or time-delay control. The pattern may also vary depending on the amplitude control of each element. The phased array antenna pattern is described in [115] as

$$\text{Array Pattern} = \text{Element Pattern} \times \text{Array Factor} \quad (4.1)$$

Namely,

$$E(\theta, \phi) = f(\theta, \phi) \sum_{m,n} |a_{m,n}| \exp(jk_0(m d_x(u - u_0) + n d_y(v - v_0))) \quad (4.2)$$

where $k_0 = 2\pi/\lambda_0$ (λ_0 is wavelength), $f(\theta, \phi)$ is the element pattern in elevation θ and azimuth ϕ directions, $|a_{m,n}|$ is the weight amplitudes, m, n, d_x, d_y are element numbers and cell distances in two axes respectively, $u = \sin\theta \cos\phi$, $v = \sin\theta \sin\phi$ are direction cosines, and u_0 and v_0 are the beam direction in u-v space when beam peak is placed at (θ_0, ϕ_0) [115]. Figure 4.1 shows the simulated 32×32 element rectangular phased array antenna patterns with various steering angles. The element factor is defined as cosine shape and element spacing and frequency are configured as $\lambda_0/2$ and 14GHz. As can be observed in Figure 4.1, the basic behavior of a phased array antenna is that the beamwidth and scan loss of the main lobe becomes larger as the steering angle is increased [116]. Also, the beam becomes narrower when the number of elements is increased. The precision of pointing is affected by the number of elements, spacing, and bits utilized for phase shifting [117].

Due to the impracticality of creating reference data for all possible radiation patterns, the method used in FORTE cannot accurately measure depointing. Currently, there is no existing solution for evaluating phased array antenna depointing for SOTM applications, and the patterns are often evaluated using antenna simulators. Therefore, a new approach for measuring depointing that does not rely on reference data is necessary.

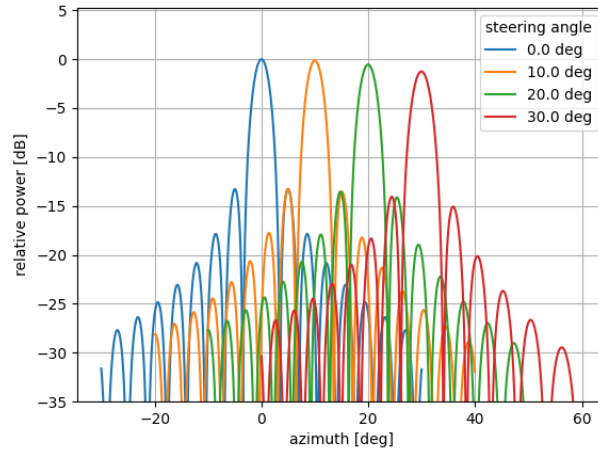


Figure 4.1: Simulated phased array antenna with various steering angles

4.2.3 LSTM-TD3_{1halhc}

The developed solution utilizes an RL algorithm. RL is one of the branches of machine learning techniques where an agent learns the optimal policy by interacting with the environment instead of requiring a model of the environment. The environment provides the state s_t and reward r_t at timestep t . The agent generates action a_t based on the state s_t . The reward r_t is the evaluation of the action when the agent is at the state s_t . When the true state is not available due to disturbances or limited sensors, the problem is referred to as a Partially Observable Decision Making Process (POMDP). In POMDPs, it is common to use sequential observation as input to the policy through an enhanced Recurrent Neural Network (RNN), such as Long-Short-Term-Memory (LSTM) [51] or Gated Recurrent Unit (GRU) [52]. When the environment dynamically changes due to non-static disturbances to observations or by its inherent nature like this problem, it is beneficial to consider sequential actions as well as sequential observations as inputs to the network. By considering both sequences, the causality between them can be internally identified, and the agent can dynamically adjust its policy to the environment. LSTM-TD3_{1halhc}[94] was developed to provide robustness in such situations. Unlike conventional test facilities, the UAV-based measurement system allows the UAVs to dynamically change their positions during the measurement. The proposed method tracks depointing by positioning UAVs with RF sensors around the boresight, as will be discussed in Section 4.3. To realize this approach, LSTM-TD3_{1halhc}[94] is implemented as a guidance algorithm for the UAVs. As it was described, there is an existing transition pattern of the radiation pattern in a phased array antenna. Also, it may be possible to construct a dynamic model of the boresight, even though the steering angle is digitally controlled. In this kind of environment, LSTM-TD3_{1halhc} could internally model these transitions from the sequential information and guide the UAV to the boresight angle.

4.3 Method

In this section, considered test setup and coordinate system are introduced to define the problem. Then, the problem is conceptually analyzed to define RL environment.

4.3.1 Measurement setup

The considered real-world measurement setup comprises UAVs with a receiving RF payload, an AUT, and a motion emulator, as shown in Fig. 4.2a. The angular motion of the intended vehicle is generated at the motion emulator based on the motion profile [7], while the AUT steers its beam towards the target direction to compensate for the motion. We introduce coordinate systems defined by azimuth and elevation [118]. Fig. 4.2b shows a projection of the elements on a measurement surface, illustrating the considered coordinate systems in a Cartesian coordinate system. The reference coordinate system with origin O_R defines the target direction, the realized boresight angles, the angular position of the agents, and the angular motion caused by the motion emulator. A typical phased array antenna has an operation range of steering angles, illustrated as a gray shadow in the figure. The second coordinate system is the AUT coordinate system with origin O_{AUT} , which is attached to the motion emulator. The angle towards the target direction seen from this coordinate defines the ideal steering angle for the AUT. If the angular motion of the motion emulator and steering angle are zero, the AUT's boresight is projected onto the origin O_R .

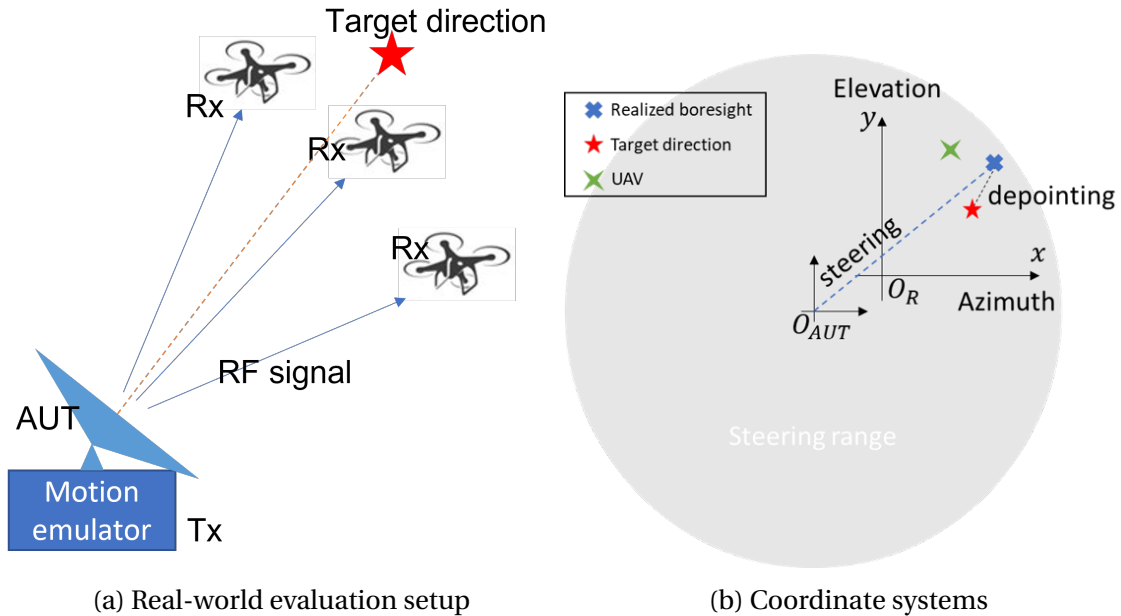


Figure 4.2: Measurement setup

4.3.2 Objective of guidance

The objective of this application is to estimate the direction of the boresight angle. When only the signal strength is available from the environment and the measurement

model is not, the estimation cannot rely on time-series information via a recursive update (e.g., the Kalman filter) but must solely depend on the current observation. Depointing may be estimated by two methods: either (a) applying the observed received signal strength and its measured position to a mathematical model and obtaining depointing from the constructed model, or (b) using the agent's ideal position. In this work, a polynomial model is selected for method (a), and is referred to as "polynomial estimation". The accuracy of the polynomial estimation is sensitive to the position of the sensors and the number of samples simultaneously collected at different positions. Method (b) does not rely on an explicit model like method (a). Instead, the estimation is obtained based on the ideal position of a selected agent, which is the waypoint generated from the RL. The realized position of the agent might contain errors due to disturbances. The term "ideal" position differentiates it from the realized UAV position. This approach is named "positional estimation" in this paper. The positional estimation requires one of the agents to be located at the boresight angle. Additionally, for polynomial estimation, the estimation accuracy should be improved if the measurement is taken at the boresight angle. Therefore, the objective of the guidance should be to dynamically place one of the agents at the moving boresight angle.

When the agent's waypoint at timestep $t + 1$ is set to the estimated boresight angle at timestep t , based on the measurement at t , there is a delay in estimation at $t + 1$. It would be beneficial if the guidance policy could predict the boresight angle at $t + 1$ based on the current measurement and generate the waypoint to the predicted angular position of the boresight. Hence, the objective of the guidance is restated to place one of the agents at the predicted boresight angle at $t + 1$.

The agent intended to be on the boresight angle is called a central agent. The other agents, referred to as subset agents, are placed around the central agent and collect measurements. With all the collected data, the estimation is performed, and the guidance algorithm generates trajectories.

4.3.3 Problem analysis in POMDP framework

To analyze the problem, let us consider a simple dynamic model of boresight angle and measurement model at the origin of the reference coordinate system O_R . The dynamic model could be assumed as (4.3).

$$\begin{bmatrix} \theta_{t+1} \\ \dot{\theta}_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_t \\ \dot{\theta}_t \end{bmatrix} + \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{bmatrix} \ddot{\theta}_t \quad (4.3)$$

where Δt is the sampling interval time and θ and $\dot{\theta}$, $\ddot{\theta}$ represent angle, angular velocity and angular acceleration respectively. The measurement model at O_R is defined as (4.4) since the available measurement from the environment is only signal strength and the angular velocity does not contribute to the signal strength.

$$Pr_t = [f, 0] \begin{bmatrix} \theta_t \\ \dot{\theta}_t \end{bmatrix} \quad (4.4)$$

where Pr_t is the received signal strength and f is a hypothetical function related to (4.2). If we analyze the observability of this system, it is found that this system is not

observable. Therefore, the prediction of the boresight at $t + 1$ cannot be formulated explicitly.

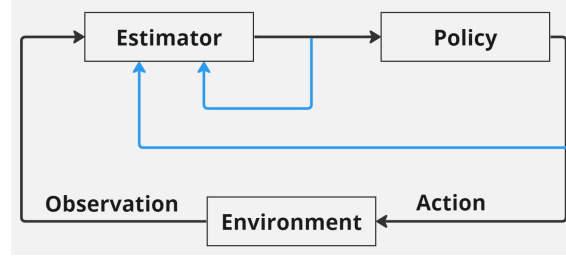
The system under consideration is not observable and the measurements contain noise in this problem. When the agent has no access to the true state of the system, the problem is categorized as POMDP. The standard framework to approach POMDP is illustrated in Fig. 4.3a [49, 70]. The estimation is recursively updated based on the current observation, previous estimation, and action, utilizing measurement and dynamic models of the system. However, antenna models are hard to obtain and the states of the system are not observable. Therefore, the external estimator utilizing recursive update in Fig. 4.3a is not available for this problem.

Some available strategies, in this case, are depicted in Fig. 4.3b and Fig. 4.3c as "framework1" and "framework2", respectively. It might be possible to predict the boresight angle at $t + 1$ by processing the available information (i.e., observation, action, and output of the estimator) from the previous timesteps at the policy. Framework1 more closely resembles the standard POMDP as it contains an external estimator. In framework1, the policy considers all available information from the current and previous timesteps including the output from the estimator. Provided the estimation is sufficiently accurate, taking the values from the estimator might add valuable information. However, due to the limited number of sensors, obtaining accurate estimations might be challenging. Therefore, we propose the structure seen in Fig. 4.3c, which omits an external estimator, and action is generated solely based on the observations from the previous and current timesteps and previous actions. LSTM-TD3_{1ha1hc}[94] utilizes l length of past sequential information from $t - l$ to $t - 1$ and the available information at t . By considering the sequential information, which includes disturbances from the noise, LSTM-TD3_{1ha1hc} should find the causality between them and learn to identify the pattern of the transition of the radiation pattern and boresight dynamic to conduct targeted guidance.

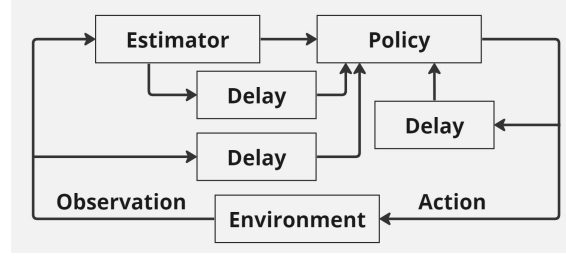
4.3.4 RL setup

Based on the above discussions, POMDP tuple to implement RL is defined in this section.

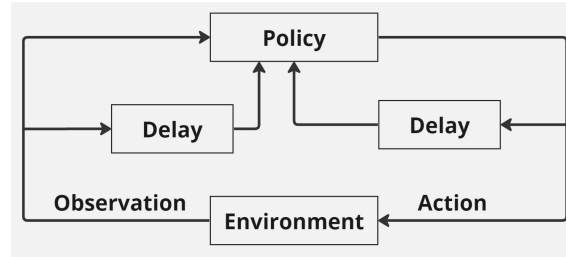
State and observation In RL, there are two types of "state": the "environment state" and the "agent state" [119]. The environment state is the private representation of the environment in which the observation and reward are determined. In the problem under consideration, the environmental state could include the state of the AUT, AUT configurations, and the radiation patterns of the AUT which are not visible to the RL agent and may not be relevant to decision-making for the defined guidance objective. The agent state is the internal representation of the agents and the action is based on the agent state. In POMDPs, the action is generated from the observation, which can be disturbed or limited, rather than the agent state. Since the guidance objective is to place the central agent at the boresight angle, the state must have a positional interpretation between the agent and the boresight angle. The agent position on the reference coordinate system is independent of the AUT coordinate system. Therefore, the agent's position in this coordinate system does not contribute to decision-making. The rela-



(a) Standard POMDP framework



(b) framework1



(c) framework2

Figure 4.3: Structure diagram

tive position between the estimated boresight angle from the estimator could provide useful information if it is accurate. The received signal itself also has an implicit relationship between the agent position and the boresight angle as found in (4.2). Therefore, the elements of the observation should either be defined as the relative angular state of the agents with respect to the estimated boresight angle and received signal strength, or only signal strength.

As such, when the framework1 with an external estimator illustrated in Fig. 4.3b is selected, the observation is defined as $\left[\Delta\hat{d}_t, P_t^{0:N}\right]$, where $\Delta\hat{d}_t$ is a realized vector between the angular position of the central agent and the estimated center, and $P_t^{0:N}$ is the received signal strength from N number of agents. On the other hand, the observation of the algorithm using the structure without an external estimator (Fig. 4.3c) is set to be $\left[P_t^{0:N}\right]$. Although the range of received signal strength varies depending on the test configuration (e.g., transmission power level and distance between the sensor and the AUT), the received signal strength for RL can be normalized at steering angle zero since it is valid to assume that the signal strength at the boresight angle with steering angle zero is easily obtained in this scenario.

Reward Since the state and observation have the interpretation of the positional relationship, the reward is also defined as (4.5).

$$r_t = -\Delta d_t - 0.01\Delta a_t \quad (4.5)$$

where Δd_t is the angular distance between the central agent and realized boresight angle and Δa_t is the displacement of the agent to prevent aggressive and inefficient maneuvers.

Action The action is configured as the displacement of the agents. This approach makes the developed application applicable to the NGSO scenario by avoiding the utilization of absolute position values. In this work, a fixed formation is considered. Therefore, both the central and subset agents receive the same action from the policy.

4.3.5 Imaginary agent

The agent, in the above descriptions, is understood as UAVs. However, the agent could also be an entity that does not perform the measurement but is involved as a part of the RL system. In such a case, the agent is referred to as an "imaginary agent" in this work, and a UAV is not assigned to this agent.

It is preferable to use a smaller number of UAVs for the practicality of the application. The shape of the beam around the beam peak (i.e., boresight) is generally flat and not distinctive. Especially when the depointing is estimated by the "positional estimation" method from framework2 (Fig. 4.3c), it might be able to maintain the estimation accuracy without assigning a UAV at the central agent because the obtained signal strength at the central agent would be less informative due to its lack of distinctiveness, as discussed. Therefore, it might be feasible to designate the central agent as an imaginary agent that does not provide information. In this case, observation is replaced by $P_t^{1:N}$, labeling the central agent as agent 0.

4.4 Experiment

4.4.1 Environment

Antenna pattern of AUT The RL system is trained with the antenna model, as defined in Section 4.2, as illustrated in Fig. 4.1. In a real-world evaluation setup, a motion emulator simulates the angular motion of the intended vehicle based on its motion profile (Fig. 4.2a). The AUT compensates for the angular motion by steering the beam. Depointing is caused by a misalignment between the target direction and the actual beam direction. To simulate this test setup, the dynamics of the AUT are decomposed into two parts: base angle and depointing. The base angle represents the angular state of the motion emulator, and depointing represents the misalignment angle. Therefore, the actual steering angle of the AUT at the AUT coordinate is given by

$$\text{steering angle} = \text{boresight angle} = -\text{base angle} + \text{depointing}. \quad (4.6)$$

Based on the calculated steering angle, the radiation pattern at that moment is simulated. Furthermore, the signal strength at each UAV is calculated at the transformed angular position of the UAV position on the AUT coordinate. The observed signal strength is disturbed by simulated RF noise defined by SNR with respect to the beam peak when the steering angle is zero.

AUT dynamic model In the simulation, the motion of the base angle and depointing is simulated by generating random acceleration. The dynamic model of the AUT is governed by (4.3), with $\Delta t = 0.1$, and the angular acceleration is randomly generated as per (4.7).

$$\ddot{\theta}_t \sim U(a, b) \quad (4.7)$$

where $[a, b] = [-1.0, 1.0]$ for base angle and $[a, b] = [-0.1, 0.1]$ for depointing.

UAV dynamic model In this work, UAVs are assumed to be replaced to the guided position instantaneously as described in (4.8).

$$\begin{bmatrix} \theta_{t+1}^{UAV} \end{bmatrix} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_t \\ a_t \end{bmatrix} \quad (4.8)$$

where θ^{UAV} is the angular position of UAV and a_t is the generated angular displacement.

Initial condition and test length The AUT is initialized with a random base angle and zero depointing. Also, the angular velocity and acceleration terms for the AUT are set to zero. The initial position of the central agent, or the imaginary central agent, is set to the beam peak. These initial conditions represent a measurement scenario where the initial states of the boresight are stationary and are perfectly known to the test system. Hence, the initial angular position of the UAV is aligned with the boresight angle. The SNR is randomly configured between 60 to 100 dB at the start of each episode. The simulation duration is set to 200 timesteps, and the AUT motion is generated during this period. The base angle is restricted between -30° and 30° , which is the typical operating range of a phased array antenna. The target direction is set to the origin of the reference coordinate O_R to simplify the simulation. It should be noted that measurement is still performed at any steering angle within the operational range since the base angle is randomly initialized.

4.4.2 Compared Algorithms

The algorithms used for comparison are summarized in Table 4.1. The "Algorithm" is the label of the algorithms used in the following discussion. The "Framework" and "Estimation method" refer to POMDP frameworks and the estimation method introduced in Section 4.3. Also, the "Policy input" shows the input elements of the policy. The algorithm "Poly" is additionally considered, which does not utilize an RL-based policy. The policy of "Poly" directly takes the vector between the current polynomial estimation and the current position of the central agent, i.e., $a_t = \Delta \hat{d}_t$. The "Poly" does

not belong to any of the frameworks and could be illustrated by removing all the delay operators from the framework1, since the policy only depends on the output of the estimator. The purpose of including "Poly" in the comparison is to evaluate whether RL-based algorithms are predicting the boresight angle of the next timestep. The algorithms "RL_{rel+Pr}", "RL^{*}_{rel+Pr}", "RL_{Pr}" and "RL^{*}_{Pr}" are the discussed algorithms in Section 4.3. "*" denotes the presence of an "imaginary agent". The expression of $t-l:t$ and $t-l-1:t-1$ describes the sequence of the element between the timestep at the left and right side of ":". The length of the taken past information l in any RL-based algorithms is set to 5.

Table 4.1: Compared algorithms

Algorithm	Framework	Policy input	Estimation method
Poly	—	$\Delta \hat{d}_t$	polynomial
RL _{rel+Pr}	1	$\Delta \hat{d}_{t-l:t}, P_{t-l:t}^{0:N}, a_{t-l-1:t-1}$	polynomial/ positional
RL _{Pr}	2	$P_{t-l:t}^{0:N}, a_{t-l-1:t-1}$	polynomial/ positional
RL [*] _{rel+Pr}	1	$\Delta \hat{d}_{t-l:t}, P_{t-l:t}^{1:N}, a_{t-l-1:t-1}$	polynomial/ positional
RL [*] _{Pr}	2	$P_{t-l:t}^{1:N}, a_{t-l-1:t-1}$	polynomial/ positional

4.4.3 Result

The performance of the trained system is evaluated from several aspects in this section based on the performance with the same environmental settings. As the AUT's motion is generated by the random angular acceleration and the pattern varies by the realized boresight angles, the agent would have different experiences from the training. Although discussing the RL algorithm itself is beyond the scope of this work, the learning trajectories of each RL algorithm can be found in Appendix 4.6.1.

Basic comparison with 4 subset agents Firstly, the result of the basic configuration, which has 4 subset UAVs (with or without an imaginary agent at the center), is compared. The angular distance between the center and the subset UAVs was set to 1° as depicted in Fig. 4.4. Table 4.2 presents the numerical performance results averaged over 1000 times of trials. "poly" and "pos" in the "Est. method" column describe "polynomial estimation" and "positional estimation". "Avg. error", "Std. of err.", "Avg. total dist." and "Std. of total dist." stand for averaged estimation error, the standard deviation of estimation error averaged total angular travel distance per episode, and its standard deviation. The last two elements of the evaluation parameter describe the stability of the trajectory.

When a central agent was not imaginary, the polynomial estimations showed slightly better estimation accuracy by comparing the results in "RL_{rel+Pr}" and "RL_{Pr}". The polynomial estimation accuracy of "RL_{rel+Pr}" and "RL_{Pr}" were improved compared with "poly" policy. As it was described, the action from "poly" policy is always one step behind because it does not consider the future state. On the other hand, RL-based policies take past information and try to place the central agent to the estimated boresight angles at the next timestep. As it was hypothesized, the proposed method could

perform in a dynamic environment by analyzing the sequential information via an internally learned model about the environments. The total distance and standard deviation of "poly" policy were also higher than the other RL-based algorithms, which reveals the instability of the polynomial estimation itself due to the delayed guidance policy. "RL_{rel+Pr}" of the positional estimation accuracy and the total distance of RL-based algorithms with a central UAV were better than "RL_{Pr}". However, the standard deviation of the total distance of "RL_{rel+Pr}" was noticeably larger. Since "RL_{rel+Pr}" takes the information based on "poly" estimation which is accurately performing, the accuracy and the distance were improved. The estimation from "poly" is independent of the time sequence and there is no restriction to produce locally scattered estimation. This might have caused the large standard deviation in traveling distance. When an imaginary agent was introduced the polynomial estimation accuracy unsurprisingly got much worse than the formation with a central UAV. Therefore "RL^{*}_{rel+Pr}" marked the worst in all evaluation parameters although the network learned to reject the unreliable information i.e., $\Delta\hat{d}$, and still somewhat could perform. On the other hand, "RL^{*}_{Pr}" resulted in comparable performance with other RL-based algorithms which assigned a UAV to the central agent regardless of the no presence of a UAV at the center in terms of positional estimation accuracy, the standard deviation of the error, and efficiency and stability of total distance. Overall, all results could satisfy the required estimation accuracy of 0.05° . From this analysis, it could be concluded that the RL algorithm with the structure without an estimator (framework2) could be a more suitable approach when the estimator is unstable, and the required number of UAVs can be reduced by introducing an imaginary agent as described in the result of "RL^{*}_{Pr}".

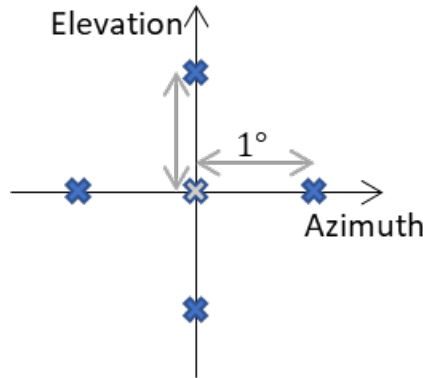
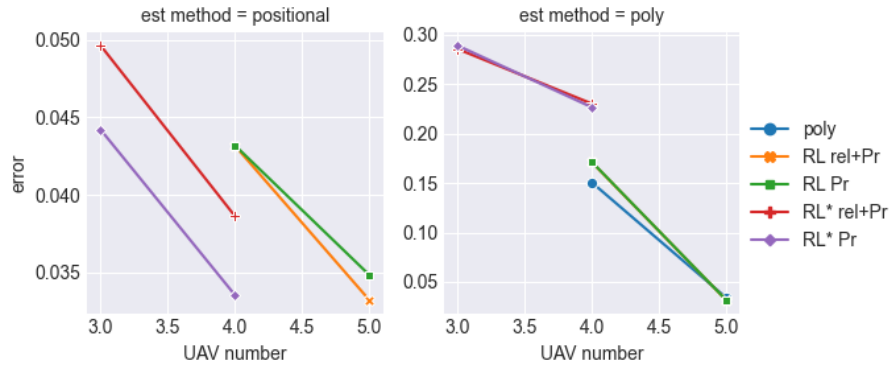


Figure 4.4: Basic formation

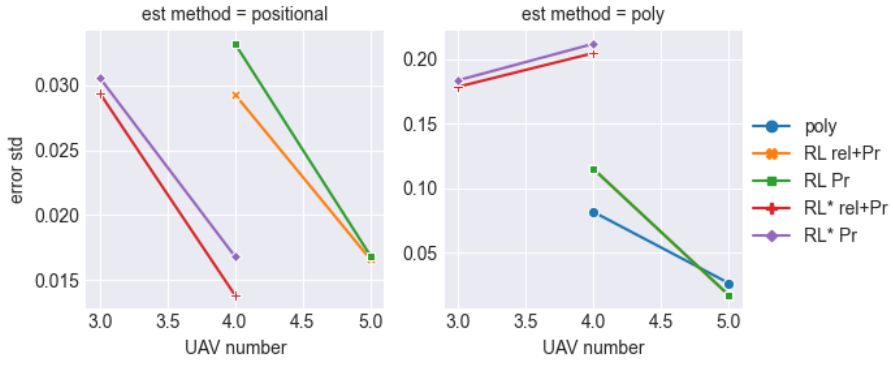
Table 4.2: Performance

Algorithm	Est. method	Avg. err.	Std. of err.	Avg. total dist.	Std. of total dist.
poly	poly	0.034	0.027	2.155	6.411
RL _{rel+Pr}	poly	0.031	0.017	1.126	0.446
	pos	0.033	0.017		
RL _{Pr}	poly	0.031	0.017	1.139	0.384
	pos	0.035	0.017		
RL [*] _{rel+Pr}	poly	0.230	0.205	2.152	1.526
	pos	0.039	0.014		
RL [*] _{Pr}	poly	0.227	0.212	1.125	0.384
	pos	0.034	0.017		

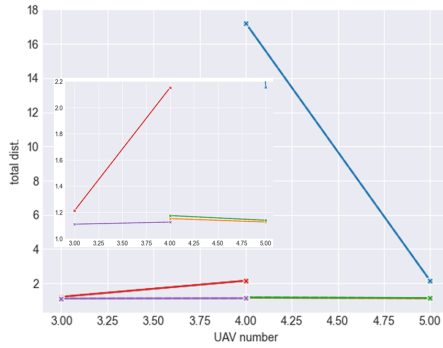
Comparison in a change of UAV numbers of each algorithm Next, the number of the subset UAVs was reduced by one and the performances of each algorithm were compared. The subset agents were distributed equally on the radius of 1.0° , as in the previous evaluation. Fig. 4.5 illustrates the results in terms of the aforementioned four evaluation parameters. Fig. 4.7a and Fig. 4.7b separately show the result of the two estimation methods. The horizontal and vertical axes show the total number of UAVs and the results of the evaluation parameters. As the number of UAVs decreases, the estimation accuracy declines. The degradation of polynomial estimation by the reduction of UAVs was more significant compared to positional estimation. The total distance and standard deviation of RL^{*}_{rel+Pr} became smaller with a lower number of UAVs since it has less weight on the estimated relative distance to the boresight angle $\Delta\hat{d}_t$. It can be visualized by the trajectory example of them shown in Fig. 4.6. The blue and red crosses in the figure indicate the start and end angular positions respectively. The figure illustrates that the trajectory was less affected by polynomial estimation. It shows that the trained agents with three subset agents had a smaller weight on $\Delta\hat{d}_t$ since it is no longer reliable and does not provide informative values. The performance of RL^{*}_{Pr} stayed comparable with the RL-based algorithms with one more UAV, as shown in Fig. 4.5.



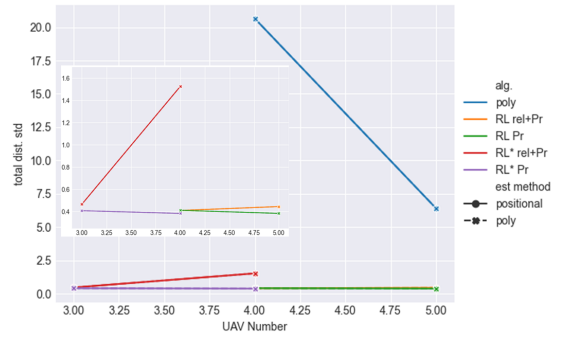
(a) Average error



(b) Error std

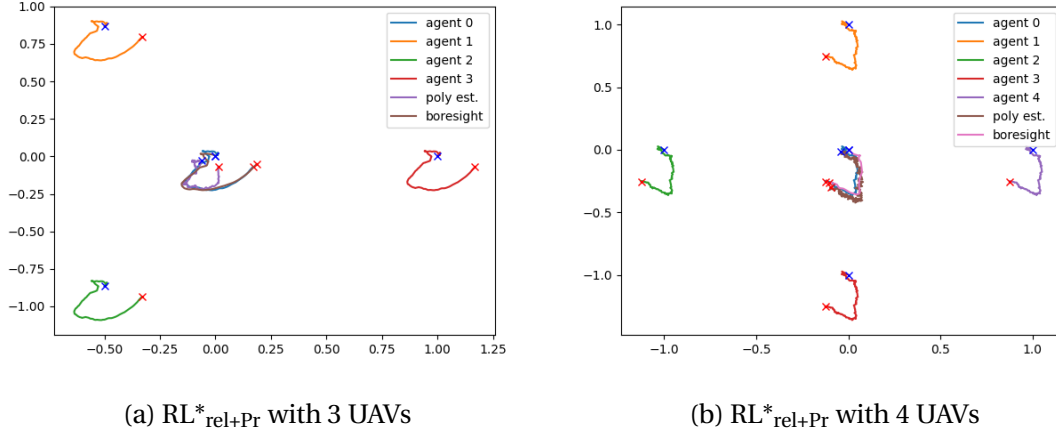


(c) Average displacement

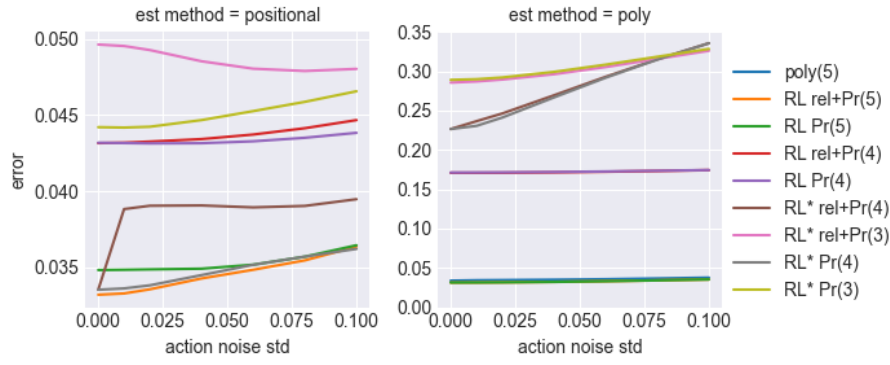


(d) Displacement std

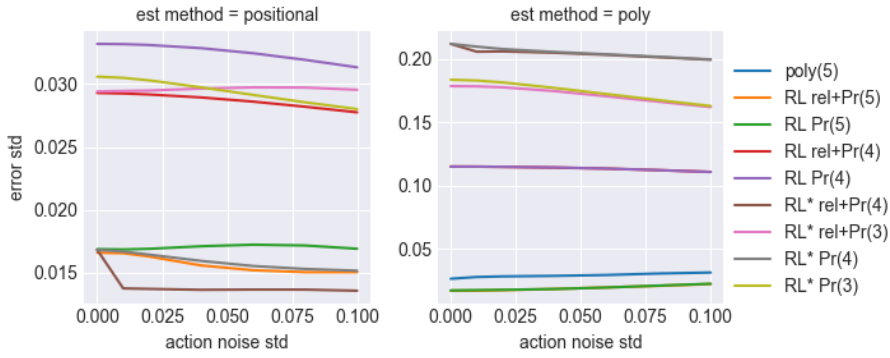
Figure 4.5: Performance comparison in total UAV numbers of each algorithm

Figure 4.6: Trajectory example of RL^*_{rel+Pr} of 3, 4 subset UAVs

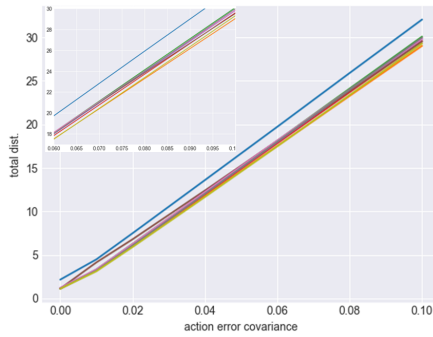
The effect of action noise In this section, the performance under the presence of action noise is analyzed. The additive noise was defined as a zero-mean Gaussian with various standard deviations. The horizontal and vertical axes show the standard deviation of the action noise and the results of the evaluation parameters respectively. The labels represent the corresponding algorithms. The numbers in brackets in the labels describe the total number of UAVs including central and subset. It should be noted the positional estimation is based on the generated waypoints which are not affected by the positional noise. In Fig. 4.7, as the scale of the noise increased, the total distance and its standard deviation were naturally increased and the positional estimation accuracy degraded except for RL^*_{rel+Pr} with three UAVs although it performed the worst as shown in Fig. 4.7a. Since it is hard to identify the error of polynomial estimation and compare it with positional estimation from Fig. 4.7a, Fig. 4.8 also shows the comparison of the estimation error of the algorithms with five UAVs without an imaginary agent with the vertical axis on the shared scale. The polynomial estimation by RL-based algorithms was better than positional estimation although the degradation ratio against noise scale was larger. Hence, polynomial estimation would be a better estimation method when 5 UAVs are available. Unsurprisingly, by comparing results based on the number of UAVs between individual algorithms, the accuracy and its standard deviation of the system were improved by utilizing more UAVs and this trend did not change by action noise scale. Considering the target accuracy of 0.05° , RL^*_{Pr} with three UAVs satisfies the criteria in any action noise with the minimum number of UAVs in the compared algorithms.



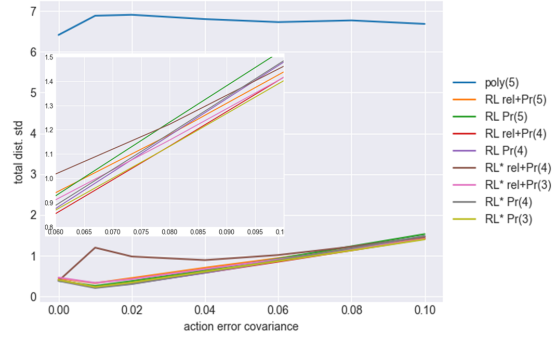
(a) Average error



(b) Error std



(c) Average displacement



(d) Displacement std

Figure 4.7: Performance comparison in action noise

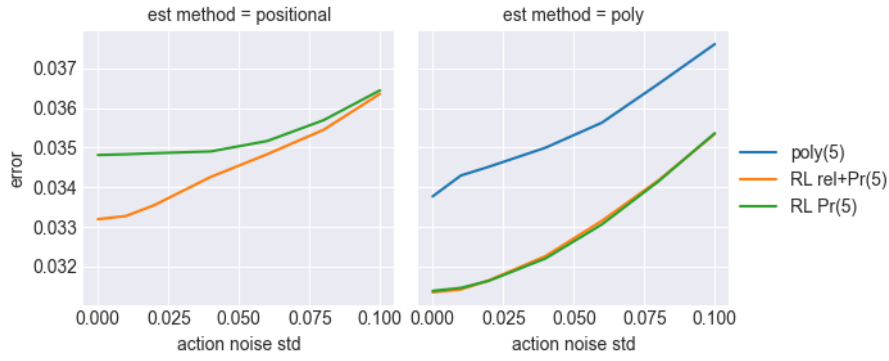
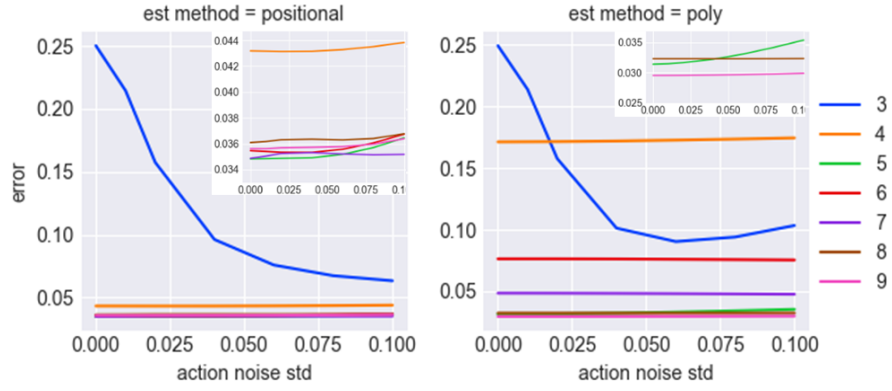


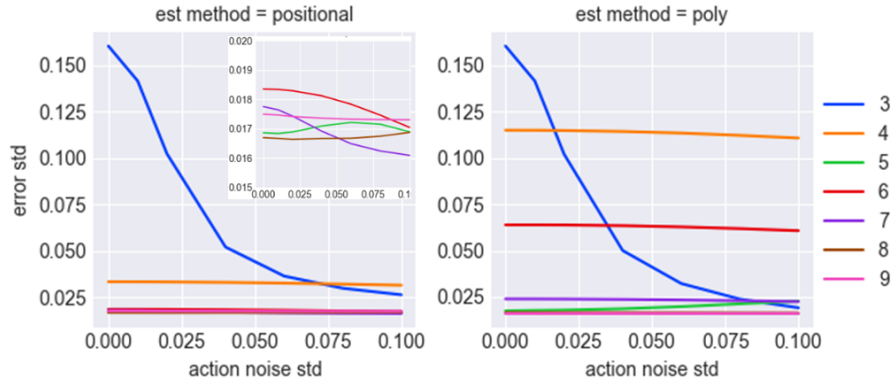
Figure 4.8: Estimation method accuracy comparison

The effect of action noise and UAV numbers The effect of the UAV numbers under the presence of various action noises was investigated to address the minimum required number of UAVs with this formation and find whether the performance converges by increasing the number of UAVs. RL_{Pr} was selected in this analysis to evaluate the polynomial estimation accuracy as well, although RL^*_{Pr} showed comparable performance with fewer UAVs in the above results.

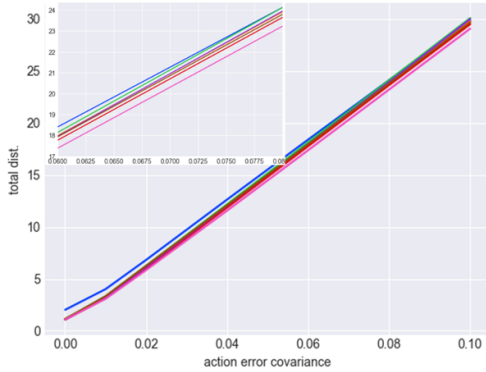
The subset UAVs were equally distributed in terms of phase between them on the circle of which the radius is 1° . Fig. 4.9 shows the comparison of performance in the change of the number of UAVs. The label indicates the used number of UAVs. When the number of UAVs is above 5, there was no tendency in positional estimation error and its standard deviation with respect to UAV numbers. The performance of positional estimation randomly varies with more than 5 UAVs. However, the result shown in this figure is depending on the final state of the trained network. The recorded total reward on the learning trajectory also fluctuates (Appendix 4.6.1). For example, the formation with 7 UAVs, which showed the best performance, did not always keep the best total reward on the learning trajectory. From this observation, it is concluded that increasing the number of UAVs with this formation does not necessarily improve the performance of positional estimation beyond 5 UAVs. Contrarily, the estimation error and its standard deviation from polynomial estimation have a clearer trend since they improved as the number of the UAVs increased except for the case with 5 UAVs. Clearly, the trend of the algorithm with 3 UAVs is different from the others. Since the subset UAVs were distributed in one line, the data in the orthogonal direction might not have been obtained and this degraded the estimation error.



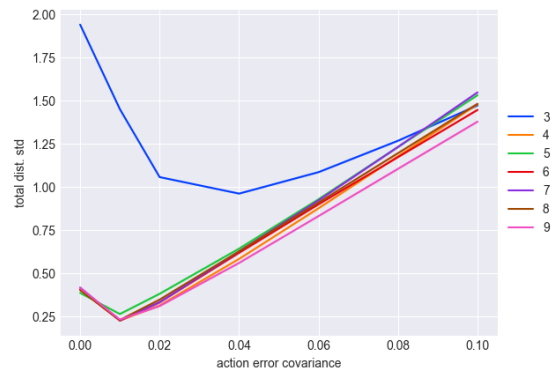
(a) Average error



(b) Error std



(c) Average displacement



(d) Displacement std

Figure 4.9: Performance comparison in action noise and UAV numbers

The effect of formation Next, let us evaluate the formation to the positional estimation of UAVs using RL_{Pr} with 5 UAVs. The tested formations are listed as follows:

- V shape formation: The 5 UAVs were placed in V-shaped formation. The angles of each line were 45° and 135° from the azimuth axis and the separation between each UAV was 0.5° in angular distance.
- I shape formation: The 5 UAVs were placed in I-shaped formation. All UAVs were

aligned on one line and the separation between each UAV was 0.5° .

- Radius 0.5: Same formation as the basic formation. The separation from the central UAV was decreased to 0.5° .
- Radius 1.5: Same as above. The separation from the central UAV was increased to 1.5° .

Fig. 4.10 represents the performance comparison when the formation was changed. Symmetrical formation in one axis (V shape and I shape) showed worse performance. Particularly, the performance degradation of I shape was distinctive and it has a similar trend with 3 UAVs in Fig. 4.9. Hence it reinforces the hypothesis that, with the formation in one line, the information related to the pattern is available only for one dimension and the orthogonal motion of the boresight angle to the line of UAVs is hard to detect. The transition of the radiation pattern is symmetrical and therefore, the relation between observation and action cannot be mapped uniquely in the network. By comparing the formation with different radii (rad 0.5, 1.0, 1.5), the formation with a larger radius resulted in better performance. As observed in imaginary agent cases, the data obtained near the beam center is less useful to guide UAVs since it is less distinctive. However, the RF measurements are affected by RF noise if the received signal strength becomes lower which is also as the FORTE method suggests. The design of the radius should have a relation with SNR and beam shape and the optimal radius might dynamically change during the measurement as the beam shape changes in the case of phased array antenna evaluation.

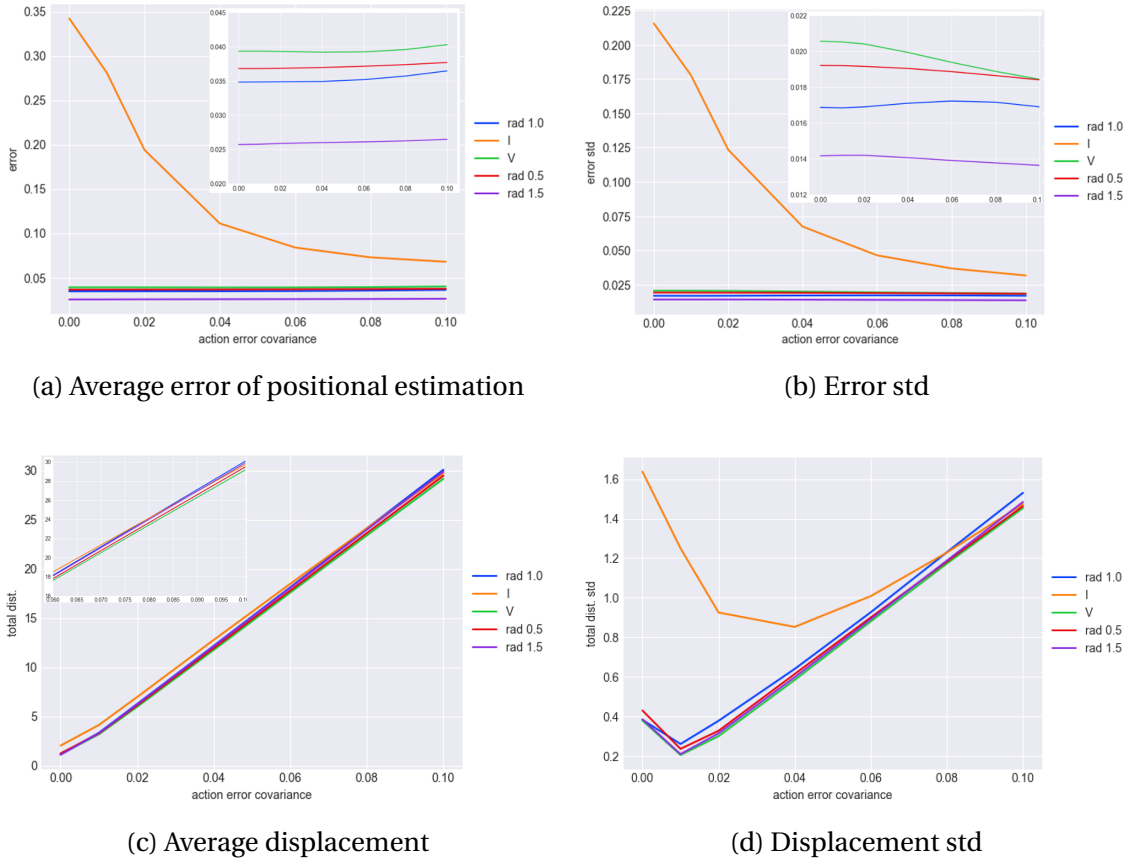


Figure 4.10: Performance comparison in action noise

The effect of past information Finally, the effect of past information inclusion on the inputs of RL_{Pr} algorithm is evaluated. Two new algorithms were introduced without changing RL elements definition: (a) RL_{Pr} without past action and (b) RL_{Pr} without any past information. It can be conceptually understood that the delay operator connected to action is eliminated for (a) from Fig. 4.3c and the ones connected to action and observation were deleted for (b). To provide higher contrast, the SNR was set to 40dB in observed signal strength and the performance of these three algorithms is compared in Table 4.3. As the result illustrates, the positional estimation and total distance were better with past information which confirms the existence of causality between sequential action and observation, and hence, RL_{Pr} improved the performance by identifying the pattern between them. Also, it would be worth pointing out that the algorithms that utilize the past information (RL_{Pr} and RL_{Pr} without past action) showed much more sample-efficient as observed in learning trajectory Appendix 4.6.1.

Table 4.3: The effect of past information in SNR 40dB

Algorithm	Est. method	Avg. err.	Std. of err.	Avg. total dist.	Std. of total dist.
RL _{Pr}	poly	0.057	0.022	8.138	0.850
	pos	0.049	0.013		
RL _{Pr} without past action	poly	0.057	0.022	11.566	1.387
	pos	0.056	0.013		
RL _{Pr} without past information	poly	0.057	0.022	11.645	0.802
	pos	0.057	0.014		

4.5 Conclusion

In this paper, the RL-based application was proposed to evaluate SOTM phased array depointing by utilizing UAVs. The conventional method is not applicable and there is currently no solution to measure depointing in SOTM scenario when AUT is a phased array due to the change of the pattern. The proposed method guides the central agent to the boresight angle and existing depointing can be estimated by either constructing a mathematical model or a UAV position. The former could estimate the depointing accurately if there were more than 5 sensors from the numerical simulation results. The latter is a more robust method if the number of sensors is limited. The proposed structure of the RL system does not require an explicit estimator, unlike the standard POMDP structure. By using this structure and utilizing an imaginary agent, the minimum number of UAVs is effectively reduced without compromising measurement accuracy.

In the training and evaluation, a single configuration of a phased array antenna model was simulated. In our future work, we will increase the types of AUTs to evaluate the method in a wider scope of phased array antennas and investigate whether the system can perform in more general environments. Also, the dynamic model of UAVs was not considered in this work in detail. A development that takes the dynamics of UAVs into consideration would be needed to apply this system in a real-world application. Finally, the result showed that the formation of the UAVs affected the performance. Thus, further development to optimize the formation should be continued. The formation might be dynamic because it might have a relation to the shape of the beam and the beam shapes of phased array antennas dynamically change.

4.6 Appendix

4.6.1 Learning trajectory

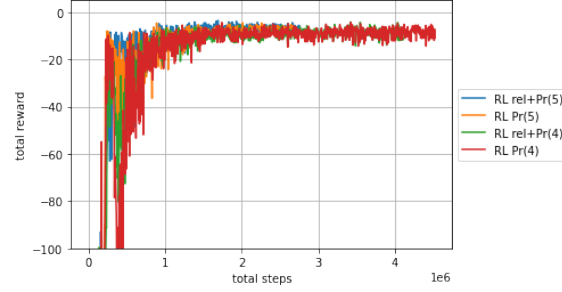


Figure 4.11: Learning trajectory of without imaginary agents

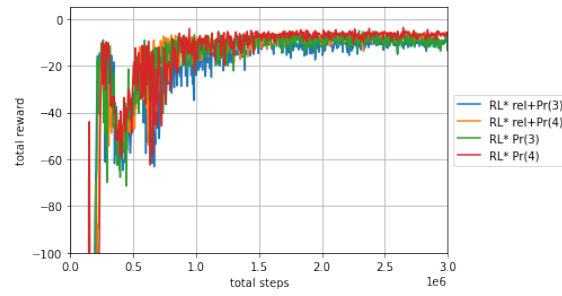


Figure 4.12: Learning trajectory of imaginary agents

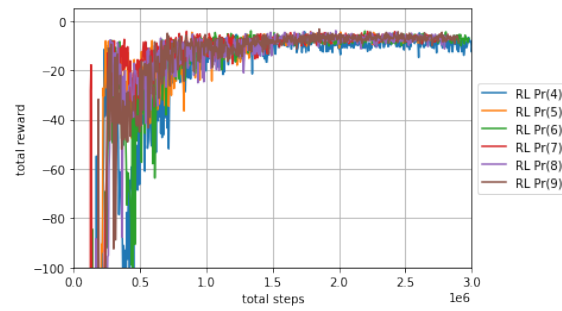


Figure 4.13: Learning trajectory of various number agents

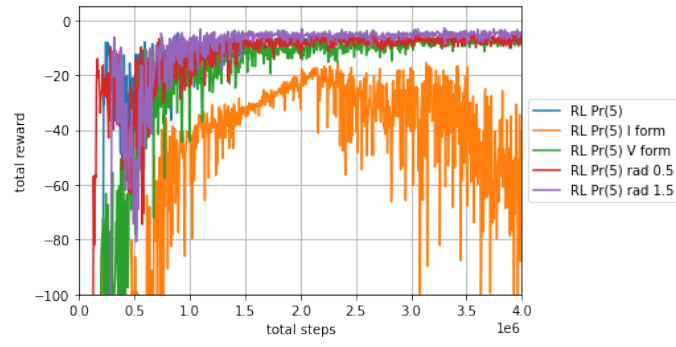


Figure 4.14: Learning trajectory of various formation

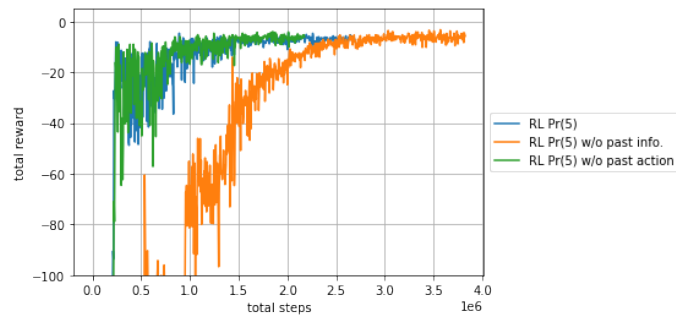


Figure 4.15: Learning trajectory of algorithm without past information

Chapter 5

Overall Discussion

Overall discussion on the contributions contained in Chapter 2, Chapter 3 and Chapter 4 is found in this chapter in terms of the objectives, hypothesis, and aim of the thesis stated in Chapter 1. Future work, possible extended scenarios, and implementation plan are also discussed.

5.1 Accomplishments of Objectives

Objective1: Develop a deep-RL algorithm with improved robustness of optimality which can dynamically adapt to the POMDP environment under the presence of non-static disturbance and which provides an end-to-end solution

In Chapter 2, a deep-RL was developed to pursue the robustness of optimality in POMDP environment. The challenge of POMDP is that the agent has no access to the true state of the system. We considered the scenarios where the observation was disturbed by adding stochastic noise or dynamic noise to the observation, or by hiding one of the elements of the observation. To cope with this situation, the necessity of identifying the causal effect of action and observation and recognizing the pattern of the transition of the environment was addressed. In the standard framework of POMDP, the belief state is updated via Bayesian inference where action contributes to the belief state update. In model-free RL, the model is internally constructed in the neural network. Therefore, it was hypothesized that including sequential action information as a part of the input would help improve the robustness of optimality. The LSTM network implemented in the neural network architecture processes the sequential information and generates internal state representations. The disturbances were analyzed into two categories based on whether the disturbance had a temporal correlation with one another. It was discussed that the role of LSTM was different depending on this category. If there is a temporal correlation in the disturbance, the neural network would model the transition function including the dynamics of the disturbance, whereas LSTM would learn to exclude the disturbance from the observation if it does not exist. The hypothesis regarding the effect of the inclusion of the past action sequence was confirmed in the classical control "Pendulum" environment under the presence of different types of disturbances by using an existing algorithm, called LSTM-TD3. LSTM-TD3 had a network structure to treat the current information and past sequential information separately.

Hence, it was proposed to modify LSTM-TD3 structure to process the information from the past to the current in one stream so that the flow of processing information follows the belief state update. With this structure, we also proposed another algorithm where the LSTM network in the Critic network takes the generated hidden state from the behavior actor as its initial hidden state and cell state to decrease the time of training by avoiding to repeat the trajectory in several networks in TD3. The numerical simulation result showed that the robustness of optimality was improved by modifying the structure and the computational time was drastically decreased by sharing the hidden state from the behaviour actor. The generalizability of the algorithms was also discussed. Although the type of disturbance was different between the training environment and the evaluation environment, the agent could still perform when the disturbances in both environments were in the same category. This work contributes to enhance the generalizability of general RL algorithms trained in a simulation environment toward the implementation in real-world.

Objective2: Develop a boresight angle localization application using the developed RL algorithm

In Chapter 3, the application to localize the boresight angle was developed. Firstly, the range of beamwidth of VSAT antenna and the effect of the positioning noise in angles at different distances between the AUT and the UAV were analyzed for the existing technology. The meta-RL concept was introduced since the application was intended to perform in the environments covering entirety of the range of the beam shape of VSAT terminal antenna. When multiple MDP exists, it also requires recognizing the MDP pattern based on the causality between action and observation. The application comprised two components. The first component was a guidance algorithm to drive the UAV to the boresight angle for desired accuracy. Another component was in charge of the numerical estimation of the boresight angle. One of the developed algorithms, LSTM-TD3_{1ha1hc} developed in Chapter 2 was utilized to guide the UAV to the boresight angle on the two-dimensional surface. The data, which can be obtained from the sensors, include the received signal strength and the position on GPS coordinates in the currently available measurement system in general. To implement the RL algorithm, the problem was analyzed and found that these measured values could not be used directly, otherwise, the state space became massive and the learning would become difficult. To resolve this issue, it was proposed to utilize the differentials of the obtainable information. As the action was defined to be the displacement from the current position, the differential of the position was already a part of the input element of LSTM-TD3_{1ha1hc}. Hence, the observation in the RL algorithm was defined as only the differential of the received signal strength. Based on the definition of the observation, state, reward, and termination conditions were defined for the RL settings. Even though RL settings were defined to reduce the size of the state space, the range of the beam shape addressed is quite large, and the agent would struggle to cover the entire range. To mitigate this challenge, a scaling function was introduced in the sequence of the observation. With this method, the range of the observation was fixed while the ability to identify the pattern's characteristics was still maintained, based on the sequential action and the scaled sequential observation. For the second component of the application, it was proposed to localize the boresight angle by using Gaussian Pro-

cess (GP) regression based on the collected data along the trajectory generated by the proposed guidance algorithm. The mathematical modeling of the radiation pattern to estimate the boresight was critical in order to improve the robustness of the estimation accuracy. The proposed guidance algorithm was evaluated by comparing it to the gradient ascent method in the numerical simulations. The success criteria of the mission was defined to reach the boresight angle within the accuracy of 0.05° to meet the industrial requirement. The algorithms were evaluated in terms of success rate, total distance in successful cases, and standard deviations of the total distance in successful cases. Overall, the proposed guidance algorithms demonstrated a success rate of over 90%, while the success rate of the gradient ascent was much lower and it was highly dependent on the parameters. The simulation results were closely analyzed by changing positioning noise, SNR, beamwidth, and the initial angular distance between the UAV and the boresight angle. The analysis highlighted that the developed guidance algorithm with scaling observation sequence could provide 100% success rate with efficient and robust trajectories in any shape of the beam as long as its initial position was not at the edge between the main beam and the side lobes while the developed algorithms without scaling function showed lower success rates when the beams were wide. The developed guidance algorithms were also tested in empirically obtained radiation patterns. The results proved the proposed method was applicable to realistic radiation patterns and particularly the algorithm with scaling showed a 100% success rate again in this environment. Also, the localization method was evaluated by comparing the polynomial optimization method and a method simply taking the angular position where the maximum signal strength was observed on the trajectory. The results confirmed that GP regression method was the most suitable approach in this scenario because the polynomial optimization method assumes IID condition which is not assured by the developed guidance algorithm, although it could potentially obtain higher accuracy. Finally, the overall performance of the proposed application was compared with the Bayesian Optimization method and showed that the trajectory was much more efficient. The final estimation error from the proposed application was around 0.004° which greatly satisfies the required measurement accuracy of 0.05° defined in Chapter 1.

Objective3: Develop a depointing measurement application for SOTM phased array antenna using the developed RL algorithm

In Chapter 4, the application of depointing measurement for SOTM phased array antenna has been developed. The purpose of this application is to evaluate whether the AUT can keep tracking the target direction when it is on the move. The work aimed to develop a method that can estimate the boresight angle with the minimum number of UAVs while satisfying the estimation accuracy requirement of 0.05° . The radiation patterns of phased array antennas vary depending on the steering angle. The trend of the pattern transition is that the beamwidth becomes wider and the scan loss gets larger as the steering angle is increased gradually. In the case of reflector antennas of which the radiation pattern is static, depointing can be measured by taking the correlation between measured signal strength from the sensors and reference radiation pattern. However, this approach is applicable only when the radiation pattern is static. As the radiation patterns of phased array antennas change, it is not feasible to collect refer-

ence data for all possible radiation patterns and utilize them to measure depointing. Also, since the specific measurement model is not available and the angular velocity of boresight angle is not observable, the estimator, that performs stochastic recursive updates, e.g., Kalman filter, is not available. One of the advantages of UAV measurement is that the UAVs with RF payload can dynamically change their positions during the measurement. By taking advantage of this aspect, novel two methods to estimate the depointing were proposed. The first estimation method, called "polynomial estimation", estimates the depointing by constructing a polynomial model based on the measured signal strength and estimating the boresight angle based on the obtained model. The second method, named "positional estimation", takes the ideal position of one of the RL agents to estimate the boresight angle. In both cases, it was identified that the objective of guidance should be placing one of the agents at the boresight angle to improve the estimation accuracy. The POMDP framework was also discussed since the standard POMDP framework was not available, two other frameworks were configured. As such, a guidance algorithm was designed to guide the UAVs in a formation to the dynamically moving boresight angle by utilizing LSTM-TD3_{1halhc} developed in Chapter 2. The roles of the UAVs in fixed formation were that the central agent was in charge of tracking the boresight angle and data was collected by subset agents, which were equally distributed around the central agent at a specific distance from the central agent. The observation of the RL element in the proposed application required only received signal strength which was indirectly interpreted to the UAV position in the AUT coordinate system. The imaginary agent was introduced to reduce the number of required UAVs. It was hypothesized that the measurement from the agent in the center did not provide informative information because the radiation pattern at the beam center is less distinctive. In the numerical simulation, the performance was evaluated in terms of accuracy and trajectory stability. The result indicated that the proposed method could predict the future state of the depointing angle from the sequential information and place the agents at the accurately predicted position. The result also revealed that the minimum required number of the subset UAVs was three to meet the required accuracy and the performance converged utilizing after utilizing more than four subset UAVs. By comparing the performance of the system with and without an imaginary agent, the result represented that the performance was comparable if the number of the subset UAVs was the same. The simulation result showed that the system robustly met the required accuracy even when positioning noise was added. In the evaluation, the framework was also contrasted and it was addressed that utilizing an external estimator degrades the entire performance when the polynomial estimation became unreliable. Also, the effect of the inclusion of the past action and past information (action and observation) was investigated and the result represented that the positioning estimation accuracy was better when past information was utilized and the trajectory was more stable which confirmed the discussion addressed in Chapter 2 in a totally different environment.

5.2 Response to Hypothesis

The thesis was developed on the hypothesis stated in Chapter 1 which was:

Accurate and robust UAV antenna evaluation applications can be designed by utilizing an algorithm which performs in a POMDP environment under the presence of disturbances without the specific knowledge of models of AUT.

The hypothesis was proven in Chapter 2, Chapter 3 and Chapter 4. In Chapter 2, deep-RL algorithms were discussed and developed to cope with POMDP environment with disturbance by adding past action sequences as input of the neural network. By analyzing the causal effect of the available information, the developed algorithms recognize the pattern of the environment and dynamically adapt its policy. Although the RL algorithm was developed in a classic control problem, it was applicable for the guidance problem in different environments since both require decision making in dynamic POMDP environments.

The objective of the RL algorithm in boresight localization application developed in Chapter 3, was to maximize the difference of received signal strength between the first step and the final step in minimum steps. The observed signal was disturbed by stochastic noise, the effect of which is exaggerated when the sampled signal strength is weaker at the lower gain of the AUT if it was measured in dB units. In this environment, the scale of the noise to the observation varies during the episode. Also, it was necessary to learn to perform a large number of MDPs generated by changing meta-parameters of the environment. In meta-RL, the environment does not change during the episode but varies between the episodes. Hence, the developed RL algorithm learned to identify the pattern of MDP and excluded the effect of noise from the sequential information to produce the action to achieve the objective.

On the other hand, the application of depointing measurement in Chapter 4 required the RL algorithm to identify the transition of radiation pattern and boresight angle and predict the depointing angle of the next timestep. The pattern transition is caused by a dynamically changing steering angle. Therefore, it has the same characteristic as when the disturbance has temporal correlations. Although this transition is not a disturbance, the RL algorithm learns to model the transition of the radiation pattern. Also, the observation was disturbed by RF noise and also the angular velocity term of the system was hidden. Hence, it could be also seen that the disturbance of this environment was the combination of "noise" and "hidden" from the scenarios tested in Chapter 2.

Both applications commonly require the recognition of dynamic patterns to execute their respective missions, the RL algorithm was developed to perform in dynamic POMDP environments. Therefore, the developed RL algorithm could address the challenges of these applications.

One might wonder if the boresight angle localization application is applicable to the depointing measurement application. However, this would not be feasible as the former algorithm was trained to identify the MDP and does not anticipate a dynamic environment in which both the pattern and the boresight angle change during the measurement. Furthermore, to identify the movement of the boresight angle from the transition of the radiation pattern, multiple samples from different positions are required. Oppositely, it would be possible to utilize the depointing measurement application for boresight localization as it can be considered as the special case of depointing measurement where velocity and acceleration are zero. However, there are differences in

their conditions. In the depointing measurement, the range of received power can be expected because it can utilize the information of antenna patterns of AUT, whereas it cannot in the boresight localization application since it is assumed that there is no information on the AUT based on the measurement. This was the reason why the differentials of received signal strength were used as observation in the boresight localization application. The considered condition about pre-known information is different and additional development would be needed to utilize the depointing measurement application for the boresight localization application.

Also, the observation and the reward of the boresight angle localization application were defined using the differentials of measured signal strength, whereas these elements of the depointing angle application were set as directly measured signal strength. The reason why differentials were used in the former application was that it was not possible to define the expected range of the received signal strength during the measurement without knowing the radiation pattern of AUT. In contrast, the latter application assumed that the radiation pattern of AUT was known at steering angle 0° and measured signal strength can be internally processed to normalized value based on this knowledge. Hence, the range of observation can be operated to some extent. The rewards were defined to reflect the defined states. The former application maps the reward from the scaled sequence of differential transition of the signal strength. In the latter application, the received signal strength has direct interpretations regarding the position of the UAV in the AUT coordinate system. Therefore, the reward, defined as the distance to the target angle, can be mapped from the observation.

5.3 Response to Aim

The aim of the thesis was stated in Chapter 1 was as follows:

To extend knowledge and develop applications of autonomous UAV-based measurement system for satellite terminal antenna evaluation.

The thesis was motivated by the urgent need within the satellite industry to verify antennas, which have undergone drastic changes due to industry and technological transformation. As utilizing NGSO has become more popular and the number of SOTM applications is rapidly increasing, the risk of interference continues to soar. The currently required antenna measurement system is the solution to efficiently verify a multitude of AUTs and evaluate the risk of interference caused by mobile terminals utilizing NGSO satellite communication and SOTM. The UAV-based measurement can offer much more flexible scenarios and can be integrated into dynamic measurement, unlike the conventional test facilities. However, the current research trends in UAV-based measurement do not seem to be focusing on these aspects. The efficiency of the measurement can be enhanced by an autonomous system, which is a novel concept in antenna measurement. The autonomous system, which performs under the restrictions of flight time, UAV numbers and positioning error, is a popular topic in GNC, however, UAV-based measurement is a minor topic in this field. The thesis demonstrated the development of autonomous systems for UAV-based antenna measurement and its efficiency to solve practical problems. As will be discussed later in Sec. 5.5, the appli-

cations developed can be extended to other scenarios, including NGSO communication. From this point, the thesis contributed to extending the knowledge of UAV-based antenna measurement and to responding to the urgent needs of applications to the satellite industry.

5.4 Future Work from the objectives

Dynamic RL algorithm Toward the implementation of RL systems in the real world, as discussed in Chapter 2, it is important to increase the robustness of the trained system rather than focusing on increasing the fidelity of the simulator. Regardless of the effort to increase the fidelity of the simulator, it is not possible to perfectly model the real world, and discrepancies between the simulator and the real world would remain. The robust system with generalizability should be able to keep the performance when it is encountered unexpected conditions. Chapter 2 worked on improving the robustness of optimality and developed algorithms for the environment which contains non-static disturbances toward building such a robust system. The analyses showed that the trained system could perform in an environment with disturbances which had never been experienced during the training if both disturbances belonged to the same category. The application developed in Chapter 3 showed that the developed RL algorithm can perform in the environment where the scale of the stochastic noise dynamically changes during the episode and meta-parameters of the environment are different between episodes. The RL algorithm in the application developed in Chapter 4 was trained in the environment where the change of the environment was temporally correlated and the state of the system was not fully observable. Hence, it is hypothesized that RL systems trained in the environment where combined disturbances exist can simultaneously learn eliminating the effect of uninformative data and modeling the dynamics of the disturbances or the change of the environment from the sequential information. The work in Chapter 2 proposed the representative disturbances. This hypothesis should be tested in the environment under these combined disturbances. If the hypothesis holds, the general robustness of the RL systems would be improved.

Boresight angle localization The trajectory generation in the application in Chapter 3 only relied on RL-based policy which was intended to guide the UAV to near the boresight angle. The trajectory, after reaching the region of the boresight angle, could be optimized for efficiency, thereby enhancing the boresight angle localization accuracy. This could be realized by handing over the execution to another policy, by developing a strategy to trigger the transition. As discussed in Chapter 3, the policy could employ a raster scan to construct a polynomial model or utilize Bayesian Optimization. The following policy can be executed in much smaller search areas compared with the case in which they have to search the entire initial search domain. The estimation based on the polynomial model could provide better accuracy than Bayesian Optimization from the results. However, IID condition should be ensured to apply this method. A continued evaluation of the efficiency by configuring the number of samples and size of the search range is required together with the development of the triggering mechanism of the second policy.

Also, the developed application requires an initial condition where the UAV is placed on the beam. Ensuring this condition, however, requires further development. If there is no information about the radiation pattern, it might require searching the entire region equally since Bayesian Optimization would still miss the main beam when it is very narrow. For example, the search area could be split into a grid. The grid shape could be defined based on the expected first null beamwidth obtained from the diameter of the AUT and the operation frequency. By collecting the information equally from each grid, the cell where the main beam exists could be identified. However, the required trajectory could be drastically increased with this method. Antennas might have a general relationship between sidelobes and a main beam. It would be possible to construct a stochastic model incorporating the antenna theory to estimate the main beam position without visiting all cells.

Thirdly, a two-dimensional surface was considered during the development phase. The method could be expanded to provide guidance in three-dimensional space, allowing for the localization of the boresight to be completed when UAV reaches the evaluation range. As such, the continued antenna evaluation could be executed instantly. It would significantly reduce the time required for antenna evaluation processes. The effect of the UAV's positioning error on the measurement becomes smaller as the distance between the UAV and the AUT is increased. On the other hand, the traveling distance to cover the same angular range gets larger as the distance is increased. Extending to three-dimensional space would entail addressing the balance between uncertainty and trajectory efficiency.

Depointing measurement for phased array antenna The application developed in Chapter 4 was trained and evaluated in only one example of the phased array antenna. The application should be further evaluated with various phased array antennas and be tested if it can be a general solution for any phased array antenna SOTM depointing measurement. However, a large number of parameters to simulate a phased array antenna exists and it is difficult to get information to define the general industrial configuration or range of the parameters as the development in Chapter 3 since phased array antenna is emerging technology. Further extensive research on literature and increasing the understanding of phased array antenna using sophisticated antenna simulators might help the continued development.

Also, it was assumed that the UAV can be instantly moved to the guided waypoint in the development. If the depointing angle varies faster than AUT, the proposed method would not perform well. The evaluation including the appropriate UAV dynamic model and study on a dynamic model of boresight angle for phased array antenna are required in the future.

Finally, the current development was performed with fixed formation of the UAVs. However, the simulation result found that the accuracy of the measurement had a dependency on the formation. Thus, the formation should be optimized in continued research. Particularly, when the subset UAVs were placed on the different radii of circles, the performance varied depending on the radius. The optimal distance from the boresight angle may have a relationship with the beam shape and SNR as identified in [111]. Since the beam shape varies in a phased array antenna, the optimal formation would need to dynamically change.

The development of both applications assumed that the source of uncertainty was SNR to the signal and positioning error of the UAV. Further development might consider other error factors such as gimbal stability and reflection from the ground to evaluate the applications in a more realistic environment.

5.5 Extended scenarios of the applications

To discuss the industrial impact of the developed applications and new potentials of UAV-based measurement, a few examples of extended scenarios are introduced.

Partial sidelobe characterisation As described in Chapter 1, one of the evaluation criteria is sidelobe characteristic by introducing a regulatory mask. However, sampling data to construct radiation patterns can be time-consuming in general and collecting all possible radiation patterns for the phased array antenna is not feasible. The point which has the largest potential to violate the mask would be the peak of the first sidelobes. Hence, just revealing the characteristics of the first sidelobe peaks in comparison to the main beam peak would provide useful information as an initial investigation of interference. The boresight angle localization application in Chapter 3 can be also applied for sidelobe localization. By simultaneously operating multiple UAVs with the developed application, a new framework to evaluate the sidelobe characteristics is proposed. The framework for a phased array antenna could follow the procedure below:

1. Find the main beam and the first sidelobes to track.
2. Apply the proposed application to find the beam center of each beam
3. Record the power level and the angles of the beam centers
4. Slightly steer the beam
5. Repeat 2 to 4 till covering the operation steering angles

This new framework would drastically decrease the time for the evaluation of a phased array antenna and provides only critical information regarding the relationship between the main beam and sidelobes. This framework could be also applied to examine beam squint and grating lobe tracking which are also special characteristics of phased array antennas.

Evaluation of SOTM phased array antenna in NGSO communication The test scenario in SOTM depointing measurement application developed in Chapter 4 can be extended to include SOTM depointing measurement for NGSO communication. The considered scenario in Chapter 4 was that the target satellite of the AUT is a GEO satellite and the target direction was fixed during the measurement. On the contrary, the target direction constantly moves in NGSO communication. The application should be able to measure depointing without additional amendment to the algorithm to include the trajectory information of the NGSO satellite because the additional maneuver introduced by NGSO scenario to steering angle does not need to be differentiated from

experienced motion during the training and since the output from the guidance algorithm is the displacement from the current position, the application is not constrained by space. Furthermore, satellite handovers are required for terminals in NSGO. The application could be also extended to the scenario where a handover is executed by introducing another team of UAVs operated with this application. The first team measures the depointing before the handover execution and the second team awaits at the direction where the communication is transferred next.

Development of VSAT mobile terminal evaluation The VSAT mobile terminal evaluation has not been developed in this thesis. However, the developed method to define the RL environment in the boresight angle localization application in Chapter 4 should be applicable in VSAT scenario. The radiation pattern of the VSAT antennas is static and hence, the problem is simpler than in a phased array antenna case as the agent need not construct the transition model of the AUT pattern. The required accuracy of the measurement of 0.05° might be achievable with less number of UAVs compared with the case in which the AUT is a phased array antenna.

5.6 Implementation plan

The implementation plan towards autonomous pointing accuracy evaluation is discussed in this section. The implementation of the applications does not require any new hardware development since they were built based on the existing technology. Also, since the applications do not generate direct control input to the UAVs but generate waypoints, it is less likely to generate dangerous maneuvers to make them crash. Table 5.1 shows the milestones toward implementation. MS and PA in the table refer to milestone and phased array terminal. The milestones are defined by the type of terminals, the orbit of communicating satellite, and whether it is SOTM application. The

Table 5.1: Implementation plan

MS	Terminal type	Orbit	SOTM
MS1	VSAT/PA	GEO	No
MS2	VSAT	NSGO	No
MS3	VSAT	GEO	Yes
MS4	VSAT	NSGO	Yes
MS5	PA	NSGO	No
MS6	PA	GEO	Yes
MS7	PA	NSGO	Yes

required equipment is listed as follows:

- UAV / multiple-UAVs with RF payload
- AUTs

- GCS
- Motion emulator (SOTM only)

Additionally, the antenna simulator would help the continued development of phased array antenna scenarios in MS5-MS7. The GCS should be able to receive the data of the UAV position and measured signal strength and send the new waypoints to the UAV(s) in real time. The plan would be executed by accomplishing the stated future work at the related milestones. The proposed tasks for each milestone are described as follows:

MS1 The developed beam localization application is verified in MS1 for a static environment. In this milestone, the tasks could be decomposed as follows:

- Investigation of the effect of other uncertainty factors, e.g., gimbal stability and reflection from the ground in simulation and empirical experiments.
- Development of an algorithm to find the approximate position of the beam, if applicable from future work.
- Implementation of the boresight localization algorithm developed in Chapter 3.
- Evaluation of accuracy by comparing the results obtained from high-resolution raster scans.

MS2 In this milestone, the VSAT-NGSO scenario is tested. The tasks of this milestone would be:

- Extension of the algorithm for VSAT depointing measurement from the future work and its evaluation in a simulation environment.
- Analysis of the UAV to obtain a dynamic model.
- Feasibility test of the application in simulation environment considering the dynamic model and uncertainty factors from MS1.
- Implementation of the application and test in the real world.
- Evaluation of the results by comparing them with those obtained by using the method in FORTE [111] and simulating the NGSO trajectory with UAVs.

MS3 VSAT-GEO-SOTM scenario is tested in this milestone. The approach is distinguished from FORTE although the scenario is identical. The tasks for MS3 would be:

- Feasibility analysis of the application extended in MS2 for this scenario, considering the dynamic model and uncertainty factors in a simulation environment.
- Implementation of the application and test in the real world.
- Evaluation of the results by comparing them with those obtained using the method in FORTE.

MS4 VSAT-NSGO-SOTM scenario is tested in this milestone. The tasks for MS4 would be identical to MS3.

MS5 PA-NSGO scenario is tested in this milestone. The antenna simulator would support achieving the milestones from this point. The tasks could comprise as follows:

- Research to increase the knowledge of the standard phased array in the industry as mentioned in future work.
- Evaluation of the developed depointing measurement application in Chapter 4 in simulation environment considering the dynamic model, uncertainty factors, and configuration of the phased array antenna.
- Extension and evaluation of the developed depointing measurement application for general phased array antenna from the future work.
- Implementation of the application and test in the real world.
- Evaluation of the results, which could be obtained from the simultaneous reconstruction of the pattern. The data for this reconstruction should be obtained from a number of sensors. If this isn't available, a less precise evaluation could be conducted by measuring the signal strength at the target position and assessing its consensus with the simulated result from the simulator, using the estimated steering angle by the developed application.

MS6 PA-GEO-SOTM scenario is tested in this milestone. The tasks for MS6 could consist of:

- Evaluation of the developed and extended depointing measurement application in MS5 in simulation environment considering the dynamic model, uncertainty factors and configuration of the phased array antenna also from MS5.
- Preparation of standard simple motion profiles with various speeds such as circles and spirals for verification purposes.
- Implementation of the application and test with the simple motion profile in the real world.
- Evaluation of the results as MS5.
- Implementation of the application and test with the realistic motion profile in the real world.
- Evaluation of the results as MS5.

MS7 PA-NSGO-SOTM scenario is tested in this milestone. The suggested tasks are identical to the tasks in MS7.

Chapter 6

Conclusion

Due to the industrial and technological transformation in the satellite industry, the conventional approach to verifying antennas is almost obsolete. It is urgently required to develop new methods to address this issue. The core contribution of this thesis was extending the knowledge of UAV-based measurement to respond to these urgent needs, and to indicate possibilities to explore new aspects of antenna measurement systems that contain the GNC perspective by showcasing two applications. The efficiency, robustness, and accuracy can be improved by introducing autonomous systems. Autonomous UAV-based measurement, which flexibly adapts to emerging complex scenarios, would support the transition of the satellite industry and accelerate the deployment of new technology for a better communication environment.

In the thesis, the development of deep-RL algorithms was motivated based on the analysis of the environments of the considered applications. The deep-RL algorithm with improved robustness of optimality for POMDP was developed in Chapter 2 and it was effectively employed by the two applications for satellite antenna evaluation in Chapter 3 and Chapter 4. The targeted environments to apply the applications were where the specific models were not available and the observations were disturbed by noise. Therefore, the agent had to dynamically adapt to the environment due to the existence of multiple MDPs or dynamically changing patterns and boresight angles during the measurement. The developed applications demonstrated that they could solve each problem with the required accuracy in simulation environments.

The UAV-based antenna measurement allows more flexibility and control in measurement scenarios and conditions compared to conventional antenna measurement methods. Further developments of autonomous systems for UAV-based antenna measurement would make the verification more efficient and reliable.

References

1. Euroconsult. Ground Segment Market Prospects, 2nd edition A Euroconsult Report. Paris, France: Euroconsult; 2020.
2. Satellite Operator's Minimum Antenna Performance(SOMAP) Group, a GVF-MRA sub-group. SOMAP – Comms-On-The-Move Interference; 2019. Available at: https://gsoasatellite.com/wp-content/uploads/SOMAP-COTM-Interference_2019.pdf. Accessed on July 23rd, 2023.
3. Maral G, Bousquet M, Sun Z. Satellite communications systems: systems, techniques and technology. John Wiley & Sons; 2020.
4. GVF. Technology Working Group;. Available at: <https://gsoasatellite.com/working-groups/technology-working-group/>. Accessed on July 23rd, 2023.
5. GVF. SOMAP Matrix Ka-Band; 2019. Available at: https://gsoasatellite.com/wp-content/uploads/SOMAP-Matrix-Ka-Band_Aug-2019.pdf. (Accessed 25 July 2023).
6. eutelsat. Antenna and VSAT Type Approval/Characterization ESOG 120 – Issue 8 - Rev. 1; 2021. Available at: <https://www.eutelsat.com/files/contributed/satellites/pdf/esog120.pdf>. (Accessed 25 July 2023).
7. GVF. PERFORMANCE AND TEST GUIDELINES FOR TYPE APPROVAL OF 'COMMS ON THE MOVE' MOBILE SATELLITE COMMUNICATIONS TERMINALS GVF-105; 2016. Available at: <https://gsoasatellite.com/wp-content/uploads/gvf-105.pdf>. (Accessed 27 July 2023).
8. Fritzel T, Steiner H, Habersack J, Schippers H, Ferrante J. Antenna pattern measurements of full-size air vehicles with an airborne near-field test facility (ANTF). European Test & Telemetry Conference, ETTC; 2005. .
9. ITU. Airborne verification of antenna patterns of broadcasting stations: Report ITU-R SM.2056-1. ITU-International Telecommunication Union; 2014.
10. Manton R, Beeke K. HF antenna radiation patterns over real terrain. 1988 International Broadcasting Convention, IBC 1988. IET; 1988. p. 143-7.
11. Steele J. Measurement of antenna radiation patterns using a tethered balloon. IEEE Transactions on Antennas and Propagation. 1965;13(1):179-80.

12. Üstüner F, Aydemir E, Güleç E, İlarıslan M, Çelebi M, Demirel E. Antenna radiation pattern measurement using an unmanned aerial vehicle (UAV). 2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS). IEEE; 2014. p. 1-4.
13. Virone G, Lingua AM, Piras M, Cina A, Perini F, Monari J, et al. Antenna pattern verification system based on a micro unmanned aerial vehicle (UAV). IEEE Antennas and Wireless Propagation Letters. 2014;13:169-72.
14. Picar AM, Marqué C, Anciaux M, Lamy H, Ranvier S. Antenna pattern calibration of radio telescopes using an UAV-based device. 2015 International Conference on Electromagnetics in Advanced Applications (ICEAA). IEEE; 2015. p. 981-4.
15. Schreiber J. Antenna pattern reconstitution using unmanned aerial vehicles (UAVs). 2016 IEEE Conference on Antenna Measurements & Applications (CAMA). IEEE; 2016. p. 1-3.
16. Miller JM, Decrossas E. Using small unmanned aerial systems and helium aerostats for far-field radiation pattern measurements of high-frequency antennas. 2018 IEEE Conference on Antenna Measurements & Applications (CAMA). IEEE; 2018. p. 1-4.
17. Culotta-López C, Gregson S, Buchi A, Rizzo C, Trifon D, Skeidsvoll S, et al. On the uncertainty sources of drone-based outdoor far-field antenna measurements. 2021 Antenna Measurement Techniques Association Symposium (AMTA). IEEE; 2021. p. 1-6.
18. Culotta-López C, Skeidsvoll S, Buchi A, Espeland J. Environmental reflections diagnostics through radius flights using uass. 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE; 2022. p. 1-4.
19. Geise R, Vaske A, Fritzel T, Strauß R. Indoor Antenna Calibration with UAVs—On Time Domain Gating and the Influence of Motion. 2023 17th European Conference on Antennas and Propagation (EuCAP). IEEE; 2023. p. 1-5.
20. Steiner HJ, Fritzel T, Strauß R. Aspects of near-field antenna measurement technology when using UAV's. 2021 15th European Conference on Antennas and Propagation (EuCAP). IEEE; 2021. p. 1-5.
21. Virone G, Paonessa F, Ciorba L, Lumia M, Addamo G, Peverini O, et al. Effect of conductive propellers on VHF UAV-based antenna measurements: Simulated results. 2021 15th European Conference on Antennas and Propagation (EuCAP). IEEE; 2021. p. 1-4.
22. Henkel P, Lamm M, Mittmann U, Fritzel T, Strauß R, Steinert HJ, et al. Verification of RTK Positioning of UAVs with High-Precision Laser Tracker. 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE; 2022. p. 1-5.
23. Salazar-Cerreno JL, Jehangir SS, Aboserwal N, Segales A, Qamar Z. An uav-based polarimetric antenna measurements for radar and communication systems from

- 3 ghz to 32 ghz. 2021 IEEE Conference on Antenna Measurements & Applications (CAMA). IEEE; 2021. p. 55-60.
24. Culotta-López C, Skeidsvoll S, Costin V, Buchi A, Espeland J. On gain measurements using the substitution method with UASs. 2021 IEEE Conference on Antenna Measurements & Applications (CAMA). IEEE; 2021. p. 51-4.
 25. Geise A, Fritzel T, Paquay M. Ka-band measurement results of the irregular near-field scanning system PAMS. 2017 Antenna Measurement Techniques Association Symposium (AMTA). IEEE; 2017. p. 1-6.
 26. Fritzel T, Strauß R, Steiner HJ, Eisner C, Eibert T. Introduction into an UAV-based near-field system for in-situ and large-scale antenna measurements. 2016 IEEE Conference on Antenna Measurements & Applications (CAMA). IEEE; 2016. p. 1-3.
 27. Punzet S, Faul FT, Mittereder T, Oetl C, Ganser M, Häusler M, et al. Fully coherent UAV-based near-field measurement and transformation of the s67-15 m ground station antenna at the german space operations center in weilheim. 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE; 2022. p. 1-5.
 28. García-Fernández M, López YÁ, Arboleya A, González-Valdés B, Rodríguez-Vaqueiro Y, Gómez MEDC, et al. Antenna diagnostics and characterization using unmanned aerial vehicles. IEEE Access. 2017;5:23563-75.
 29. Faul FT, Kornprobst J, Fritzel T, Steiner HJ, Strauß R, Weiß A, et al. Near-field measurement of continuously modulated fields employing the time-harmonic near-to far-field transformation. Advances in Radio Science. 2019;17:83-9.
 30. Gueuning Q, Craeye C, de Lera Acedo E, Brown AK. Improving the accuracy of a model-based approach for the near-field measurement of antenna arrays with UAVs. 2023 17th European Conference on Antennas and Propagation (EuCAP). IEEE; 2023. p. 1-3.
 31. Paulus A, Knapp J, Eibert TF. Phaseless near-field far-field transformation utilizing combinations of probe signals. IEEE Transactions on Antennas and Propagation. 2017;65(10):5492-502.
 32. Garcia-Fernandez M, Lopez YA, Andres FLH. Unmanned aerial system for antenna measurement and diagnosis: evaluation and testing. IET Microwaves, Antennas & Propagation. 2019;13(13):2224-31.
 33. Virone G, Matteoli S, Paonessa F, Bolli P, Wijnholds S, Ciorba L, et al. Preliminary results on the verification of the LOFAR-HBA with a flying test source. 2021 15th European Conference on Antennas and Propagation (EuCAP). IEEE; 2021. p. 1-4.
 34. de Lera Acedo E, Bolli P, Paonessa F, Virone G, Colin-Beltran E, Razavi-Ghods N, et al. SKA aperture array verification system: electromagnetic modeling and beam pattern measurements using a micro UAV. Experimental Astronomy. 2018;45:1-20.

35. Dünner R, Fluxá J, Best S, Carrero F, Boettger D. Drone-based polarization calibration source for mm-wave telescopes. 2021 15th European Conference on Antennas and Propagation (EuCAP). IEEE; 2021. p. 1-5.
36. Henault S. Verification of a Multi-Beam HF Radar-Embedded Communication System Using a Drone. 2023 17th European Conference on Antennas and Propagation (EuCAP). IEEE; 2023. p. 1-5.
37. Henault S. OTHR phase-only beamformer drone-based measurements. 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE; 2022. p. 1-5.
38. Horsmanheimo S, Tuomimäki L, Semkin V, Mehnert S, Chen T, Ojennus M, et al. 5G communication QoS measurements for smart city UAV services. 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE; 2022. p. 1-5.
39. Heikkilä M, Malkamo V, Eteläaho P, Kippola T, Koskela M. Latency Validation Method for 3D 5G Networks' URLLC Applications. 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE; 2022. p. 1-5.
40. Dewan R, Rahman KE. A Survey on Applications of Unmanned Aerial Vehicles (UAVs). Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 2. Springer; 2022. p. 95-110.
41. Skorobogatov G, Barrado C, Salamí E. Multiple UAV systems: A survey. Unmanned Systems. 2020;8(02):149-69.
42. Liu J, Shi Y, Fadlullah ZM, Kato N. Space-air-ground integrated network: A survey. IEEE Communications Surveys & Tutorials. 2018;20(4):2714-41.
43. Swinney CJ, Woods JC. RF detection and classification of unmanned aerial vehicles in environments with wireless interference. 2021 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE; 2021. p. 1494-8.
44. Ezuma M, Erden F, Anjinappa CK, Ozdemir O, Guvenc I. Detection and classification of UAVs using RF fingerprints in the presence of Wi-Fi and Bluetooth interference. IEEE Open Journal of the Communications Society. 2019;1:60-76.
45. Xiao Y, Zhang X. Micro-UAV detection and identification based on radio frequency signature. 2019 6th international conference on systems and informatics (ICSAI). IEEE; 2019. p. 1056-62.
46. Shakhathreh H, Sawalmeh AH, Al-Fuqaha A, Dou Z, Almaita E, Khalil I, et al. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. Ieee Access. 2019;7:48572-634.
47. GVF. SOMAP Matrix Ku-Band; 2019. Available at: https://gsoasatellite.com/wp-content/uploads/SOMAP-Matrix-Ku-Band_Aug-2019.pdf. (Accessed 25 July 2023).

48. OpenAI. OpenAI Gym;. Available at: <https://gymnasium.farama.org/>. Accessed on January 14, 2023.
49. Kurniawati H. Partially observable markov decision processes (pomdps) and robotics. arXiv preprint arXiv:210707599. 2021.
50. Hernandez-Leal P, Kartal B, Taylor ME. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*. 2019;33(6):750-97.
51. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*. 1997;9(8):1735-80.
52. Chung J, Gulcehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. *International conference on machine learning*. PMLR; 2015. p. 2067-75.
53. Miki T, Lee J, Hwangbo J, Wellhausen L, Koltun V, Hutter M. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*. 2022;7(62):eabk2822.
54. Li X, Li L, Gao J, He X, Chen J, Deng L, et al. Recurrent reinforcement learning: a hybrid approach. arXiv preprint arXiv:150903044. 2015.
55. Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps. 2015 aaai fall symposium series; 2015. .
56. Singla A, Padakandla S, Bhatnagar S. Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge. *IEEE Transactions on Intelligent Transportation Systems*. 2019;22(1):107-18.
57. Janner M, Li Q, Levine S. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*. 2021;34:1273-86.
58. Chen L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, et al. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*. 2021;34:15084-97.
59. Hasani R, Lechner M, Amini A, Rus D, Grosu R. Liquid time-constant networks. *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35; 2021. p. 7657-66.
60. Shen B, Xia F, Li C, Martín-Martín R, Fan L, Wang G, et al. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2021. p. 7520-7.
61. Freeman CD, Frey E, Raichuk A, Girgin S, Mordatch I, Bachem O. Brax—A Differentiable Physics Engine for Large Scale Rigid Body Simulation. arXiv preprint arXiv:210613281. 2021.

62. Truong J, Rudolph M, Yokoyama N, Chernova S, Batra D, Rai A. Rethinking sim2real: Lower fidelity simulation leads to higher sim2real transfer in navigation. arXiv preprint arXiv:220710821. 2022.
63. Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.
64. Ravichandiran S. Hands-On Deep Learning Algorithms with Python: Master deep learning algorithms with extensive math by implementing them using TensorFlow. Packt Publishing Ltd; 2019.
65. Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods. International conference on machine learning. PMLR; 2018. p. 1587-96.
66. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:150902971. 2015.
67. Hasselt H. Double Q-learning. Advances in neural information processing systems. 2010;23.
68. Meng L, Gorbet R, Kulić D. Memory-based deep reinforcement learning for pomdps. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2021. p. 5619-26.
69. Pearl J, et al. Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress. 2000;19(2).
70. Bertsekas D. Dynamic programming and optimal control: Volume I. vol. 1. Athena scientific; 2012.
71. Hauskrecht M. Planning and control in stochastic domains with imperfect information. Massachusetts Institute of Technology; 1997.
72. Hauskrecht M. Value-function approximations for partially observable Markov decision processes. Journal of artificial intelligence research. 2000;13:33-94.
73. Gaudet B, Linares R. Adaptive guidance with reinforcement meta-learning. arXiv preprint arXiv:190104473. 2019.
74. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. International conference on machine learning. PMLR; 2016. p. 1928-37.
75. Heess N, Hunt JJ, Lillicrap TP, Silver D. Memory-based control with recurrent neural networks. arXiv preprint arXiv:151204455. 2015.
76. Bakker B. Reinforcement learning with long short-term memory. Advances in neural information processing systems. 2001;14.
77. Harmon ME, Baird III LC. Multi-player residual advantage learning with general function approximation. Wright Laboratory, WL/AACE, Wright-Patterson Air Force Base, OH. 1996:45433-7308.

78. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *nature*. 2015;518(7540):529-33.
79. Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. *International conference on machine learning*. PMLR; 2014. p. 387-95.
80. Malikopoulos AA. On Separation Between Learning and Control in Partially Observed Markov Decision Processes. *arXiv preprint arXiv:221114972*. 2022.
81. Gasse M, Grasset D, Gaudron G, Oudeyer PY. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:210614421*. 2021.
82. Chakrabarti S. How many satellites are orbiting Earth?; 2021. Available at: <https://www.space.com/how-many-satellites-are-orbiting-earth>. Accessed on May 29th, 2023.
83. GVF. Mutual Recognition of Performance Measurement Guidelines and Procedures for Satellite System Operator Type Approvals GVF-101 Rev E; 2014. Available at: <https://gsoasatellite.com/wp-content/uploads/gvf-101.pdf>. (Accessed 27 July 2023).
84. ITU. ITU-R S.728-1. Maximum permissible level of off-axis e.i.r.p. density from very small aperture terminals (VSATs). *International Telecommunication Union*; 2002.
85. GVF. SOMAP – SATELLITE OPERATOR’S MINIMUM ANTENNA PERFORMANCE; 2019. Available at: <https://gvf.org/working-groups/mutual-recognition-arrangement-working-group-mra-wg/somap-satellite-operators-minimum-antenna-performance/>. Accessed on May 23rd, 2023.
86. Chang C, Monstein C, Refregier A, Amara A, Glauser A, Casura S. Beam calibration of radio telescopes with drones. *Publications of the Astronomical Society of the Pacific*. 2015;127(957):1131-43.
87. Duthoit S, Salazar JL, Doyle W, Segales A, Wolf B, Fulton C, et al. A new approach for in-situ antenna characterization, radome inspection and radar calibration, using an unmanned aircraft system (UAS). *2017 IEEE Radar Conference (Radar-Conf)*. IEEE; 2017. p. 0669-74.
88. Steiner HJ, Strauß R, Lamm M, Henkel P. UAV-Based Near-Field Antenna Measurements at 20 MHz of a Reference Ground-Plane Antenna. *2021 IEEE Conference on Antenna Measurements & Applications (CAMA)*. IEEE; 2021. p. 48-50.
89. Heikkilä M, Koskela M, Kippola T, Kocak M, Erkkilä J, Tervonen J. Using unmanned aircraft systems for mobile network verifications. *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE; 2018. p. 805-11.

90. Microwave Journal. QuadSAT Emulates LEO/MEO Passes for Highly Dynamic Testing; 2023. Available at: <https://www.microwavejournal.com/articles/40286-quadsat-emulates-leo-meo-passes-for-highly-dynamic-testing>. (Accessed 21 July 2013).
91. Gawronski W, Craparo EM. Antenna scanning techniques for estimation of spacecraft position. *IEEE Antennas and Propagation Magazine*. 2002;44(6):38-45.
92. Cho G, Kim GT. Monopulse tracking performance of a satcom antenna on a moving platform. *Journal of electromagnetic engineering and science*. 2017;17(3):120-5.
93. Owais M, Midtiby H, Trifon D, Hasan A. Autonomous Drone-Based Antenna Radiation Pattern Characterization. 2022 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE; 2022. p. 207-13.
94. Omi S, Shin HS, Cho N, Thourdos A. Dynamic deep-reinforcement-learning algorithm in Partially Observed Markov Decision Processes. *arXiv preprint arXiv:arXiv:230715931*. 2023.
95. Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, et al. Learning to reinforcement learn. *arXiv preprint arXiv:161105763*. 2016.
96. Duan Y, Schulman J, Chen X, Bartlett PL, Sutskever I, Abbeel P. RL²: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:161102779*. 2016.
97. Goodfellow I, Bengio Y, Courville A. *Deep learning*. MIT press; 2016.
98. Frazier PI. A tutorial on Bayesian optimization. *arXiv preprint arXiv:180702811*. 2018.
99. Orfanidis S. *Electromagnetic waves and antennas*; sophocles j. Orfanidis: Roma, Italy. 1999.
100. Nocedal J, Wright SJ. *Numerical optimization*. Springer; 1999.
101. Martinez-Cantin R, De Freitas N, Brochu E, Castellanos J, Doucet A. A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots*. 2009;27:93-103.
102. Pawar AA, Warbhe U. Optimizing Bayesian acquisition functions in Gaussian Processes. *arXiv preprint arXiv:211104930*. 2021.
103. authors TG. GPyOpt: A Bayesian Optimization framework in python; 2016. <http://github.com/SheffieldML/GPyOpt>. Accessed on May 23rd, 2023.
104. Catalani A, Toso G, Angeletti P, Albertini M, Russo P. Development of enabling technologies for Ku-band airborne SATCOM phased-arrays. *Electronics*. 2020;9(3):488.

105. Shen X, Chang R, Yuan D. Pseudo-monopulse tracking method for low profile mobile satellite antenna system. *AEU-International Journal of Electronics and Communications*. 2019;101:160-7.
106. Nistal-González I, Otto S, Litschke O, Bettray A, Wunderlich L, Gieron R, et al. Planar phased array antenna for nomadic satellite communication in Ka-band. 2014 11th European Radar Conference. IEEE; 2014. p. 396-9.
107. microwave journal. 93% of industry suffers from satellite interference; September 12, 2013. Available at: <https://www.microwavejournal.com/articles/20554-of-industry-suffers-from-satellite-interference>. Accessed on June 21st, 2023.
108. Federal Communications Commission. 47 CFR 25.115 (last amendment 24 July 2023); 2023. Available at: <https://www.ecfr.gov/current/title-47/section-25.114>. (Accessed 27 July 2023).
109. International telecommunication Union. Recommendation S.1857-0 Methodologies to estimate the off-axis e.i.r.p. density levels and to assess the interference towards adjacent satellites resulting from pointing errors of vehicle mounted earth stations in the 14 GHz frequency band; 2010. Available at: <https://www.itu.int/rec/R-REC-S.1857/en>. (Accessed 27 July 2023).
110. European Telecommunications Standards Institute. ETSI EN 302 977 V2.1.1 Satellite Earth Stations and Systems (SES); Harmonised Standard for Vehicle-Mounted Earth Stations (VMES) operating in the 14/12 GHz frequency bands covering the essential requirements of article 3.2 of the Directive 2014/53/EU ; 2016. Available at: https://www.etsi.org/deliver/etsi_en/302900_302999/302977/02.01.01_60/en_302977v020101p.pdf. (Accessed 27 July 2023).
111. Elkhoully A. Standardized testing conditions for satellite communications on-the-move (SOTM) terminals. PhD Thesis. TECHNISCHE UNIVERSITÄT ILMENAU; 2019.
112. Fritzel T, Steiner HJ, Strauß R. Advances in the development of an industrial UAV for large-scale near-field antenna measurements. 2019 13th European Conference on Antennas and Propagation (EuCAP). IEEE; 2019. p. 1-3.
113. Culotta-López C, Skeidsvoll S, Buchi A, Espeland J. Use of UASs for Outdoor Diagnostics of Large Antennas. 2022 Antenna Measurement Techniques Association Symposium (AMTA). IEEE; 2022. p. 1-4.
114. Chung SJ, Paranjape AA, Dames P, Shen S, Kumar V. A survey on aerial swarm robotics. *IEEE Transactions on Robotics*. 2018;34(4):837-55.
115. Mailloux RJ. Phased array antenna handbook. Artech house; 2017.
116. Ahn B, Hwang IJ, Kim KS, Chae SC, Yu JW, Lee HL. Wide-angle scanning phased array antenna using high gain pattern reconfigurable antenna elements. *Scientific reports*. 2019;9(1):18391.

117. Kavya K, Kotamraju SK, Kumar BN, Mounika M, Singh S, SIDDA A. Beam pointing accuracy of phased arrays for satellite communication. *Journal of Theoretical and Applied Information Technology*. 2017;95(10):2170-81.
118. Masters GF, Gregson SF. Coordinate system plotting for antenna measurements. *AMTA Annual Meeting & Symposium*. vol. 32; 2007. .
119. Silver D. Lectures on Reinforcement Learning; 2015. Available at: <https://www.davidsilver.uk/teaching/>. Accessed on June 21st, 2023.

Appendix A

Ethical Approval Letter



16 September 2020

Dear Ms Omi ,

Reference: CURES/12217/2020

Title: Development of the performance evaluation of SOTM satellite antenna systems using an autonomous vehicle swarm

Thank you for your application to the Cranfield University Research Ethics System (CURES).

We are pleased to inform you your CURES application, reference CURES/12217/2020 has been reviewed. You may now proceed with the research activities you have sought approval for.

If you have any queries, please contact CURES Support.

We wish you every success with your project.

Regards,

CURES Team