CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

PhD Thesis

Deep Reinforcement Learning for Joint Cruise Control and Intelligent Data Acquisition in UAVs-Assisted Sensor Networks

Yousef Emami

CISTER-TR-231101

2023/11/08

Deep Reinforcement Learning for Joint Cruise Control and Intelligent Data Acquisition in UAVs-Assisted Sensor Networks

Yousef Emami

CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: emami@isep.ipp.pt

https://www.cister-labs.pt

Abstract

Unmanned aerial vehicle (UAV)-assisted sensor networks (UASNets), which play a crucial role in creating new opportunities, are experiencing significant growth in civil applications worldwide. UASNets provide a range of new functionalities for civilian sectors. Just as UASNets have revolutionized military operations with improved surveillance, precise targeting, and enhanced communication systems, they are now driving transformative change in numerous civilian sectors. For instance, UASNets improve disaster management through timely surveillance and advance precision agriculture with detailed crop monitoring, thereby significantly transforming the commercial economy. UASNets revolutionize the commercial sector by offering greater efficiency, safety, and cost-effectiveness, highlighting their transformative impact. A fundamental aspect of these new capabilities and changes is the collection of data from rugged and remote areas. Due to their excellent mobility and maneuverability, UAVs are employed to collect data from ground sensors in harsh environments, such as natural disaster monitoring, border surveillance, and emergency response monitoring. One major challenge in these scenarios is that the movements of UAVs affect channel conditions and result in packet loss. Fast movements of UAVs lead to poor channel conditions and rapid signal degradation, resulting in packet loss. On the other hand, slow mobility of a UAV can cause buffer overflows of the ground sensors, as newly arrived data is not promptly collected by the UAV.

Our proposal to address this challenge is to minimize packet loss by jointly optimizing the velocity controls and data collection schedules of multiple UAVs. The states of ground sensors include battery level, data queue length, and channel quality. In the absence of up-to-date knowledge of ground sensors 19 states, we propose a multi-UAV deep reinforcement learning-based scheduling algorithm (MADRL-SA). This algorithm allows UAVs to asymptotically minimize packet loss due to buffer overflows and poor channel conditions, even in the presence of outdated knowledge of the network states at individual UAVs.

Furthermore, in UASNets, swift movements of UAVs result in poor channel conditions and fast signal attenuation, leading to an extended age of information (AoI). In contrast, slow movements of UAVs prolong flight time, thereby extending the AoI of ground sensors. Additionally, the UAVs should consider the movements of other UAVs to minimize the average AoI by coordinating their velocities. Hence, finding an equilibrium solution among UAVs to optimize velocity and reduce the average AoI becomes crucial.

To address this challenge, we propose a new mean-field flight resource allocation optimization to minimize the AoI of sensory data. Balancing the trade-off between UAV movements and AoI is formulated as a mean-field game (MFG). We introduce a new mean-field hybrid proximal policy optimization (MF-HPPO) scheme to handle the expanded solution space of MFG optimization. This scheme minimizes the average AoI by optimizing the UAV trajectories and ground sensor data collection schedules, considering mixed continuous and discrete actions. Additionally, we incorporate a long short-term memory (LSTM) in MF-HPPO to predict the time-varying network state and stabilize the training.

Deep Reinforcement Learning for Joint Cruise Control and Intelligent Data Acquisition in UAVs-Assisted Sensor Networks

Yousef Emami

Supervisor: Dr. Kai Li

Co-Supervisor: Prof. Dr. Eduardo Tovar

Co-Supervisor: Prof. Dr. Mario Sousa

Programa Doutoral em Engenharia Electrotécnica e de Computadores

February 2023

Faculdade de Engenharia da Universidade do Porto

Deep Reinforcement Learning for Joint Cruise Control and Intelligent Data Acquisition in UAVs-Assisted Sensor Networks

Yousef Emami

Dissertation submitted to Faculdade de Engenharia da Universidade do Porto
to obtain the degree of

Doctor Philosophiae in Electronic & Computer Engineering

President: Prof. Dr. Nuno Fidalgo
Referee: Prof. Dr. Xiaoming Fu
Referee: Prof. Dr. Nuno Lau
Referee: Prof. Dr. Rosario Pinho
Referee: Dr. Pedro Santos
Supervisor: Dr. Kai Li

February 2023

To my loving parents,
who have always supported and encouraged me
throughout my academic journey.

Abstract

Unmanned aerial vehicle (UAV)-assisted sensor networks (UASNets), which play a crucial role in creating new opportunities, are experiencing significant growth in civil applications worldwide. UASNets provide a range of new functionalities for civilian sectors. Just as UASNets have revolutionized military operations with improved surveillance, precise targeting, and enhanced communication systems, they are now driving transformative change in numerous civilian sectors. For instance, UASNets improve disaster management through timely surveillance and advance precision agriculture with detailed crop monitoring, thereby significantly transforming the commercial economy. UASNets revolutionize the commercial sector by offering greater efficiency, safety, and cost-effectiveness, highlighting their transformative impact. A fundamental aspect of these new capabilities and changes is the collection of data from rugged and remote areas. Due to their excellent mobility and maneuverability, UAVs are employed to collect data from ground sensors in harsh environments, such as natural disaster monitoring, border surveillance, and emergency response monitoring. One major challenge in these scenarios is that the movements of UAVs affect channel conditions and result in packet loss. Fast movements of UAVs lead to poor channel conditions and rapid signal degradation, resulting in packet loss. On the other hand, slow mobility of a UAV can cause buffer overflows of the ground sensors, as newly arrived data is not promptly collected by the UAV.

Our proposal to address this challenge is to minimize packet loss by jointly optimizing the velocity controls and data collection schedules of multiple UAVs. The states of ground sensors include battery level, data queue length, and channel quality. In the absence of up-to-date knowledge of ground sensors' states, we propose a multi-UAV deep reinforcement learning-based scheduling algorithm (MADRL-SA). This algorithm allows UAVs to asymptotically minimize packet loss due to buffer overflows and poor channel conditions, even in the presence of outdated knowledge of the network states at individual UAVs.

Furthermore, in UASNets, swift movements of UAVs result in poor channel conditions and fast signal attenuation, leading to an extended age of information (AoI). In contrast, slow movements of UAVs prolong flight time, thereby extending the AoI of ground sensors. Additionally, the UAVs should consider the movements of other UAVs to minimize the average AoI by coordinating their velocities. Hence, finding an equilibrium solution among UAVs to optimize velocity and reduce the average AoI becomes crucial.

To address this challenge, we propose a new mean-field flight resource allocation optimization to minimize the AoI of sensory data. Balancing the trade-off between UAV movements and AoI is formulated as a mean-field game (MFG). We introduce a new mean-field hybrid proximal policy optimization (MF-HPPO) scheme to handle the ex-

panded solution space of MFG optimization. This scheme minimizes the average AoI by optimizing the UAV trajectories and ground sensor data collection schedules, considering mixed continuous and discrete actions. Additionally, we incorporate a long short-term memory (LSTM) in MF-HPPO to predict the time-varying network state and stabilize the training.

Keywords: UAVs, Mean-field game, Age of information, Proximal policy optimization, Long short term memory, Communication scheduling, Velocity control, Deep Q-Network.

Resumo

As redes de sensores assistidas por veículos aéreos não tripulados (UAV) (UASNets), que desempenham um papel crucial na criação de novas oportunidades, estão a experimentar um crescimento significativo em aplicações civis em todo o mundo. As UASNets fornecem uma gama de novas funcionalidades para setores civis. Assim como os UASNets revolucionaram as operações militares com vigilância aprimorada, direcionamento preciso e sistemas de comunicação aprimorados, elas agora estão a conduzir mudanças transformadoras em vários setores civis. Por exemplo, as UASNets melhoram a gestão de desastres por meio de vigilância oportuna e avançam na agricultura de precisão com monitoramento detalhado de culturas, transformando significativamente a economia comercial. As UASNets também revolucionam o setor comercial ao oferecer maior eficiência, segurança e economia, destacando o seu impacto transformador. Um aspeto fundamental desses novos recursos e mudanças é a coleta de dados de áreas acidentadas e remotas. Devido à sua excelente mobilidade e capacidade de manobra, os UAVs são empregados para coletar dados de sensores terrestres em ambientes hostis, como monitoramento de desastres naturais, vigilância de fronteiras, e monitoramento de resposta a emergências. Um grande desafio nesses cenários é que os movimentos dos UAVs afetam as condições do canal e resultam em perda de pacotes. Movimentos rápidos de UAVs levam a más condições do canal e rápida degradação do sinal, resultando em perda de pacotes. Por outro lado, a mobilidade lenta de um UAV pode causar estouros de buffer dos sensores de solo, pois os dados recém-chegados não são coletados prontamente pelo UAV.

A nossa proposta para enfrentar esse desafio é minimizar a perda de pacotes otimizando conjuntamente os controlos de velocidade e os cronogramas de coleta de dados de vários UAVs. Os estados dos sensores de solo incluem o nível da bateria, o comprimento da fila de dados e a qualidade do canal. Na ausência de conhecimento atualizado dos estados dos sensores de solo, propomos um algoritmo de programação baseado em aprendizado de reforço profundo multi-UAV (MADRL-SA). Esse algoritmo permite que os UAVs minimizem de modo assintótico a perda de pacotes devido a estouros de buffer e más condições do canal, mesmo na presença de conhecimento desatualizado dos estados da rede em UAVs individuais. Além disso, em UASNets, movimentos rápidos de UAVs resultam em más condições de canal e rápida atenuação de sinal, levando a uma idade da informação (AoI). Em contraste, os movimentos lentos dos UAVs prolongam o tempo de voo, estendendo assim o AoI dos sensores terrestres.

Além disso, os UAVs devem considerar os movimentos de outros UAVs para minimizar o AoI médio coordenando as suas velocidades. Portanto, encontrar uma solução de equilíbrio entre os UAVs para otimizar a velocidade e reduzir o AoI médio torna-se crucial.

Para enfrentar esse desafio, propomos uma nova otimização de alocação de recursos de voo de campo médio para minimizar o AoI dos dados sensoriais. Equilibrar o compromisso entre os movimentos do UAV e AoI é formulado como um jogo de campo médio (MFG). Introduzimos um novo esquema de otimização de política proximal híbrida de campo médio (MF-HPPO) para lidar com o espaço de solução expandido da otimização de MFG. Este esquema minimiza o AoI médio otimizando as trajetórias do UAV e os cronogramas de coleta de dados do sensor de solo, considerando ações mistas contínuas e discretas. Além disso, incorporamos uma memória de longo prazo (LSTM) no MF-HPPO para prever o estado da rede variável no tempo e estabilizar o processo de treino.

Palavras-chave: UAVs, Mean-field game, Age of Information, políticas de otimização por proximidade, memória de longo e curto prazo, escalonamento de comunicações, controlo de velocidade, Deep Q-Network.

Acknowledgments

I would like to express my heartfelt thanks to my parents for their unwavering support and the sacrifices they have made to help me reach this point in my academic journey. Your constant encouragement and belief in me have been instrumental in my success. To my brothers and sisters, I am grateful for the courage and confidence you have instilled in me. Your unwavering support and motivation have pushed me to strive for excellence.

I am deeply thankful to my Ph.D. advisor, Kai Li, for his invaluable support and meticulous guidance throughout my research. Under his supervision, I have developed a new scientific character, and I will forever be grateful for his mentorship. I am committed to following the path he has paved for me.

I would like to acknowledge my co-supervisor and CISTER director, Eduardo Tovar, who is a distinguished scientist and an exceptional manager. His erudition and leadership have greatly influenced my academic journey. I also would like to acknowledge my co-supervisor, Mario Sousa for his support and generosity.

I am grateful to Luis Almeida, the director of PDEEC and vice director of CISTER, who is a distinguished scientist. His paternal support helped me complete this journey

I extend my gratitude to the members of the jury who took the time from their busy schedules to evaluate my dissertation.

I would like to thank my colleagues at CISTER for their support and companionship. In particular, I am thankful to Cristiana, Inês, Sandra, and Marwin for their immense support and kindness.

Special thanks to Mohammad Nassiri, who imparted the fundamental knowledge of programming and laid a strong foundation for my scientific career. I am also grateful to Manijeh Keshtgari for her contributions to my scientific growth.

I want to express my appreciation to Ali Shanezan Zadeh for mentoring our innovative activities at the Islamic Azad University of Dezful and for his unwavering support in my life. I am also grateful to Rahim, Vahid, Hamid Reza, Omid Reza, and Filipa for their sincere support and kindness.

I dedicate this work to the late Dr. Ehsan Malekian, whose passion for computer networks served as my greatest motivation, and to my late cousin Mehdi.

Yousef Emami

Publications

Journals

1. **Y. Emami**, B. Wei, K. Li, W. Ni and E. Tovar, *Joint Communication Scheduling and Velocity Control in Multi-UAV-Assisted Sensor Networks: A Deep Reinforcement Learning Approach*, in IEEE Transactions on Vehicular Technology, vol. 70, no. 10, pp. 10986-10998, Oct. 2021, doi: 10.1109/TVT.2021.3110801. Impact Factor: 6.239. [Chapter 3]

2. **Y. Emami**, H. Gao, K. Li, L. Almeida, E. Tovar, and Z. Han, *Age of Information Minimization using Multi-agent UAVs based on AI-Enhanced Mean Field Resource Allocation*, IEEE Transactions on Vehicular Technology, 2023, under review. [Chapter 4]

3. K. Li, W. Ni, **Y. Emami** and F. Dressler, *Data-Driven Flight Control of Internet-of-Drones for Sensor Data Aggregation Using Multi-Agent Deep Reinforcement Learning*, in IEEE Wireless Communications, vol. 29, no. 4, pp. 18-23, August 2022, doi: 10.1109/MWC.002.2100681. Impact Factor: 12.777.

4. K. Li, **Y. Emami**, W. Ni, E. Tovar and Z. Han, *Onboard Deep Deterministic Policy Gradients for Online Flight Resource Allocation of UAVs*, in IEEE Networking Letters, vol. 2, no. 3, pp. 106-110, Sept. 2020, doi: 10.1109/LNET.2020.3002341. Impact Factor: 4.18.

5. K. Li, W. Ni, **Y. Emami**, Y. Shen, R. Severino, D. Pereira, and E. Tovar. 2019. *Design and Implementation of Secret Key Agreement for Platoon-based Vehicular Cyber-physical Systems*. ACM Trans. Cyber-Phys. Syst. 4, 2, Article 22 (April 2020), 20 pages. https://doi.org/10.1145/3365996. Impact Factor: 3.08.

Conference and Workshops

1. **Y. Emami**, B. Wei, K. Li, W. Ni and E. Tovar,*Deep Q-Networks for Aerial Data Collection in Multi-UAV-Assisted Wireless Sensor Networks*, International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 669-674, doi: 10.1109/IWCMC51323.2021.9498726. [Chapter 3]

2. **Y. Emami**, K. Li and E. Tovar, *Buffer-Aware Scheduling for UAV Relay Networks with Energy Fairness*, IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020, pp. 1-5, doi: 10.1109/VTC2020-Spring48590.2020.9128998.

3. K. Li, **Y. Emami**, and E. Tovar. 2019. *Privacy-preserving control message dissemination for PVCPS: poster abstract*. In Proceedings of the 18th International Conference on Information Processing in Sensor Networks (IPSN '19). Association for Computing Machinery, New York, NY, USA, 301–302. https://doi.org/10.1145/3302506.3312599

4. **Y.Emami**, K. Li, Y. Niu and E. Tovar, *AoI Minimization Using Multi-Agent Proximal Policy Optimization in UAVs-Assisted Sensor Networks*, ICC 2023-IEEE International Conference on Communications, Rome, Italy, 228-233. doi: 10.1109/ICC45041.2023.10278748

Contents

List of Figures

List of Tables

List of Abbreviations

UAV	Unmanned Aerial Vehicle
DRL	Deep Reinforcement Learning
DQN	Deep Q-Network
DDPG	Deep Deterministic Policy Gradient
PPO	Proximal Policy Optimization
MFG	Mean Field Game
RL	Reinforcement Learning
MDP	Markov Decision Process
MMDP	Multi-Agent Markov Decision Process
AoI	Age of Information
LSTM	Long Short Term Memory
UASNets	UAVs-Assisted Sensor Networks
FPK	Fokker-Planck-Kolmogorov
LoS	Line-of-Sight
QoS	Quality of Service
UMi	Urban Micro
UMa	Urban Macro
SWAP	Size, Weight, and Power
TRPO	Trusted Region Policy Optimization
MARL	Multi-Agent Reinforcement Learning
Dec-POMDP	Decentralized Partially Observable Markov Decision Process
MADRL-SA	Multi-UAV Deep Reinforcement Learning based Scheduling Algorithm
MF-HPPO	Mean Field Hybrid Proximal Policy Optimization
WSNs	Wireless Sensor Networks.
ABS	Aerial Base Station
DoF	Degree of Freedom
ML	Machine Learning
IEEE	Institute of Electrical and Electronics Engineers
RNN	Recurrent Neural Network

6G	6 Generation
5G	5 Generation
eMBB	Enhanced Mobile Broadband
URLLC	Ultra Reliable Low Latency Communications
mMTC	Massive Machine Type Communications
EU	European Union
IoT	Internet of Things
UAS	Unmanned Aerial Systems
US	United States
DNN	Deep Neural Network

Chapter 1

Introduction

Unmanned aerial vehicles (UAVs) have become indispensable in today's technological advancements, bringing about significant changes in various fields. They have revolutionized sectors such as agriculture, public safety, environmental monitoring, and security. In the realm of agriculture, UAVs hold great potential for precision farming, aligning with the European Union's focus on sustainable and environmentally friendly agricultural practices. Additionally, UAVs have proven their worth in assessing hazardous situations, conducting search and rescue missions, gathering evidence for investigations, and detecting potential threats Undertaking et al. (2017). Furthermore, UAVs play a crucial role in the development of 5th generation (5G) networks, contributing to the realization of 5G's goals, including enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC), and massive machine-type communications (mMTC). In the context of eMBB, UAVs provide high data rates, particularly in densely populated or remote areas. They can act as aerial base stations (ABS) or relays, supporting URLLC and reducing latency for real-time communication. Moreover, UAVs facilitate mMTC by enabling the deployment of Internet of Things (IoT) devices in challenging environments and optimizing network resources to handle a large number of connections. Looking ahead, UAVs are expected to play a pivotal role in 6th generation (6G) networks, enabling improved data collection and analysis. Enhanced data collection techniques allow for real-time capture of a wider range of data, thereby enhancing decision-making processes. This opens up opportunities for precise environmental monitoring, real-time traffic analysis, and prompt disaster response through immediate aerial assessments Li et al. (2018), JIANG et al. (2022).

UAVs have the capability to operate in challenging and remote environments, making them ideal for aerial data collection. The integration of UAVs into sensor networks for this purpose is known as UAVs-assisted sensor networks (UASNets). UAVs can serve

as aerial base stations (ABS) or relays to extend the coverage and connectivity of sensor networks Mozaffari et al. (2019). The advancement in UAV manufacturing and the miniaturization of communications equipment have made it possible to incorporate compact and lightweight wireless transceivers into UAVs, enabling efficient aerial data collection. Commercial wireless transceivers suitable for UAV installation with moderate payloads are already available in the market. Compared to traditional terrestrial communications that rely on fixed gateway locations, UASNets offer several advantages. Firstly, aerial data collectors can be rapidly deployed, making them particularly beneficial for harsh and remote areas. Secondly, due to their high altitude, UAVs have a higher probability of establishing line-of-sight (LoS) connections with ground sensors, resulting in more reliable communication links. Thirdly, the mobility of UAVs provides an additional degree of freedom (DoF) for optimizing communication performance by dynamically adjusting their positions in three dimensions to meet the communication demands on the ground.

Integrating UAVs into wireless sensor networks (WSNs) presents new design opportunities but also brings challenges. UASNets differ significantly from terrestrial networks due to factors such as the high altitude and mobility of UAVs, the likelihood of LoS channels between UAVs and ground sensors, varying quality of service (QoS) requirements for payload and non-payload data, strict size, weight, and power (SWAP) constraints of UAVs, and the need to jointly optimize UAV mobility control and communication scheduling/resource allocation to maximize system performance.

- High altitude: UAV data collectors are positioned at much higher altitudes compared to traditional terrestrial gateways. While terrestrial gateways are typically located at around 10m for urban micro deployment and 25m for urban macro deployment, UAVs can fly as high as 122m under current regulations. This higher altitude enables UAV data collectors in UASNets to achieve wider ground coverage compared to their terrestrial counterparts.

- Higher channel gain: The air-ground channels experienced by UAVs exhibit distinct characteristics due to their high altitude. Unlike terrestrial channels that suffer from low channel gain due to shadowing and multipath fading, UAV ground sensor channels generally have limited scattering and primarily rely on LoS links, resulting in higher channel gain. This LoS-dominant air-ground channel offers more reliable link performance between UAVs and associated ground sensors.

- Controlled mobility: Unlike fixed terrestrial gateways, UAVs possess the capability to move at high speeds in three-dimensional space, allowing for controlled mobility. While this mobility introduces time-varying channels with ground sensors, it

also opens up new design opportunities for communication-aware control of UAV mobility. UAVs can optimize their position, altitude, velocity, heading direction, and trajectories to adapt to communication objectives and improve overall network performance.

- SWAP constraints: UAVs face significant SWAP constraints, which limit their endurance, computational capacity, and communication capabilities. Unlike terrestrial communications systems that benefit from stable power supplies at fixed gateways, UAVs must operate within these constraints, requiring efficient power management, lightweight hardware, and optimized communication protocols Wu et al. (2020).

Meanwhile, UASNets enhance the decision-making process through their advanced data collection capabilities. By gathering comprehensive and real-time information, they provide a rich and accurate basis for decision-making. These networks combine the agility and adaptability of UAVs with the extensive data collection capabilities of ground sensors, creating a system that not only collects valuable data but also reacts quickly and adjusts to changing environmental conditions. This adaptability makes UASNets highly effective in dealing with different situations. The following reasons highlight the importance of UASNets as a significant research area: (i) UASNets can cover large areas and collect high-quality data in real time. This makes them valuable in various fields such as environmental monitoring, wildlife protection, and infrastructure inspection. (ii) UASNets play a critical role in providing vital information to first responders during natural disasters. This information enriches the decision-making process and allows for more efficient resource allocation. (iii) Farmers can efficiently monitor crop health, soil conditions, and water needs using UASNets. This improves agricultural quality, increases productivity, and reduces environmental impact. (iv) UASNets enable agencies to regularly inspect critical infrastructure such as bridges, dams, and power lines. This proactive approach reduces the risk of catastrophic failures. In summary, research on UASNets contributes to the development of innovative solutions for practical problems, improves quality of life, and promotes sustainable development. The EU recognizes the importance of UASNets and their integration into various disciplines. The following are reasons highlighting the importance of UASNets in the EU: The EU is committed to sustainable development and environmental preservation. UASNets play a crucial role in providing important data for monitoring air pollutants, identifying their sources, and assessing ecosystem health. This aligns with the EU's goals of reducing emissions and preserving the environment. The EU has an aging infrastructure network that requires regular inspection and maintenance. UASNets can help take proactive measures by identifying potential problems, enabling

timely maintenance, and reducing the risk of catastrophic failures. This contributes to ensuring public safety. With a projected fleet of approximately 50,000 UAVs, UASNets can support public safety missions. UASNets can be utilized in the energy sector for performing preventative maintenance inspections and mitigating risks to personnel and infrastructure. It is estimated that around 10,000 UAVs will be used in this sector, contributing to efficient and safe energy operations. The EU envisions a fleet of 400,000 UAVs for civil applications by 2050. By leveraging UASNets, the EU can address various challenges while promoting sustainability, economic growth, and improved quality of life for its citizens.

Europe is not the only region making intensive efforts to utilize UAVs. The United States (US) and China, two major countries, are investing significantly in technology and innovative companies, surpassing the level of European investments. Specifically, the US and China are leaders in the production of defense and civil UAVs Undertaking et al. (2017). This emphasizes the transformative potential of UASNets in addressing technological complexities and economic constraints.

A practical example of UASNets can be observed in precision agriculture. The role of agriculture is of paramount importance to the European economy, with food security being a top priority. UASNets can optimize agricultural practices, minimize waste, and enhance crop productivity, thereby contributing to the overall goals of the European agricultural sector. It is predicted that in the agriculture sector, more than 100,000 UAVs will enable precision agriculture to achieve the necessary increase in productivity.

Fig. 1.1 shows a typical UASNets setup where ground sensors monitor farmland. The ground sensors generate sensory data, which is stored in a data queue for later transmission to the UAVs. The UAVs hover over the farmland, approaching each ground sensor closely to collect data over short distances. In this scenario, a large farm is equipped with soil sensors that continuously monitor various parameters, including soil moisture, temperature, and nutrient levels. These ground sensors consistently gather data, providing farmers with information about irrigation, fertilization, and crop protection

UAVs are employed as aerial data collectors, patrolling over the farmland and utilizing LoS communications. They manage their mobility to approach ground sensors and collect sensory data. UAVs can improve overall network coverage and performance, enabling farmers to access comprehensive and accurate data for optimizing their farming practices.

The remainder of this chapter is organized as follows: Section 1.1 presents the motivation for this work. Section 1.2 outlines the thesis and research questions. Section 1.3 presents the methodology. Section 1.4 outlines the structure of the thesis.

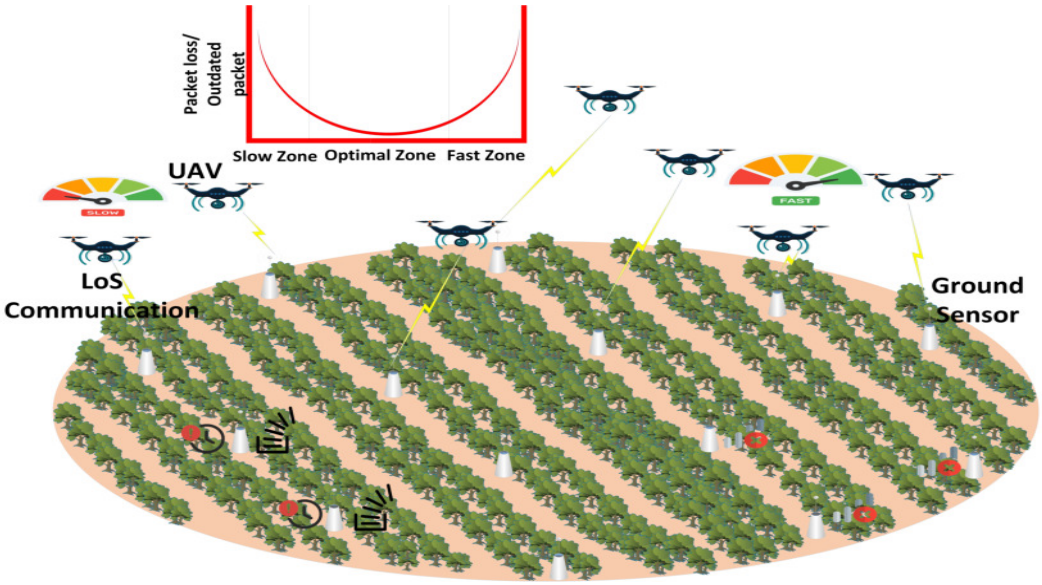


Figure 1.1: An overview of UASNets for precision agriculture.

1.1 Motivation

Thanks to their exceptional mobility and maneuverability, UAVs are utilized in various civil and commercial applications, including weather monitoring, traffic control, package delivery Shakhatreh et al. (2019a), and crop monitoring Kim et al. (2019). They are also employed as data relays for ground sensors in challenging environments such as natural disaster monitoring Zhao et al. (2019), border surveillance Shakhatreh et al. (2019b), and emergency assistance Gao et al. (2020). In scenarios where ground sensors are deployed beyond the reach of terrestrial gateways and lack a consistent power supply, UAVs can physically approach each individual ground sensor. The short-range LoS communication link between a UAV and a ground sensor exhibits significant channel gain, enabling high-speed data transmission. By utilizing UAVs for data collection, network throughput can be improved, and coverage can be extended beyond terrestrial gateways. Moreover, UASNets offer several advantages for data collection in remote and inhospitable environments. UAVs can access areas that are challenging for humans to reach, resulting in more efficient and cost-effective data collection. This approach reduces safety risks as the use of UAVs eliminates the need for human intervention in hazardous environments. Due to their mobility, UAVs have the capability to cover vast areas, thereby reducing the time and resources required for data collection.

In UASNets, ground sensors are exposed to random data inputs as data generation is influenced by unpredictable variations in temperature and humidity. As depicted in Fig. 1.1, UAVs are deployed to hover over farmland, allowing close proximity to ground sen-

sors and utilizing short LoS communication links for data collection. However, selecting a ground sensor for data collection may lead to buffer overflows for other sensors if their buffers are already full while new data continues to arrive. Moreover, transmissions from ground sensors located far away from the UAVs, experiencing poor channel conditions, are susceptible to errors at the UAVs. The slow mobility of a UAV can contribute to buffer overflows in ground sensors as newly arrived data is not promptly collected by the UAV. Properly scheduling data collection, taking into account the onboard velocity of the UAVs, is crucial to avoid data queue overflow and communication failures. Additionally, coordination between participating UAVs is necessary for joint velocity control and sensor selection. However, real-time sharing of velocities and selected sensors among UAVs is challenging due to limited radio coverage and the fast movements of UAVs.

In summary, the effective management of joint communication scheduling and velocity control is crucial to minimize packet loss, preventing buffer overflows and communication failures in UASNets. However, it is important to note that ensuring the freshness and relevance of collected data is also essential. To achieve this, minimizing the age of information (AoI) becomes necessary in UASNets.

In UASNets, the AoI is commonly used to measure the freshness of sensory data Kaul et al. (2012) collected at ground sensors and received by the UAVs. It represents the time elapsed between data generation at a ground sensor and its receipt at the UAV, accounting for transmission time and network delays. When the UAV's flight is not properly controlled, it can move away from the ground sensor, increasing the AoI and causing data to expire. Additionally, different ground sensors may have varying AoI due to the impact of monitored natural conditions on data generation Li et al. (2022a). The optimization of UAV cruise control and communication schedules to minimize AoI becomes challenging because the UAV has limited knowledge of ground sensors' data generation rate and channel conditions. Swift movements of the UAV result in poor channel conditions and frequent data retransmissions, leading to a prolonged AoI. Conversely, slow UAV movements extend flight time and increase the AoI of ground sensors. Furthermore, the UAV needs to consider the movements of other UAVs to minimize the average AoI by coordinating their velocities, highlighting the importance of finding an equilibrium solution.

Decentralized approaches are relevant when UAVs have limited information about each other's actions, such as trajectory, flight speed, and scheduled ground sensors. Game theory can be applied to design decentralized control and determine equilibria in UAV networks Mkiramweni et al. (2019). However, traditional game theory approaches become computationally intractable with a large number of UAVs. Mean-field game (MFG), on the other hand, offers a scalable framework to address the joint optimization of cruise control and communication schedules. MFG approximates the interactive behavior of a large

number of UAVs using a continuum or mean field, significantly reducing computational complexity. It enables UAVs to make decisions based on the overall swarm behavior rather than individual actions.

1.2 Thesis Statement and Research Questions

In this research, our proposed solutions aim to address the challenges faced by UASNets. In this framework, the thesis statement is as follows:

> *We postulate that incorporating cruise control and data collection scheduling into UASNets can effectively alleviate the impact of channel conditions and unlock the advantages of timeliness and resource utilization in UASNets.*

Based on this thesis, our research focuses on investigating and enhancing the performance of data collection in UASNets. We envision that by adopting this new paradigm, real-time decision-making can be facilitated, leading to improved resource utilization. Consequently, we anticipate advancements in QoS, overall system reliability, and productivity of UASNets. However, achieving this goal entails addressing several challenging scientific problems, which we formulate as the following two research questions.

(RQ_1). **Research Question 1:** How can we develop a joint communication scheduling and velocity control mechanism for data collection in UASNets to minimize packet loss and mitigate the impact of UAV movement on data transmission, while addressing the challenges posed by limited radio coverage and rapid movement?

> How the joint communication scheduling and velocity control mitigate the effects of UAVs' movement on data transmission in the presence of fast movements.

(RQ_2). **Research Question 2:** In the presence of a large number of UAVs, the challenge lies in developing cruise control mechanisms that minimize the AoI and mitigate the impact of UAVs' movements on AoI. Additionally, how can we find an equilibrium solution and capture the temporal dependencies of cruise control?

How can we develop cruise control and mitigate the impact of UAV move ments on AoI.

1.3 Methodology

This work focuses on improving the performance of data collection in UASNets, particularly in lossy channels. The main objective is to minimize packet loss and AoI in order to enhance the efficiency of data collection. To achieve this goal, the study explores the application of deep reinforcement learning (DRL). By leveraging DRL algorithms, the research aims to develop intelligent and adaptive solutions that can learn from the environment and determine the optimal policy.

The use of DRL-based techniques is expected to provide valuable insights and effective approaches to significantly enhance the performance of data collection in UASNets. The ultimate aim is to contribute to the advancement of this emerging field by proposing novel solutions that leverage DRL for improved data collection performance.

In this thesis, the joint communication schedule and velocity control of multiple UAVs are formulated as a multi-agent Markov decision process (MMDP) to minimize packet loss caused by buffer overflows and communication failures. The ground sensor keeps a record of the visit time whenever a UAV schedules data transmission from the sensor. Furthermore, the visiting records of the sensor are shared with the UAV, serving as evidence of other UAVs' communication schedules. The network state in the MMDP includes battery levels and data queue lengths of the ground sensors, channel conditions, visit time, and waypoints along the UAVs' trajectories. The UAVs take actions such as selecting ground sensors for data transmission, determining modulation schemes, and adjusting patrol velocities. In practical scenarios, the UAVs lack real-time knowledge of the battery level and data queue length of the ground sensors. Thus, multi-UAV Q-learning can be employed to train the UAVs' actions. However, since each UAV's trajectory may have a large number of waypoints, controlling the velocities of the UAVs along the trajectories results in a vast state and action space, making multi-UAV Q-learning complex.

In our MFG approach, the optimal velocities of the UAVs are determined by solving a Fokker–Planck–Kolmogorov (FPK) equation. This equation describes the evolution of the mean field to achieve an equilibrium of the optimal velocities of the UAVs. However, in practical scenarios, the proposed MFG solution is challenging to implement online due to the lack of instantaneous knowledge of the UAV's cruise control decisions and AoI. To address this, we formulate the flight resource allocation optimization problem in the

MFG framework as an MMDP. The network states in the MMDP consist of the AoI of the ground sensors and the waypoints of the UAV swarm. The action space in the MMDP includes continuous variables such as waypoints and velocities, as well as discrete variables representing transmission schedules. To tackle this complex problem, we propose a solution called the mean-field hybrid proximal policy optimization (MF-HPPO). MF-HPPO aims to optimize both the cruise control of the UAVs and the transmission schedules of the ground sensors in a coordinated manner, leveraging the advantages of the mean-field approximation.

The research topics of this thesis can be summarized as follows:

- Joint velocity control and communication scheduling to minimize packet loss and using DRL to find the optimal policy.
- Cruise control based on MFG to minimize AoI and using DRL to find the mean field equilibrium.

1.4 Contributions

In this section, we summarize the main findings of our research in relation to the research questions outlined in Section 1.2 and discuss the contribution of our work to the existing body of knowledge in the field of data collection.

1. This contribution addresses RQ1. The problem of joint velocity control and data collection scheduling is formulated as an MMDP to minimize packet loss caused by buffer overflow and channel fading. To handle the large state and action spaces, we propose the multi-UAV DRL-based scheduling algorithm (MADRL-SA) using Deep-Q-Networks (DQN) to optimize the selection of ground sensors, instantaneous patrol velocities of UAVs, and modulation schemes. The inclusion of experience replay enhances the learning efficiency of the algorithm by reducing sample correlations.

   The mentioned contribution is of utmost importance as it addresses the challenges faced by modern UAV networks in handling complex dynamic environments, including UAV movement. The proposed methodology showcases the potential of DRL in solving complex problems. It also contributes to the development of intelligent, adaptive, and autonomous systems capable of self-optimization. The use of DRL in conjunction with experience replay enhances the system's ability to learn and evolve, leading to improved performance.

2. This contribution addresses RQ1. : In practice, the online decisions of UAVs during flight are unknown to each other, which can hinder the training of MADRL-SA. To address this, a local action recording process is developed where ground sensors record historical visits of all UAVs. The UAV scheduling a ground sensor receives these records, providing information about the past scheduling decisions of other UAVs.

   The introduction of the local action recording process in this contribution is an important step in addressing the practical challenges associated with UASNets. In practical scenarios, UAVs are unaware of each other's decisions, leading to uncertainty and potentially incomplete training of MADRL-SA. This situation can result in suboptimal decisions and degrade network performance. By incorporating a local action recording process, the algorithm's effectiveness is ensured even under realistic operating conditions. This approach promotes a more collaborative environment among UAVs, allowing them to adjust their actions based on the observed behavior of other UAVs in the network.

3. This contribution addresses RQ2. A novel formulation of MFG optimization with a large number of UAVs is proposed to address the trade-off between UAV cruise control and AoI. Due to the computational complexity of MFG, the MF-HPPO algorithm is introduced to minimize average AoI. The algorithm learns state dynamics and optimizes UAV actions in a mixed discrete and continuous action space.

   This contribution represents a significant advancement in the field of UASNets with a large number of UAVs. By leveraging MFG optimization, we effectively address the challenges associated with managing such complex systems. Our approach focuses on the collective behavior of UAVs, leading to improved resource allocation and overall performance. MF-HPPO efficiently optimizes both continuous and discrete actions to minimize the average AoI. This ability to optimize UAV actions in a mixed-action space highlights their versatility and adaptability, enabling them to meet diverse network conditions and requirements. The proposed method underscores the importance of advanced optimization techniques in solving complex real-world problems. Moreover, this contribution pushes the boundaries of UASNets and highlights the wider applicability of MFG optimization in addressing complex challenges across various domains.

4. This contribution addresses RQ2. To capture temporal dependencies in cruise control and improve learning convergence, a new long short-term memory (LSTM)

layer is developed within the MF-HPPO algorithm. This LSTM layer predicts time-varying network states, such as AoI and UAV waypoints.

By incorporating the LSTM layer, our contribution tackles the issue of capturing temporal dependencies in cruise control, which is crucial for efficiently managing and optimizing UASNets. This development emphasizes the significance of combining advanced machine learning (ML) techniques to create more robust and adaptable algorithms.

The above contributions are presented in the following publications:

- **Y. Emami**, B. Wei, K. Li, W. Ni and E. Tovar, *Deep Q-Networks for Aerial Data Collection in Multi-UAV-Assisted Wireless Sensor Networks*, 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 669-674, doi: 10.1109/IWCMC51323.2021.9498726.

- **Y. Emami**, B. Wei, K. Li, W. Ni and E. Tovar, *Joint Communication Scheduling and Velocity Control in Multi-UAV-Assisted Sensor Networks: A Deep Reinforcement Learning Approach*, in IEEE Transactions on Vehicular Technology, vol. 70, no. 10, pp. 10986-10998, Oct. 2021, doi: 10.1109/TVT.2021.3110801.

- **Y. Emami**, H. Gao, K. Li, L. Almeida, E. Tovar, and Z. Han, *Age of Information Minimization using Multi-agent UAVs based on AI-Enhanced Mean Field Resource Allocation*, IEEE Transactions on Vehicular Technology, 2023, under review.

1.5 Thesis Structure

The rest of this document is structured as follows.

- Chapter 2 discusses background and related work. In this chapter, we present background on DRL and then delve into the existing literature to explore related work on flight resource allocation and scheduling. Our objective is to gain a comprehensive understanding of the relevant research in order to comprehend the joint communication scheduling and velocity control in UASNets. Additionally, we review the literature on mean-field flight resource allocation and time-critical flight resource allocation to analyze their strengths and weaknesses. This analysis serves as a foundation for developing a cruise control system based on MFG theory that minimizes AoI.

- Chapter 3 formulates the joint communication scheduling and velocity control problem as an MMDP to minimize packet loss resulting from communication failures and buffer overflows. Given the large state and action spaces, we employ DRL techniques to discover the optimal policy for this problem.
- Chapter 4 formulates cruise control based on MFG theory to minimize AoI. Solving the MFG problem online poses challenges, hence we formulate it as an MMDP, encompassing both continuous and discrete actions. To address this MMDP formulation, we propose the MF-HPPO algorithm, which optimizes actions in a mixed-action space.

Chapter 2

Background and Related work

UASNets have emerged as an innovative technology that offers enhanced data collection capabilities for various applications, including environmental monitoring, disaster management, and surveillance. In UASNets, UAVs play a critical role in gathering sensory data. However, a significant challenge in UASNets is the dynamic nature of UAV movements, which greatly impacts channel conditions and gives rise to issues such as packet loss and outdated packets. The rapid movements of UAVs can result in unfavorable channel conditions and quick signal degradation, requiring frequent data retransmissions. Conversely, slow movements prolong the flight time, leading to delays in collecting newly arrived data by the UAV. To tackle these challenges, one potential strategy is to perform joint cruise control and communication scheduling in the presence of lossy channels, aiming to minimize packet loss and AoI. One effective approach for addressing the challenges in UASNets is cruise control and data collection scheduling. In the following section, we present background information on DRL then we provide a review of the existing literature on this problem. The relevant state-of-the-art works in this area can be classified into three categories: i) DRL-aided flight resource allocation and scheduling, ii) DRL-aided flight resource allocation using MFG, and iii) DRL-aided flight resource allocation for data freshness.

2.1 Deep Reinforcement Learning

DRL is a prominent branch of ML in which an agent learns to interact within an environment by taking actions and observing the resulting outcomes Sutton and Barto (2018).

DRL is particularly useful for solving MMDPs that have unknown transition probabilities. During the DRL process, an agent observes its current state, selects an action, and receives immediate feedback in the form of a cost or reward, along with the new state.

The observed information, such as the immediate cost and new state, is then utilized to adjust the agent's policy. This iterative process continues until the agent's policy converges toward the optimal policy Luong et al. (2019). DRL can be applied to UASNets for the following reasons: (a) UAVs may face challenges in implementing mathematical models of the complex environment or may not have access to such models. (b) The mobility feature of UAVs leads to large state spaces and action spaces. (c) UAVs often lack up-to-date knowledge about the status of ground sensors, including battery, energy, and channel conditions.

Formally, DRL can be described as an MMDP, which includes the number of agents, state, action, shared cost function, and transition probability. An MMDP is a mathematical framework used to model decision-making in situations where multiple agents interact with each other in an uncertain environment. In an MMDP, the action taken by each agent not only determines the future state but also affects the actions of other agents. Furthermore, in an MMDP, a shared cost function is employed. This shared cost function considers the joint action of agents and provides feedback that is common to all agents. The objective is to encourage collaboration among the agents toward a shared goal, rather than individual goals that may conflict with each other. Well-designed shared cost functions can foster collaboration among agents and lead to more favorable outcomes for the entire team. The MMDP framework finds applications in various domains, including UAV swarm control and multiplayer games..

Q-learning, due to the exponential growth of states and actions caused by the mobility of UAVs, is unable to handle the resource allocation problem in UASNets. This issue, commonly referred to as the curse of dimensionality, poses a significant challenge. However, DQN offers a solution to overcome this challenge. In the context of UASNets, DQNs play a crucial role in optimizing various operational aspects, including flight trajectory, cruise control, and data collection scheduling. They employ deep neural networks to represent the action-value function of each agent. The state information captured by the DQN can encompass the UAV's current location, the status of sensor nodes, and the traffic conditions within the network. The available actions can involve adjusting parameters such as speed, trajectory, or data transmission schedules. DQN incorporates the use of a target network and experience replay for each UAV to ensure stability in the learning process. Experience replay is utilized in DQN to eliminate correlations in the observation sequence and mitigate abrupt changes in the data distribution by randomizing states and actions within the MMDP. Consequently, DQN aids in forming a policy that minimizes the cost function and enhances the overall performance of UASNets.

DQN is primarily designed to optimize discrete actions and is limited in its ability to handle continuous actions. To address this limitation and optimize both discrete and

continuous actions, proximal policy optimization (PPO) can be employed. In the context of UASNets, PPO plays a crucial role in enhancing the operational efficiency of these networks. As a type of policy gradient method, PPO enables the optimization of UAV trajectories (continuous action space) and data collection schedules (discrete action space). The algorithm maintains a delicate balance between exploring new strategies and exploiting the current strategy, which is particularly advantageous in complex and dynamic environments like UASNets.

PPO achieves stability and efficient learning by ensuring only a small deviation from the previous strategy during each update, thereby mitigating the risk of detrimental updates. The objective of PPO is to minimize costs, which can be tailored to reflect real-time data collection requirements. By optimizing both trajectories and data collection schedules, PPO facilitates the development of robust policies that enhance the overall performance of UASNets. PPO has two primary variants: PPO-penalty modifies the hard constraint of TRPO by incorporating it as a penalty in the objective function. On the other hand, PPO-clip does not impose a constraint but utilizes clipping techniques to bind the changes in the policy.

2.2 DRL-aided Flight resource allocation and scheduling

The work in Wu et al. (2018) develops a framework for trajectory control, user association, and power control in multi-UAV enabled wireless networks. Communication throughput gains can be obtained by mobile UAVs over static UAVs/fixed terrestrial base stations, by exploiting the design DoF via UAV trajectory adjustment. A general mixed integer nonlinear program formulation for a multi-UAV network is presented in Thammawichai et al. (2017) to adjust the communication and the computational energy. Chen (2020) explores a multi-UAV-aided relaying system, where UAV relays aim to establish communication between senders and receivers and to improve the rate between the pair of sender and receiver, the UAVs' positions are adjusted, and resource allocations are conducted. In Sharma et al. (2016), a cooperative framework designed which allowed the formation of a network between the aerial and the ground nodes. Their approach provides continuous connectivity, enhanced lifetime, and improved coverage in the UAV coordinated WSNs and laid the foundation of guided network formations between the UAVs and the ad hoc networks on the ground. A framework is developed in Albu-Salih and Seno (2018) to improve energy efficiency in deadline-based WSN data collection with multiple UAVs. In Zhan and Zeng (2019), the mission completion time is adjusted for multi-UAV-enabled data collection. An energy-efficient transmission scheduling scheme of UAVs in a coop-

erative relaying network is developed in Li et al. (2015) such that the maximum energy consumption of all the UAVs is minimized, in which an applicable sub-optimal solution is developed and the energy could be saved up to 50% via simulations. In Samir et al. (2020) a UAV is used to collect data from time-constrained Internet of Things (IoT) devices. The UAV trajectory and radio resource allocation are adjusted to collect data from IoT devices, adapting to their deadline.

InLi et al. (2019), a single-agent DQN for UAV-assisted online power transfer and data collection is developed. However, in most situations, multiple UAVs are needed to interact with each other to solve a resource allocation problem. In Li et al. (2020a), online velocity control and data capture are studied in UAV-enabled IoT networks. DQN is developed in the presence of outdated knowledge to determine the patrolling velocity and data transmission schedule of the IoT node. In Li et al. (2020b), the joint flight cruise control and data collection scheduling in the UAV-aided IoT network is formulated as a POMDP to minimize the data lost due to buffer overflows at the IoT nodes and fading airborne channels. A UAV-assisted IoT communication is investigated in Munaye et al. (2020) where by applying multi-agent DRL a resource allocation scheme adapting to bandwidth, throughput, and interference is obtained. A wireless powered communication network is developed in Tang et al. (2020) where multiple UAVs provide energy supply and communication services to IoT devices. They used a multi-UAV DQN based approach to improve throughput by jointly adjusting UAVs' path design and time resource assignment. They follow an independent learner approach without cooperation between UAVs. In Zhang et al. (2017), the authors consider long-term, long-distance sensing tasks in a smart city scenario where UAVs make decisions based on DQN for energy-efficient data collection. An energy-saving DRL-based UAV control strategy is developed in Liu et al. (2018) to enhance energy efficiency and communication coverage. They used deep deterministic policy gradient (DDPG) method and take into account communications coverage, fairness, energy consumption, and connectivity. In Wang et al. (2019), the dueling DQN is employed to adjust the UAV deployment in the multi-UAV wireless networks so that downlink capacity is to be enhanced while covering all ground terminals. They modeled the problem as a constrained MDP problem.

The MARL framework is developed in Cui et al. (2020) to investigate the dynamic resource allocation problem in UAV networks. A Q-learning based algorithm is developed to enhance the long-term rewards where each UAV runs Q-learning algorithm and automatically selects its communication mode, power levels and sub-channels in concurrent manner. Shamsoshoara et al. (2019) studies spectrum sharing among a network of UAVs. A relaying service is realized by team of UAVs to serve primary users on the ground aiming to gain spectrum access consequently. The gained spectrum belongs to

not only UAV relay, but also other UAVs that perform the sensing task. The problem is formulated as deterministic MMDP and distributed Q-learning is utilized to solve it. Challita et al. (2018) develops the DRL algorithm based on echo state network cells to find an interference-aware path and allocate resources to the UAVs. The developed scheme reduces wireless latency and improves energy efficiency. The work in Liu et al. (2019) adjusts trajectory and power control in multiple UAVs scenarios to enhance the users' throughput and satisfying the users' rate requirement.

2.3 DRL-aided flight resource allocation using mean field game

In Chen et al. (2020), the authors explore energy-efficient control strategies for UAVs that provide fair communication coverage for ground users. The UAV control problem is modeled as an MFG and a mean-field TRPO algorithm is studied to design the UAVs' trajectories. In Li et al. (2020c), the authors apply the MFG theory to the downlink power control problem in ultra-dense UAV networks to improve the network's energy efficiency. Due to the complexity of the MFG, a DRL-MFG algorithm is developed to learn the optimal power control strategy. Shi et al. (2020) studies the task allocation in cooperative mobile edge computing and a mean field guided Q-function is formulated to reduce the network latency. MFG and DRL are integrated to guide the learning process of DRL according to the equilibrium of MFG. In Sun et al. (2020), the authors model the trajectory planning and power control for heterogeneous UAVs as an MFG, aiming to reduce energy consumption. A mean field Q-learning is studied to find the optimal solution. In Wang et al. (2021b), the authors study UAV-assisted ultra-dense networks, where each UAV can adjust its location to reduce the AoI. They formulate the problem as an MFG and apply a DDPG-MFG algorithm to find the mean field equilibrium. In Li et al. (2020c), downlink power control for a large number of UAVs is suggested to enhance the energy efficiency by learning the optimal power control policy. MFG is used to model the power control problem of the UAV network, where each UAV tries to enhance the energy efficiency by adjusting its transmit power. Then, due to the complexity of solving the formulated MFG, an effective DRL-MFG algorithm is suggested to learn the optimal power control strategy.

Although, DRL-based solutions are mainly used, the following works adopt numerical solutions. In Xue et al. (2018), the focus is on adaptive coverage problem in emergency communication system, where multiple UAV act as aerial base stations to serve randomly distributed users. The problem is formulated using discrete MFG, each UAV aims to reduce its flight energy consumption and increase the number of users it can serve. Finally,

optimal control and state of each UAV are computed. In Xu et al. (2018), a discrete MFG is formulated to address joint adjustment of power and velocity for a large number of UAVs that act as aerial base stations. Decentralized control laws are developed, and mean field equilibrium is analyzed. In Gao et al. (2022), the authors present an energy-efficient velocity control algorithm for a large number of UAVs based on the MFG theory. The velocity control of the UAVs is modeled using a differential game in which energy and delay are balanced by using an original double mixed gradient method.

2.4 DRL-aided flight resource allocation for data freshness

In Oubbati et al. (2022), the authors consider ground sensors with limited energy and apply airborne base stations to collect sensory data. Each UAV's task is decomposed into energy transfer and fresh data collection. A centralized multi-agent DRL based on DDPG is developed to adjust the UAV trajectories in a continuous action space, to reduce the AoI of the ground sensors. In Chi et al. (2022), the authors study UAV-assisted sensor networks where multiple UAVs cooperatively conduct the data collection to reduce the AoI. The trajectory planning is formulated as a decentralized partially observable markov decision process (Dec-POMDP). A multi-agent DRL is studied to find the optimal strategy. In Hu et al. (2019) and Hu et al. (2020b), the authors develop the trajectory planning for multiple UAVs that perform cooperative sensing and transmission, aiming to reduce the AoI. In Samir et al. (2022), ground sensors sample and upload data in a UAV-assisted IoT network. PPO is used to explore the optimal scheduling policy and altitude control for the UAV to reduce the AoI. In Sun et al. (2021), a data collection scheme characterized by AoI and energy consumption in a UAV-assisted IoT network is investigated. The average AoI, and energy consumption of propulsion and communication are reduced by adjusting the UAV flight speed, hovering waypoints, and bandwidth allocation for data collection using a TD3-based approach.

Although DDPG and PPO are used to adjust continuous and discrete actions to reduce AoI, the following works use DQN to adjust discrete actions. In Eldeeb et al. (2022), the authors investigate UAV-assisted IoT networks where multiple UAVs relay data between sensors and base station. A DQN-based trajectory planning algorithm is presented to reduce the AoI. In Abd-Elmagid et al. (2019), ground sensors with limited energy are used to observe various physical processes in the context of a UAV-assisted wireless network. The trajectory and scheduling policy are adjusted to reduce the weighted sum of AoI, and a DQN-based solution is applied to obtain the best strategy. In Zhou et al. (2019), trajec-

tory planning of the UAV is performed to reduce the AoI in a UAV-assisted IoT network. The problem is formulated as an MDP, and a DQN-based algorithm is studied to find the optimal trajectories of the UAV. In Tong et al. (2020), a UAV-assisted data collection for ground sensors is studied, where the UAV with limited energy is dispatched to collect sensory data. The UAV's trajectory is adjusted to reduce the average AoI and keep the packet loss rate low. The trajectory planning is formulated as an MDP while DQN is applied to design the UAV's trajectory. In Liu et al. (2021a), a UAV-assisted wireless network with an energy supply is used, where the UAV performs wireless energy transmission to ground sensors, and the sensors transmit data to the UAV using the harvested energy. A DQN-based trajectory planning algorithm is presented to reduce the average AoI by adjusting the trajectory, transmission schedule, and harvested energy.

2.5 Research Opportunity

The works by Li et al. (2020a) and Li et al. (2020b) address the optimization of velocity control and data collection schedules to minimize packet loss. However, these works are formulated for a single agent scenario. In contrast, our proposed approach, MADRL-SA, differs from the MARL framework introduced by Cui Cui et al. (2020). In MARL, UAVs operate based on an independent learner paradigm, whereas MADRL-SA promotes cooperation among UAVs to minimize packet loss. Additionally, MADRL-SA is specifically designed for practical scenarios and utilizes DQN, unlike MARL, which relies on Q-learning. The work by Li et al. (2019) adopts a single UAV approach, while MADRL-SA adopts a multi-UAV approach, offering advantages in terms of scalability and robustness. Our focus is on minimizing packet loss and providing velocity control, whereas the work by Zhang et al. (2017). (2017) prioritizes energy efficiency while neglecting velocity control. Furthermore, in the reviewed literature, the UAVs act independently without any explicit strategy for collaboration among them.

The existing literature in the fields of UASNets and DRL has yielded promising outcomes in tackling various challenges. Nevertheless, based on our current knowledge, no work has specifically focused on jointly optimizing cruise control and communication scheduling in the presence of multiple UAVs using DRL techniques. This presents an intriguing opportunity to explore innovative approaches to tackle this intricate problem.

Most of the works in Section 3.2 formulate MFGs to address energy efficiency in UASNets. For example, Wang et al. (2021b) propose an MFG formulation to minimize AoI and suggest the use of DDPG-MFG in a continuous action space to find the optimal solution. On the other hand, the works in Section 2.4 investigate resource allocation

to reduce AoI, however, the actions are adjusted either in continuous or discrete action spaces. For instance, Samir et al. (2022) formulate a resource allocation problem to reduce AoI and employ PPO in a discrete action space to find the optimal solution.

Existing literature in the field of UASNets and AoI has shown promising results in addressing various challenges related to trajectory optimization and communication scheduling. However, most of the existing work focuses on single UAV scenarios, where actions are optimized in either continuous or discrete action spaces. On the other hand, research on multi-UAV systems using MFG formulations primarily targets energy efficiency. This presents an opportunity to explore novel approaches that utilize MFG formulations and optimize actions in mixed-action spaces to minimize AoI.

This thesis addresses the problem of joint velocity control and data collection scheduling in UASNets by formulating it as an MMDP to minimize overall packet loss caused by buffer overflow and channel fading. To handle the large state and action spaces, we propose MADRL-SA, which is based on DQN and enables the optimization of ground sensor selection, UAVs' patrol velocity, and modulation scheme. Additionally, collaboration among UAVs is facilitated by allowing each ground sensor to maintain a history of UAV visits and share this information with other UAVs. Furthermore, this thesis formulates cruise control for multiple UAVs based on MFG to minimize the average AoI. We introduce MF-HPPO as a method to optimize the actions of UAVs in a mixed discrete and continuous action space. To capture temporal dependencies in the cruise control problem, we leverage an LSTM layer. By adopting these approaches, we aim to enhance the performance and efficiency of UASNets.

Chapter 3

Joint communication scheduling and velocity control in UAVs-assisted sensor networks: A deep reinforcement learning approach

In this chapter, we address the joint optimization of communication scheduling and velocity control for multiple UAVs in UASNets. We formulate this problem as an MMDP aiming to minimize packet loss caused by buffer overflows and communication failures. The MMDP network state comprises battery levels, data queue lengths of ground sensors, channel conditions, visit times, and waypoints along the trajectories of the UAVs. UAVs take actions to schedule ground sensors for data transmissions, determine modulation schemes, and adjust patrol velocities. Ground sensors record and share visit times with UAVs as evidence of other UAVs' communication schedules. The rest of this chapter is organized as follows. Section 3.1 dedicates to the problem statement, where the system model is presented and the joint optimization of the velocity control and communication schedule is formulated. In Section 3.2, multi-UAV DQN is developed and a new MADRL-SA scheme is designed to optimize the decision process of the MMDP, thereby optimizing the patrol velocities as well as the transmission schedule of the ground sensors. Performance evaluation is presented in Section 3.3. This paper is concluded in Section 3.4.

Most material included in this chapter is derived from the following scientific publications:

- **Y. Emami**, B. Wei, K. Li, W. Ni and E. Tovar, *Joint Communication Scheduling and Velocity Control in Multi-UAV-Assisted Sensor Networks: A Deep Reinforcement*

Learning Approach, in IEEE Transactions on Vehicular Technology, vol. 70, no. 10, pp. 10986-10998, Oct. 2021, doi: 10.1109/TVT.2021.3110801. Impact Factor: 6.239.

- **Y. Emami**, B. Wei, K. Li, W. Ni and E. Tovar,*Deep Q-Networks for Aerial Data Collection in Multi-UAV-Assisted Wireless Sensor Networks*, 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 669-674, doi: 10.1109/IWCMC51323.2021.9498726.

3.1 Problem Statement

3.1.1 System Model

The network contains J ground sensors and I UAVs. Our study focuses on the joint velocity control and communication scheduling under preconfigured UAV trajectories. The UAVs fly along pre-determined trajectories which consist of a large number of waypoints to cover all the ground sensors in the field. The trajectories of the UAVs can be predesigned according to the required network capacity Choi et al. (2014), coverage Li et al. (2019), or the UAVs' propulsion energy consumption Zeng and Zhang (2017). The optimization of UAV trajectories has been widely studied in the literature Zhao et al. (2020), Hu et al. (2020a), Wang et al. (2021a). The proposed MADRL-SA is generic to any given trajectory. The channel coefficient between the UAV i ($\in [1,I]$) and device j ($\in [1,J]$) at t is $h^i_j(t)$, which can be known by channel reciprocity. The modulation scheme of device j at t is denoted by $\phi_j(t)$. In particular, $\phi_j(t)$= 1, 2, and 3 indicates binary phase-shift keying (BPSK), quadrature-phase shift keying (QPSK), and 8 phase-shift keying (8PSK), respectively, and $\phi_j(t) \geq 4$ provides $2^{\phi_j(t)}$ quadrature amplitude modulation (QAM).

Let $h^i_j(t)$ denote channel gain between ground sensor j and UAV i. The transmit power of the ground sensor, denoted by $P^i_j(t)$, isLi et al. (2020a)

$$P^i_j(t) = \frac{ln\frac{k_1}{\varepsilon}}{k_2 h^i_j(t)^2}(2^{\phi_j(t)} - 1) \tag{3.1}$$

where k_1 and k_2 are channel constants, and ε denotes the required bit error rate (BER) of the channel. We consider that UAV i moves in low attitude for data collection, where the probability of LoS communication between UAV i and ground sensor j can be

$$Pr_{LoS}(\varphi^i_j) = \frac{1}{1 + aexp(-b[\varphi^i_j - a])} \tag{3.2}$$

Table 3.1: Notation and Definition

Notation	Definition
J	number of ground sensors
I	number of UAVs
a_u^{t-1}	past actions of other UAVs on a ground sensor
a_i	action of UAV i
$S_{\alpha,i}$	state of UAV i
$S_{\beta,i}$	next state of UAV i
$P_j^i(t)$	transmit power between device j and UAV i
$h_j^i(t)$	channel gain between device j and UAV i
$\zeta_i(t)$	location of the UAV on its trajectory
$v(t)$	velocity of the UAV
v_{max}, v_{min}	the maximum and minimum velocity of the UAV
$e_j(t)$	battery level of device j
$q_j(t)$	queue length of device j
TVR_p	Time of each visiting record
D	maximum queue length of ground sensor
$\phi_j(t)$	modulation scheme of device j
γ	discount factor for future states
θ	learning weight in deep Q-network

where a and b are constants, and φ_j^i denotes the elevation angle between UAV i and ground sensor j. Furthermore, path loss of the channel between UAV i and device j can be obtained by

$$\gamma_j^i = Pr_{LoS}(\varphi_j^i)(\eta_{LoS} - \eta_{NLoS}) + 20log(r\sec(\varphi_j^i)) + 20log(\lambda) + 20log(\frac{4\pi}{v_c}) + \eta_{NLoS} \quad (3.3)$$

where r denotes the radius of the radio coverage of UAV i, λ is the carrier frequency, and v_c is the speed of light. η_{LoS} and η_{NLoS} represent the excessive path losses of LoS or non-LoS, respectivelyAl-Hourani et al. (2014). Please See Appendix A.

3.1.1.1 Communication Protocol

Fig. 3.1 shows the data collection protocol for the UASNets. Specifically, the proposed MADRL-SA operates onboard at the UAVs to determine their velocities and sensor selection and allocate the modulation scheme for the selected sensors. The details of MADRL-SA will be provided in the next section. Next, the UAV broadcasts a short beacon message which contains the ID of the selected sensor. Upon the receipt of the beacon message, the selected sensor transmits its data packets to the UAV, along with the state information

of $e_j(t)$, $q_j(t)$, and TVR_p in the control segment of the data packet. Once the data is correctly received, the UAV sends an acknowledgment to the ground sensor.

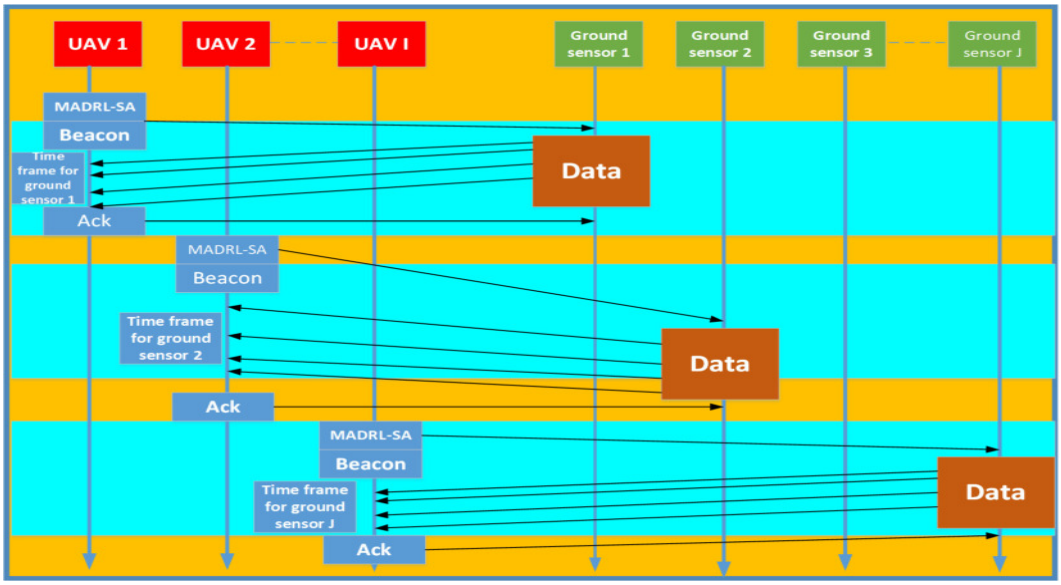


Figure 3.1: Data communication protocol for UASNets. MADRL-SA conducts velocity determination, sensor selection, and modulation scheme allocation in each communication frame

3.1.2 Problem Formulation

In this section, we present the problem formulation.

3.1.2.1 Optimization Formulation

Let $\kappa_j^i(t)$ be the binary indicator of ground sensor j being selected by UAV i for data transmission at time t. If ground sensor j is scheduled by UAV i at time t, $\kappa_j^i(t) = 1$; otherwise, $\kappa_j^i(t) = 0$. The joint optimization of UAV velocity and communication schedule aims to minimize the packet loss of all the ground sensors, as given by
Optimization problem:

$$\min_{\kappa_j^i(t), v_i(t), P_j^i(t)} \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij}(\kappa_j^i(t), v_i(t), P_j^i(t)) + \sum_{j=1}^{J} g_j(\kappa_j^i(t))$$

subject to:

$$0 \le P_j^i(t)\kappa_j^i(t) \le P_{max}, \tag{3.4}$$

where

$$f_{ij}(\kappa_j^i(t), v_i(t), P_j^i(t)) = \begin{cases} 1, & \text{if } (\kappa_j^i(t) = 1) \ \& \ (h_j^i(t) \le h_{th}) \ \& \ (v_i(t) \le v_{max}); \\ 0, & \text{otherwise,} \end{cases} \tag{3.5}$$

and

$$g_j(\kappa_j^i(t)) = \begin{cases} 1, & \text{if } (q_j(t) > D) \ \& \ (\kappa_j^i(t) = 0); \\ 0, & \text{otherwise,} \end{cases} \tag{3.6}$$

Constraint (3.4) ensures that the transmit power of the scheduled ground sensor does not exceed the maximum transmit power P_{max}.

3.1.2.2 MMDP Formulation

MMDP can be defined by the tuple $\{I, \{S_{\alpha,i}\}, \{a_i\}, C\{S_\beta \mid S_\alpha, a\}, \Pr\{S_\beta \mid S_\alpha, a\}\}$

1. I is the number of agents, i.e., UAVs.

2. $S_{\alpha,i}$ is the network state observed by agent i ($i \in I$). $S_{\alpha,i}$ comprises: channel quality $h_j^i(t)$, battery level $e_j(t)$, queue length $q_j(t)$, visit time TVR_p, and the location of UAV $\zeta_i(t)$, i.e., $S_{\alpha,i} = \{(h_j^i(t), e_j(t), q_j(t), TVR_p, \zeta_i(t)), i = 1, 2, \ldots, I\}$.

   In particular, each ground sensor maintains a list of visiting time of the agents. Joint state of all the agents is denoted S_α, where S_α=$S_{\alpha,1}$×.....×$S_{\alpha,I}$.

3. a_i represents the action of agent i. a_i is to schedule one sensor to transmit data to the UAV, determine the modulation and the instantaneous patrol velocity of the UAV, i.e., a_i=$\{(j, \phi_j(t), v(t)), i = 1, 2, \ldots.I\}$. Joint action a which consists of the actions of all the agents is a=a_1×.....×a_I. The size of action space is $J\Phi \mid v(t) \mid$, where Φ is the highest modulation order and $\mid v(t) \mid$ stands for the cardinality of the set $[v_{min}, v_{max}]$.

4. C$\{S_\beta | S_\alpha, a\}$ is the network cost yielded when joint action a is taken at joint state S_α and the following joint state changes to S_β. The network cost is the packet loss of the ground sensors.

5. $\Pr\{S_\beta \mid S_\alpha, a\}$ denotes the transition probability from joint state S_α to joint state S_β when joint action a is taken.

3.1.2.3 Transition Probability

The transition probability of the MMDP, from S_α to S_β can be given by

$$\Pr\{S_\beta \mid S_\alpha\} = \prod_{i=1}^{I} \left(\Pr\{(e_{\beta,j}, q_{\beta,j}, h_{\beta,j}, \zeta_{\beta,j}) \mid (e_{\alpha,j}, q_{\alpha,j}, h_{\alpha,j}, \zeta_{\alpha,j}), j \in a_i\}_i\right) \times \prod_{k=1}^{K} \left(\Pr\{(e_{\beta,k}, q_{\beta,k}, h_{\beta,k}, \zeta_{\beta,k}) \mid (e_{\alpha,k}, q_{\alpha,k}, h_{\alpha,k}, \zeta_{\alpha,k}), k \neq a_i; i \in [1,I]\}_i\right) \quad (3.7)$$

Specifically, the state transition probability presented in (3.7) consists of two parts. The first part, i.e., $\Pr\{(e_{\beta,j}, q_{\beta,j}, h_{\beta,j}, \zeta_{\beta,j}) | (e_{\alpha,j}, q_{\alpha,j}, h_{\alpha,j}, \zeta_{\alpha,j}), j \in a_i\}$is the state transition probability from S_α to S_β in terms of the selected ground sensor ($j \in a_i$). Let K denote the total number of unselected ground sensors. The second part, i.e.,

$\Pi_{k=1}^{K} \Pr\{(e_{\beta,k}, q_{\beta,k}, h_{\beta,k}, \zeta_{\beta,k}) | (e_{\alpha,k}, q_{\alpha,k}, h_{\alpha,k}, \zeta_{\alpha,k}, k \neq a_i; i \in [1,I]\}$

is the probability from S_α to S_β in terms of the unselected ground sensors, where $k \neq a_i; i \in [1,I]$ indicates the sensors that are not selected by any of the I agents.

Let $d_{i,j}$ denote the distance between ground sensor j and UAV i, $v(t)$ is velocity of the UAV , $R(t)$ is the data rate of the ground sensor and λ is the packet arrival probability. The state transition probability of the selected sensor j, which is specified in (3.8), depends on the following possible transitions.

1. Packet transmission is successful due to the good channel quality, i.e., $h_{\beta,j} > h_{\alpha,j}$ and low velocity. There is no packet arrival, the data queue of the selected node decreases, i.e., $q_{\beta,j} = q_{\alpha,j} - 1$. The state transition probability is $(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}}(1-\lambda)$.

2. Packet transmission is failed due to the poor channel quality, i.e., $h_{\beta,j} < h_{\alpha,j}$ and high velocity. A new data packet is generated and buffered, the data queue of the selected node increases, i.e., $q_{\beta,j} = q_{\alpha,j} + 1$. The state transition probability is $(1-(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}})\lambda$.

3. Packet transmission is successful due to the good channel quality, i.e., $h_{\beta,j} > h_{\alpha,j}$ and low velocity. A new data packet is generated and buffered, the data queue of the selected node remains unchanged, i.e., $q_{\beta,j} = q_{\alpha,j}$. The state transition probability is $(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}}\lambda$.

$$\Pr\{(e_{\beta,j},q_{\beta,j},h_{\beta,j},\zeta_{\beta,j}) \mid (e_{\alpha,j},q_{\alpha,j},h_{\alpha,j},\zeta_{\alpha,j}), j \in a_i\} = \begin{cases} (1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}}(1-\lambda) & \text{if } e_{\beta,j} = e_{\alpha,j} - \Delta e \text{ and } q_{\beta,j} = q_{\alpha,j} - 1 \\ & \text{and } h_{\beta,j} > h_{\alpha,j} \\ (1-(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}})\lambda & \text{if } e_{\beta,j} = e_{\alpha,j} - \Delta e \text{ and } q_{\beta,j} = q_{\alpha,j} + 1 \\ & \text{and } h_{\beta,j} < h_{\alpha,j} \\ (1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}}\lambda & \text{if } e_{\beta,j} = e_{\alpha,j} - \Delta e \text{ and } q_{\beta,j} = q_{\alpha,j} \\ & \text{and } h_{\beta,j} > h_{\alpha,j} \\ (1-(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}})(1-\lambda) & \text{if } e_{\beta,j} = e_{\alpha,j} - \Delta e \text{ and } q_{\beta,j} = q_{\alpha,j} \\ & \text{and } h_{\beta,j} < h_{\alpha,j} \end{cases} \tag{3.8}$$

$$\Pr\{(e_{\beta,k},q_{\beta,k},h_{\beta,k},\zeta_{\beta,k}) | (e_{\alpha,k},q_{\alpha,k},h_{\alpha,k},\zeta_{\alpha,k}, k \neq a_i; i \in [1,I]\} = \begin{cases} \lambda & \text{if } e_{\beta,k} = e_{\alpha,k} \text{ and } q_{\beta,k} = q_{\alpha,k} + 1 \\ 1-\lambda & \text{if } e_{\beta,k} = e_{\alpha,k} \text{ and } q_{\beta,k} = q_{\alpha,k} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

4. Packet transmission is failed due to the poor channel quality, i.e., $h_{\beta,j} < h_{\alpha,j}$ and high velocity. There is no packet arrival, the data queue of the selected node remains unchanged, i.e., $q_{\beta,j} = q_{\alpha,j}$. The state transition probability is $(1-(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}})(1-\lambda)$.

Due to the packet transmission, the battery level of the selected sensor decreases by Δe.

(3.9) corresponds to the unselected sensors with two different cases. The first case corresponds to the case when queue of the ground sensor increases, i.e., $q_{\beta,k} = q_{\alpha,k} + 1$ due to a new packet arrival, i.e., λ. The second case gives that the data queue remains unchanged, i.e., $q_{\beta,k} = q_{\alpha,k}$ since there is no packet arrival, i.e., $(1-\lambda)$.

By solving the formulated MDP, e.g., by using dynamic programming techniques, the optimal solution with complete states could be achieved, which could be used for performance benchmarking in multi-UAV-assisted wireless sensor networks. Unfortunately, dynamic programming (and the MDP formulation) suffers from the well-known curse-of-dimensionality, and incurs a prohibitive complexity and intractability, which can be noted in Appendix B. Please See Appendix B.

3.2 Proposal

Algorithm 1 MADRL-SA

1.Initialize:

Randomly initialize the networks

$Q_i\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}; \theta^{Q_i}\}$ with θ^{Q_i}

Initialize target networks Q_i' with weights $\theta^{Q_i'} = \theta^{Q_i}$

$\forall i \in (1, I)$

2.Learning:

for *episode=1 to M* **do**

  Obtain state $S_{\alpha,i}$

  for *t=1 to T* **do**

    if(Probability ε)

    Select a random action a_i

    else

    $a_i = argmin_{a_i} Q_i\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}; \theta^{Q_i}\}$

    end

    Execute action a_i in the environment

    Receive the visiting record

    for *p=1 to I* **do**

      if(i==p)

      $\delta[p]$=t

      else

      $\delta[p]$=t - TVR_p

      end

    end for

    Obtain the cost function $C_{t,i} = \{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}\}$ and the next state $S_{\beta,i}$ at $t+1$

    Store Transition $(S_{\alpha,i}, S_{\beta,i}, a_i, C_{t,i})$

    Sample random minibatch $(S_{\alpha,b}, S_{\beta,b}, a_b, C_{t,b})$

    $y_i = C\{S_{\beta,b} \mid S_{\alpha,b}, a_b, a_{ub}\} + \gamma min_{a_b'} Q_i'\{S_{\beta,b'} \mid S_{\beta,b}, a_b', a_{ub}'; \theta^{Q_i'}\}$

    Derive the loss function

    $\Gamma(\theta_i^Q) = y_i - Q_i\{S_{\beta,b} \mid S_{\alpha,b}, a_b, a_{ub}; \theta^{Q_i}\}$

    Update the target networks.

    $\theta^{Q_i'} = \theta^{Q_i}$

    $S_\alpha = S_\beta$

  end for

end for

3.2.1 Proposed MADRL-SA

We present a multi-UAV version of DQN called MADRL-SA, MADRL-SA realizes co-operation between UAVs, by enabling them to learn the scheduling decisions of each other.

According to Fig. 3.2, MADRL-SA has three UAVs, and each UAV is equipped with a classical DQN algorithm and learns through interaction by environment. As can be seen in Fig.3.2, UAV 3 performs its action and schedules a ground sensor, then receives its visiting record and consequently calculates the time differences $\delta[]$ between its visiting time(t) and TVR_p. $\delta[]$ is augmented to state and utilized in the learning process. Therefore, each UAV learns to coordinate its action. The UAVs that visited the same ground sensor would learn to improve their scheduling process based on computed timing information. For example, if the computed time differences are large the UAV is encouraged to schedule the ground sensor for the next time. Overall, our goal is to allow different UAVs schedule different ground sensors (other ground sensors may have buffer overflow probability) and if a ground sensor recently visited by an UAV no other UAV visits that ground sensor. The proposed scheme is described in Algorithm 1, which optimizes the actions based on the multi-UAV DQN to solve the online resource allocation problem.

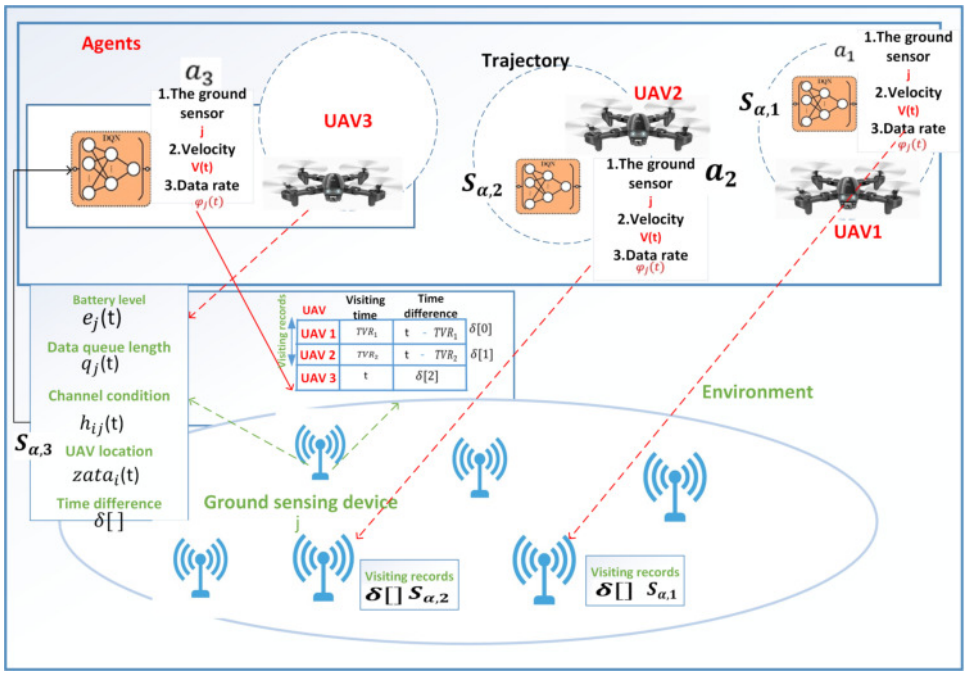


Figure 3.2: Overview of MADRL-SA: UAVs observe the current environment state, follow their policy, and take actions

Overall, two separate Q-networks are maintained with each UAV, Q-network: $Q_i\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}; \theta^{Q_i}\}$ and target network: $Q_i'\{S_{\beta,i'} \mid S_{\beta,i}, a_i', a_u'; \theta^{Q_i'}\}$, with weights θ^{Q_i} and $\theta^{Q_i'}$

respectively. At first step, Q-network and associated target of each UAV are initialized and then learning is ignited. Each UAV samples its state and computes its local state $S_{\alpha,i}$ including $\delta[]$. Each UAV receives the local state $S_{\alpha,i}$ and selects a random action with probability ε or exploits its knowledge and produce its action. Each UAV executes the selected action and computes the vector of δ using t and TVR_p; then corresponding cost and next state including $\delta[]$ are sampled. Then the associated transition $(S_{\alpha,i}, S_{\beta,i}, a_i, C)$ is stored. θ^{Q_i} is learned by sampling batches of transitions from the replay memory and minimizing the squared temporal difference error:

$$\Gamma(\theta^{Q_i}) = y_i - Q_i\{S_{\beta,b} \mid S_{\alpha,b}, a_b, a_{ub}; \theta^{Q_i}\} \tag{3.10}$$

where

$$y_i = C\{S_{\beta,b} \mid S_{\alpha,b}, a_b, a_{ub}\} + \gamma min_{a'_b} Q'_i\{S_{\beta,b'} \mid S_{\beta,b}, a'_b, a'_{ub}; \theta^{Q'_i}\} \tag{3.11}$$

finally for each agent the parameters of a Q-network θ^{Q_i} copied into those of target network $\theta^{Q'_i}$ after a constant number of iterations. The proposed MADRL-SA can be readily repurposed to support different objective functions. For example, it can be potentially repurposed to maximize the energy efficiency, which is the ratio of network throughput to the energy consumption.

3.2.2 Energy and Feasibility

UAVs are becoming increasingly less restrictive in terms of energy due to new advancements of battery and energy harvesting technologies. For example, Atlantik Solar has developed an autonomous, solar-powered drone (UAV) that can fly up to 10 days continuously. A ground sensor can be equipped with solar panels, wind power generators or other energy harvesting mechanisms to harvest renewable energy from ambient resources and recharge its battery.

The UAVs select the optimal sensors to transmit data and allocate their modulation schemes, by learning the states of the ground sensors. The selected sensor uses the allocated modulation to transmit data to the UAV, while updating the visiting time of the UAV. In particular, the historical record of the visiting time typically has a small size. Consider 100 UAVs, the size of the historical record at the sensor is just seven bits. The time for updating the record is negligible. Also, the sensors only need to synchronize with the UAVs the recent historical record of visits. The overhead is small. Therefore,

the proposed deep reinforcement learning based data collection requires a small amount of computation at the sensors, which is feasible and practical in real-world UASNets

3.2.3 Complexity of MADRL-SA

The time complexity for training each network Q_i that has Z layers with z_i neurons per layer is given by,

$$\mathcal{O}(MT \times (\Sigma_{i=1}^{Z-1} z_i z_{i+1})) \tag{3.12}$$

where M is the number of episodes and T is the number of iterations. Therefore, the time complexity of MADRL-SA with I networks of Q_i is given by

$$\mathcal{O}(I \times MT \times (\Sigma_{i=1}^{Z-1} z_i z_{i+1})) \tag{3.13}$$

The case of an equal number of neurons in each layer, the time complexity can be written as

3.3 Evaluation

3.3.1 Implementation of MADRL-SA

J number of ground sensors are randomly deployed, where J increases from 20 to 120. Each ground sensor has the maximum discretized battery capacity 50 Joules, the highest modulation = 5, and the maximum transmit power 100 milliwatts. For calculating $P_j^i(t)$ of the ground sensor, the two channel constants, k_1 and k_2 are set to 0.2 and 3, respectively. The required BER is 0.05, and the carrier frequency is 2000 MHz. ε is set to 0.05. However, the value of ε can be configured based on the traffic type and quality-of-service (QoS) requirement of the user's data, as well as the transmission capability of the UAV. Other simulation parameters are listed in Table 3.2. Moreover, the region of interest is set to be a square area with a size of 1000 x 1000 meters, where the ground sensors are distributed in the targeted region. MADRL-SA is implemented in Python 3.5 using Pytorch (the Python deep learning library). A Lenovo Workstation running 64-bit Ubuntu 16.04 LTS, with Intel Core i5-7200U CPU @ 2.50GHz × 4 and 8 G memory is used for the PyTorch setup. DRL trains MADRL-SA for 1000 episodes. The discount factor and learning rate are set to 0.99 and 0.001, respectively. We use 2-layer fully connected neural network for each agent, which includes 400 and 300 neurons in the first and second layers, respectively. We utilize the rectified linear unit (ReLU) function for the activation function. The experience replay memory with the size of 10^6 is created for each agent to

store the learning outcomes in the format of a quadruplet <state, action, cost, next state>. The memory is updated by calling the function replay buffer$_i$.add((state, action, cost, next state)), and retrieves the experiences by using replay buffer$_i$.sample(batch size).

Table 3.2: PyTorch Configuration

Parameters	Values
Number of ground sensors	20-120
Queue length	40
Energy levels	50
Discount factor	0.99
Learning rate	0.001
Replay memory size	10^6
Batch size	100
Number of episodes	1000

3.3.2 Baseline Description

For performance evaluation, the proposed MADRL-SA is compared with Random scheduling policy (RSA), Channel scheduling policy (CHSA) and DRL-SA Li et al. (2019) algorithms.

- RSA randomly determines the velocities of the UAVs at each waypoint, and one of the ground sensors within the communication range of the UAV is randomly selected to transmit data. The velocity control and sensor selection are independent of the batteries, data queue lengths of the ground sensors, channel variation, and UAVs' positions.

- CHSA allows the UAVs to move with the minimum velocity and schedule the ground sensors based on their channel quality. Each UAV sends beacons along the trajectory. Based on the sensors' replies to the beacons, the UAV measures the channel gains. The ground sensor with the highest channel gain is selected to transmit.

- DRL-SA enables a single-agent DQN, where each UAV leverages DQN to learn the optimal velocity control and sensor selection strategy based on the data queue length, energy level, channel variation and UAV's positions. The selection of the ground sensor, modulation scheme, and velocity of the UAV is jointly optimized (independently of the rest of the UAVs).

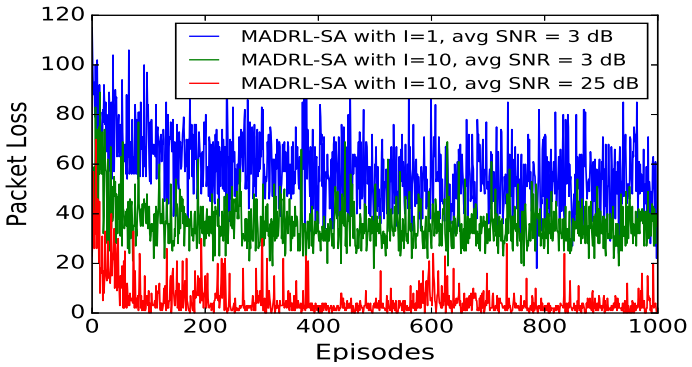


Figure 3.3: Network cost at each episode of MADRL-SA with I = 10 and DRL-SA. .

3.3.3 Performance Analysis of MADRL-SA

Fig. 3.3 depicts the convergence of MADRL-SA with I=10 for low and high SNR cases and DRL-SA. MADRL-SA with I=10 and high SNR show the best performance since it reduce the overflow cost as well as the fading cost due to good SNR. MADRL-SA with I=10 and low SNR outperform the DRL-SA which has the highest network cost. The reason is that when multiple UAVs act it results in the reduction of overflow cost.

Fig. 3.4 depicts the network cost of MADRL-SA (data queue length=40) and the baselines in term of ground sensors. MADRL-SA with I=5 and I=10 achieves a lower network cost in comparison to CHSA. The network cost of MADRL-SA with I=5 is lower than that of CHSA . Overall, MADRL-SA with I=5 and I=10 outperforms CHSA. Particularly, when J=100 the packet loss of MADRL-SA with I=5 and I=10 is lower than CHSA by around 21% and 40%, respectively.

Fig. 3.5 shows the trade-off between the number of ground sensors and UAVs. Specifically, a large number of ground sensors expedites the buffer overflows in UASNets and in

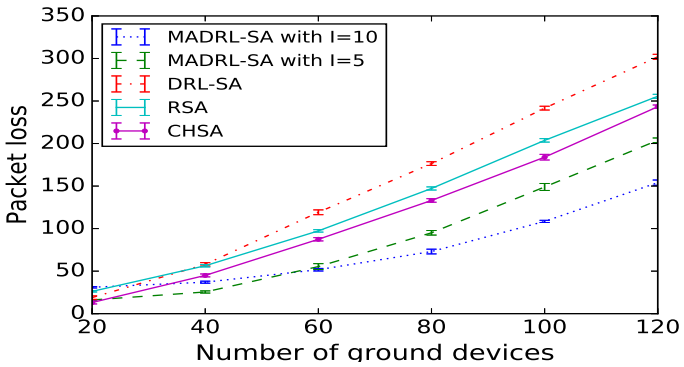


Figure 3.4: Comparison of packet loss between MADRL-SA and the baselines in terms of ground sensors.

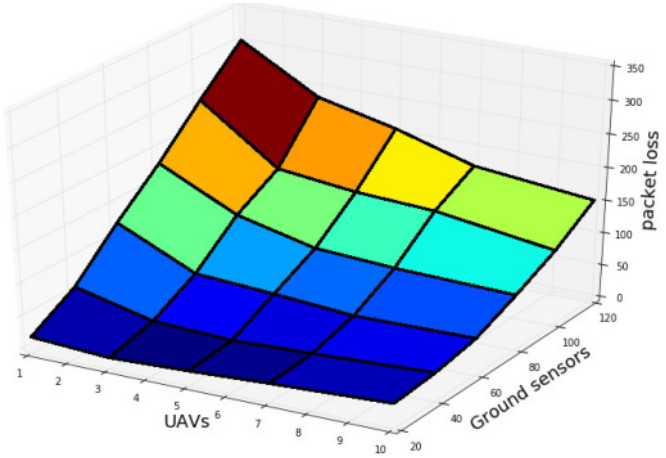


Figure 3.5: Trade-off between the number of UAVs and ground sensors.

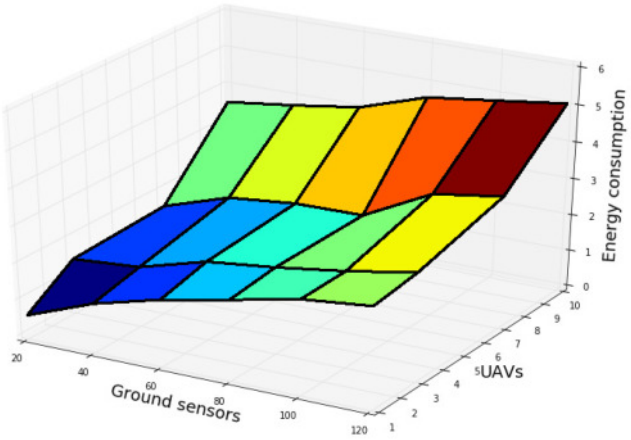


Figure 3.6: Energy consumption of ground sensors.

turn, increases the packet loss. On the other hand, increasing the number of UAVs allows the ground sensors to be scheduled in parallel, hence reducing the buffer overflow. A balance needs to be struck between the numbers of UAVs and ground sensors to minimize the packet loss.

Fig. 3.6 shows the energy consumption of the ground sensors by varying the number of ground sensors and UAVs. For a given number of UAVs, the energy consumption of the network increases with the number of ground sensors. On the other hand, the increasing number of UAVs helps increase the number of ground sensors scheduled to transmit data, hence raising the energy consumption of the ground sensor network.

Fig. 4.5 show the velocities and trajectories of different UAVs for the MADRL-SA with I=7. Fig. 4.5(a) demonstrates the velocity of 7 UAVs given 20 waypoints. The color bar shows the range of values for velocity and color map shows the actual velocity of each

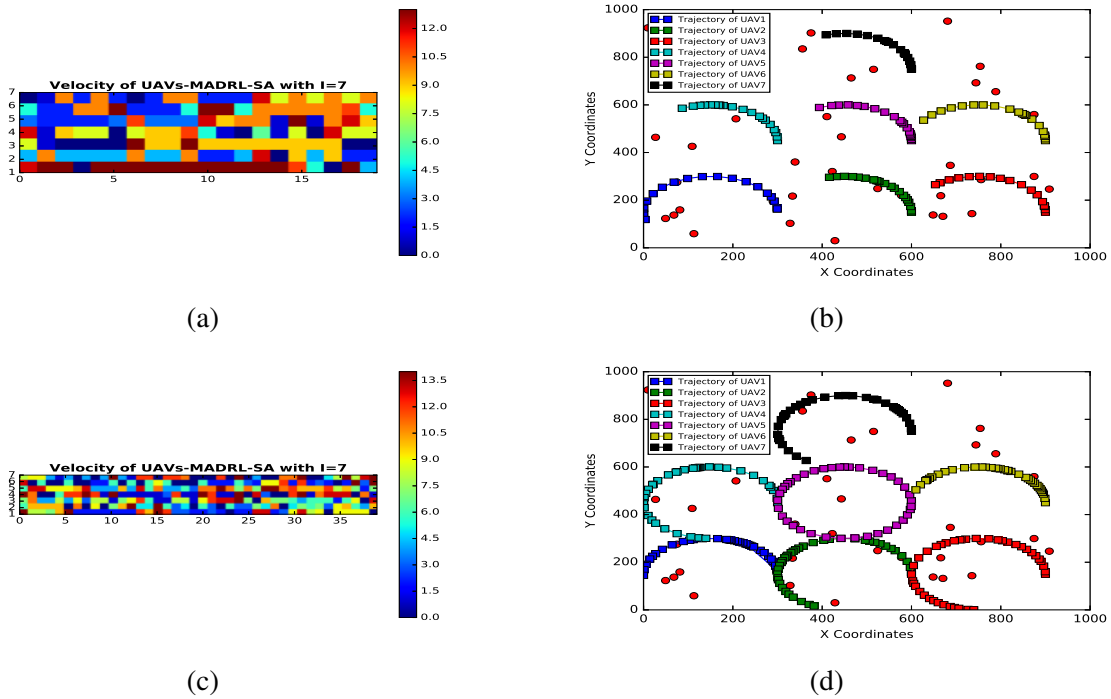


Figure 3.7: Velocities and trajectories of MADRL-SA with I=7.(a) and (b)velocity and trajectory given number of waypoints=20. (c) and (d) velocity and trajectory given number of waypoints=40

UAV for each waypoint in color format. As can be seen UAV 2 moves with the lowest velocity as confirmed by its small trajectory in Fig. 4.5(b). In contrast, UAV 1 moves with the highest velocity as confirmed by its trajectory. Overall, for waypoints 1-12, UAV 3-7 move with the lowest velocity witnessing subtle changes. After these waypoints the velocity of these UAVs is increasing.

Fig. 4.5(c) is similar to Fig. 4.5(a) except that number of waypoints is increased to 40. Overall, the pattern for all UAVs except UAV 5 is almost similar and all of them move with low or moderate velocity witnessing high velocity at some points, this can be confirmed by their associated trajectories in Fig. 4.5(d). UAV 5 moves smoothly before waypoint 20. After this point its velocity start increasing and hence a full trajectory is shaped as can be seen in Fig. 4.5(d).

Fig. 3.8 evaluates the network cost with the increasing number of UAVs, where the buffer size of MADRL-SA is set to 20 or 40 and the number of ground sensors is 40. For MADRL-SA with buffer size of 40, increasing the number of UAVs from 3 to 10 leads to a reduction of the packet loss by 68%. In contrast, when the buffer size is 20, a reduction of 77% in the packet loss is witnessed. Fig. 3.8 also shows that MADRL-SA significantly outperforms RSA by 80% when the buffer size is 40, and by 34% when the buffer size is

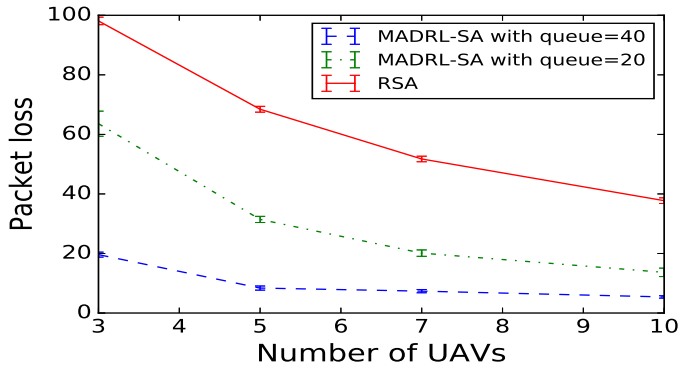


Figure 3.8: Network cost with an increasing number of UAVs, where the data queue length of MADRL-SA is set to 20 and 40 and number of ground sensors as 40.

20.

Fig. 3.9 demonstrates the training performance with varied learning rates(lr). After few episodes in the beginning, the network cost have an obvious tendency to decrease and converge in the case of lr=1e-3 and lr=5e-4. Nevertheless, the algorithm may converge to a local optimum in case of large learning rate, this situation can be seen in the case of lr=1e-1 and lr=1e-2.

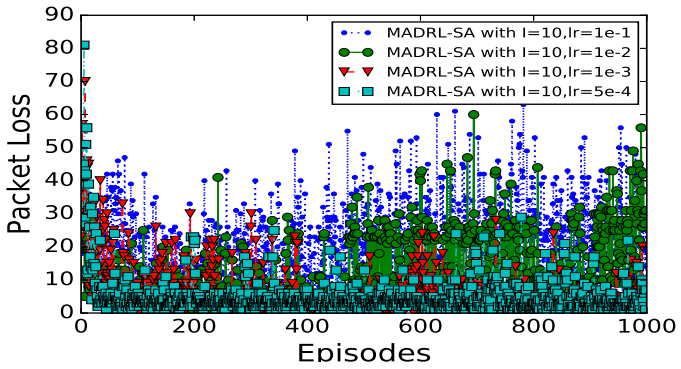


Figure 3.9: Training performance with varied learning rates.

3.4 Summary

In the first chapter, we study the joint flight cruise control and data collection scheduling in the UASNets. We formulate the problem using MMDP to minimize the packet loss due to buffer overflows at the ground sensors and fading airborne channels. We propose MADRL-SA to solve the formulated MMDP, where all UAVs utilize DQN to conduct respective decisions. In MADRL-SA, the UAVs acting as agents learn the underlying patterns of the data and energy arrivals at all the ground sensors as well as the scheduling decisions of the other UAVs. We conduct simulation using PyTorch deep learning library and results reveal that the proposed MADRL-SA for UASNets reduces packet loss by up to 54% and 46%, as compared to the single agent case and existing non-learning greedy algorithm, respectively.

Appendix A

The path loss of the LoS link is given by

$$PL_{LOS} = 20\log d + 20\log f + 20\log(\frac{4\pi}{c}) + \eta_{LOS} \tag{3.14}$$

The path loss of the non-LoS link is given by

$$PL_{NLOS} = 20\log d + 20\log f + 20\log(\frac{4\pi}{c}) + \eta_{NLOS} \tag{3.15}$$

The LoS probability is given by

$$Pr_{LOS} = \frac{1}{1 + aexp(-b[\varphi_j^i - a])} \tag{3.16}$$

Then, the NLoS probability is

$$Pr_{NLOS} = 1 - Pr_{LOS} \tag{3.17}$$

The expectation of the path loss between UAV i and device j can be obtained by

$$\gamma_j^i = Pr_{LOS} \times PL_{LOS} + Pr_{NLOS} \times PL_{NLOS} \tag{3.18}$$

By substituting (4.23) into (4.25), we have

$$\gamma_j^i = Pr_{LOS}(PL_{LOS} - PL_{NLOS}) + PL_{NLOS} \tag{3.19}$$

Substituting (3.14),(4.22),(4.24) into (3.19) leads to

$$\gamma_j^i = \frac{(\eta_{LOS} - \eta_{NLOS})}{1 + aexp(-b[\varphi_j^i - a])} + 20\log d + 20\log f + 20\log(\frac{4\pi}{c}) + \eta_{NLOS} \tag{3.20}$$

Rewriting 3.20 in term of φ_j^i and r, we finally obtain

$$\gamma_j^i = \frac{(\eta_{LOS} - \eta_{NLOS})}{1 + aexp(-b[\varphi_j^i - a])} + 20log(r\sec(\varphi_j^i)) + 20\log(\lambda) + 20\log(\frac{4\pi}{c}) + \eta_{NLOS} \tag{3.21}$$

Appendix B

Let ε denote the bit error rate, L denote the data packet length and λ denote the packet arrival probability. Depending on the transmission status and arrival pattern, four transitions may happen as presented in (3.8):

1. In the first case, the packet transmission is successful $(1-\varepsilon)^L$ and there is no packet arrival $(1-\lambda)$. The probability of such transition is $(1-\varepsilon)^L \times (1-\lambda)$. Given L=R(t)*T where T is the conversation time of UAV i and ground sensor j, and $T = \frac{2d_{i,j}}{v(t)}$. We have $L = \frac{2d_{i,j}R(t)}{v(t)}$ by substituting T into L. Therefore, the transition probability of the first case is $(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}}(1-\lambda)$.

2. In the second case, the packet transmission is not successful $(1-(1-\varepsilon)^L)$ and there is packet arrival λ. The probability of such transition is $(1-(1-\varepsilon)^L) \times \lambda$. By

substituting T into L, we have $L = \frac{2d_{i,j}R(t)}{v(t)}$. Therefore, the transition probability of the second case is $(1-(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}})\lambda$.

3. In the third case, the packet transmission is successful $(1-\varepsilon)^L$ and there is packet arrival λ. The probability of such transition is $(1-\varepsilon)^L \times \lambda$. By substituting T into L, we have $L = \frac{2d_{i,j}R(t)}{v(t)}$. Therefore, the transition probability of the third case is $(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}}\lambda$.

4. In the fourth case, the packet transmission is not successful $(1-(1-\varepsilon)^L)$ and there is no packet arrival $(1-\lambda)$. The probability of such transition is $(1-(1-\varepsilon)^L) \times (1-\lambda)$. We have $L = \frac{2d_{i,j}R(t)}{v(t)}$. Therefore, the transition probability of the fourth case is $(1-(1-\varepsilon)^{\frac{2d_{i,j}R(t)}{v(t)}})(1-\lambda)$.

(3.9) investigates the transmission probabilities for unselected ground sensors. These ground sensors do not transmit data. In this case, the ground sensors either receive packet with transition probability λ or no packet is received with transition probability $1-\lambda$.

Chapter 4

Age of Information Minimization using Multi-agent UAVs based on AI-Enhanced Mean Field Resource Allocation

In this chapter, we introduce a cruise control approach based on MFG theory to minimize the AoI, while balancing the trade-off between UAVs' movements and AoI. This method reduces the complexity of the cruise control problem and enhances optimization of UAVs' movements. However, in practice, obtaining instantaneous knowledge of the UAV's cruise control decision and AoI is challenging, making the proposed MFG difficult to solve online. We formulate MMDP, with network states comprising the AoI of ground sensors and waypoints of the UAV swarm. The MMDP action space includes continuous waypoints and velocities, as well as discrete transmission schedules. We propose a mean field hybrid proximal policy optimization (MF-HPPO) approach. The rest of this chapter is organized as follows: In Section 4.1, we present the system model in which the channel model as well as the AoI in the UASNets is formulated. Moreover, we formulate the flight resource allocation of the UAV swarm as the MFG to minimize the AoI. Section 4.2 develops the proposed MF-HPPO, to jointly optimize the cruise control of multiple UAVs and data collection scheduling. Section 4.3 presents the implementation of the proposed MF-HPPO in Pytorch as well as performance evaluation. Finally, Section 4.4 concludes this paper.

The techniques in this chapter have been discussed in the following papers.

- **Y. Emami**, H. Gao, K. Li, L. Almeida, E. Tovar, and Z. Han, *Age of Information Minimization using Multi-agent UAVs based on AI-Enhanced Mean Field Resource*

Allocation, IEEE Transactions on Vehicular Technology, 2023, under review.

- **Y.Emami**, K. Li, Y. Niu and E. Tovar, *AoI Minimization Using Multi-Agent Proximal Policy Optimization in UAVs-Assisted Sensor Networks*, ICC 2023-IEEE International Conference on Communications, Rome, Italy, 228-233. doi: 10.1109/ICC 45041.2023.10278748

4.1 Problem Statement

4.1.1 System Model

In this section, we present the system model of the considered UAVs-assisted sensor network. Notations used in this paper are summarized in Table 4.1. The system consists of I UAVs, $i \in [1,I]$ and J ground sensors, $j \in [1,J]$ in which the ground sensors are deployed in a target region. The UAVs are employed to patrol in the target zone while collecting the sensory data. Fig. 4.1 depicts an example of UASNets along with mean field representation. With the increase in the number of UAVs in Fig. 4.1 the interactions between them become complex and can dominate the overall behavior of the system. MFG designed to deal with the optimal control problem involving a large number of players. It has unique characteristics suitable for UAV swarm and modelling these interactions. Each UAV seeks to minimize the AoI according to the actions of other agents surrounded. As depicted, the UAV consider the mean field effect of the other UAVs, which represents the collective behavior of the UAVs in the system. The coordinates (x_i, y_i, z_i) and $(x_j, y_j, 0)$ represent the position of UAV i and ground sensor j, respectively. The UAVs fly to the ground sensors, collect sensory data, and then their operation is terminated. The UAVs fly at a constant altitude, represented by $\zeta_i(t) = (x_i, y_i, z)$. The distance between ground sensor j and UAV i is $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + z^2}$. For the safety of the UAV during flight by preventing it from exceeding the maximum safe speed or stalling, we denote the maximum and minimum velocity of the UAV as v_{max} and v_{min}, respectively.

We consider that UAV i moves in low attitude for data collection, where the probability of LoS communication between UAV i and ground sensor j is given by Al-Hourani et al. (2014)

$$\Pr_{LoS}(\varphi_j^i) = \frac{1}{1 + a\exp(-b[\varphi_j^i - a])} \tag{4.1}$$

where a and b are constants, and φ_j^i denotes the elevation angle between UAV i and ground sensor j. Moreover, path loss of the channel between UAV i and device j can be modeled

by

$$\gamma_j^i = \Pr_{LoS}(\varphi_j^i)(\eta_{LoS} - \eta_{NLoS}) + 20\log\left(r\sec(\varphi_j^i)\right) + 20\log(\lambda) + 20\log\left(\frac{4\pi}{v_c}\right) + \eta_{NLoS} \quad (4.2)$$

where r is the radius of the radio coverage of UAV i, λ is the carrier frequency, and v_c is the speed of light. η_{LoS} and η_{NLoS} are the excessive path losses of LoS or non-LoS, respectively.

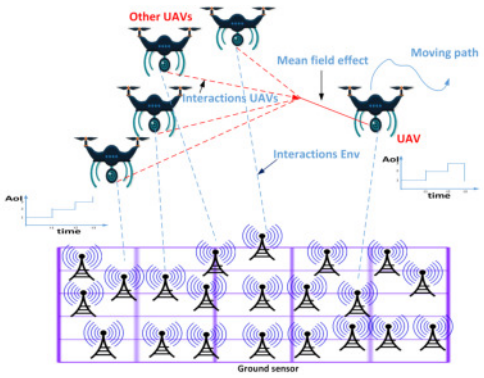


Figure 4.1: Mean field representation of UASNets.

To characterize the freshness of the collected sensory data at the UAV, AoI is defined as the time that has passed since ground sensor generates the latest information. The AoI of ground sensor j that generated a data packet at t_j and collected by UAV i at t_i is given by

$$AoI_j^i(t) = t_i - t_j. \quad (4.3)$$

According to (4.3), it can be also known that maintaining a low $AoI_j^i(t)$ is critical for improving the effectiveness and timeliness of the sensory data, reducing the response time, and providing real-time information for decision-making at the UAVs.

4.1.2 Problem Formulation

In this section, we formulate the MFG optimization with a large number of UAVs to address the trade-off between the cruise control of the UAVs and AoI. We also explore the FPK equation to determine the optimal velocities of the UAVs while characterizing the collective behavior of the UAVs. We begin with optimal control formulation in Section 4.1.2.1 and then proceed with MFG formulation in Section 4.1.2.2.

Table 4.1: Notation and Definition

Notation	Definition
J	number of ground sensors
I	number of UAVs
$h_j^i(t)$	channel gain between device j and UAV i
$\zeta_i(t)$	location of the UAV on its trajectory
$v_i(t)$	velocity of UAV i
v_{max}, v_{min}	the maximum and minimum velocity of UAV i
M	number of episodes
L	length of each episode
γ	discount factor
η	learning rate
D	buffer size
B	mini-batch size
a_i	action of UAV i
o_i	mean field of UAV i
a_i^c	continuous action of UAV i
a_i^d	discrete action of UAV i
$s_{\alpha,i}$	state of UAV i
$E[..]$	mathematical expectation
A	advantage function
θ	network parameter
π	policy
π^c	continuous policy
π^d	discrete policy
σ	diffusion coefficient
W	weiner process
H	entropy

4.1.2.1 Optimal Control Formulation

We derive the state dynamics and cost function, then we formulate the velocity control problem using the optimal control theory.

1. Time-varying Dynamics of Network States: Let $\zeta_i(t)$ denote the position of the UAV i at time t and $v_i(t)$ denotes the velocity. According to Newton's laws of motion Waldrip et al. (2013), the location dynamics of UAV i can be expressed by

$$d\zeta_i(t) = v_i(t)dt + \sigma dW_i(t) \tag{4.4}$$

where $W_i(t)$ is a standard Wiener process Mörters and Peres (2010) with a diffusion coefficient σ.

2. Cost Function: Each UAV intends to optimize its velocity to minimize the cost function. Our cost is defined as the average AoI of all ground sensors. The average AoI can be computed as:

$$c(t) = \frac{1}{IJ}\Sigma_{j=1}^{J}\Sigma_{i=1}^{I} AoI_j^i(t). \tag{4.5}$$

3. Velocity Control Problem Formulation: Given a period of time T regarding the data collection, the velocity of UAV i at t, denoted as $v_i^*(t)$, is optimally controlled to minimize $c(t)$, which gives:

$$v_i^*(t) = \arg\min_{v_i(t)} E\left[\int_0^T c(t)dt)\right], \tag{4.6}$$

$$s.t.\ (4.4).$$

To determine $v_i^*(t)$ in (4.6), classical game theories, such as differential game, fails to capture the aggregate behavior of all the UAVs. Differential game assumes each agent's movement is independent of others. This assumption fails to capture the fact that a large number of UAVs' trajectories decisions are influenced by the aggregate behavior of all the UAVs, thus hardly minimizing the average AoI, $c(t)$.

We novelly extend MFG to capture the impact of the aggregate behavior of the UAVs, in terms of cruise control. The MFG models the aggregate decision of UAVs as a probability distribution, rather than focusing on the actions of individual UAVs. This recognizes that the cruise control of each UAV is influenced by the behavior of all other UAVs. Moreover, the formulated MFG is defined to minimize $c(t)$ given a large number of UAVs, which classical game theory struggles with due to the computational complexity of solving for the equilibrium.

4.1.2.2 MFG Problem Formulation

We reformulate the optimal cruise control problem in (4.6) into a cooperative MFG problem. The computational complexity of the system is greatly reduced by formulating an MFG, since a large number of interactions with other agents is converted into an interaction with the mass. The interaction between each UAV with the other UAVs is modeled as a mean-field term, which is denoted by $m(\zeta(t))$. The mean-field term is the distribution over agents´ state space or control to model the overall state and control of them. We can measure the state and control of all agents in an MFG using the mean-field term.

Given dynamics, $\zeta_i(t)$, the mean-field term of $m(\zeta(t))$ can be denoted by

$$m(\zeta(t)) = \lim_{I\to\infty} \frac{1}{I} \Sigma_{i=1}^{I} \mathbb{1}\{\zeta_i(t) = \zeta(t)\}, \tag{4.7}$$

where $\mathbb{1}$ is an indicator function which returns 1 if the given condition is true, or 0, otherwise.

Given $m(\zeta(t))$, the state dynamics, cost function and FPK equation can be defined as:

- **State dynamics:** The state dynamics of each UAV can be expressed by

$$d\zeta(t) = v(t)dt + \sigma dW(t). \tag{4.8}$$

- **Cost function:** The mean-field term affects the running cost function of each UAV. The average AoI of the all UAVs is computed by

$$c(v(t), m(\zeta(t))) = \int c(v(t)) \cdot m(\zeta(t)) d\zeta. \tag{4.9}$$

  Mathematically, the cost function can be written by

$$J(v(t), m(\zeta(t)))) = \int_{t=0}^{T} c(v(t), m(\zeta(t)) dt. \tag{4.10}$$

  If the UAV move quickly, lead to poor channel condition and retransmissions thereby AoI prolongs. In contrast, slow movement of the UAV, may prolong the AoI of the ground sensors because the data are not collected in time. The cost function addresses these trade-offs and find the optimal velocity to balance these objectives.

- **Focker-Planck equation:** Based on (4.8) we develop the FPK equation. The FPK equation governs the evolution of the mean field function of UAVs and given by:

$$\partial_t m(\zeta(t)) + \nabla_\zeta m(\zeta(t)) \cdot v(t) - \frac{\sigma^2}{2} \nabla_\zeta^2 m(\zeta(t)) = 0. \tag{4.11}$$

  See *Appendix.*

After deriving the state dynamics, cost function, and FPK equation, we now proceed to present the MFG.

To summarize, the cooperative MFG problem is given by

$$\min_{v,m} J(v(t), m(\zeta(t))) \tag{4.12}$$

$s.t.$ (4.11).

4.2 Proposal

4.2.1 Proposed MF-HPPO

In this section, we describe the MFG as an MMDP in Section 4.2.1.1 so that the optimal actions of UAVs can be learned by the proposed MF-HPPO. MF-HPPO is presented in Section 4.2.1.2, which employs onboard PPO to minimize the average AoI of the ground sensors. The trajectory and instantaneous speed of the UAVs, and the selection of the ground sensors are optimized in a mixed action space. In Section 4.2.1.3, an LSTM layer is developed with MF-HPPO to capture the long-term dependency of data.

4.2.1.1 MMDP Formulation

We reformulate the MFG using MMDP framework to enable the application of PPO for optimizing the actions and minimizing average AoI. By adapting the MMDP framework to our problem, we define the relevant state space, action space, transition probabilities, policy and cost function, thus facilitating an effective solution approach based on MF-HPPO. We define our MMDP as follows.

- *Agents*: the number of agents, i.e., UAVs is denoted by I.
- *State*: A state s_α of the MMDP consists of the positions of UAV i, the AoI of ground sensors, i.e, s_α=$\{\zeta_i(t), AoI^i_j(t) : i \in [1,I], j \in [1,J]\}$. All states of the MMDP constitute the state space.
- *Action*: Each UAV i takes an action a_i that schedules a ground sensor for data transmission and determines the flight trajectory and velocity, i.e, $a_i = \{k^i_j, v_i(t), \zeta_i(t)\}$
- *Policy*: Policy π_i is the probability of taking each action of agent i.
- *State Transition*: The current state s_α transit to a new state s_β according to probability $P(s_\beta \mid s_\alpha, a)$, where a indicates a joint action set that includes the actions of all the UAVs.
- *Cost*: The immediate cost of the UAVs is $\frac{1}{IJ}\Sigma^J_{j=1}\Sigma^I_{i=1}AoI^i_j(s_\alpha, a)$.

4.2.1.2 MF-HPPO

The proposed MF-HPPO operates onboard at the UAVs to determine their trajectories and sensor selection. The UAV chooses a sensor and moves to it, then sends out a short beacon message with the ID of the chosen sensor. Upon the receipt of the beacon message, the selected sensor transmits its data packets to the UAV, along with the state information of $AoI_j^i(t)$ in the control segment of the data packet. After the UAV correctly receives the data, it sends an acknowledgement to the ground sensor.

The following equation highlights the mean field idea of MF-HPPO Yang et al. (2018):

$$Q_i(s_{\alpha,i}, a) = \frac{1}{N_i}\Sigma_{k\in N(i)} Q_i(s_{\alpha,i}, a_i, a_k) = Q_i(s_{\alpha,i}, a_i, o_i). \quad (4.13)$$

Here, Q_i is the Q value of agent i, a represents the joint action of all agents. The neighbor agents of agent i are characterized by N_i. o_i is an indicator of the mean field. In essence, in multi-agent systems the Q value of an agent is computed based on the current state and joint action, but when we have a large number of agents computing joint action is impractical, therefore (4.13) allow an agent to compute its Q value just based on the mean field of its neighbors.

Fig. 2 shows the proposed MF-HPPO with LSTM layer, where each UAV equipped with the MF-HPPO to minimize the average AoI by optimizing the trajectory and data collection schedule. The use of the LSTM layer, continuous and discrete actors, and the objective function of PPO, are the features of the MF-HPPO in this diagram. As shown, The decision-making component of each agent consists of two actors and a critic, which is preceded by the LSTM layer to draw conclusions based on experience. The actor for continuous action spaces outputs continuous values for cruise control, such as position and velocity, and the actor for discrete action spaces outputs a categorical value that can be used to select one of the ground sensors. Each agent samples the actions and performs in the environment. The rollout buffer is filled with data generated by these interactions such as, state, mean field, action, cost and policy. As can be seen, we use Generalized Advantage Estimate (GAE) Schulman et al. (2015a) as a sample-efficient method to estimate the advantage function. As depicted, based on the RolloutBuffer, mini-batches are then formed to train the LSTM and the actors and critics so that the agent can continuously improve its policies. The definition of the objective function of PPO is the total of actor losses and critic loss subtracted by entropy, as depicted in the diagram. The actor loss is inputted by the ratio of old policy and current policy and the advantage value. The critic loss is inputted by the critic's output and the return value. The policy is designed to encourage the agent to take advantageous actions, while punishing

actions that deviate from the current policy.

Algorithm 2 summarizes the MF-HPPO with the LSTM-based characterization layer. In the initialization step, Input and Output are characterized; the algorithm receives parameters like Clip threshold, discount factor and mini-batch size as input and specify its output as trajectory and scheduling policy of UAV i. Next, the actor π_i and critic w_i are initialized with random weights for each agent. The number of training episodes is M, where the length of each episode is L. Each agent is trained using a predetermined set of iterations throughout the learning phase. Sampling and optimization constitutes the learning phase. In the beginning of learning, the state $s_{\alpha,i}$ and mean field o_i are randomly initialized for each agent. With the start of the sampling policy, UAV i samples its action based on the policy θ^i_{old}. The sampled action represents sensor selection, velocity and locations, and executed in the environment to obtain the cost, new state and new mean field. Consequently, trajectories (i.e., sequence of states, actions, policy, mean field, and costs) are gathered and stored in the RolloutBuffer. In addition, GAE is applied to calculate the advantage that is used in (4.14). In the optimization step, the policies are optimized. In the optimization step, the policy parameter is updated for each epoch. The PPO objective is computed in each epoch according to the following equation:

$$L^{clip}(\theta^i) = \min\Big(\frac{\pi_{\theta^i}(a_i|s_{\alpha,i},o_i)}{\pi_{\theta^i_{old}}(a_i|s_{\alpha,i},o_i)} A_{\pi_{\theta^i_{old}}}(s_{\alpha,i},o_i,a_i), \\ g(\varepsilon, A_{\pi_{\theta^i_{old}}}(s_{\alpha,i},o_i,a_i))\Big) \tag{4.14}$$

where

$$\pi_{\theta^i}(a_i|s_{\alpha,i},o_i) = \pi^c_{\theta^i}(a^c_i|s_{\alpha,i},o_i)\pi^d_{\theta^i}(a^d_i|s_{\alpha,i},o_i). \tag{4.15}$$

Here a^c_i and a^d_i correspond to actions in continuous and discrete spaces. In (4.15), to obtain the hybrid policy $\pi_{\theta^i}(a_i|s_{\alpha,i},o_i)$, we multiply the policies for continuous and discrete actions Neunert et al. (2020). Meanwhile, we assume that wireless radio of the UAV can cover the whole field.

Continuous policy $\pi^c_{\theta^i}$ is modeled using multivariate normal distribution and discrete policy $\pi^d_{\theta^i}$ is modeled using categorical distribution. In the next step, the overall objective function is optimized according to the following equation:

$$L^{total}(\theta^i) = L^{clip}(\theta^i) - K_1 L^{VF}(\theta^i) + K_2 * H. \tag{4.16}$$

Here, $L^{VF}(\theta^i)$ is the critic loss and H acts as a regularizer encourages the agent to execute actions more unpredictably for exploration and guard against the policy being overly

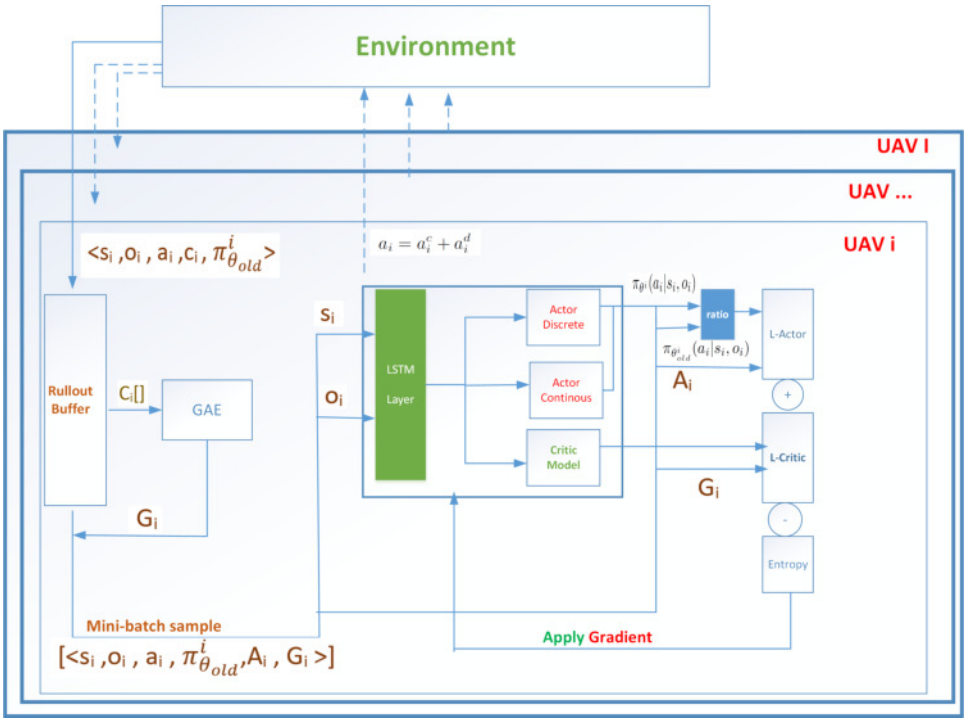


Figure 4.2: Overview of MF-HPPO: Each UAV equipped with LSTM layer to optimize discrete and continuous actions using hybrid policy .

deterministic. The entropy for continuous and discrete actions is computed based on the actions' distribution. We obtain the entropy by multiplication of the entropy of continuous and discrete action spaces to enable enforcing consistent regularization to both continuous and discrete action spaces. K_1 balances the importance of the critic loss and the actor loss, and K_2 coefficient controls the amount of entropy in the policy.

Finally, the sampling policy $\pi_{\theta^i_{old}}$ is updated with the policy π_{θ^i}, and the stored data are dropped. The next iteration then begins.

4.2.1.3 LSTM Layer

We further develop an LSTM layer in the proposed MF-HPPO, which captures long-term dependencies of time-varying network state s_α. Cell memory and the gating mechanism are main components of LSTM. Cell memory is responsible to store the summary of the past input data and the gating mechanism regulates the information flow between the input, output, and cell memory. The network states are fed into LSTM one by one (one at each step). The last hidden state κ_i^{hidd} is returned as the output of the state characterization layer. Each agent uses an LSTM layer to predict their respective hidden states. The hidden

states κ_i^{hidd} are calculated by the following composite function:

$$\kappa_i^{hidd} = out_i tanh(C_i), \tag{4.17}$$

$$out_i = \sigma(W_0 \cdot [C_i, \kappa_{i-1}^{hidd}, A_i] + e_i), \tag{4.18}$$

$$C_i = F_i C_{i-1]} + p_i tanh(W_c . [\kappa_{i-1}^{hidd}, A_i] + e_c), \tag{4.19}$$

$$F_i = \sigma(W_f \cdot [\kappa_{i-1}^{hidd}, C_{i-1}, A_i] + e_f), \tag{4.20}$$

$$p_i = \sigma(W_f \cdot [\kappa_{i-1}^{hidd}, C_{i-1}, A_i] + e_p), \tag{4.21}$$

where the output gate, cell activation vectors, forget gate, and input gate of the LSTM layer are denoted by out_i, C_i, F_i, and p_i, respectively. σ and tanh correspond to logistic sigmoid function and the hyperbolic tangent function, respectively. W_0, W_c, W_f, Wp are the weight matrix, and e_0, e_c, e_f, e_p are the bias matrixLi et al. (2022b), Zheng et al. (2022).

4.2.2 Complexity and Convergence of MF-HPPO

The overall complexity of MF-HPPO is calculated as follows, $O(I \cdot ML \cdot (\Sigma_{g=1}^{G} n_{g-1}.n_g))$ where n_g is the number of neural units in the g-th hidden layer. In this work, the PPO architecture is built with the same n_g in all hidden layers. Therefore, the PPO complexity can be reduced to $O(I \cdot ML \cdot (g-1) \cdot n_g^2) = O(I \cdot ML \cdot n_g^2)$. The convergence analysis is proved by simulation results (see Fig. 4.4).

Algorithm 2 MF-HPPO Characterized by LSTM Layer

1.Initialize

Input: Clip threshold ε, discount factor γ, learning rate η, buffer size D, mini-batch size B

Output: The scheduled ground sensor j and trajectory ζ_i of UAV i

Randomly initialize the Actors π_i and Critics w_i with networks parameters θ^i

The LSTM layer with $\{W_o, W_c, W_f, W_p\}$ and $\{e_o, e_c, e_f, e_p\}$.

Initialize the sampling policy $\pi_{\theta^i_{old}}$ with $\theta^i_{old} \leftarrow \theta^i$.

$\forall i \in (1, I)$

2.Learning

for *episode=1* **to** M **do**

  Randomly obtain the initial state $s_{\alpha,i}$

  for t = 1 **to** L **do**

    The sampling phase

    Sample: Sample action $a_i \sim \pi_{\theta^i_{old}}(a_i|s_{\alpha,i}, o_i, \theta^i)$;

    Execute the action a_i that specifies the scheduled ground sensor j and trajectory ζ_i of UAV i.

    Obtain the cost and new state $s_{\beta,i}$ and new mean field $_i(t+1)$. RolloutBuffer: store the trajectory $(s_{\alpha,i}, a_i, c, o_i, \pi_{\theta^i_{old}}(a_i|s_{\alpha,i}, o_i, \theta^i))$

    $s_{\alpha,i} = s_{\beta,i}$

  end for

  Compute the advantage using GAE

  for *epoch* = 1 **to** P **do**

    The optimization phase

    Sample the RolloutBuffer

    Compute the PPO-Clip objective function using (4.14)

    Compute the critic loss.

    Optimize the overall objective function using (4.16)

  end for

  Synchronize the sampling policy $\pi_{\theta^i_{old}} \leftarrow \pi_{\theta^i}$

  Drop the stored data in RolloutBuffer.

end for

4.3 Evaluation

4.3.1 Implementation of MF-HPPO

MF-HPPO is implemented in Python 3.8 using Pytorch (the Python deep learning library). A Predator Workstation running 64-bit Ubuntu 20.04 LTS, with Intel Core i7-11370 H CPU @ 3.30 GHz 8 and 16 GB memory is used for the Pytorch setup. Table 4.2 clearly outlines the different considered simulation parameters. MF-HPPO algorithm is trained over 3000 episodes with 40 steps each. The discount factor and learning rate are set to 0.99 and 3e-4, respectively. Each agent comprises the input layer, LSTM layer, the critic and actors with fully-connected hidden layers of size 256 and output layer. Each neuron uses Rectified Linear Unit (ReLU) as an activation function. In addition, Hyperbolic tangent (tanh) and softmax are used as activation functions in the output layer of the continuous actor-network and discrete actor network. The input of each critic network is represented as a concatenation of states and mean field, and its output is a scalar that assesses the states according to the global policy. The total log probability of the hybrid policy is the sum of the log probabilities of the continuous and discrete action spaces. This log probability would be used as part of the calculation of the objective function in MF-HPPO, along with the estimated cost and the entropy regularization term.

4.3.2 Baseline Description

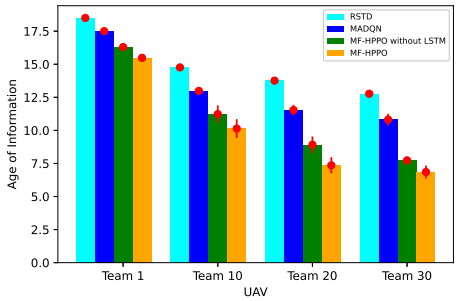
The MF-HPPO characterized with LSTM layer is compared by single-agent PPO, random scheduling and trajectory design (RSTD), multi-agent DQN (MADQN) and MF-HPPO without LSTM Layer. A brief introduction of the four benchmarks is given below

1. PPO, in this algorithm single-agent running PPO to optimize trajectory and transmission scheduling.

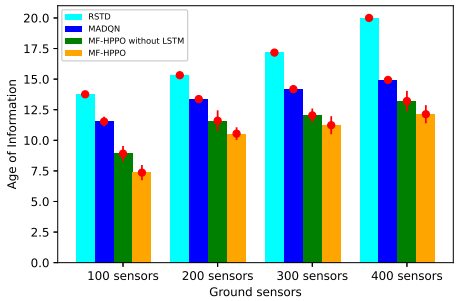
2. RSTD, in this algorithm transmission scheduling and trajectory design, are randomly designed.

3. MADQN, in this algorithm, each agent running DQN cooperate to reduce average AoI following circular trajectories.

4. MF-HPPO without LSTM Layer, the structure of this algorithm is same as MF-HPPO but without LSTM layer.



(a) Evaluation of MF-HPPO's performance with a variable number of UAVs in comparison to RSTD, MADQN and MF-HPPO without LSTM



(b) Evaluation of MF-HPPO's performance with a variable number of ground sensors in comparison to RSTD, MADQN and MF-HPPO without LSTM

Figure 4.3: Performance evaluation of MFFPO by changing the number of UAVs and ground sensors

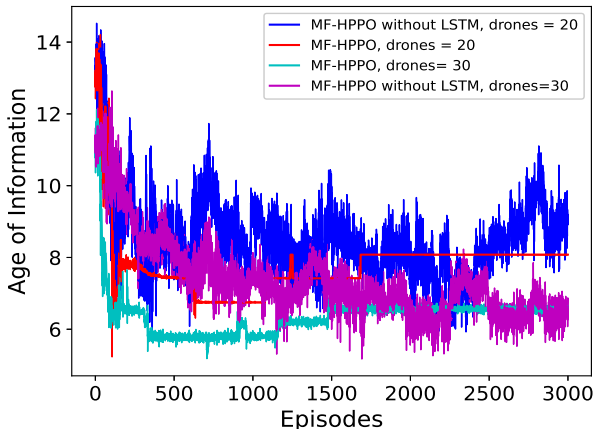


Figure 4.4: The network cost for each episode of MF-HPPO with I=30 and benchmarks

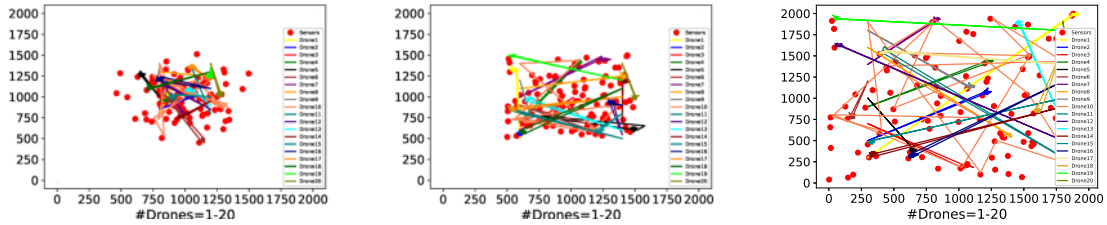
Table 4.2: PyTorch Configuration

Parameters	**Values**
Number of ground sensors	100
Number of UAVs	30
Geographical area size [m]	1,000*1,000
Altitude of the UAVs	120 m
Activation Function for Hidden Layers	Relu
Activation Function for Continuous Action	Tanh
Activation Function for Discrete Action	Softmax
Critic Network Learning Rate	3e-4
Actor Network Learning Rate	3e-4
Number of Hidden Layers for Networks	2
Number of Neurons	256
Loss Coefficients for K_1 and K_2	0.2 and 3
Optimizer Technique	Adam
Clip Fraction	0.2

4.3.3 Performance analysis of MF-HPPO

Fig. 4.3 depicts the performance evaluation of MF-HPPO in comparison to the baselines by changing the number of UAVs and ground sensors. Fig. 4.3a shows the impact of the number of UAVs on the AoI. Overall, the AoI decreases when more UAVs are deployed because time efficiency increases and more ground sensors can be operated in less time. Increasing the number of UAVs from 1 to 30 result in a 61% decrease in the average AoI for MF-HPPO, while that of MADQN is 37%. The reason is that MF-HPPO performs the optimization in a mixed action space with higher training stability than MADQN with circular trajectories. Fig. 4.3b evaluates the average AoI given 20 UAVs and groups of 100, 200, 300, and 400 ground sensors. The MADQN and the RSTD are used as baselines. Overall, increasing the number of ground sensors results in a uniform increase in the average AoI, since more sensor data should be collected. In particular, when the number of ground sensors is 400, the proposed MF-HPPO outperforms the RSTD by 38% and the MADQN by 17%.

We obtain the convergence trend of MF-HPPO in Fig. 4.4 by deploying 20 UAVs serving 100 ground sensors. In general, the proposed MF-HPPO (I=30) achieves the lowest AoI compared to MF-HPPO without LSTM layer (I= 20 and 30) with a gain of 33% and 66%, respectively. Since the trajectories and scheduling of data collection for multiple UAVs are optimized with better time efficiency. At the same time, the LSTM layer enables better exploration as agents use experience to guide their actions. Moreover, thanks to the LSTM layer, convergence is accelerated and stabilized. The peak AoI of the proposed MF-HPPO drops significantly from 14 seconds to 6 seconds in the first 1,000 episodes. From episode 1,500 to episode 3,000, the AoI stabilizes at 7 seconds with minimal fluctuations.



(a) Normal Distribution (b) Square Distribution (c) Uniform Distribution

Figure 4.5: MF-HPPO trajectory distributions for various UAV counts and ground sensor distributions.

MF-HPPO-generated trajectories for 20 UAVs are shown in Fig. 4.5, where the ground sensor distribution patterns are uniform, square, or normal ones. When designing trajectories for AoI minimization, the UAVs' trajectories are impacted by the distribution

of the ground sensors. The UAV needs to approach to the location of each scheduled sensor to collect the data and update its AoI. Fig. 5(a), refer to the normal distribution and shows trajectories for 20 UAVs, focusing on the center area of the ground sensors and less on the corners. The normal distribution of the ground sensors can affect the UAVs' trajectories by determining which ground sensors are prioritized for data collection. For example, as can be seen, most ground sensors are centered and their data may become stale, in this case, the UAVs'trajectories are designed to visit these ground sensors more frequently to minimize the average AoI. Figs. 5(b) is related to the square distribution. As can be seen, the ground sensors are less centered. This cause diverse set of ground sensors in wider range to be covered in comparison to normal distribution. Fig. 5(c) refer to the uniform distribution. As can be seen, the UAVs design wide-area trajectories due to the wider distribution of ground sensors covering the entire area and the AoI requirements of the scattered ground sensors.

Fig. 4.6 demonstrates the convergence figures for two variants of MF-HPPO by changing the clip threshold. PPO uses the clip threshold, commonly referred to as epsilon, to regulate the amount of policy updating. A larger clip threshold allows

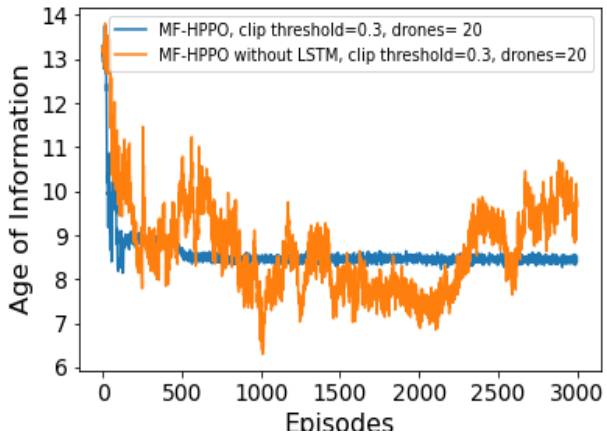


Figure 4.6: Performance evaluation of MF-HPPO by changing clip threshold

for more aggressive updating, while a smaller clip threshold restricts updating more severely, resulting in less policy change. The blue curve shows the MF-HPPO with LSTM layer and a clip threshold of 0.3 outperforming the MF-HPPO without LSTM layer clip threshold 0.3. The latter shows a deviating behavior due to the influence of the clip threshold, while the blue curve shows an absolutely stable trend despite the same value of the clip threshold thanks to the LSTM layer. Overall, adding the LSTM layer to MF-HPPO can stabilize the training and prevent divergence of the strategies.

4.4 Summary

In this chapter, we propose a mean field flight resource allocation to model velocity control for a swarm of UAVs, in which each UAV minimizes the average AoI by considering the collective behavior of others. Due to the high computational complexity of MFG, we leverage AI and propose MF-HPPO characterized with an LSTM layer to optimize the UAV trajectories and data collection scheduling in mixed action space. Simulation results based on PyTorch deep learning library show that the proposed MF-HPPO for UASNets reduces average AoI by up to 57% and 45%, as compared to existing non-learning random algorithm and MADQN method (which performs the action of trajectory planning in the discrete space), respectively. This confirms the AI-enhanced mean field resource allocation is a practical solution for minimizing AoI in UAV swarms.

Proof of FPK Equation for Cruise Control

We derive the mean field via an arbitrary test function $g(\zeta)$, which is a twice continuously differentiable compactly supported function of the state space. The integral of $m(\zeta)g(\zeta)d\zeta$ can be considered as the continuum limit of the sum $g(\zeta(t))$, where $\zeta(t)$ is the UAV's state at time t. It is known that,

$$\int m(\zeta(t))g(\zeta)d\zeta = \frac{1}{N}\Sigma_{i=1}^{N} g(\zeta(t)). \tag{4.22}$$

At time t, the first-order differential function with regard to time t is derived to check how this integral varies in time. By utilizing the chain rule, we can derive the heuristic formula as

$$\int \partial_t m(\zeta(t))g(\zeta)d\zeta = \frac{1}{N}\Sigma_{i=1}^{N} \partial_t \zeta(t)\nabla g(\zeta(t)) + \partial_t^2 \zeta(t)\nabla^2 g(\zeta(t)). \tag{4.23}$$

Taking the limit of the right side of the above equation when N tends to infinity, we get

$$\int [\partial_t m(\zeta(t)) + \nabla_\zeta m(\zeta(t)) \cdot \frac{\partial \zeta}{\partial t} - \frac{\eta^2}{2}\nabla_\zeta^2 m(\zeta(t))]g(\zeta(t))d\zeta = 0, \tag{4.24}$$

for any test function g through integration by parts. Then the above equation leads to the following equation:

$$\partial_t m(\zeta(t)) + \nabla_\zeta m(\zeta(t)) \cdot v(t) - \frac{\sigma^2}{2} \nabla_\zeta^2 m(\zeta(t)) = 0. \tag{4.25}$$

which correspond to FPK equation defined in (4.11).

Chapter 5

Conclusions and Future Work

5.1 Summary

We employed UAVs for data collection from ground sensors in harsh environments, such as crop monitoring. The use of UAVs for data collection offers advantages such as improved network throughput and extended coverage range beyond terrestrial gateways. However, a major challenge arises from the impact of UAV movements on channel conditions, leading to packet loss or outdated packets. To address this challenge, we proposed a joint optimization approach to minimize packet loss by controlling the velocities of multiple UAVs and optimizing their data collection schedules. Our proposed solution, MADRL-SA, enables UAVs to asymptotically minimize packet loss even when they have outdated knowledge of the network states. Furthermore, we introduced a novel mean-field flight resource allocation optimization method to minimize the AoI for sensory data. This involved formulating the trade-off between UAV movements and AoI as an MFG. To tackle practical scenarios, we proposed the MF-HPPO scheme, which optimizes UAV trajectories and data collection scheduling for ground sensors using a combination of continuous and discrete actions. Additionally, we incorporated LSTM to predict the time-varying network state and enhance training stability in MF-HPPO. We conducted extensive simulations to evaluate the effectiveness of our proposed approaches. The results demonstrated that MADRL-SA reduced packet loss by up to 54% and 46% compared to existing solutions involving single UAV with DRL and non-learning greedy heuristics, respectively. Similarly, the simulation results showed that MF-HPPO reduced the average AoI by up to 45% and 57% compared to the MADQN method and non-learning random algorithm, respectively.

5.2 Future Works

MADRL-SA and MF-HPPO can be enhanced with explainable AI and human-in-the-loop mechanisms to significantly improve their effectiveness and usability. By enriching these approaches, we can leverage human expertise, provide transparent explanations for decisions, enhance performance, foster user trust, and promote better collaboration between humans and AI systems in UASNets. These enhancements have the potential to improve the efficiency and reliability of communication scheduling and cruise control, ultimately enhancing the overall operation of UASNets. One approach to achieve this is by using feature importance techniques to identify and quantify the contribution of input features to the decisions made by DRL algorithms. By developing visualizations that depict the relationships between input features, intermediate algorithm states, and output decisions, we can provide users with a better understanding of the model's decision-making process. Another avenue is to incorporate human feedback to shape the cost function utilized by DRL algorithms. By shaping the cost function based on human preferences, we can guide the algorithms to make decisions that align better with human values and expectations. Furthermore, adopting an interactive ML paradigm allows human experts to interact with DRL algorithms during the training process. Human feedback, in the form of instructions or corrections, can be integrated into the learning process to improve the model's performance. In conclusion, the integration of explainable AI and human-in-the-loop reinforcement learning in this thesis can significantly enhance the performance and usability of the proposed algorithms in UASNets.

Bibliography

"PyTorch," http://pytorch.org/, accessed: 2022-10-19.

M. A. Abd-Elmagid and H. S. Dhillon, "Average peak age-of-information minimization in uav-assisted iot networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 2003–2008, 2018.

M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon, and W. Saad, "Deep reinforcement learning for minimizing age-of-information in uav-assisted networks," in *IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, Dec 2019. Cited on page 18.

A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, 2014. Cited on pages 23 and 42.

A. Al-Hourani, S. Chandrasekharan, S. Kandeepan, and A. Jamalipour, "7 - aerial platforms for public safety networks and performance optimization," in *Wireless Public Safety Networks 3*, D. Câmara and N. Nikaein, Eds. Elsevier, 2017, pp. 133 – 153. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978178548053950007X

A. T. Albu-Salih and S. A. H. Seno, "Energy-efficient data gathering framework-based clustering via multiple uavs in deadline-based wsn applications," *IEEE Access*, vol. 6, pp. 72 275–72 286, 2018. Cited on page 15.

L. Buşoniu, R. Babuška, and B. De Schutter, *Multi-agent Reinforcement Learning: An Overview*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 183–221. [Online]. Available: https://doi.org/10.1007/978-3-642-14435-6_7

L. Buşoniu, R. Babuška, and B. De Schutter, *Multi-agent Reinforcement Learning: An Overview*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 183–221. [Online]. Available: https://doi.org/10.1007/978-3-642-14435-6_7

U. Challita, W. Saad, and C. Bettstetter, "Deep reinforcement learning for interference-aware path planning of cellular-connected uavs," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–7.

U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected uavs: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, April 2019.

U. Challita, W. Saad, and C. Bettstetter, "Deep reinforcement learning for interference-aware path planning of cellular-connected uavs," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–7. Cited on page 17.

D. Chen, Q. Qi, Z. Zhuang, J. Wang, J. Liao, and Z. Han, "Mean field deep reinforcement learning for fair and efficient uav control," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 813–828, Jul 2020. Cited on page 17.

Q. Chen, "Joint position and resource optimization for multi-uav-aided relaying systems," *IEEE Access*, vol. 8, pp. 10 403–10 415, 2020. Cited on page 15.

K. Chi, F. Li, F. Zhang, M. Wu, and C. Xu, "Aoi optimal trajectory planning for co-operative uavs: A multi-agent deep reinforcement learning approach," in *IEEE International Conference on Electronic Information and Communication Technology (ICEICT)*, Hefei, China, Aug 2022, pp. 57–62. Cited on page 18.

D. H. Choi, S. H. Kim, and D. K. Sung, "Energy-efficient maneuvering and communication of a single uav-based relay," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 3, pp. 2320–2327, 2014. Cited on page 22.

C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multi-agent systems," *AAAI/IAAI*, vol. 1998, no. 746-752, p. 2, 1998.

J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for uav networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 729–743, 2020. Cited on pages 16 and 19.

S. Ecke, J. Dempewolf, J. Frey, A. Schwaller, E. Endres, H.-J. Klemmt, D. Tiede, and T. Seifert, "Uav-based forest health monitoring: A systematic review," *Remote Sensing*, vol. 14, no. 13, p. 3205, 2022.

E. Eldeeb, D. E. Pérez, J. Michel de Souza Sant'Ana, M. Shehab, N. H. Mahmood, H. Alves, and M. Latva-Aho, "A learning-based trajectory planning of multiple uavs for aoi minimization in iot networks," in *Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, Grenoble, France, Jun 2022, pp. 172–177. Cited on page 18.

Y. Emami, K. Li, and E. Tovar, "Buffer-aware scheduling for uav relay networks with energy fairness," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

A. Ferdowsi, M. A. Abd-Elmagid, W. Saad, and H. S. Dhillon, "Neural combinatorial deep reinforcement learning for age-optimal joint trajectory and scheduling design in uav-assisted networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1250–1265, 2021.

H. Gao, W. Lee, Y. Kang, W. Li, Z. Han, S. Osher, and H. V. Poor, "Energy-efficient velocity control for massive numbers of uavs: A mean field game approach," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6266–6278, Mar 2022. Cited on page 18.

Y. Gao, X. Chen, J. Yuan, Y. Li, and H. Cao, "A data collection system for environmental events based on unmanned aerial vehicle and wireless sensor networks," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, 2020, pp. 2175–2178. Cited on page 5.

A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS J. on Computing*, vol. 21, no. 2, p. 178–192, Apr. 2009. [Online]. Available: https://doi.org/10.1287/ijoc.1080.0305

A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

S. Guan, Z. Zhu, and G. Wang, "A review on uav-based remote sensing technologies for construction and civil applications," *Drones*, vol. 6, no. 5, p. 117, May 2022.

Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge University Press, 2011.

A. Heidari, N. Jafari Navimipour, M. Unal, and G. Zhang, "Machine learning applications in internet-of-drones: Systematic review, recent deployments, and open issues," *ACM Comput. Surv.*, vol. 55, no. 12, Mar 2023.

S. A. Hoseini, J. Hassan, A. Bokani, and S. S. Kanhere, "Trajectory optimization of flying energy sources using q-learning to recharge hotspot uavs," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 683–688.

J. Hu, H. Zhang, K. Bian, L. Song, and Z. Han, "Distributed trajectory design for cooperative internet of uavs using deep reinforcement learning," in *IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, Dec 2019. Cited on page 18.

J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of uavs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6807–6821, 2020. Cited on page 22.

J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of uavs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6807–6821, Aug 2020. Cited on page 18.

M. Hua, Y. Wang, Z. Zhang, C. Li, Y. Huang, and L. Yang, "Power-efficient communication in uav-aided wireless sensor networks," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1264–1267, June 2018.

M. Hüttenrauch, S. Adrian, G. Neumann *et al.*, "Deep reinforcement learning for swarm systems," *Journal of Machine Learning Research*, vol. 20, no. 54, pp. 1–31, 2019.

X. JIANG, M. SHENG, N. ZHAO, C. XING, W. LU, and X. WANG, "Green uav communications for 6g: A survey," *Chinese Journal of Aeronautics*, vol. 35, no. 9, pp.

19–34, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1000936121001801 Cited on page 1.

S. Jordan, J. Moore, S. Hovet, J. Box, J. Perry, K. Kirsche, D. Lewis, and Z. T. H. Tse, "State-of-the-art technologies for uav inspections," *IET Radar, Sonar & Navigation*, vol. 12, no. 2, pp. 151–164, 2018.

S. Kandeepan, K. Gomez, T. Rasheed, and L. Reynaud, "Energy efficient cooperative strategies in hybrid aerial-terrestrial networks for emergencies," in *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep. 2011, pp. 294–299.

S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proceedings IEEE INFOCOM*, Mar 2012, pp. 2731–2735. Cited on page 6.

J. Kim, S. Kim, C. Ju, and H. I. Son, "Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications," *IEEE Access*, vol. 7, pp. 105 100–105 115, 2019. Cited on page 5.

H. Kurunathan, H. Huang, K. Li, W. Ni, and E. Hossain, "Machine learning-aided operations and communications of unmanned aerial vehicles: A contemporary survey," *arXiv preprint arXiv:2211.04324*, 2022.

J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese journal of mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

B. Li, Z. Fei, and Y. Zhang, "Uav communications for 5g and beyond: Recent advances and future trends," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2241–2263, 2018. Cited on page 1.

K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "Energy-efficient cooperative relaying for unmanned aerial vehicles," *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1377–1386, June 2016.

K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep q-network for uav-assisted online power transfer and data collection," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 215–12 226, Dec 2019. Cited on pages 16, 19, 22, and 32.

K. Li, Y. Emami, W. Ni, E. Tovar, and Z. Han, "Onboard deep deterministic policy gradients for online flight resource allocation of uavs," *IEEE Networking Letters*, vol. 2, no. 3, pp. 106–110, 2020. Cited on page 22.

K. Li, W. Ni, E. Tovar, and M. Guizani, "Joint flight cruise control and data collection in uav-aided internet of things: An onboard deep reinforcement learning approach," *IEEE Internet of Things Journal*, pp. 1–1, 2020. Cited on pages 16 and 19.

K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "Energy-efficient cooperative relaying for unmanned aerial vehicles," *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1377–1386, 2015. Cited on page 16.

K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "Energy-efficient cooperative relaying for unmanned aerial vehicles," *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1377–1386, Aug 2016.

K. Li, W. Ni, E. Tovard, and A. Jamalipour, "Online velocity control and data capture of drones for the internet of things: An onboard deep reinforcement learning approach," *IEEE Vehicular Technology Magazine*, vol. 16, no. 1, pp. 49–56, 2020. Cited on pages 16 and 19.

K. Li, W. Ni, and F. Dressler, "Lstm-characterized deep reinforcement learning for continuous flight control and resource allocation in uav-assisted sensor network," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4179–4189, Aug 2022. Cited on page 51.

K. Li, W. Ni, A. Noor, and M. Guizani, "Employing intelligent aerial data aggregators for the internet of things: Challenges and solutions," *IEEE Internet of Things Magazine*, vol. 5, no. 1, pp. 136–141, Mar 2022. Cited on page 6.

L. Li, Q. Cheng, X. Tang, T. Bai, W. Chen, Z. Ding, and Z. Han, "Resource allocation for noma-mec systems in ultra-dense networks: A learning aided mean-field game approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1487–1500, 2020.

L. Li, Q. Cheng, K. Xue, C. Yang, and Z. Han, "Downlink transmit power control in ultra-dense uav network based on mean field game and deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 594–15 605, Dec 2020. Cited on page 17.

C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-uav navigation for long-term communication coverage by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1274–1285, 2020.

C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, 2018. Cited on page 16.

C. H. Liu, Z. Dai, Y. Zhao, J. Crowcroft, D. Wu, and K. K. Leung, "Distributed and energy-efficient mobile crowdsensing with charging stations by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 130–146, 2021.

L. Liu, K. Xiong, J. Cao, Y. Lu, P. Fan, and K. B. Letaief, "Average aoi minimization in uav-assisted data collection with rf wireless power transfer: A deep reinforcement learning scheme," *IEEE Internet of Things Journal*, vol. 9, no. 7, pp. 5216–5228, 2021. Cited on page 19.

X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-uav networks: Deployment and movement design," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8036–8049, 2019.

X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-uav assisted wireless networks: A machine learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7957–7969, 2019. Cited on page 17.

N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019. Cited on page 14.

C. Mao, J. Liu, and L. Xie, "Multi-uav aided data collection for age minimization in wireless sensor networks," in *2020 International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2020, pp. 80–85.

M. E. Mkiramweni, C. Yang, J. Li, and W. Zhang, "A survey of game theory in unmanned aerial vehicles communications," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3386–3416, May 2019. Cited on page 6.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

P. Mörters and Y. Peres, *Brownian motion*. Cambridge University Press, 2010, vol. 30. Cited on page 44.

M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on uavs for wireless networks: Applications, challenges, and open problems," *IEEE communications surveys & tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019. Cited on page 2.

Y. Y. Munaye, R. T. Juang, H. P. Lin, and G. B. Tarekegn, "Resource allocation for multi-uav assisted iot networks: A deep reinforcement learning approach," in *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, 2020, pp. 15–22. Cited on page 16.

M. J. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 2915–2934, July 2006.

M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan and Claypool Publishers, 2010.

M. Neunert, A. Abdolmaleki, M. Wulfmeier, T. Lampe, T. Springenberg, R. Hafner, F. Romano, J. Buchli, N. Heess, and M. Riedmiller, "Continuous-discrete reinforcement learning for hybrid control in robotics," in *Conference on Robot Learning*. PMLR, May 2020, pp. 735–751. Cited on page 49.

O. S. Oubbati, M. Atiquzzaman, A. Lakas, A. Baz, H. Alhakami, and W. Alhakami, "Multi-uav-enabled aoi-aware wpcn: A multi-agent reinforcement learning strategy," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2021, pp. 1–6.

O. S. Oubbati, M. Atiquzzaman, H. Lim, A. Rachedi, and A. Lakas, "Synchronizing uav teams for timely data collection and energy transfer by deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, pp. 1–1, Apr 2022. Cited on page 18.

Q. Pan, X. Wen, Z. Lu, L. Li, and W. Jing, "Dynamic speed control of unmanned aerial vehicles for data collection under internet of things," *Sensors*, vol. 18, no. 11, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/11/3951

P. M. Pappachan, "An mdp-based policy for stochastic multi-agent domains," in *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028)*, vol. 5, 1999, pp. 464–468 vol.5.

W. Qiang and Z. Zhongli, "Reinforcement learning model, algorithms and its application," in *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, 2011, pp. 1143–1146.

H. Qie, D. Shi, T. Shen, X. Xu, Y. Li, and L. Wang, "Joint optimization of multi-uav target assignment and path planning based on multi-agent reinforcement learning," *IEEE Access*, vol. 7, pp. 146 264–146 272, 2019.

M. Samir, S. Sharafeddine, C. M. Assi, T. M. Nguyen, and A. Ghrayeb, "Uav trajectory planning for data collection from time-constrained iot devices," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 34–46, 2020. Cited on page 16.

M. Samir, C. Assi, S. Sharafeddine, D. Ebrahimi, and A. Ghrayeb, "Age of information aware trajectory planning of uavs in intelligent transportation systems: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 12 382–12 395, 2020.

M. Samir, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Online altitude control and scheduling policy for minimizing aoi in uav-assisted iot wireless networks," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2493–2505, Dec 2020.

M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3978–3983, 2021.

M. Samir, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Online altitude control and scheduling policy for minimizing aoi in uav-assisted iot wireless networks," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2493–2505, Dec 2022. Cited on pages 18 and 20.

C. Savaglio, P. Pace, G. Aloi, A. Liotta, and G. Fortino, "Lightweight reinforcement learning for energy efficient communications in wireless sensor networks," *IEEE Access*, vol. 7, pp. 29 355–29 364, 2019.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.

J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, Jun 2015. Cited on page 48.

H. Shakhatreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019. Cited on page 5.

H. Shakhatreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019. Cited on page 5.

H. Shakhatreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019.

A. Shamsoshoara, M. Khaledi, F. Afghah, A. Razi, and J. Ashdown, "Distributed cooperative spectrum sharing in uav networks using multi-agent reinforcement learning," in *2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2019, pp. 1–6. Cited on page 16.

V. Sharma, I. You, and R. Kumar, "Energy efficient data dissemination in multi-uav coordinated wireless sensor networks," *Mobile Information Systems*, vol. 2016, 2016. Cited on page 15.

D. Shi, H. Gao, L. Wang, M. Pan, Z. Han, and H. V. Poor, "Mean field game guided deep reinforcement learning for task placement in cooperative multiaccess edge computing," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9330–9340, Mar 2020. Cited on page 17.

D. Simões, N. Lau, and L. P. Reis, "Multi-agent double deep q-networks," in *Progress in Artificial Intelligence*, E. Oliveira, J. Gama, Z. Vale, and H. Lopes Cardoso, Eds. Cham: Springer International Publishing, 2017, pp. 123–134.

M. Sun, X. Xu, X. Qin, and P. Zhang, "Aoi-energy-aware uav-assisted data collection for iot networks: A deep reinforcement learning method," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17 275–17 289, May 2021. Cited on page 18.

Y. Sun, L. Li, Q. Cheng, D. Wang, W. Liang, X. Li, and Z. Han, "Joint trajectory and power optimization in multi-type uavs network with mean field q-learning," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, Dublin, Ireland, Jul 2020, pp. 1–6. Cited on page 17.

R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018. Cited on page 13.

J. Tang, J. Song, J. Ou, J. Luo, X. Zhang, and K. Wong, "Minimum throughput maximization for multi-uav enabled wpcn: A deep reinforcement learning method," *IEEE Access*, vol. 8, pp. 9124–9132, 2020.

J. Tang, J. Song, J. Ou, J. Luo, X. Zhang, and K.-K. Wong, "Minimum throughput maximization for multi-uav enabled wpcn: A deep reinforcement learning method," *IEEE Access*, vol. 8, pp. 9124–9132, 2020. Cited on page 16.

M. C. Tatum and J. Liu, "Unmanned aircraft system applications in construction," *Procedia Engineering*, vol. 196, pp. 167–175, 2017, creative Construction Conference 2017, CCC 2017, 19-22 June 2017, Primosten, Croatia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877705817330461

M. Thammawichai, S. P. Baliyarasimhuni, E. C. Kerrigan, and J. B. Sousa, "Optimizing communication and computation for multi-uav information gathering applications," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 2, pp. 601–615, 2017. Cited on page 15.

P. Tong, J. Liu, X. Wang, B. Bai, and H. Dai, "Deep reinforcement learning for efficient data collection in uav-aided internet of things," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, Dublin, Ireland, Jun 2020, pp. 1–6. Cited on page 19.

C. Torresan, A. Berton, F. Carotenuto, S. F. D. Gennaro, B. Gioli, A. Matese, F. Miglietta, C. Vagnoli, A. Zaldei, and L. Wallace, "Forestry applications of uavs in europe: a review," *International Journal of Remote Sensing*, vol. 38, no. 8-10, pp. 2427–2447, 2017.

D. C. Tsouros, S. Bibi, and P. G. Sarigiannidis, "A review on uav-based applications for precision agriculture," *Information*, vol. 10, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2078-2489/10/11/349

S. J. Undertaking *et al.*, "European drones outlook study: unlocking the value for europe." 2017. Cited on pages 1 and 4.

D. K. Villa, A. S. Brandao, and M. Sarcinelli-Filho, "A survey on load transportation using multirotor uavs," *Journal of Intelligent & Robotic Systems*, vol. 98, pp. 267–296, Oct 2020.

B. Waldrip, V. Prain, and P. Sellings, "Explaining newton's laws of motion: Using student reasoning through representations to develop conceptual understanding," *Instructional Science*, vol. 41, pp. 165–189, Mar 2013. Cited on page 44.

C. Wang, J. Wang, X. Zhang, and X. Zhang, "Autonomous navigation of uav in large-scale unknown complex environment with deep reinforcement learning," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2017, pp. 858–862.

L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and L. Hanzo, "Multi-agent deep reinforcement learning-based trajectory planning for multi-uav assisted mobile edge computing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 73–84, 2021. Cited on page 22.

M. Wang, L. Li, W. Lin, B. Wei, W. Chen, and Z. Han, "Uav position optimization based on information freshness: A mean field game approach," in *13th International Conference on Wireless Communications and Signal Processing (WCSP)*, Changsha, China, Dec 2021, pp. 1–5. Cited on pages 17 and 19.

Q. Wang, W. Zhang, Y. Liu, and Y. Liu, "Multi-uav dynamic wireless networking with deep reinforcement learning," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2243–2246, 2019. Cited on page 16.

X. Wang, Z. Mi, H. Wang, and N. Zhao, "Performance test and analysis of multi-hop network based on uav ad hoc network experiment," in *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2017, pp. 1–6.

F. Wu, H. Zhang, J. Wu, Z. Han, H. V. Poor, and L. Song, "Uav-to-device underlay communications: Age of information minimization by multi-agent deep reinforcement learning," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4461–4475, 2021.

Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-uav enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, March 2018. Cited on page 15.

Q. Wu, Y. Zeng, and R. Zhang, *Overview*. John Wiley Sons, Ltd, 2020, ch. 1, pp. 1–16. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119575795.ch1 Cited on page 3.

T. Wu, J. Liu, J. Liu, Z. Huang, H. Wu, C. Zhang, B. Bai, and G. Zhang, "A novel ai-based framework for aoi-optimal trajectory planning in uav-assisted wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 4, pp. 2462–2475, 2022.

Z. Xiong, Y. Zhang, W. Y. B. Lim, J. Kang, D. Niyato, C. Leung, and C. Miao, "Uav-assisted wireless energy and data transfer with deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, 2020.

Y. Xu, L. Li, Z. Zhang, K. Xue, and Z. Han, "A discrete-time mean field game in multi-uav wireless communication systems," in *IEEE/CIC International Conference on Communications in China (ICCC)*, Beijing, China, Feb 2018, pp. 714–718. Cited on page 18.

K. Xue, Z. Zhang, L. Li, H. Zhang, X. Li, and A. Gao, "Adaptive coverage solution in multi-uavs emergency communication system: a discrete-time mean-field game," in *International Wireless Communications & Mobile Computing Conference (IWCMC)*, Limassol, Cyprus, Aug 2018, pp. 1059–1064. Cited on page 17.

J. Yang, X. Ye, R. Trivedi, H. Xu, and H. Zha, "Learning deep mean field games for modeling large population behavior," *arXiv preprint arXiv:1711.03156*, 2017.

Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5571–5580. Cited on page 48.

M. Yi, X. Wang, J. Liu, Y. Zhang, and B. Bai, "Deep reinforcement learning for fresh data collection in uav-assisted iot networks," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 716–721.

Y. Zeng and R. Zhang, "Energy-efficient uav communication with trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3747–3760, 2017. Cited on page 22.

C. Zhan, Y. Zeng, and R. Zhang, "Energy-efficient data collection in uav enabled wireless sensor network," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 328–331, June 2018.

C. Zhan and Y. Zeng, "Completion time minimization for multi-uav-enabled data collection," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4859–4872, 2019. Cited on page 15.

B. Zhang, C. H. Liu, J. Tang, Z. Xu, J. Ma, and W. Wang, "Learning-based energy-efficient data collection by unmanned vehicles in smart cities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1666–1676, 2017. Cited on pages 16 and 19.

Y. Zhang, C. Yang, J. Li, and Z. Han, "Distributed interference-aware traffic offloading and power control in ultra-dense networks: Mean field game with dominating player," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8814–8826, 2019.

N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F. R. Yu, and K. Wong, "Uav-assisted emergency networks in disasters," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 45–51, 2019. Cited on page 5.

N. Zhao, Z. Liu, and Y. Cheng, "Multi-agent deep reinforcement learning for trajectory design and power allocation in multi-uav networks," *IEEE Access*, vol. 8, pp. 139 670–139 679, 2020. Cited on page 22.

J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Exploring deep-reinforcement-learning-assisted federated learning for online resource allocation in privacy-preserving edgeiot," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 099–21 110, May 2022. Cited on page 51.

C. Zhou, H. He, P. Yang, F. Lyu, W. Wu, N. Cheng, and X. Shen, "Deep rl-based trajectory planning for aoi minimization in uav-assisted iot," in *International Conference on Wireless Communications and Signal Processing (WCSP)*, Xi'an,China, Oct 2019, pp. 1–6. Cited on page 18.