

# AI-Driven Multi-objective Decision-Making With Applications to IoT

Shubham Vaishnav





# AI-Driven Multi-objective Decision-Making With Applications to IoT

Shubham Vaishnav

Academic dissertation for the Degree of Doctor of Philosophy in Computer and Systems Sciences at Stockholm University to be publicly defended on Friday 29 May 2026 at 13.00 in Small Auditorium NOD (Lilla hörsalen), plan 2, Borgarfjordsgatan 12.

## Abstract

Artificial intelligence (AI) plays an increasingly central role in enabling autonomous decision-making in complex, uncertain environments. Many modern systems must optimise multiple, often conflicting objectives while operating under dynamic resource constraints and incomplete knowledge of system dynamics. Classical approaches such as dynamic programming, constrained stochastic optimisation, and static multi-objective scalarisation provide principled solutions when accurate models are available. However, in distributed and stochastic environments such as the Internet of Things (IoT), system dynamics are often unknown, non-stationary, and resource-limited, making purely model-based methods difficult to apply.

Reinforcement learning (RL) and online learning offer an alternative by enabling policy adaptation through interaction rather than relying on explicit system models. Within multi-objective settings, existing approaches often assume fixed scalarisation weights or externally specified preferences and typically focus on learning policies for given trade-offs. In dynamic IoT systems, however, both resource constraints and preference parameters may vary over time, requiring algorithms that can adapt efficiently without repeated retraining or centralised coordination.

This thesis investigates how AI-based methods, with a primary focus on reinforcement learning and complementary distributed learning techniques, can support adaptive multi-objective decision-making under communication constraints, explicit resource limitations, and dynamically changing trade-offs. The research is organised around three themes. First, communication-efficient distributed learning methods are developed to balance model accuracy and communication cost in federated learning through adaptive sparsification. Second, constrained bandit and reinforcement learning formulations are proposed to incorporate explicit and time-varying resource constraints while maintaining theoretical performance guarantees. Third, multi-objective reinforcement learning methods are designed to adapt routing decisions in distributed IoT systems under dynamically changing energy–reliability trade-offs without retraining.

Overall, the thesis demonstrates that integrating communication awareness, constraint handling, and preference adaptation directly into learning algorithms is essential for reliable AI-based decision-making in IoT environments. The results provide both algorithmic advances and a conceptual framework for designing autonomous systems that operate robustly under dynamic objectives and limited resources.

**Keywords:** *Artificial Intelligence, Multi-objective, Internet of Things, Federated Learning, Reinforcement Learning.*

Stockholm 2026  
<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-254165>

ISBN 978-91-8107-602-8  
ISBN 978-91-8107-603-5  
ISSN 1101-8526



Stockholm  
University

Department of Computer and Systems Sciences

Stockholm University, 164 07 Kista



AI-DRIVEN MULTI-OBJECTIVE DECISION-MAKING WITH  
APPLICATIONS TO IOT

Shubham Vaishnav





Stockholm  
University

# AI-Driven Multi-objective Decision-Making With Applications to IoT

Shubham Vaishnav

©Shubham Vaishnav, Stockholm University 2026

ISBN print 978-91-8107-602-8

ISBN PDF 978-91-8107-603-5

ISSN 1101-8526

Cover image: The cover image was generated using Google Gemini, a generative artificial intelligence model. The image is an original, AI-generated illustration created for this thesis and was not derived from copyrighted or identifiable source material.

Printed in Sweden by Universitetservice US-AB, Stockholm 2026

# Abstract

Artificial intelligence (AI) plays an increasingly central role in enabling autonomous decision-making in complex, uncertain environments. Many modern systems must optimise multiple, often conflicting objectives while operating under dynamic resource constraints and incomplete knowledge of system dynamics. Classical approaches such as dynamic programming, constrained stochastic optimisation, and static multi-objective scalarisation provide principled solutions when accurate models are available. However, in distributed and stochastic environments such as the Internet of Things (IoT), system dynamics are often unknown, non-stationary, and resource-limited, making purely model-based methods difficult to apply.

Reinforcement learning (RL) and online learning offer an alternative by enabling policy adaptation through interaction rather than relying on explicit system models. Within multi-objective settings, existing approaches often assume fixed scalarisation weights or externally specified preferences and typically focus on learning policies for given trade-offs. In dynamic IoT systems, however, both resource constraints and preference parameters may vary over time, requiring algorithms that can adapt efficiently without repeated retraining or centralised coordination.

This thesis investigates how AI-based methods, with a primary focus on reinforcement learning and complementary distributed learning techniques, can support adaptive multi-objective decision-making under communication constraints, explicit resource limitations, and dynamically changing trade-offs. The research is organised around three themes. First, communication-efficient distributed learning methods are developed to balance model accuracy and communication cost in federated learning through adaptive sparsification. Second, constrained bandit and reinforcement learning formulations are proposed to incorporate explicit and time-varying resource constraints while maintaining theoretical performance guarantees. Third, multi-objective reinforcement learning methods are designed to adapt routing decisions in distributed IoT systems under dynamically changing energy–reliability trade-offs without retraining.

Overall, the thesis demonstrates that integrating communication awareness, constraint handling, and preference adaptation directly into

learning algorithms is essential for reliable AI-based decision-making in IoT environments. The results provide both algorithmic advances and a conceptual framework for designing autonomous systems that operate robustly under dynamic objectives and limited resources.

"I am seated in everyone's heart, and from  
Me come remembrance, knowledge, and  
forgetfulness."

— *Bhagavad Gītā* 15.15



# Sammanfattning

Artificiell intelligens (AI) spelar en allt viktigare roll för att möjliggöra autonomt beslutsfattande i komplexa och osäkra miljöer. Många moderna system måste optimera flera, ofta motstridiga mål samtidigt som de verkar under dynamiska resursbegränsningar och med ofullständig kunskap om systemets dynamik. Klassiska metoder såsom dynamisk programmering, begränsad stokastisk optimering och statisk multiobjektiv skalärisering erbjuder principiella lösningar när exakta modeller finns tillgängliga. I distribuerade och stokastiska miljöer, såsom Internet of Things (IoT), är dock systemdynamiken ofta okänd, icke-stationär och resursbegränsad, vilket gör rent modellbaserade metoder svåra att tillämpa.

Förstärkningsinlärning (reinforcement learning, RL) och onlineinlärning erbjuder ett alternativ genom att möjliggöra policyanpassning genom interaktion snarare än genom explicita systemmodeller. Inom multiobjektiva problem antar befintliga metoder ofta fasta skaläriseringsvikter eller externt specificerade preferenser och fokuserar främst på att lära policyer för givna avvägningar. I dynamiska IoT-system kan dock både resursbegränsningar och preferensparametrar förändras över tid, vilket kräver algoritmer som kan anpassa sig effektivt utan upprepade ominlärning eller centraliserad styrning.

Denna avhandling undersöker hur AI-baserade metoder, med särskilt fokus på förstärkningsinlärning och kompletterande distribuerade inlärningsmetoder, kan stödja adaptivt multiobjektivt beslutsfattande under kommunikationsbegränsningar, explicita resurskrav och dynamiskt föränderliga avvägningar. Forskningen är organiserad kring tre huvudteman. För det första utvecklas kommunikationseffektiva distribuerade inlärningsmetoder som balanserar modellnoggrannhet och kommunikationskostnad i federerad inlärning genom adaptiv sparsifiering. För det andra föreslås begränsade bandit- och förstärkningsinlärningsformuleringar för att hantera explicita och tidsvarierande resursbegränsningar med teoretiska prestandagarantier. För det tredje utvecklas multiobjektiv förstärkningsinlärning för att möjliggöra adaptiv routing i distribuerade IoT-system under dynamiskt föränderliga avvägningar mellan energiförbrukning och tillförlitlighet, utan behov av ominlärning.

Sammanfattningsvis visar avhandlingen att det är avgörande att explicit integrera kommunikationsmedvetenhet, begränsningshantering och preferensanpassning i inlärningsalgoritmer för att uppnå tillförlitligt AI-baserat beslutsfattande i IoT-miljöer. Resultaten bidrar med både algoritmiska framsteg och en konceptuell grund för att utforma autonoma system som kan verka robust under dynamiska mål och begränsade resurser.

# Acknowledgements

A Ph.D. journey is quite a memorable and budding phase of an academic career. It involves growing from an academic child to an independent adult researcher. They say, “It takes a village to raise a child,” and it definitely holds for my Ph.D. journey as well. I can only attempt to thank all the people who make up this “village” for me. I am fortunate to have so many well-wishers and contributors. I will not be able to mention all the names here, but I am grateful to everyone who has directly or indirectly contributed.

Let me begin with the person who has contributed the most directly—my main supervisor, *Sindri Magnússon*. I am deeply grateful for his continuous guidance, encouragement, and unwavering support throughout my doctoral studies. At times, I felt he acted like a therapist who never let my spirits fall. I am also deeply thankful to my co-supervisor, *Praveen Kumar Donta*, for his consistent support. His contribution to my academic career is foundational—from supervising my master’s thesis at IIT (ISM) Dhanbad, to introducing me to this Ph.D. program, acting as my co-supervisor during my Ph.D., and now further guiding and helping me with my career ahead.

Thanks to the DEMOCRITUS project for funding my research, *Jelena Zdravkovic*, and to the department (DSV) for all the administrative support. I am grateful to my unit head, *Panagiotis Papapetrou*, for fostering a friendly and supportive research environment at DSV. He also served on my internal examining committees for my predoc and defence. Special thanks to *Ali Ramezani-kebrya* for acting as the external examiner for my halftime seminar, predoc, and final thesis defence. Their feedback improved the clarity and quality of this thesis. I also thank my other Ph.D. grading committee members, including *Salman Toor*, *Ming Xiao*, *Yuhong Li*, and *Per Gösta Andersson*, for thoroughly evaluating my thesis. Their feedback will provide deeper insights and will guide my future work.

I must also thank my former unit head, *Tony Lindgren*, for facilitating my China trip, where I delivered an invited talk at IEEE WF-IoT 2025. I would further like to thank my co-authors, *Sarit Khirirat*, *Sindri Magnússon*, *Praveen Kumar Donta*, and *Maria Efthymiou*, for their fruitful collaborations, thoughtful feedback, and meaningful contribu-

tions to the work presented in this thesis. Working with them has been both intellectually rewarding and motivating. I would also like to acknowledge my fellow “academic siblings,” *Ali, Guilherme, Mohsen, and Alireza*, for being a friendly support system at DSV—especially *Ali*, for being an expert in everything and helping in all kinds of ways, both personal and professional. *Guilherme’s* thoughtful reading of my work and constructive feedback have also helped improve this thesis. I was touched by how *Sayeh* hugged and accepted me when I came out as my authentic self.

I would also like to appreciate the academics outside of DSV who became my friends during this Ph.D. journey. My recent research visit to *Prof. Rajkumar Buyya* in Melbourne was especially rewarding. I was impressed by his simplicity, hard work, and deep spiritual roots. It opened opportunities for collaboration with his wonderful teammates, particularly *Murtaza* and *Abhishek*. I am grateful to *Krishnendu* from KTH, who was a great travel partner during GLOBECOM 2025 and also a guide for my future career options.

Stability in professional life is a reflection of how nourished one is in one’s personal life and relationships. I must give a shout-out to my parents, *Mr. Krishna Das* and *Mrs. Anita*, who have been immensely encouraging throughout my life. Furthermore, my senior uncle, *Mr. Pawan*, and my sister, *Ms. Vishakha*, have been such supportive pillars in my life, without whose support in India my stay in Sweden would not have been as smooth. My grandfather, *Sri Doongar Das ji*, laid the spiritual and moral foundation of my life. Thanks to all my uncles, aunts, and cousins, who have always cheered for my growth.

I am most grateful to the Almighty and all-loving God, known as *Krishna* in my tradition, for providing me with inner guidance and strength. I cannot be grateful enough to my spiritual teachers, *Srila Prabhupada* and *H.H. Radhanath Swami*, whose wisdom and compassion form the foundation and backbone of all my efforts. I am deeply thankful for my association with *Tore Karlsen (Tapas Prabhu)*, who, before leaving his body, was my everyday buddy, philosopher, and guide in ISKCON Stockholm. I am deeply grateful to all my guides and friends from the ISKCON family, including *Manoj (Maharaas Pr)*, *Ferdinando*, *Lenka*, *Udaranga*, *Arthur*, *Ciranjiva*, *Nikolai*, *Shreekara*, *Tuva*, *Neeraj*, *Geetesh*, *Caroline*, *Avelo*, *Shubham*, *Sooraj*, *Beenoo*, *Vasanth*, *Divyarka*, *Manoj*, *Abhay*, *Shantanu*, *Ashwini–Tilotma*, *Maulik–Anuradha*, *Naresh–Soniya*, *Nikhil–Charu*, *Prerna*, *Pavitra*, *Sujata*, *Shilpa*, *Twinkle*, *Leena*, *Kamlesh*, and *Kalpesh*. I am extremely grateful to all those who supported my initiative, *Bhakti Yoga Society*, at SU—especially *Tuva*, *Marie*, *Kira*, *Lucia*, *Akshat*, *Gabriel*, and others from Stockholms universitets studentkår (SUS). This student association will remain a lifelong cherished memory for me.

# List of Papers

The following papers [1–6], referred to in the text by their Roman numerals, are included in this thesis.

**PAPER I: Energy-Efficient and Adaptive Gradient Sparsification for Federated Learning**

Shubham Vaishnav, Maria Efthymiou, Sindri Magnússon,  
In *ICC 2023-IEEE International Conference on Communications*, pages 1256–1261 (2023).

DOI: 10.1109/ICC45041.2023.10278999

**PAPER II: Communication-Adaptive Gradient Sparsification for Federated Learning with Error Compensation**

Shubham Vaishnav, Sarit Khirirat, Sindri Magnússon,  
*IEEE Internet of Things Journal*, pages 1137–1152 (2024).

DOI: 10.1109/JIOT.2024.3490855

**PAPER III: Adaptive Budgeted Multi-armed Bandits for IoT with Dynamic Resource Constraints**

Shubham Vaishnav, Praveen Kumar Donta, Sindri Magnússon,  
In *Globecom 2025-IEEE International Global Communications Conference* (2025).

DOI: 10.1109/GLOBECOM59602.2025.11432479

**PAPER IV: Multi-objective and Constrained Reinforcement Learning for IoT**

Shubham Vaishnav, Sindri Magnússon,  
In *Learning Techniques for the Internet of Things*, Springer, Cham, pages 153–170 (2023).

DOI: [https://doi.org/10.1007/978-3-031-50514-0\\_8](https://doi.org/10.1007/978-3-031-50514-0_8)

**PAPER V: Intelligent Processing of Data Streams on the Edge Using Reinforcement Learning**

Shubham Vaishnav, Sindri Magnússon,  
In *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1265–1270 (2023).

DOI: 10.1109/ICCWorkshops57953.2023.10283692

PAPER VI: **Dynamic and Distributed Routing in IoT Networks Based on Multi-objective Q-Learning**

Shubham Vaishnav, Praveen Kumar Donta, Sindri Magnússon,

*IEEE Internet of Things Journal* (2026).

DOI: 10.1109/JIOT.2026.3666236

---

Reprints were made with permission from the publishers.

# Author's Contribution

In all papers included in this thesis, I took the leading role in the research. This involved defining the research problems, developing the proposed methods and algorithms, and implementing them in code. I designed and carried out the experiments, analysed the results, and prepared all figures and visualisations. I was also responsible for writing the manuscripts, revising them based on reviewer comments, and managing the submission process. The co-authors contributed through technical discussions, feedback on the ideas and presentation, and guidance in their supervisory roles.



# Contents

<b>Abstract</b>	<b>i</b>
<b>Sammanfattning</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>List of Papers</b>	<b>ix</b>
<b>Author's Contribution</b>	<b>xi</b>
<b>Abbreviations</b>	<b>xvii</b>
<b>List of Figures</b>	<b>xix</b>
<b>List of Tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>23</b>
1.1 Multi-objective Decision-Making Under Dynamic Constraints and Preferences . . . . .	23
1.2 Background . . . . .	25
1.2.1 AI-Based Decision-Making . . . . .	26
1.2.2 Resource Costs of Learning and Communication in IoT Systems . . . . .	26
1.2.3 Communication-Dominated Operating Regime . . . . .	27
1.2.4 Constraint-Aware Decision-Making . . . . .	29
1.2.5 Multi-objective Decision-Making in IoT Systems . . . . .	29
1.3 Research Aims . . . . .	30
1.4 Research Questions . . . . .	31
1.5 Contributions . . . . .	31
1.6 Summary of Contributions . . . . .	34
1.7 Outline . . . . .	36
<b>2 Extended Background</b>	<b>37</b>
2.1 AI-Based Decision-Making Under Uncertainty . . . . .	37

2.1.1	Multi-objective Formulations Beyond Linear Scalarisation . . . . .	38
2.1.2	Online Decision-Making and Regret . . . . .	39
2.2	Multi-armed Bandits . . . . .	39
2.2.1	Constrained and Budgeted Bandits . . . . .	40
2.3	Reinforcement Learning . . . . .	40
2.3.1	Markov Decision Processes . . . . .	41
2.3.2	Value Functions and Policy Learning . . . . .	42
2.3.3	Exploration–Exploitation Trade-Off . . . . .	42
2.3.4	AI-Based Reinforcement Learning in IoT Systems . . . . .	42
2.3.5	Limitations of Standard Reinforcement Learning . . . . .	43
2.3.6	Relevance to the Thesis Contributions . . . . .	43
2.4	Constrained Reinforcement Learning . . . . .	43
2.5	Multi-objective Decision-Making . . . . .	44
2.5.1	Multi-objective Optimisation . . . . .	44
2.5.2	Pareto Optimality and Pareto Fronts . . . . .	45
2.5.3	Limitations of Scalarisation-Based Approaches . . . . .	45
2.5.4	Multi-objective Reinforcement Learning . . . . .	46
2.5.5	MORL in IoT Systems . . . . .	47
2.5.6	Relation to the Thesis Contributions . . . . .	47
2.6	Distributed and Federated Learning . . . . .	47
2.6.1	Federated Learning in Resource-Constrained IoT Systems . . . . .	48
2.7	Rationale for Problem Formulations and Objectives . . . . .	49
2.7.1	Federated Learning in Distributed IoT Systems . . . . .	49
2.7.2	Channel Selection and Resource Allocation Using Multi-armed Bandits . . . . .	50
2.7.3	Conceptual Foundations for IoT Decision-Making . . . . .	50
2.7.4	Data Stream Processing and Offloading . . . . .	50
2.7.5	Routing in IoT Networks . . . . .	51
2.7.6	Summary of Objectives Across Problem Domains . . . . .	52
2.8	Summary . . . . .	52
<b>3</b>	<b>Methodology</b> . . . . .	<b>55</b>
3.1	Methodological Approach . . . . .	55
3.1.1	Design Science Research in Computer Science . . . . .	56
3.1.2	Design Science Research Process . . . . .	56
3.1.3	Application of DSR Across the Thesis Papers . . . . .	57
3.2	Data Sources . . . . .	59
3.2.1	Datasets for Federated Learning Studies . . . . .	60
3.2.2	Simulation Models for Constrained Multi-armed Bandits . . . . .	60

3.2.3	Simulation Models for Reinforcement Learning and Data Stream Processing . . . . .	61
3.2.4	Simulation Environment for Multi-objective Routing . . . . .	61
3.2.5	Reproducibility and Experimental Control . . . . .	62
3.3	Ethical Considerations . . . . .	63
3.3.1	Data Access, Privacy, and Integrity . . . . .	63
3.3.2	Manipulation, Security, and Robustness of Decision-Making . . . . .	63
3.3.3	Bias and Fairness in Algorithmic Decision Systems . . . . .	64
3.3.4	Autonomy, Accountability, and Human Oversight . . . . .	64
3.3.5	Ethical Framework and Positioning of the Thesis . . . . .	65
3.4	Societal Implications . . . . .	65
<b>4</b>	<b>Summary of Papers</b>	<b>69</b>
4.1	Results Outline . . . . .	69
4.2	Aim 1: Communication-Efficient Federated Learning . . . . .	69
4.2.1	Motivation . . . . .	69
4.2.2	Setup . . . . .	70
4.2.3	Approach and Methods . . . . .	71
4.2.4	Results . . . . .	74
4.3	Aim 2: Constraint-Aware Online Learning . . . . .	80
4.3.1	Motivation . . . . .	81
4.3.2	Setup . . . . .	82
4.3.3	Approach and Methods . . . . .	83
4.3.4	Results . . . . .	86
4.4	Aim 3: Preference-Adaptive Multi-objective Reinforcement Learning . . . . .	90
4.4.1	Motivation . . . . .	91
4.4.2	Setup . . . . .	92
4.4.3	Approach and Methods . . . . .	93
4.4.4	Results . . . . .	94
4.5	Summary . . . . .	100
<b>5</b>	<b>Concluding Remarks</b>	<b>103</b>
5.1	Discussion . . . . .	103
5.1.1	Aim I: Communication-Efficient Federated Learning . . . . .	103
5.1.2	Aim II: Constraint-Aware Online Learning . . . . .	104
5.1.3	Aim III: Preference-Adaptive Multi-objective Reinforcement Learning . . . . .	105
5.2	Conclusion . . . . .	105

5.2.1	Answering RQ1: How can communication and energy costs in distributed learning systems, such as federated learning, be reduced while preserving convergence and model performance? . . . .	106
5.2.2	Answering RQ2: How can online learning methods handle explicit and time-varying constraints in a principled and theoretically grounded manner? 107	
5.2.3	Answering RQ3: How can multi-objective reinforcement learning adapt routing decisions in IoT systems under dynamically changing trade-offs between energy consumption and reliability? . . .	107
5.2.4	Design Guidelines for AI-Driven IoT Systems Under Resource Constraints . . . . .	107
5.3	Limitations . . . . .	112
5.4	Future Directions . . . . .	113
5.4.1	Beyond Image Data: Streaming and Medical Time Series . . . . .	113
5.4.2	Other Future Work . . . . .	114

**References**

# Abbreviations

<b>AI</b>	Artificial Intelligence
<b>CNN</b>	Convolutional Neural Network
<b>CTS</b>	Constant- $T$ Sparsification
<b>DSR</b>	Design Science Research
<b>FL</b>	Federated Learning
<b>FL-CAT</b>	Communication-Adaptive Training for Federated Learning
<b>FL-CTS</b>	Federated Learning with Constant- $T$ Sparsification
<b>FL-TBS</b>	Federated Learning with Threshold-Based Sparsification
<b>FL-WS</b>	Federated Learning Without Sparsification
<b>GIP</b>	Grid-Interpolated Policy
<b>IID</b>	Independent and Identically Distributed
<b>IoT</b>	Internet of Things
<b>MAB</b>	Multi-armed Bandit
<b>MDP</b>	Markov Decision Process
<b>MEC</b>	Multi-access Edge Computing
<b>MNIST</b>	Modified National Institute of Standards and Technology Dataset
<b>MORL</b>	Multi-objective Reinforcement Learning
<b>PDR</b>	Packet Delivery Ratio
<b>Q-learning</b>	Action-Value Reinforcement Learning Algorithm
<b>R-learning</b>	Average-Reward Reinforcement Learning Algorithm
<b>RL</b>	Reinforcement Learning
<b>RLO</b>	Reinforcement Learning-based Offloading
<b>SNR</b>	Signal-to-Noise Ratio
<b>UCB</b>	Upper Confidence Bound
<b>vq</b>	Virtual Queue



# List of Figures

1.1	Illustrative summary of thesis structure, research questions, aims, and papers. . . . .	32
2.1	Reinforcement learning cycle. . . . .	41
2.2	Illustration of a Pareto front in a two-objective optimisation problem, showing trade-offs between energy consumption and communication reliability. Adapted from the book chapter (Paper IV) [4]. . . . .	46
3.1	Markov chain model used to generate the simulated input data stream for data stream processing experiments, adapted from Paper V. . . . .	61
4.1	High-level overview of communication-adaptive federated learning with sparsification and error compensation. Adapted from <b>Paper II</b> . . . . .	72
4.2	Total energy consumption under the affine communication cost model ( <b>Paper I</b> ). . . . .	75
4.3	Accuracy–energy trade-off and total data transfer under the affine communication cost model ( <b>Paper I</b> ). . . . .	76
4.4	Minimising communicated bits ( <b>Paper II</b> , Exp. 1: MNIST + logistic regression, convex). . . . .	78
4.5	Minimising energy consumption ( <b>Paper II</b> , Exp. 3: MNIST + logistic regression, convex). . . . .	79
4.6	Minimising communication time ( <b>Paper II</b> , Exp. 5: MNIST + logistic regression, convex). . . . .	79
4.7	Scalability and generalisability ( <b>Paper II</b> , Exp. 7–9: Fashion-MNIST + CNN, varying clients). . . . .	80
4.8	Performance evaluation of Budgeted UCB under randomly varying energy constraints ( <b>Paper III</b> ). . . . .	87

4.9	$\lambda$ -plot showing the trade-off between average energy consumption and average unprocessed data as the priority parameter $\lambda$ varies from 0 to 1. Results are shown for two real data streams and one simulated data stream (adapted from <b>Paper V</b> ). . . . .	89
4.10	Objectives in some common optimisation problems in the IoT. Adapted from <b>Paper IV</b> . . . . .	91
4.11	Overall reward (accumulated return) under sequential exploration–exploitation in <b>Paper VI</b> . . . . .	97
4.12	PDR performance (cumulative packets delivered) under simultaneous exploration–exploitation in <b>Paper VI</b> . . . . .	97
4.13	Energy consumption (cumulative/total) under simultaneous exploration–exploitation in <b>Paper VI</b> . . . . .	98
5.1	Conceptual overview of the thesis contributions, showing the connection between the research question, the three research aims, the associated papers, and their resulting contributions to adaptive multi-objective decision-making in the IoT. . . . .	106

# List of Tables

2.1	Overview of AI-based decision-making problems, modelling frameworks, and objectives considered in the thesis.	53
4.1	Sensitivity results for window length $W = 200$ (mean $\pm$ std).	98
4.2	Multi-objective optimisation problems in the IoT, representative MORL approaches, and their objectives (adapted from <b>Paper IV</b> ).	100



# 1. Introduction

This chapter introduces the scope, motivation, and structure of the thesis. It situates reinforcement learning (RL) and related online learning methods within the broader machine learning landscape, emphasising their role in sequential decision-making under uncertainty. While early learning approaches often focused on single-agent settings, practical systems increasingly require learning in distributed environments, where communication constraints, resource limitations, and system heterogeneity pose additional challenges. Against this background, the chapter formulates the overall research aim of the thesis and presents three specific research aims together with corresponding research questions. It further summarises the six papers that constitute the thesis and explains how they collectively address the stated aims. Finally, the chapter outlines the structure of the remainder of the thesis to guide the reader through the subsequent chapters.

## 1.1 Multi-objective Decision-Making Under Dynamic Constraints and Preferences

Modern engineered systems increasingly operate under multiple, simultaneously active objectives. Energy efficiency, latency, reliability, throughput, fairness, and privacy are often required to be optimised together rather than in isolation. These objectives are typically conflicting, and their relative importance (or preferences) may vary over time. In addition, such systems are subject to dynamic constraints. Resource budgets may tighten or relax, traffic patterns may change, environmental conditions may fluctuate, and user requirements may evolve. As a result, decision-making must simultaneously account for uncertainty, multiple objectives, and time-varying constraints.

In the context of this thesis, uncertainty refers to several distinct but interacting factors. First, decision-makers often operate with incomplete knowledge of system behaviour, including unknown state transitions, stochastic rewards, and uncertain cost signals. Second, constraints are not only dynamic but may themselves be uncertain or revealed online, for example, through time-varying resource budgets or feasibility limits that can only be observed through interaction. Third,

in multi-objective settings, uncertainty also arises from externally changing preference parameters that alter the relative importance of competing objectives. Finally, distributed operation introduces additional uncertainty due to heterogeneous devices, partial observability, and unreliable communication. The methods studied in this thesis are designed to operate under these combined sources of uncertainty.

In many real-world systems, particularly distributed and resource-constrained networks, full system models are unavailable, transition probabilities are unknown, and environmental statistics are non-stationary. Under such conditions, purely model-based dynamic programming becomes impractical, and static optimisation techniques fail to adapt to evolving trade-offs [7]. Internet of Things (IoT) systems represent a canonical example of this setting. Devices operate with limited energy, bandwidth, and computational capacity while serving heterogeneous and dynamically changing application requirements. Decisions such as routing, offloading, scheduling, and resource allocation must balance performance objectives against strict resource limitations.

Classical approaches to multi-objective optimisation include scalarisation methods and Pareto-front analysis from vector/multi-objective optimisation [8; 9], stochastic control [10], constrained Markov decision processes [11], and dynamic programming [7; 12]. When system dynamics are fully known, dynamic programming provides optimal policies through Bellman recursion [7; 12]. Constrained stochastic optimisation methods use Lagrangian relaxation or dual-variable techniques to balance competing objectives while ensuring constraint satisfaction [13]. Other classical multi-objective optimisation methods, such as weighted-sum and  $\epsilon$ -constraint formulations [14; 15], as well as evolutionary approaches like NSGA-II and MOEA-D [16; 17], typically assume fixed trade-off structures or predefined search preferences.

AI-based methods have shown considerable success in sequential decision-making; however, their effectiveness is often limited in scenarios with time-varying objectives, dynamic constraints, stochasticity, and partially known system behaviour. Some of these approaches, such as multi-task learning framed as multi-objective optimisation [18] and conflict-averse gradient-based multi-objective learning [19], typically assume static task preferences or a fixed conflict structure during training. Existing multi-objective reinforcement learning approaches typically rely on fixed preferences or stationary reward trade-offs, including scalarisation-based and Pareto-front learning methods [20; 21]. Constrained reinforcement learning methods are similarly grounded in stationary constrained Markov decision process (CMDP) formulations or fixed safety and resource thresholds. Some examples of these methods include classical CMDP models [11], policy optimisation ap-

proaches such as constrained policy optimisation (CPO) and reward-constrained policy optimisation [22; 23], and primal–dual methods [24]. Similar static assumptions also appear in distributed and federated learning, where the trade-off between learning performance and communication cost is typically governed by static compression or communication rules, such as deep gradient compression [25], sparsified stochastic gradient descent (SGD) with memory [26], error-feedback compression [27], and fixed local-update strategies [28; 29]. As a result, current AI techniques struggle to adapt when objectives, constraints, and system conditions evolve simultaneously. This limits their applicability in dynamic and resource-constrained environments.

This thesis formulates multi-objective decision-making under dynamic constraints and evolving preferences as its central research problem. Reinforcement learning, online learning, and bandit-based methods are adopted as suitable approaches for environments in which objectives, constraints, and system dynamics change over time and explicit system models are unavailable or impractical. The focus is not on learning preference representations themselves, but on designing learning algorithms that can efficiently adapt to changing preferences and constraints. The proposed methods enable rapid adjustment to new trade-offs without retraining from scratch while preserving theoretical guarantees and computational tractability.

## 1.2 Background

Autonomous decision-making is a central requirement in many modern systems, where agents must repeatedly select actions while operating under uncertainty. Artificial intelligence (AI) offers a broad range of approaches for enabling such behaviour, including optimisation, rule-based systems, and machine learning.

Within this landscape, reinforcement learning (RL) plays a distinct role by focusing on learning decision-making policies through interaction with the environment rather than relying on pre-collected training data. This interaction-based learning paradigm makes RL particularly relevant for sequential decision-making problems in dynamic settings where system models are incomplete or stochastically evolve over time.

Internet of Things (IoT) and wireless systems exemplify such environments. Devices are resource constrained, operate under uncertainty, and must often balance multiple, potentially conflicting objectives while adapting to changing conditions. These characteristics motivate the use of AI-based decision-making methods that can operate online, respect resource limitations, and adapt to evolving system requirements. This thesis investigates AI-based approaches for multi-

objective decision-making under such constraints, with IoT and wireless systems serving as a primary application domain.

### 1.2.1 AI-Based Decision-Making

AI-based decision-making methods aim to improve decisions over time by leveraging feedback from past actions. Supervised and unsupervised learning approaches have been highly successful for prediction and pattern recognition tasks, but they typically assume access to representative training data prior to deployment. In many real-world systems, such assumptions are difficult to satisfy due to non-stationarity, incomplete information, or the cost of data collection.

Reinforcement learning and online learning methods address these challenges by learning directly from interaction with the environment. Multi-armed bandits, online learning, and RL algorithms are designed for sequential decision-making under uncertainty, making them well suited for problems where system dynamics and reward structures are not fully known in advance.

Despite their success, existing RL-based methods often rely on assumptions that limit their applicability in dynamic and partially observable environments. Many reinforcement learning and online learning approaches presume stationary reward structures, even though real-world systems frequently exhibit changing objectives. These limitations motivate the development of AI-based decision-making methods that can adapt to evolving objectives and constraints while operating under incomplete knowledge of system dynamics.

### 1.2.2 Resource Costs of Learning and Communication in IoT Systems

In practical systems, learning and decision-making are subject to resource limitations. IoT and wireless devices often operate under strict constraints on energy, communication bandwidth, and computational capacity. These limitations affect not only the decisions being made but also the learning process itself.

In distributed learning settings such as federated learning (FL), multiple devices collaboratively train a shared model by exchanging updates over a network. As models increase in size and complexity, communication becomes a dominant cost. In wireless systems, transmitting data can consume significantly more energy than local computation [30]. This creates a trade-off between learning performance and resource consumption, motivating the development of communication- and energy-aware learning algorithms.

While resource-aware learning has received increasing attention, most existing AI-based approaches treat resource constraints as fixed external limitations rather than as dynamic and uncertain quantities. Communication-efficient and energy-aware learning methods commonly rely on static budgets and communication costs, predefined compression levels, or offline tuning. In practice, however, resource availability and communication costs may fluctuate over time in a stochastic manner. This highlights the need for learning algorithms that explicitly account for the cost of learning and decision-making under dynamic resource constraints and incomplete system information.

The IoT systems considered in this thesis consist of battery-powered devices operating under strict communication and energy limitations. Typical examples include wireless sensors, wearables, and industrial monitoring nodes using low-power wireless technologies such as BLE, Zigbee, LoRaWAN, or NB-IoT.

Although the nominal physical-layer data rates of such technologies may range from a few kbit/s to a few hundred kbit/s, the effective application-layer throughput is significantly lower in practice. Medium access overhead, duty cycling, acknowledgments, retransmissions, and multi-year lifetime requirements reduce the usable uplink capacity per device. In realistic deployments, the effective per-node uplink throughput often lies in the range of approximately 100–5,000 bit/s.

### 1.2.3 Communication-Dominated Operating Regime

**When does communication energy dominate computation?** To make the assumed operating regime explicit, let us consider the representative orders of magnitude for low-power IoT platforms. For typical microcontrollers (e.g. ARM Cortex-M class devices), the energy per elementary operation (e.g. integer or floating-point multiply–accumulate) is on the order of picojoules to low nanojoules when these microcontrollers are operated at low supply voltages and modest clock frequencies [31–33]. In contrast, the energy required to transmit one bit over a low-power wireless link (e.g. BLE, Zigbee, LoRaWAN, NB-IoT) is typically several orders of magnitude larger, often in the range of hundreds of nanojoules to microjoules per bit when accounting for protocol overhead, medium access, and retransmissions [34].

Under such conditions, transmitting a packet containing hundreds or thousands of bits may consume energy comparable to performing millions of local arithmetic operations. In this regime, communication dominates whenever

$$N_{\text{bits}}E_{\text{bit}} \gg N_{\text{ops}}E_{\text{op}},$$

which is typical for battery-powered IoT nodes with low-rate radios and lightweight local processing. Another example of these energy dynamics can be seen in wireless communications, in general, where typically sending a single bit of data consumes an amount of energy at least 480 times that required for executing one CPU addition instruction [30].

The algorithms developed in this thesis are designed for this communication-dominated regime. This assumption is consistent with the empirical findings in Papers I, II, V, and VI, where reducing the number of transmitted bits yields substantial energy savings, while the additional local computation required for sparsification, buffering, or learning has a negligible impact on overall energy consumption [26; 35–37]. For example, in the federated learning experiments of Papers I and II, transmitting dense, full-precision model updates typically requires on the order of  $10^6$ – $10^7$  bits per communication round per client, whereas adaptive sparsification reduces this to  $10^4$ – $10^5$  bits, thus saving approximately two to three orders of magnitude (i.e. a  $10^2$ – $10^3 \times$  reduction) of transmitted data while preserving the learning performance. By contrast, the additional local computation involved consists primarily of simple vector operations and bookkeeping, incurring only marginal overhead.

A representative class of real-world applications that motivate this communication-dominated setting arises in cross-device federated learning deployments on user devices, such as mobile keyboards and other on-device personalisation applications. In these scenarios, machine learning models are trained collaboratively across large numbers of smartphones using local user data, for tasks such as text prediction and user-specific model adaptation, without transferring raw data to a central server. Because these devices operate under strict battery constraints and rely on wireless communication, they are highly sensitive to communication cost and energy consumption. Prior studies have repeatedly shown that transmitting model updates between devices and a server dominates energy usage in such settings, whereas performing additional local computation on the device incurs relatively little overhead [35–37].

The communication cost models and resource budgets used throughout this thesis reflect this constrained setting. Papers I and II restrict the number of transmitted gradient or model elements per federated learning round. Paper III introduces explicit dynamic resource budgets. Paper V models per-offload communication costs and limited transmission capacity. Paper VI captures energy scarcity through per-hop transmission costs in multi-hop routing. Although they are expressed in normalised units, these models correspond to low-throughput, energy-constrained IoT environments where communication is the dominant limitation.

### 1.2.4 Constraint-Aware Decision-Making

Many IoT decision-making problems involve explicit operational constraints, such as energy budgets, communication limits, or safety requirements. While temporary violations of such constraints may be acceptable during early learning phases, long-term operation typically requires sustained constraint satisfaction.

Learning under constraints introduces additional challenges beyond unconstrained optimisation. Algorithms must balance exploration and performance improvement with adherence to constraints over time. Bandit-based methods and reinforcement learning provide principled frameworks for addressing such problems by incorporating constraint handling directly into the learning process.

Constrained reinforcement learning and CMDP-based formulations provide principled methods for handling explicit constraints, but they typically assume stationary constraint thresholds. Many methods rely on accurately estimated constraint models, which are often unavailable in practice. In real systems, constraints such as energy budgets or communication limits can vary unpredictably and be observed only indirectly. These challenges motivate constraint-aware learning approaches that can adapt online to stochastically time-varying constraints.

### 1.2.5 Multi-objective Decision-Making in IoT Systems

IoT systems often require balancing multiple objectives simultaneously. Examples include routing, task processing, and resource allocation. In routing, for instance, decisions may involve trade-offs between energy consumption and communication reliability. Improving reliability may require additional transmissions and higher energy usage, while conserving energy may reduce the packet delivery performance.

Multi-objective reinforcement learning (MORL) provides a framework for studying such problems by learning policies that adapt to changing trade-offs between competing objectives, rather than optimising a single metric. Although multi-objective reinforcement learning offers a natural framework for IoT decision-making, most existing approaches assume fixed objective preferences or require retraining when trade-offs change. In dynamic IoT environments, however, priorities such as energy efficiency, reliability, and latency may unpredictably evolve over time. This limits the practical applicability of current MORL methods and motivates the development of preference-adaptive decision-making frameworks that can respond to changing objectives and constraints without relying on model retraining.

This thesis focuses on AI-based methods that can handle such multi-objective decision-making problems in a dynamic and resource-constrained

setting. The six papers included in this thesis address these challenges using AI-based approaches, spanning constrained bandit models, multi-objective reinforcement learning, and communication-aware distributed learning. Together, they form a coherent body of work on adaptive and resource-aware decision-making for IoT and wireless systems.

### 1.3 Research Aims

The overall aim of this thesis is to develop AI-based methods for adaptive multi-objective decision-making under resource constraints, with applications to IoT and wireless systems. In this thesis, the IoT is viewed as an application-level concept in which physical devices are equipped with sensing, computation, and actuation capabilities to enable monitoring, control, and data-driven services. These devices rely on wireless systems and networks to communicate, exchange data, and coordinate actions. Examples include wireless sensor networks and technologies such as Wi-Fi, Bluetooth, cellular networks, and low-power wide-area networks. In this sense, wireless systems provide the underlying communication infrastructure, while the IoT represents a class of applications that use this infrastructure. Many IoT deployments can therefore be viewed as application-driven instances of wireless and sensor networks, often operating under tight constraints on energy, bandwidth, and computation.

This overall aim is addressed through the following specific aims:

- **Aim I: Communication-Efficient Federated Learning:** This aim addresses the trade-off between the communication cost and learning performance in distributed systems such as federated learning. Transmitting large model updates improves global accuracy but increases communication overhead, while aggressive compression conserves bandwidth at the expense of model quality. The challenge lies in identifying an adaptive balance between accuracy and the communication cost, which may vary over time and across heterogeneous nodes. The objective is to design distributed learning mechanisms that dynamically regulate communication intensity while preserving convergence and performance guarantees.
- **Aim II: Constraint-Aware Online Learning** This aim focuses on decision-making under explicit and time-varying operational constraints, such as energy budgets, resource limits, or safety thresholds. In many IoT settings, temporary constraint violations may be tolerable during exploration, but long-term operation requires

sustained compliance. The challenge is to integrate constraint awareness directly into the learning process without reducing performance. The goal is to develop online learning algorithms that balance performance optimisation with principled and theoretically grounded constraint handling.

- **Aim III: Preference-Adaptive Multi-objective Reinforcement Learning:** This aim investigates MORL as a framework for handling dynamically changing trade-offs in IoT systems. In such environments, the relative importance of objectives such as energy consumption, reliability, or latency may shift due to evolving system requirements. The focus is not on learning preference parameters but on designing MORL algorithms that can efficiently adapt to externally changing preference vectors without retraining. This framework is studied in the context of distributed IoT routing, where dynamic adaptation to energy–reliability trade-offs is essential for maintaining stable and efficient network performance.

## 1.4 Research Questions

This thesis addresses the following main research question:

**Main RQ:** How can AI-based methods support adaptive multi-objective decision-making under dynamic preferences and resource constraints in IoT and wireless systems?

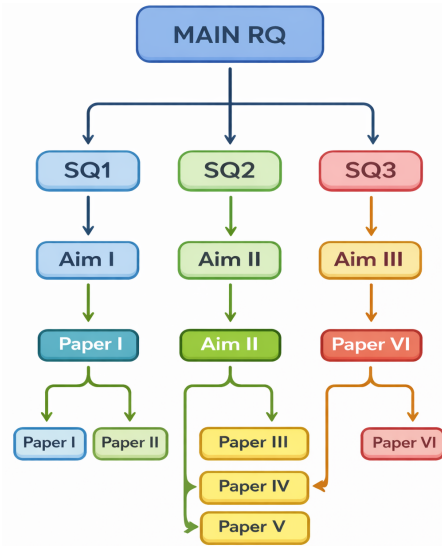
To answer this main question, the thesis focuses on the following sub-questions (SQs):

1. How can communication and energy costs in distributed learning systems, such as federated learning, be reduced while preserving convergence and model performance?
2. How can online learning methods handle explicit and time-varying constraints in a principled and theoretically grounded manner?
3. How can multi-objective reinforcement learning adapt routing decisions in IoT systems under dynamically changing trade-offs between energy consumption and reliability?

## 1.5 Contributions

This thesis consists of six peer-reviewed papers that collectively address the main research question (RQ) and the three sub-questions (SQ1–SQ3) formulated above. Each paper focuses on a specific AI-based

decision-making problem arising in IoT and wireless systems while contributing to the overarching goal of adaptive multi-objective decision-making under resource and constraint limitations. Fig. 1.1 shows the correlation between the RQ, SQs, aims, and papers.



**Figure 1.1:** Illustrative summary of thesis structure, research questions, aims, and papers.

**Paper I: Energy-Efficient and Adaptive Gradient Sparsification for Federated Learning** This paper focuses on reducing communication and energy costs in distributed learning systems. It proposes an adaptive gradient sparsification approach that dynamically selects the sparsification level by optimising the trade-off between information content and energy consumption. The method adapts online to data characteristics and communication costs, removing the need for manual hyperparameter tuning. This work primarily addresses **SQ1** by demonstrating how communication and energy costs can be reduced in federated learning while preserving the learning performance and contributes to **RQ** by enabling scalable learning under resource constraints.

**Paper II: Communication-Adaptive Gradient Sparsification for Federated Learning with Error Compensation** Building on the previous paper, this work incorporates error compensation mechanisms into communication-adaptive gradient sparsification to improve convergence and robustness. The proposed method explicitly accounts for heterogeneous communication conditions and provides theoretical convergence guaran-

tees under adaptive sparsification. This paper further strengthens the contribution to **SQ1**, showing how resource-aware learning algorithms can be designed to operate reliably in practical distributed settings, and thus supports the main research question **RQ**.

**Paper III: Adaptive Budgeted Multi-armed Bandits for IoT with Dynamic Resource Constraints** This paper addresses online decision-making under explicit and time-varying constraints. It proposes a constrained stochastic bandit framework with a dynamically decaying violation budget, allowing controlled constraint violations during early learning while ensuring long-term constraint satisfaction. The paper introduces the Budgeted UCB algorithm and establishes theoretical guarantees on regret and constraint violations. This work primarily contributes to answering **SQ2** by demonstrating how online learning methods can handle explicit and dynamic constraints in a principled manner and thereby supports the overarching research question **RQ**.

**Paper IV: Multi-objective and Constrained Reinforcement Learning for IoT** This book chapter provides a conceptual and methodological foundation for multi-objective and constrained reinforcement learning in IoT systems. It reviews key challenges, formalises representative IoT decision-making problems, and discusses how multi-objective and constrained reinforcement learning frameworks can be applied in practice. While this chapter does not introduce a new algorithm, it contextualises and unifies the learning paradigms explored in the thesis. As such, it provides supporting contributions to **RQ** and the sub-questions (**SQ2** and **SQ3**) by framing constrained, multi-objective, and resource-aware learning as central themes in IoT optimisation.

**Paper V: Intelligent Processing of Data Streams on the Edge Using Reinforcement Learning** This paper studies decision-making for data stream processing at the network edge, where nodes must decide whether to process data locally, offload it, or defer processing under resource constraints. The problem is formulated as a reinforcement learning task, and an RL-based framework is proposed to learn energy-efficient processing policies online. This work contributes to **RQ** by demonstrating AI-based decision-making under resource constraints in a realistic IoT edge-computing scenario and complements **SQ2** by extending constraint-aware learning beyond bandit-based formulations.

**Paper VI: Dynamic and Distributed Routing in IoT Networks Based on Multi-objective Reinforcement Learning** This paper studies rout-

ing in IoT networks as a multi-objective decision-making problem, focusing on the trade-off between energy consumption and communication reliability. A distributed multi-objective reinforcement learning framework is proposed that enables nodes to adapt routing decisions online as objective preferences change over time. The paper shows how policies can generalise across different preference settings without retraining. This work directly addresses **SQ3** by demonstrating how multi-objective reinforcement learning can be used to adapt routing decisions under dynamically changing trade-offs and contributes to **RQ** by illustrating adaptive decision-making in distributed IoT systems.

## 1.6 Summary of Contributions

The six papers included in this thesis collectively address the main research question (RQ) by investigating AI-based decision-making under resource and constraint limitations from complementary perspectives. Rather than focusing on a single learning paradigm, the thesis spans communication-aware distributed learning, constrained online learning, and multi-objective reinforcement learning, thereby providing a coherent contribution to adaptive multi-objective decision-making in IoT and wireless systems.

Papers I and II primarily address **SQ1**, which concerns reducing communication and energy costs in distributed learning systems. Both papers focus on federated learning as a representative distributed learning framework and propose adaptive gradient sparsification methods that explicitly account for communication and energy constraints. Paper I introduces an energy-efficient and adaptive sparsification strategy that dynamically balances learning performance and resource consumption, while Paper II extends this approach by incorporating error compensation mechanisms and accounting for heterogeneous communication conditions.

Papers III and V mainly contribute to **SQ2**, which focuses on learning-based decision-making under explicit and dynamic constraints. Paper III introduces a constrained bandit framework with a dynamically decaying violation budget, enabling principled exploration while ensuring long-term constraint satisfaction. Paper V extends constraint-aware learning to an edge computing scenario, where reinforcement learning is used to manage data stream processing decisions under limited resources. Together, these papers show how learning algorithms can be designed to respect operational constraints over time while remaining adaptive to changing system conditions and application requirements.

Paper VI directly addresses **SQ3** by studying multi-objective reinforcement learning for routing in IoT networks. It demonstrates how

routing decisions can adapt online to dynamically changing trade-offs between energy consumption and communication reliability in a distributed setting. Rather than relying on fixed scalarisation or repeated retraining, the proposed approach enables policies to generalise across different objective preferences, highlighting the suitability of multi-objective reinforcement learning for dynamic IoT environments.

Paper IV complements and unifies the technical contributions of the thesis by providing a broader conceptual and methodological foundation for multi-objective and constrained reinforcement learning in IoT systems. While it does not introduce a new algorithm, it contextualises the learning paradigms explored in the other papers and frames constrained, multi-objective, and resource-aware learning as central challenges in IoT optimisation. As such, it provides supporting contributions to the main research question and to the sub-questions (SQ2 and SQ3).

**On communication and energy model abstraction.** Across the six papers, communication and energy costs are modelled at an abstract level. In Papers I–III, the proposed algorithms operate on scalar communication or cost signals and do not assume a specific physical-layer or protocol model. In particular, the packet-based and affine communication models used in Papers I and II serve as illustrative instantiations of the cost term rather than structural assumptions of the method. The algorithms optimise a utility–cost trade-off and remain applicable under arbitrary monotonic communication cost functions.

In Papers V and VI, additive cost structures are adopted for tractability, where communication and computation costs are modelled as separable components. While these instantiations assume additive per-action or per-hop costs, the underlying reinforcement learning frameworks operate on abstract reward and cost signals and are not restricted to a specific hardware platform or radio technology. Mapping the abstract cost terms to a particular hardware platform requires only instantiating the relevant cost parameters (e.g. energy per bit or per operation), without modifying the learning algorithms. Thus, the contributions of this thesis are formulated at the algorithmic and system level rather than at the physical-layer modelling level.

Taken together, the papers in this thesis provide complementary contributions that collectively answer the main research question. They demonstrate that AI-based learning methods can support adaptive multi-objective decision-making in IoT and wireless systems by explicitly accounting for resource constraints, dynamic objectives, and the cost of learning itself. The combination of theoretical analysis, algorithm design, and application-driven evaluation highlights both the generality

of the proposed approaches and their relevance for real-world, resource-constrained systems.

## 1.7 Outline

The remainder of the thesis is organised to guide the reader from background and methodological foundations to a synthesis of the included research contributions and their implications. Chapter 2 provides an extended background on AI-based decision-making, covering online learning, reinforcement learning, multi-objective optimisation, and resource-aware learning concepts that underpin the work presented in the thesis. Chapter 3 describes the methodological framework adopted across the included studies, including the problem formulation, algorithm design principles, and evaluation methodology. Chapter 4 summarises the six papers included in the thesis and synthesises their results with respect to the research aims and sub-questions, highlighting the key ideas, methodological contributions, and findings of each study. Finally, Chapter 5 concludes the thesis by reflecting on how the included papers collectively address the main research question, discussing the limitations of the current work, and outlining directions for future research. The appended part of the thesis contains the full versions of Papers I–VI in their published form, allowing the reader to examine each contribution in detail.

## 2. Extended Background

This chapter provides a detailed background on the AI-based decision-making paradigms that underpin the contributions of this thesis. The focus is on sequential decision-making under uncertainty, with particular emphasis on multi-armed bandits, reinforcement learning, constrained and multi-objective optimisation, and distributed and federated learning. These paradigms are central to the design of adaptive decision-making algorithms for IoT and wireless systems, where devices operate under limited resources, partial information, and dynamically (potentially stochastically) changing conditions.

While these paradigms have achieved significant success in controlled and well-modelled settings, many existing methods rely on assumptions such as stationary environments, fixed objectives, or known system models. In realistic IoT and wireless deployments, where resources, constraints, and operating conditions evolve over time, and information is incomplete, these assumptions often limit the practical applicability and adaptability of current state-of-the-art solutions. The purpose of this chapter is to establish a common conceptual and mathematical foundation for the proposed methods and applications presented in the subsequent chapters.

### 2.1 AI-Based Decision-Making Under Uncertainty

IoT and wireless systems are inherently uncertain environments. Devices operate with incomplete information about network conditions, traffic patterns, and resource availability, and must make decisions sequentially as the system evolves over time. AI-based decision-making addresses such problems by enabling agents to adapt their behaviour through interaction with the environment and feedback from observed outcomes, rather than relying on fixed models or offline optimisation [38; 39].

In sequential decision-making problems, an agent selects an action at each time step and observes a reward or cost signal that depends on unknown system dynamics. The objective is typically to optimise long-term performance metrics, such as the cumulative reward or average cost, rather than immediate outcomes. This perspective naturally leads

to online learning, bandit models, and reinforcement learning formulations, which explicitly capture uncertainty, temporal dependencies, and the exploration–exploitation trade-off.

Despite their suitability for sequential decision-making under uncertainty, many existing learning-based approaches assume stationary system behaviour, fixed performance objectives, or stable resource constraints during learning. In practice, IoT and wireless systems often exhibit evolving traffic patterns, time-varying resource availability, and changing operational requirements, which can limit the adaptability and robustness of current state-of-the-art methods.

### 2.1.1 Multi-objective Formulations Beyond Linear Scalarisation

Linear weighted-sum scalarisation is one of the most common approaches to multi-objective reinforcement learning. Given objective vector  $\mathbf{r}(s, a)$  and weight vector  $\mathbf{w}$ , the scalarised reward is  $r_{\mathbf{w}}(s, a) = \mathbf{w}^{\top} \mathbf{r}(s, a)$ . While computationally convenient, this formulation represents only one special case among several possible multi-objective decision-making frameworks.

Alternative approaches include learning value functions conditioned on preference vectors [40; 41], convex-envelope representations of optimal  $Q$ -functions [41], hypervolume-guided dynamic weight adaptation [42], and non-linear scalarisation techniques such as TOPSIS-based ranking [43] and generalised lexicographic ordering [44]. Preference inference frameworks further aim to recover implicit scalarisation coefficients from demonstrations rather than assuming they are known [45; 46].

Outside reinforcement learning, dynamic programming and constrained Markov decision processes provide classical solutions for sequential decision-making under uncertainty. When full system dynamics are known, dynamic programming yields optimal policies through Bellman recursion. However, in stochastic IoT systems, transition probabilities, arrival processes, and cost structures are often unknown or non-stationary, limiting the direct applicability of such model-based approaches.

Constrained stochastic control methods, including Lagrangian relaxation and Lyapunov drift-plus-penalty techniques, provide principled ways to convert constraints into dynamically evolving dual variables [47]. In such formulations, queue backlogs or dual multipliers effectively act as adaptive weights that balance competing objectives while ensuring long-term constraint satisfaction.

The methodological choice of this thesis is motivated by the characteristics of IoT systems, which are not adequately addressed by the

above-discussed approaches: unknown dynamics, distributed operation, and dynamically (potentially stochastically) changing objectives. Reinforcement learning and online learning methods are chosen because they enable policy adaptation through interaction without requiring full system models. Linear scalarisation is adopted as a computationally tractable and interpretable mechanism that integrates naturally with both bandit and RL formulations.

The novelty of this thesis lies not in using weighted scalarisation *per se*, but in designing learning algorithms where scalarisation coefficients are constraint-aware, dynamically adjustable, or knowledge-driven. In particular, the included papers demonstrate the following: (i) adaptive constraint handling via decaying violation budgets, (ii) distributed preference generalisation across weight vectors without retraining, and (iii) the explicit modelling of resource trade-offs in stochastic IoT environments. These contributions position scalarisation not as a static design choice but as an integral and adaptive component of the learning system.

### 2.1.2 Online Decision-Making and Regret

Online decision-making considers settings in which actions are selected sequentially and feedback is revealed incrementally. Let  $a_t$  denote the action selected at time  $t$ , and let  $r_t(a_t)$  denote the corresponding reward. A common performance metric is regret, defined as

$$R_T = \sum_{t=1}^T r_t(a^*) - \sum_{t=1}^T r_t(a_t), \quad (2.1)$$

where  $a^*$  denotes the best fixed action in hindsight.

Regret-based analysis provides a principled way to quantify how quickly an AI-based decision-making algorithm learns to perform near-optimally. Low regret guarantees imply that suboptimal decisions are made only a limited number of times, which is particularly important in IoT systems, where poor decisions may incur high energy costs or degrade system performance. Online decision-making methods are therefore well suited for environments in which system conditions evolve over time and pre-collected training data may be unavailable or unreliable [48].

## 2.2 Multi-armed Bandits

The multi-armed bandit (MAB) problem is a fundamental abstraction for online decision-making under uncertainty. In its stochastic formulation, an agent repeatedly selects one of  $K$  actions (arms), each associated

with an unknown reward distribution. The objective is to maximise cumulative reward by balancing the exploration of uncertain actions and the exploitation of actions believed to be optimal [49].

In IoT and wireless systems, bandit models have been widely used to study adaptive sampling, channel selection, and resource allocation problems, where decisions must be made sequentially under uncertainty and feedback is limited. The simplicity of the bandit framework makes it particularly attractive for resource-constrained devices, as it avoids explicit state modelling.

### 2.2.1 Constrained and Budgeted Bandits

Classical bandit models focus solely on reward maximisation, but many IoT applications involve explicit resource constraints. For example, a device may have a limited energy budget that restricts how often it can perform certain actions. Constrained bandit models extend the classical formulation by associating each action with both a reward and a cost. In a pioneering work on constrained bandits, Badanidiyuru *et al.* [50] proposed two algorithms for the Bandits with Knapsacks problem – `BalancedExploration` and `PrimalDualBwk`.

A generic constrained bandit problem can be written as

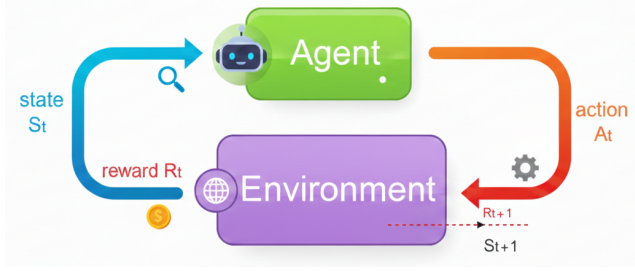
$$\max \sum_{t=1}^T r_t(a_t) \quad \text{s.t.} \quad \sum_{t=1}^T c_t(a_t) \leq B, \quad (2.2)$$

where  $c_t(a_t)$  denotes the cost incurred at time  $t$  and  $B$  is a budget.

Budgeted bandit formulations allow limited constraint violations during early learning phases while enforcing long-term feasibility. This is particularly relevant for IoT systems with dynamic resource availability, as studied in the bandit-based contribution of this thesis. Existing constrained bandit algorithms achieve near-optimal regret under fixed, known resource constraints. However, they involve an assumption of static budgets and do not cater to the needs of systems with dynamically evolving constraints.

## 2.3 Reinforcement Learning

Reinforcement learning (RL) is a central paradigm for AI-based sequential decision-making in environments where actions influence both immediate outcomes and future system states. Unlike supervised or unsupervised learning, RL does not rely on labelled datasets or static training data. Instead, an agent learns through interaction with the environment, making RL particularly suitable for dynamic and uncertain systems such as IoT and wireless networks [38; 39].



**Figure 2.1:** Reinforcement learning cycle.

In RL, the agent repeatedly observes the state of the environment, selects an action, and receives feedback in the form of a reward signal (Fig. 2.1). The goal is to learn a policy that optimises long-term performance rather than immediate reward. This long-term perspective is essential in IoT systems, where short-term gains may lead to long-term degradation, for example, through excessive energy consumption or resource depletion.

### 2.3.1 Markov Decision Processes

RL problems are commonly formalised using Markov decision processes (MDPs). An MDP is defined by a tuple

$$(\mathcal{S}, \mathcal{A}, P, r),$$

where  $\mathcal{S}$  denotes the state space,  $\mathcal{A}$  the action space,  $P(s' | s, a)$  the transition probability from state  $s$  to state  $s'$  under action  $a$ , and  $r(s, a)$  the reward function.

At each time step  $t$ , the agent observes the current state  $s_t \in \mathcal{S}$ , selects an action  $a_t \in \mathcal{A}$  according to a policy  $\pi(a | s)$ , and transitions to a new state  $s_{t+1}$  while receiving a reward  $r(s_t, a_t)$ . The objective is to learn a policy that maximises the expected discounted return,

$$J(\pi) = \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \quad (2.3)$$

where  $\gamma \in (0, 1)$  is a discount factor that controls the relative importance of future rewards [38].

The MDP framework enables the flexible modelling of IoT decision-making, where states may represent device energy levels, network conditions, or workload states, and actions correspond to routing choices, transmission decisions, or processing modes.

### 2.3.2 Value Functions and Policy Learning

A key concept in RL is the value function, which quantifies the expected return under a given policy. The state-value function is defined as

$$V^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right], \quad (2.4)$$

while the action-value function is defined as

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]. \quad (2.5)$$

Value-based methods aim to learn these functions and derive policies by selecting actions that maximise the estimated value. Policy-based methods, in contrast, directly parameterise and optimise the policy. Actor–critic methods combine both approaches by maintaining separate representations for the policy and the value function [38].

From the perspective of this thesis, the specific choice of RL algorithm is less important than the underlying modelling framework. The MDP abstraction enables the formulation of IoT decision problems in a principled manner and provides a foundation for incorporating constraints and multiple objectives.

### 2.3.3 Exploration–Exploitation Trade-Off

A fundamental challenge in RL is the exploration–exploitation trade-off. The agent must explore different actions to acquire information about the environment while exploiting known actions to achieve high reward. Excessive exploration may waste resources, whereas insufficient exploration may lead to suboptimal long-term performance.

In IoT systems, exploration can be costly. For example, exploring alternative routing paths may increase energy consumption or reduce the packet delivery performance. This makes exploration strategies particularly important in resource-constrained environments and motivates the use of structured exploration methods, as well as constraint-aware formulations.

### 2.3.4 AI-Based Reinforcement Learning in IoT Systems

Applying RL in IoT systems introduces several practical challenges. First, IoT devices often have limited computational and memory resources, restricting the complexity of models that can be deployed. Second, decision-making is frequently decentralised, with nodes relying on local observations and limited communication. Third, system

dynamics may be non-stationary due to changing network conditions, mobility, or workload variations [51].

Despite these challenges, RL has been successfully applied to a variety of IoT problems, including routing, task offloading, energy management, and adaptive resource allocation [52; 53]. AI-based RL enables devices to adapt their behaviour online without requiring accurate system models, which is particularly advantageous in complex and dynamic environments.

### 2.3.5 Limitations of Standard Reinforcement Learning

While reinforcement learning provides a powerful framework for sequential decision-making, standard RL formulations often assume a single scalar reward and unconstrained optimisation. In IoT systems, these assumptions are frequently violated. Devices must operate under explicit resource constraints and balance multiple, often conflicting objectives, whose relative importance may also change over time. Moreover, both objectives and constraints may evolve dynamically and, in some cases, stochastically due to changing network conditions, workloads, or application requirements.

For example, a routing policy that maximises the packet delivery probability may incur excessive energy consumption, while an energy-efficient policy may degrade communication reliability. Similarly, resource constraints such as energy budgets or latency limits may tighten or relax over time rather than remaining fixed. These limitations motivate extensions of RL that explicitly account for constraints and multiple objectives, such as constrained reinforcement learning and multi-objective reinforcement learning, which are discussed in subsequent sections.

### 2.3.6 Relevance to the Thesis Contributions

Reinforcement learning provides the conceptual foundation for several contributions of this thesis. It underpins the formulation of routing and edge processing problems as sequential decision-making tasks and enables AI-based adaptation to dynamic system conditions. However, the thesis goes beyond standard RL by explicitly addressing constraints, communication costs, and multiple changing objectives, reflecting the practical requirements of IoT and wireless systems.

## 2.4 Constrained Reinforcement Learning

Many IoT decision-making problems involve explicit operational constraints, such as energy budgets, latency requirements, or safety condi-

tions. These constraints must be respected over time, which introduces additional complexity beyond unconstrained optimisation.

Constrained reinforcement learning extends standard RL formulations by introducing cost functions in addition to rewards. A constrained RL problem can be written as

$$\max_{\pi} \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \quad (2.6)$$

$$\text{s.t. } \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t c_k(s_t, a_t) \right] \leq d_k, \quad (2.7)$$

where  $c_k$  denotes the  $k$ -th cost signal.

Such formulations are central to sustainable IoT operation and form the conceptual basis for the constrained decision-making problems studied in this thesis.

## 2.5 Multi-objective Decision-Making

Decision-making problems in IoT and wireless systems are rarely characterised by a single performance objective. Instead, devices and network entities must simultaneously account for multiple objectives that are often conflicting in nature. These objectives arise from physical resource limitations, application-level requirements, and system-level constraints. As a result, AI-based decision-making in IoT systems must explicitly consider trade-offs between competing objectives rather than optimising a single metric in isolation [48].

Typical objectives in IoT systems include energy consumption, communication reliability, latency, throughput, computational load, and network lifetime. For example, improving reliability often requires re-transmissions or redundant routing paths, which increases energy consumption. Similarly, reducing latency may require higher transmission power or more frequent communication, again impacting energy usage. These interdependencies make it difficult to define a single objective that accurately captures system performance under all operating conditions, motivating multi-objective formulations.

### 2.5.1 Multi-objective Optimisation

Multi-objective optimisation provides a formal framework for reasoning about problems involving multiple conflicting objectives. Instead of optimising a scalar objective function, the goal is to optimise a vector-

valued objective,

$$\max_{\pi} \mathbf{J}(\pi) = (J_1(\pi), J_2(\pi), \dots, J_M(\pi)), \quad (2.8)$$

where each component  $J_m(\pi)$  represents a distinct performance criterion.

In contrast to single-objective optimisation, multi-objective optimisation does not yield a unique optimal solution in general. Instead, it produces a set of solutions that represent different trade-offs between objectives. These solutions provide flexibility in system operation, allowing decision-makers or higher-level policies to select operating points that best match current requirements.

In IoT systems, such flexibility is essential, as operating conditions and priorities may change over time due to variations in workload, energy availability, or network conditions.

## 2.5.2 Pareto Optimality and Pareto Fronts

The concept of Pareto optimality is central to multi-objective optimisation. A policy  $\pi_1$  is said to dominate another policy  $\pi_2$  if it is at least as good in all objectives and strictly better in at least one. A policy is Pareto optimal if no other policy dominates it.

The set of all Pareto-optimal solutions forms the Pareto front, which characterises the achievable trade-offs between objectives. Points on the Pareto front represent different compromises, such as low energy consumption with reduced reliability or high reliability with increased energy usage.

In IoT decision-making, Pareto fronts provide valuable insight into the structure of trade-offs inherent in the system. Rather than committing to a single operating point, Pareto-optimal solutions allow the system to adapt dynamically to changing preferences or constraints.

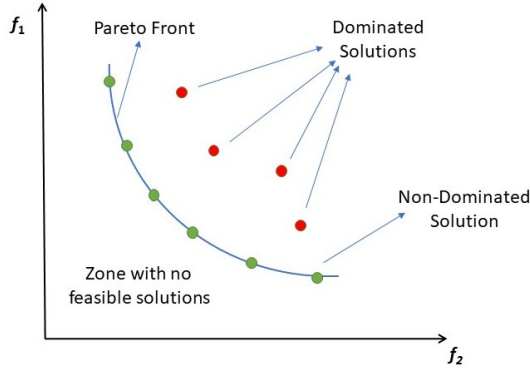
## 2.5.3 Limitations of Scalarisation-Based Approaches

A common approach to handling multiple objectives is scalarisation, where objectives are combined into a single scalar reward using a weighted sum,

$$J(\pi) = \sum_{m=1}^M w_m J_m(\pi), \quad (2.9)$$

with fixed weights  $w_m$ .

While scalarisation simplifies optimisation, it has several limitations in IoT settings. First, fixed weights assume static preferences, which



**Figure 2.2:** Illustration of a Pareto front in a two-objective optimisation problem, showing trade-offs between energy consumption and communication reliability. Adapted from the book chapter (Paper IV) [4].

may not hold in dynamic environments. For example, a device may prioritise energy efficiency when battery levels are low but shift focus to reliability during critical data transmission. Second, scalarisation may fail to capture non-convex regions of the Pareto front, leading to sub-optimal trade-off exploration [54].

These limitations motivate AI-based methods that can represent and adapt to multiple objectives explicitly.

#### 2.5.4 Multi-objective Reinforcement Learning

Multi-objective reinforcement learning (MORL) extends standard reinforcement learning by allowing reward signals to be vector-valued rather than scalar [55]. At each time step, the agent observes a reward vector,

$$\mathbf{r}(s, a) = (r_1(s, a), r_2(s, a), \dots, r_M(s, a)), \quad (2.10)$$

corresponding to multiple objectives.

A common MORL approach is to learn separate value functions for each objective,

$$\mathbf{V}^\pi(s) = (V_1^\pi(s), V_2^\pi(s), \dots, V_M^\pi(s)), \quad (2.11)$$

allowing explicit reasoning about trade-offs at decision time. Preference information, such as weight vectors or utility functions, can then be applied dynamically to select actions based on current system priorities.

This decoupling of policy learning and preference specification is particularly valuable in IoT systems, where objectives and constraints may evolve over time.

### 2.5.5 MORL in IoT Systems

MORL is well suited for IoT decision-making problems that involve persistent trade-offs between competing objectives. In routing, for example, energy consumption and communication reliability must be balanced dynamically as network conditions change. In edge computing, trade-offs arise between the local computation cost, communication overhead, and processing delay.

By learning policies that capture the structure of these trade-offs, MORL enables adaptive decision-making without retraining when preferences change. This capability is demonstrated in the routing paper included in this thesis, where distributed MORL is used to adapt routing decisions online under changing objective preferences.

### 2.5.6 Relation to the Thesis Contributions

The discussion in this section provides the conceptual foundation for the multi-objective decision-making contributions of this thesis. In particular, it motivates the use of MORL as a principled alternative to static scalarisation approaches in dynamic IoT environments. The routing contribution demonstrates how MORL can be applied in a distributed setting to enable adaptive, resource-aware decision-making under changing system conditions.

## 2.6 Distributed and Federated Learning

As IoT systems scale in size and complexity, centralised learning approaches become increasingly impractical due to communication overhead, privacy concerns, and resource limitations [56]. Devices are often geographically distributed, connected via unreliable wireless links, and constrained in terms of energy, bandwidth, and computational capacity. These characteristics motivate distributed learning paradigms in which computation is pushed closer to the data sources and learning is performed collaboratively across devices.

Federated learning (FL) has emerged as a prominent AI-based approach for distributed model training in such environments. Instead of collecting data at a central server, FL enables devices to train local models on private data and periodically communicate model updates to a central aggregator [35]. This reduces the need for raw data transmission and addresses privacy concerns while still allowing models to benefit from data distributed across the network.

### 2.6.1 Federated Learning in Resource-Constrained IoT Systems

In a standard FL setup, a global model with parameters  $\theta$  is maintained at a central aggregator. At each communication round, a subset of devices performs local training and computes model updates  $\Delta\theta_k^t$ , which are aggregated to update the global model:

$$\theta_{t+1} = \theta_t + \sum_k w_k \Delta\theta_k^t, \quad (2.12)$$

where  $w_k$  represents aggregation weights that may depend on local dataset sizes or device participation.

While this framework is conceptually simple, its deployment in IoT systems introduces significant challenges. In wireless environments, communication is often the dominant cost, frequently exceeding the cost of local computation [57]. Transmitting large model updates can rapidly deplete device energy budgets and saturate limited bandwidth. For example, in Natural Language Processing (NLP), models that are several MBs in size are common [58]. In addition, device heterogeneity, time-varying channel conditions, and intermittent connectivity complicate reliable aggregation and slow convergence.

These challenges motivate the design of communication-efficient and resource-aware FL methods that explicitly account for system constraints. Among such approaches, gradient sparsification has received significant attention. The key idea is to transmit only a subset of the most informative components of the local update, thereby reducing the communication cost. For example, devices may transmit only the top- $k$  elements of the gradient in terms of magnitude while dropping the remaining components.

However, naive sparsification can lead to biased updates and degraded convergence behaviour. To address this issue, error compensation mechanisms are often employed. In these methods, the information lost due to sparsification is accumulated locally and added back to future updates, ensuring that dropped gradient components are not permanently discarded. This allows sparsified FL algorithms to retain convergence properties comparable to full-gradient methods while significantly reducing communication overhead.

The federated learning contributions of this thesis build on these ideas by introducing adaptive gradient sparsification strategies that dynamically adjust the sparsification level based on learning dynamics and resource considerations. Instead of using fixed sparsification parameters, the proposed methods adapt to changing communication costs and data characteristics, balancing communication efficiency, energy consumption, and learning performance. Papers I and II demonstrate that such adaptive strategies can substantially reduce communication

overhead while maintaining convergence guarantees, making FL more suitable for practical IoT deployments.

## 2.7 Rationale for Problem Formulations and Objectives

The AI-based decision-making problems studied in this thesis are selected to reflect representative challenges encountered in IoT and wireless systems, where decisions must be made sequentially under uncertainty, resource constraints, and competing performance requirements. Rather than focusing on a single application domain, the thesis considers a set of complementary problem classes that arise at different layers of IoT systems, including distributed learning, resource allocation, data processing, and networking. Together, these problems provide a coherent basis for studying adaptive decision-making under multiple objectives and constraints.

A common characteristic of the selected problems is that system behaviour unfolds over time and decisions have long-term consequences. As a result, short-term optimisation or static heuristics are often insufficient. The chosen problem formulations therefore emphasise sequential decision-making models, such as reinforcement learning and multi-armed bandits, which explicitly account for uncertainty, feedback, and temporal dependencies. In addition, the objectives considered in each problem are chosen to reflect practical system-level trade-offs rather than abstract performance metrics.

### 2.7.1 Federated Learning in Distributed IoT Systems

Federated learning represents a class of distributed optimisation problems that arise when multiple IoT devices collaboratively train a shared model without sharing raw data. In wireless and energy-constrained environments, communication often becomes the dominant cost, frequently exceeding the cost of local computation. At the same time, learning performance must be preserved to ensure model accuracy and convergence.

In this thesis, federated learning is studied from a resource-aware perspective, with a focus on reducing communication and energy costs while maintaining learning performance. The primary objectives considered include communication efficiency, energy consumption, and learning accuracy. These objectives reflect practical constraints in distributed IoT deployments and motivate adaptive communication strategies, such as gradient sparsification and error compensation, which are

explored in the federated learning contributions of this thesis (Papers I and II).

### 2.7.2 Channel Selection and Resource Allocation Using Multi-armed Bandits

Channel selection and resource allocation problems arise naturally in wireless IoT systems, where devices must choose among multiple communication options under uncertainty. Channel conditions may vary due to interference, fading, or congestion, and probing or using a channel incurs energy and bandwidth costs.

In this thesis, such problems are modelled using multi-armed bandit formulations, where each arm represents a channel or communication action. Rewards capture the communication performance, through metrics such as the throughput or successful transmission probability, while costs represent resource usage, using metrics such as the energy consumption. Constrained and budgeted bandit models are employed to ensure that resource limitations are respected over time. This formulation captures the exploration–exploitation trade-off inherent in wireless resource selection and provides a lightweight decision-making model suitable for resource-constrained devices (Paper III).

### 2.7.3 Conceptual Foundations for IoT Decision-Making

In addition to application-specific problem formulations, the thesis includes a conceptual treatment of AI-based decision-making for IoT systems. This perspective is provided through a book chapter that surveys and structures multi-objective and constrained reinforcement learning approaches in the context of the IoT.

The objectives considered at this level are intentionally general and focus on characterising trade-offs between performance, constraints, and resource usage. Rather than addressing a single application, this contribution provides a unifying conceptual framework that motivates the problem formulations and modelling choices adopted throughout the thesis (Paper IV).

### 2.7.4 Data Stream Processing and Offloading

Data stream processing at the network edge represents a fundamental IoT decision-making problem, where devices must decide how to handle incoming data streams under limited computational and energy resources. At each decision point, a device may process data locally, offload it to a neighbouring node or cloud server, or defer processing.

These decisions directly affect energy consumption, the processing delay, and the quality of the processed data.

In this thesis, data stream processing and offloading are formulated as sequential decision-making problems and addressed using reinforcement learning. The primary objectives considered include minimising energy consumption while maintaining an acceptable processing performance, based on metrics such as task completion or inference accuracy. This formulation captures a common edge computing scenario in IoT systems, where devices must continuously adapt their behaviour to changing workloads and resource availability (Paper V).

### 2.7.5 Routing in IoT Networks

Routing is a core networking function in IoT systems and provides a natural application for AI-based multi-objective decision-making. Nodes must select forwarding actions based on local observations, such as the link quality or neighbour status, while adapting to time-varying network conditions. Routing decisions have long-term effects on network performance, influencing energy depletion, packet delivery success, and the overall network lifetime.

Routing in IoT networks is a representative example of sequential decision-making under uncertainty, resource constraints, and partial observability. Nodes must decide how to forward data packets based on local information, such as link quality estimates or neighbour states, while adapting to time-varying network conditions. These decisions directly affect system-level performance metrics, including energy consumption, communication reliability, and network lifetime. Traditional routing protocols for IoT and wireless sensor networks typically rely on static heuristics or predefined metrics, such as the shortest path or minimum hop count. While these approaches are simple and computationally lightweight, they often assume relatively stable network conditions and fixed optimisation objectives. In dynamic IoT environments, however, link qualities, traffic patterns, and node energy levels may change over time, making static routing strategies suboptimal.

In this thesis, the routing problem is modelled as a multi-objective sequential decision-making problem in which each forwarding action yields stochastic outcomes in terms of packet delivery success and resource consumption. Energy consumption and communication reliability form a fundamental trade-off: improving reliability often requires retransmissions or redundancy, increasing energy usage, while aggressive energy-saving strategies may degrade packet delivery performance. These conflicting objectives motivate a multi-objective formulation of routing decisions [48]. The formulation reflects widely

studied performance metrics in IoT networking and aligns with practical deployment requirements (Paper VI).

Reinforcement learning provides a natural framework for adaptive routing, as it enables nodes to learn forwarding policies from interaction with the network without requiring explicit models of link dynamics. However, standard single-objective RL formulations rely on scalar reward functions with fixed weights, which may fail to capture dynamically changing priorities, for example, when energy constraints become more critical as battery levels decrease.

Multi-objective reinforcement learning (MORL) addresses this limitation by allowing reward signals to be vector-valued, enabling the explicit representation of multiple objectives [55]. In the routing context, objectives such as energy consumption and communication reliability can be treated separately, allowing policies to adapt dynamically to changing trade-offs. The routing contribution of this thesis demonstrates how MORL can be applied in a distributed IoT setting, enabling nodes to adapt routing behaviour online without retraining as objective preferences change.

### 2.7.6 Summary of Objectives Across Problem Domains

Across the different problem domains considered in this thesis, a recurring set of objectives emerges. Energy consumption is a central concern due to the battery-powered nature of many IoT devices. Communication-related objectives, such as reliability, throughput, and bandwidth usage, play a key role in networking and distributed learning scenarios. Application-level performance metrics, including processing accuracy or task completion, capture the quality of service provided by the system.

By selecting problem formulations and objectives that are representative of real-world IoT systems, the thesis ensures that the proposed AI-based decision-making methods address practical challenges. At the same time, the diversity of problem settings highlights the generality of the underlying approaches, demonstrating how similar decision-making principles can be applied across different layers of IoT systems. Table 2.1 summarises the AI-based decision-making problems considered in this thesis, the corresponding modelling frameworks, and the primary objectives addressed in each case.

## 2.8 Summary

This chapter reviewed AI-based decision-making paradigms relevant to the thesis, including online decision-making, multi-armed bandits,

**Table 2.1:** Overview of AI-based decision-making problems, modelling frameworks, and objectives considered in the thesis.

<b>Problem Domain</b>	<b>Modelling Framework</b>	<b>Primary Objectives</b>	<b>Related Paper(s)</b>
Federated learning in distributed IoT systems	Distributed optimisation with adaptive communication	Commun. cost; energy consumption; model accuracy	Papers I, II
Channel selection and resource allocation	Constrained multi-armed bandits	Transmission success; energy budget compliance	Paper III
Conceptual foundations for IoT decision-making	Multi-objective and constrained reinforcement learning	Performance–constraints–resources trade-offs	Paper IV
Data stream processing and offloading	Reinforcement learning (R-learning)	Energy consumption; processing performance	Paper V
Routing in IoT networks	Multi-objective reinforcement learning	Energy consumption; communication reliability	Paper VI

reinforcement learning, constrained and multi-objective optimisation, and federated learning. These concepts provide the theoretical foundation for the methodology and application-driven studies presented in the subsequent chapters.

# 3. Methodology

This chapter presents the methodological foundations of the thesis and explains how the research was conducted to address the stated research questions. It describes the overarching research strategy, the types of data and evaluation settings used across the included studies, and the ethical and societal considerations relevant to the proposed AI-based decision-making methods. Given the algorithmic and systems-oriented nature of the thesis, the chapter emphasises the rationale for adopting a design-oriented research approach and clarifies how rigour and relevance are ensured through formal modelling, controlled experimentation, and systematic evaluation. Together, the sections of this chapter provide transparency concerning the research process and establish the methodological coherence of the contributions presented in the subsequent chapters.

## 3.1 Methodological Approach

This thesis adopts *Design Science Research* (DSR) as its overarching methodological approach. DSR is a problem-driven research paradigm that focuses on the purposeful design, development, and evaluation of artefacts that address identified real-world challenges [59]. Rather than aiming primarily at explanation or prediction, DSR emphasises the creation of solutions that are both practically relevant and scientifically grounded [60].

Within computer science and artificial intelligence research, DSR is commonly applied in algorithmic and systems-oriented studies, where artefacts take the form of models, methods, algorithms, or system architectures [61]. In this thesis, the artefacts are AI-based decision-making methods, including learning algorithms, optimisation frameworks, and distributed decision-making mechanisms, designed to support adaptive multi-objective decision-making under resource and constraint limitations in IoT and wireless systems.

The suitability of DSR for this thesis stems from three key characteristics of the research. First, the addressed problems are practically motivated, arising from concrete limitations in IoT systems such as energy constraints, communication overhead, and dynamic operating condi-

tions. Second, the contributions are constructive in nature, proposing novel algorithms and decision-making frameworks rather than purely analytical results. Third, the proposed artefacts are rigorously evaluated using formal analysis and controlled experimental studies, ensuring both scientific rigour and practical relevance.

### 3.1.1 Design Science Research in Computer Science

Design Science Research was originally articulated by Simon as a paradigm concerned with the design of artefacts that satisfy specified goals under constraints [62]. Subsequent work has formalised DSR within information systems and computer science by emphasising the central role of artefact construction, systematic evaluation, and iterative refinement [59; 63].

In algorithmic research, DSR treats artefact building as a theory-informed search process in a constrained design space. Design decisions are guided by established theoretical foundations, such as reinforcement learning, multi-armed bandits, and optimisation theory, while evaluation is conducted through mathematical analysis, simulation, and benchmarking. This perspective aligns naturally with the methodological practices adopted in this thesis.

### 3.1.2 Design Science Research Process

This thesis follows the DSR process as a structured approach to developing and evaluating AI-based decision-making artefacts. DSR conceptualises research as an iterative cycle of problem identification, artefact construction, and evaluation, with the goal of producing solutions that are both theoretically grounded and practically relevant.

The main stages of the DSR process, and their general meaning in the context of design-oriented research, are outlined below:

1. **Problem Identification and Motivation:** This step involves identifying a relevant and non-trivial problem grounded in real-world practice and motivating why existing approaches are insufficient. In DSR, the problem definition establishes the scope, relevance, and practical significance of the research and guides all subsequent design decisions.
2. **Definition of Objectives for a Solution:** Based on the identified problem, this stage specifies what constitutes a successful solution. Objectives are typically derived from domain requirements and theoretical considerations, and may include performance, efficiency, robustness, or constraint satisfaction criteria.

3. **Design and Development:** In this phase, one or more artefacts are constructed to address the defined objectives. Artefacts in DSR may include algorithms, models, frameworks, or methods. Design choices are informed by existing theories and prior work, and the resulting artefacts embody the proposed solution.
4. **Demonstration:** Demonstration involves showing how the developed artefact can be applied to a representative problem or scenario. This step illustrates the feasibility and applicability of the solution, often through case studies, simulations, or example applications.
5. **Evaluation:** The evaluation stage assesses how well the artefact meets the defined objectives. In algorithmic research, this typically includes theoretical analysis, simulation-based experiments, and comparison against baseline or state-of-the-art methods to establish rigour and effectiveness.
6. **Communication:** The final step concerns communicating the problem, artefact, evaluation results, and contributions to relevant audiences. In DSR, this includes disseminating findings through peer-reviewed publications and clearly articulating both theoretical and practical implications.

These stages are not strictly sequential; rather, DSR emphasises iterative refinement, where insights gained during evaluation may lead to revisions of the problem formulation or artefact design. The six papers included in this thesis collectively instantiate these stages, with each contribution addressing specific aspects of problem formulation, artefact design, and evaluation within the broader research agenda.

### 3.1.3 Application of DSR Across the Thesis Papers

The six papers included in this thesis collectively implement the DSR process. Rather than each paper independently covering all steps, different papers emphasise different phases of the process, reflecting the iterative and cumulative nature of the research.

**Problem Explication** The problem explication phase focuses on identifying and structuring the challenges associated with AI-based decision-making in IoT and wireless systems. Across the thesis, these challenges include uncertainty in system dynamics, limited energy and communication resources, explicit operational constraints, and conflicting performance objectives.

Paper IV plays a central role in this phase by synthesising existing literature on constrained and multi-objective reinforcement learning in IoT systems. It formalises representative decision-making problems and highlights the limitations of standard single-objective and unconstrained learning approaches. In addition, the problem context is further refined in Papers III, V, and VI, which focus on specific application scenarios such as budgeted decision-making, data stream processing, and routing under competing objectives.

**Definition of Objectives** Following problem explication, the thesis defines explicit and measurable objectives that guide artefact design. These objectives are chosen to reflect practical system-level trade-offs rather than abstract optimisation criteria.

Across the included studies, the objectives include energy consumption, communication efficiency, learning accuracy, reliability, and long-term constraint satisfaction. Paper III explicitly formalises objectives and constraints within a budgeted multi-armed bandit framework, while Papers I and II define objectives related to communication efficiency and convergence in federated learning. In Papers V and VI, objectives are formulated to balance energy usage against application-level performance metrics such as processing quality and communication reliability.

**Design and Development of Artefacts** The core design activity of the thesis consists of developing AI-based decision-making artefacts that address the defined objectives. These artefacts include algorithms, learning frameworks, and decision-making models.

Papers I and II design adaptive gradient sparsification mechanisms for federated learning, incorporating communication awareness and error compensation. Paper III introduces the Budgeted UCB algorithm for constrained online decision-making. Paper V develops a reinforcement learning framework for adaptive data stream processing and offloading. Paper VI designs a distributed multi-objective reinforcement learning framework for IoT routing. Although Paper IV does not introduce a new algorithm, it contributes conceptual artefacts by formalising problem classes and design principles that inform subsequent algorithmic developments.

**Demonstration** Demonstration in this thesis is achieved by applying the developed artefacts to representative IoT and wireless system scenarios. These demonstrations establish feasibility and illustrate how the proposed methods operate in practice.

For example, the federated learning algorithms in Papers I and II are demonstrated through distributed training scenarios under constrained communication budgets. The Budgeted UCB algorithm in Paper III is demonstrated in simulated IoT decision-making settings with dynamic resource constraints. Papers V and VI demonstrate reinforcement learning-based decision-making in edge computing and routing scenarios, respectively.

**Evaluation** Evaluation constitutes a central component of the thesis and is primarily quantitative in nature. The proposed artefacts are evaluated using a combination of theoretical analysis and empirical experiments.

Theoretical evaluation includes regret bounds, convergence guarantees, and constraint violation analysis, as presented in Papers II, III, and VI. Empirical evaluation is conducted through controlled simulations and benchmarks, assessing performance metrics such as energy consumption, learning accuracy, communication overhead, and reliability. Comparative evaluations against baseline and state-of-the-art methods are used to assess the effectiveness and robustness of the proposed approaches.

**Communication** The final DSR activity concerns the communication of research results. The artefacts and findings of this thesis are communicated through six peer-reviewed publications in journals, conferences, and workshops. In addition, the compilation thesis itself serves as an integrative communication artefact, synthesising the individual contributions and situating them within a coherent methodological and conceptual framework.

## 3.2 Data Sources

The research presented in this thesis focuses on the design and evaluation of AI-based decision-making algorithms rather than on data-centric modelling. Consequently, the data sources used across the included studies primarily consist of publicly available benchmark datasets and synthetically generated data obtained from controlled simulation environments. These choices enable systematic experimentation, reproducibility, and the precise analysis of algorithmic behaviour under well-defined conditions.

The specific datasets and simulation environments used in each paper are described below.

### 3.2.1 Datasets for Federated Learning Studies

The federated learning experiments presented in Papers I and II are conducted using the MNIST handwritten digit classification dataset [64]. Paper II also involves the Fashion-MNIST Dataset [65]. The dataset consists of 60,000 training images and 10,000 test images of greyscale handwritten digits. This dataset is widely used in the federated learning literature and provides a well-understood benchmark for evaluating convergence behaviour and classification accuracy in distributed learning settings.

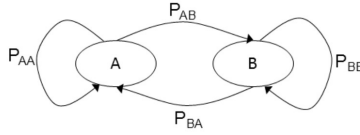
In each experimental run, the training dataset is randomly partitioned across the simulated client devices to emulate decentralised data ownership. The test dataset is shared across all simulations and is never used for training. Each experiment is repeated five times with different random data partitions, and the reported results are obtained by averaging across runs. This setup enables the robust evaluation of learning performance under varying data distributions while ensuring comparability across methods.

Communication and energy costs are modelled explicitly within the simulation framework. The cost of each communication round depends on the size of transmitted model updates, allowing the direct assessment of the impact of adaptive gradient sparsification and error compensation on communication efficiency and learning accuracy.

### 3.2.2 Simulation Models for Constrained Multi-armed Bandits

The constrained multi-armed bandit study in Paper III relies on synthetic environments designed to capture decision-making under uncertainty and dynamic resource constraints. Each arm represents a possible action, such as selecting a communication channel or operational mode, and is associated with a stochastic reward distribution and a stochastic cost distribution.

The environments are modelled using stationary or slowly varying stochastic processes, and resource constraints are implemented through time-varying budgets that limit cumulative cost. A budgeted violation model is employed, allowing limited constraint violations during early learning phases while enforcing long-term constraint satisfaction. These environments enable the controlled measurement of regret and constraint violations, supporting both theoretical analysis and empirical evaluation.



**Figure 3.1:** Markov chain model used to generate the simulated input data stream for data stream processing experiments, adapted from Paper V.

### 3.2.3 Simulation Models for Reinforcement Learning and Data Stream Processing

The experiments related to multi-objective reinforcement learning for data stream processing and offloading (Paper V) employ both simulated and real-world data streams. The simulated data stream is generated using a discrete-time Markov chain model [66], which captures temporal correlations in data arrival patterns.

The Markov chain consists of two states, denoted as  $A$  and  $B$ , as illustrated in Fig. 3.1. In state  $A$ , the incoming data size is 2 bytes, while in state  $B$ , it is 4 bytes. State transitions are governed by the probabilities  $P_{AA} = 0.3$ ,  $P_{AB} = 0.7$ ,  $P_{BA} = 0.2$ , and  $P_{BB} = 0.8$ . This model enables the controlled evaluation of decision-making behaviour under stochastic and temporally correlated workloads.

In addition to the simulated stream, two real-world data streams are used. The first is derived from the PLACES-2 dataset [67], which consists of a large-scale collection of images representing diverse scenes and provides a high-throughput data stream suitable for evaluating processing and offloading decisions. The second real-world dataset is the BATTLEDIM 2020 dataset [68], which contains sensor data streams from water distribution systems collected as part of the Battle of the Leakage Detection and Isolation Methods challenge. These datasets enable evaluation under realistic and heterogeneous workload conditions.

### 3.2.4 Simulation Environment for Multi-objective Routing

The multi-objective routing experiments presented in Paper VI are conducted using a custom simulation environment designed to model decentralised routing in resource-constrained IoT networks. The simulator abstracts key characteristics of low-power wireless networks while maintaining sufficient structure to enable the controlled evaluation of sequential and multi-objective decision-making.

The simulated network follows a grid-based topology in which nodes are arranged on a two-dimensional lattice, and communication is re-

stricted to one-hop neighbours. Each node can directly communicate only with its adjacent nodes, reflecting a unit-disc communication model commonly used to approximate short-range wireless connectivity in IoT systems. This abstraction captures local forwarding behaviour while avoiding reliance on global network knowledge.

Routing is modelled as an episodic decision-making problem. In each episode, a data packet is generated at a randomly selected source node and must be forwarded to a fixed sink node. An episode terminates either when the packet successfully reaches the sink or when forwarding fails due to unreliable nodes in the network. Unreliable nodes emulate packet loss by dropping packets with a fixed probability, allowing the systematic evaluation of communication reliability under stochastic failures.

Energy consumption is explicitly modelled at the packet-forwarding level. Nodes incur a baseline energy cost for remaining active and an additional cost for each packet transmission. Each node is initialised with a finite energy budget, and energy depletion directly affects long-term routing performance. This modelling choice enables the evaluation of trade-offs between energy consumption and communication reliability, which constitute the primary objectives in the routing problem.

This simulation environment enables the controlled analysis of multi-objective reinforcement learning for routing, isolating the effects of objective trade-offs, stochastic failures, and preference dynamics. While it is abstracted, the environment reflects key characteristics of IoT routing scenarios and provides a principled testbed for evaluating adaptive, distributed decision-making methods.

### 3.2.5 Reproducibility and Experimental Control

Across all studies, datasets, simulation models, and evaluation protocols are documented in the corresponding papers to support reproducibility. The use of publicly available datasets such as MNIST and FASHION-MNIST, together with explicitly defined stochastic and Markovian simulation environments, enables controlled experimentation and facilitates comparison with related work.

While real-world deployment is beyond the scope of this thesis, the chosen data sources and simulation environments are designed to reflect representative IoT and wireless system conditions reported in the literature. This approach balances experimental control with practical relevance and ensures that the evaluated AI-based decision-making methods yield meaningful insights for real-world applications.

### 3.3 Ethical Considerations

The integration of AI-based decision-making into IoT and wireless systems raises a range of ethical considerations related to data handling, autonomy, accountability, and societal impact. IoT systems operate at the intersection of the digital and physical worlds, often collecting, processing, and acting upon data in a continuous and autonomous manner. As a result, algorithmic decisions may have consequences that extend beyond purely technical performance, making ethical reflection an important aspect of system design.

#### 3.3.1 Data Access, Privacy, and Integrity

One of the most prominent ethical concerns in IoT systems relates to access to information, the privacy of data, and the integrity of communicated information [69]. IoT devices frequently collect sensitive data and exchange information over wireless links, making them vulnerable to unauthorised access, data breaches, or manipulation. Compromised data integrity can lead not only to privacy violations but also to incorrect system behaviour, potentially causing physical or economic harm.

Although the algorithms developed in this thesis do not directly operate on personal or sensitive data, they are intended for deployment in environments where secure and privacy-preserving data handling is essential. Papers I and II explicitly address this concern by adopting federated learning, which avoids the centralised collection of raw data and thereby reduces exposure to privacy risks. By design, only model updates are exchanged between devices and the aggregator, limiting direct access to local data. Nevertheless, these papers also acknowledge that model updates themselves may leak information if not carefully managed, highlighting the importance of communication-efficient and controlled update mechanisms.

#### 3.3.2 Manipulation, Security, and Robustness of Decision-Making

IoT systems are exposed to adversarial conditions, including malicious nodes, unreliable links, and the intentional manipulation of data or behaviour. Ethical concerns arise when AI-based systems are insufficiently robust to such conditions, as failures may propagate across distributed networks and affect system reliability or safety.

Several contributions in this thesis implicitly address these concerns through the explicit modelling of uncertainty, failures, and constraints. Paper III models decision-making under dynamic resource constraints, allowing controlled violations during early learning while enforcing long-term constraint satisfaction. This approach reflects an ethical stance

that prioritises safe long-term operation over short-term performance gains. Similarly, Paper VI incorporates unreliable nodes and stochastic packet drops into the routing environment, ensuring that routing policies are evaluated under realistic and potentially adversarial conditions rather than idealised assumptions.

### 3.3.3 Bias and Fairness in Algorithmic Decision Systems

Bias in algorithmic decision-making is a well-recognised ethical concern in machine learning systems [70]. Bias may arise from data distributions, modelling assumptions, or optimisation objectives, and can lead to systematically unfair or undesirable outcomes. In IoT systems, biased decision-making may result in uneven resource allocation, the premature depletion of certain devices, or degraded service for specific network regions.

The papers included in this thesis address bias primarily through transparent problem formulation and explicit objective modelling. In Paper V, for example, reinforcement learning is used to balance energy consumption and processing performance in data stream processing, avoiding heuristics that implicitly favour short-term throughput at the expense of long-term device sustainability. In Paper VI, the use of multi-objective reinforcement learning avoids the fixed scalarisation of objectives, allowing policies to adapt to changing preferences rather than embedding static biases into the reward structure.

### 3.3.4 Autonomy, Accountability, and Human Oversight

The increasing autonomy of IoT systems raises ethical questions concerning accountability and responsibility for system behaviour. As emphasised in early discussions on automation and cybernetics, excessive delegation of decision-making authority to machines risks obscuring human responsibility [71; 72]. In distributed IoT systems, this challenge is amplified by the absence of centralised control and the presence of many interacting autonomous agents.

The contributions of this thesis acknowledge these concerns by emphasising the explicit modelling of objectives, constraints, and system assumptions. Rather than treating AI-based decision-making as a black box, the proposed methods are designed to be analysable and interpretable at the level of objectives and trade-offs. Paper IV plays a central role in this regard by providing a conceptual framework for multi-objective and constrained reinforcement learning in IoT systems, highlighting the importance of explicit trade-off modelling as a means of retaining human oversight and control.

### 3.3.5 Ethical Framework and Positioning of the Thesis

To contextualise these considerations, this thesis draws on principles outlined in the *Ethics of Artificial Intelligence and Robotics* in the Stanford Encyclopedia of Philosophy [73]. The most relevant principles for the scope of this work include privacy and surveillance, the manipulation of data and behaviour, bias in decision systems, and the ethical implications of autonomous systems.

Rather than adopting a single abstract ethical framework, such as utilitarian or deontological ethics, the thesis takes a pragmatic approach aligned with the realities of IoT and wireless systems. Ethical considerations are addressed through careful system modelling, explicit constraint handling, robustness to uncertainty, and transparent evaluation. While application-specific ethical concerns vary widely across IoT domains, the methodological choices made throughout the thesis reflect an effort to mitigate ethical risks associated with AI-based autonomy and distributed decision-making.

## 3.4 Societal Implications

The research presented in this thesis contributes to the development of AI-based decision-making methods for IoT and wireless systems, which increasingly underpin critical digital infrastructure in domains such as sensing, communication, and distributed data processing. As such, the proposed methods have societal implications related to sustainability, reliability, privacy, and the responsible deployment of autonomous systems.

A recurring societal concern addressed across the thesis is the sustainability of large-scale IoT deployments. Many IoT systems consist of battery-powered devices that are difficult or costly to maintain. Inefficient decision-making can lead to excessive energy consumption, shortened device lifetimes, and increased electronic waste. Several contributions in this thesis directly address this issue by explicitly modelling energy consumption as a primary objective. Papers I and II reduce communication overhead in federated learning, thereby lowering the energy usage associated with wireless transmissions. Paper V similarly targets energy-efficient operation at the network edge by balancing processing decisions against energy constraints. Collectively, these contributions support more sustainable operation of IoT systems.

Another important societal implication concerns data governance and privacy. Centralised data collection in IoT systems raises concerns about surveillance, misuse of data, and loss of user control. The federated learning approaches developed in Papers I and II contribute

to addressing these concerns by enabling collaborative model training without the centralised aggregation of raw data. By keeping data local to devices, these methods align with societal expectations around data minimisation and privacy-aware system design. At the same time, the thesis acknowledges that privacy-preserving learning is not guaranteed by decentralisation alone, reinforcing the need for careful system design and evaluation.

The increasing autonomy of IoT systems also has implications for reliability, trust, and accountability in digital infrastructure. Autonomous decision-making can improve efficiency and responsiveness, but may also introduce new failure modes if systems behave unpredictably or if responsibility for outcomes becomes unclear. As emphasised by Weizenbaum, excessive delegation of decision authority to machines risks obscuring human responsibility for decisions and their consequences [74]. This concern is particularly relevant in distributed IoT systems, where decisions emerge from the interaction of many autonomous agents rather than from a single centralised controller.

Several papers in this thesis address these concerns by explicitly modelling uncertainty, constraints, and long-term system behaviour. Paper III demonstrates how online decision-making under resource constraints can be designed to avoid persistent violations, supporting predictable long-term operation. Paper VI shows how routing decisions can adapt to changing objectives without retraining, contributing to resilient and flexible network behaviour. These contributions emphasise controlled autonomy rather than unconstrained optimisation.

From a broader perspective, the thesis highlights the importance of transparency and explicit trade-off modelling in AI-based decision-making systems. While the proposed methods do not aim to provide full interpretability in the sense of explainable AI, they make objectives, constraints, and performance trade-offs explicit at the modelling level. Paper IV plays a central role in articulating this perspective by framing multi-objective and constrained reinforcement learning as tools for reasoning about trade-offs rather than hiding them within fixed heuristics or opaque reward functions.

Finally, the societal implications of this thesis extend to the role and responsibility of computer scientists. As AI-based systems become more deeply embedded in everyday infrastructure, researchers bear responsibility for ensuring that proposed methods are robust, realistic, and evaluated under conditions that reflect practical constraints. In line with the ACM Code of Ethics and Professional Conduct [75], this thesis emphasises thorough evaluation, transparency in assumptions, and awareness of potential misuse or unintended consequences. While the work focuses on algorithmic foundations rather than deployment, the

societal implications underscore the need for continued engagement with ethical and societal dimensions as AI-based IoT systems evolve.



# 4. Summary of Papers

## 4.1 Results Outline

This chapter presents a structured summary of the main theoretical and empirical contributions of the papers included in this thesis. The results are organised according to the three research aims defined in Chapter 1. Each aim is addressed by one or more papers that share a common methodological focus and collectively answer the associated research questions.

For each aim, the corresponding section follows a uniform structure comprising the following: (i) motivation, (ii) experimental and system setup, (iii) approach and methods, and (iv) results. This organisation facilitates a coherent synthesis of the findings while highlighting how individual papers contribute complementary perspectives.

Section 4.2 addresses **Aim 1** through the combined contributions of **Paper I** and **Paper II**, which focus on communication- and energy-efficient federated learning through adaptive gradient sparsification and theoretical convergence analysis.

## 4.2 Aim 1: Communication-Efficient Federated Learning

This section addresses **Aim 1** of the thesis and answers **SQ1**. The aim is to design AI-based learning methods that reduce communication and energy costs in distributed learning systems while preserving learning performance. The section synthesises the contributions of **Paper I** and **Paper II**, which study communication-adaptive gradient sparsification and error-compensated federated learning under explicit cost models.

### 4.2.1 Motivation

Federated learning (FL) enables the collaborative training of machine learning models without requiring clients to share raw data, making it well suited for IoT and wireless systems, where privacy, bandwidth, and regulatory constraints are critical [35; 76]. However, communication often becomes the dominant bottleneck in such systems [28; 35],

as modern models are high-dimensional and repeated transmissions of gradients or model updates can be costly in terms of energy, bandwidth, and latency.

A common approach to reducing communication is to transmit compressed updates, for example, through sparsification. Gradient sparsification is inspired by the rationale that not all the model updates make a significant contribution to the training process. Existing methods typically rely on fixed compression levels [77] or threshold-based rules [78; 79] that must be tuned offline. In dynamic IoT environments, however, communication conditions, energy availability, and update statistics vary over time, making static choices suboptimal [80].

The central motivation of **Paper I** and **Paper II** is to treat communication efficiency in FL as an *adaptive decision-making problem*. Instead of fixing the sparsification level in advance, the compression level is adjusted online based on the information content of the updates and the actual communication cost incurred by the system. However, gradient sparsification can slow down the convergence and reduce accuracy because of the potential loss of information in the sparsified updates. **Paper II** introduces an error compensation mechanism for adaptive gradient sparsification to counterbalance this loss of information. The goal is to correct the progressive accumulation of errors by using the memory of prior errors through an error correction step before transmitting the sparsified updates. Error compensation [26; 81; 82] and EF21 [83] are some important variants of the error correction step.

#### 4.2.2 Setup

We consider a federated learning system with  $m$  client nodes and a central server. Client  $i$  holds a local dataset  $\mathcal{D}_i$  with  $n_i$  samples, and  $n = \sum_{i=1}^m n_i$ . The global learning objective follows empirical risk minimisation:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^m f_i(x), \quad (4.1)$$

where  $x \in \mathbb{R}^d$  denotes the global model parameters and  $f_i$  is the local loss induced by  $\mathcal{D}_i$ .

Model training is coordinated by the server using standard federated optimisation schemes such as FedSGD or FedAvg, where clients perform local stochastic gradient updates and periodically communicate model updates to the server. To reduce communication overhead, both **Paper I** and **Paper II** apply sparsification to the communicated updates.

In particular, a top- $B$  sparsification operator  $S_B(\cdot)$  is used; it retains only the  $B$  largest-magnitude components of a vector:

$$[S_B(\Delta)]_j = \begin{cases} \Delta_j, & j \in I_B(\Delta), \\ 0, & \text{otherwise,} \end{cases} \quad (4.2)$$

where  $I_B(\Delta)$  denotes the index set of the  $B$  largest components.

To mitigate information loss caused by compression, **Paper II** incorporates error compensation. Each client maintains a local residual vector that accumulates the compression error and is added to future updates before sparsification. This mechanism significantly improves robustness under aggressive compression.

Communication efficiency is evaluated using explicit cost models. In energy-constrained IoT settings, communication energy is modelled as an affine function of the transmitted payload:

$$\text{EnergyCost}(B) = E_1 \cdot \text{PayLoad}(B) + E_0, \quad (4.3)$$

where  $\text{PayLoad}(B)$  depends on the sparsification level  $B$ . Similar models are used for transmitted bits and latency in **Paper II**.

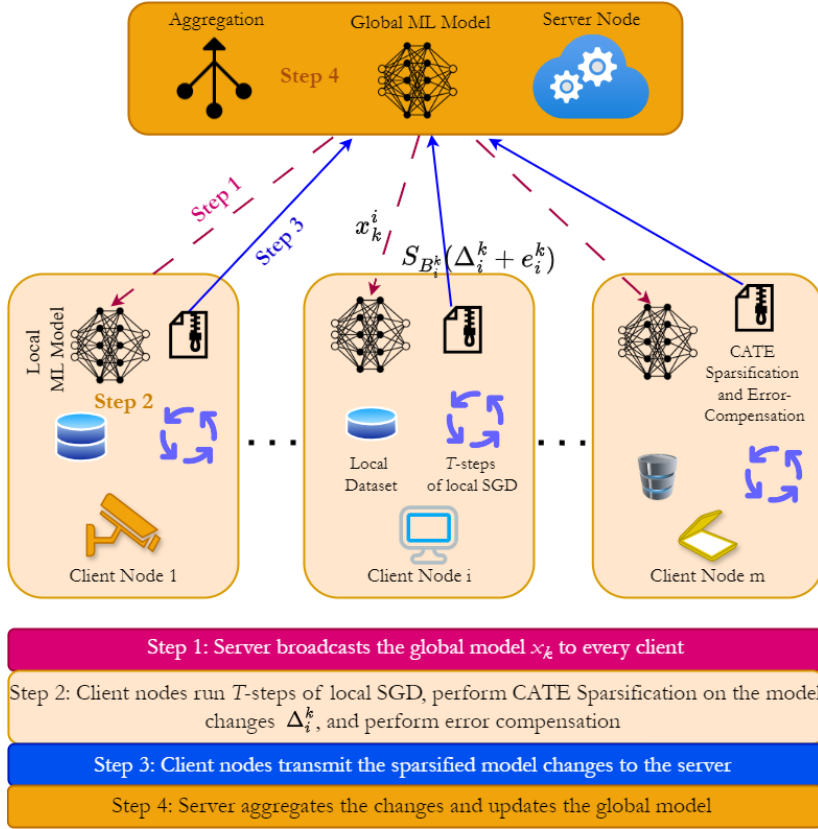
### 4.2.3 Approach and Methods

The key methodological idea shared by **Paper I** and **Paper II** is to adapt the sparsification level online based on the current update statistics and a communication cost model. Instead of fixing the sparsification budget  $B$  a priori, the proposed methods select  $B$  dynamically to balance learning progress against the communication cost.

**Approach and Methods (Paper I).** **Paper I** introduces *FL-CAT* (Federated Learning with Communication-Aware parameter Tuning), an energy-aware adaptive sparsification framework for federated learning. The central idea is to dynamically tune the sparsification level before each communication round by explicitly optimising the trade-off between learning progress and the communication energy cost, rather than fixing the sparsification budget a priori.

Let  $\Delta_i^t \in \mathbb{R}^d$  denote the local model update computed by client  $i$  at communication round  $t$ . Instead of transmitting the full update, each client applies a  $K$ -sparsification operator  $S_K(\cdot)$  that retains the  $K$  largest components in terms of magnitude. The sparsification level  $K$  is selected by maximising an *energy-efficiency* metric defined as

$$\text{Efficiency}_K(\Delta_i^t) = \frac{\text{Information}_K(\Delta_i^t)}{\text{Energy}_K(\Delta_i^t)}, \quad (4.4)$$



**Figure 4.1:** High-level overview of communication-adaptive federated learning with sparsification and error compensation. Adapted from **Paper II**.

where  $\text{Energy}_K(\Delta_i^t)$  denotes the energy required to transmit  $K$  components under a given communication model, and  $\text{Information}_K(\Delta_i^t)$  quantifies the fraction of useful information preserved by  $K$ -sparsification.

Following the CAT framework, the information content is measured using the normalised squared  $\ell_2$ -norm ratio

$$\text{Information}_K(\Delta_i^t) = \frac{\|S_K(\Delta_i^t)\|_2^2}{\|\Delta_i^t\|_2^2}, \quad (4.5)$$

which captures how much of the update energy is retained after sparsification. This metric is monotone in  $K$ , easy to compute online, and directly linked to the magnitude of the learning signal.

At each round, the client selects the sparsification level by solving the one-dimensional optimisation problem

$$K_t = \arg \max_{K \in \{1, 2, \dots, d\}} \text{Efficiency}_K(\Delta_i^t). \quad (4.6)$$

Although this is an integer optimisation problem, its structure is simple and efficient to solve in practice. Under commonly used energy models, the objective exhibits favourable properties (e.g. submodularity), enabling the fast online selection of  $K_t$ .

After selecting  $K_t$ , each client transmits the sparsified update  $S_{K_t}(\Delta_i^t)$  to the server. The server aggregates the received updates and performs a standard federated model update. By adapting the sparsification level at every communication round and for every client, FL-CAT avoids transmitting low-value update components that contribute little to learning while incurring energy costs.

Overall, the key novelty of **Paper I** lies in explicitly incorporating an energy-efficiency objective into the sparsification decision, enabling communication-efficient and adaptive federated learning without modifying the underlying optimisation algorithm.

**Error Compensation for Stable Learning (Paper II):** Strong sparsification introduces bias and information loss, which can slow convergence or destabilise learning. To address this, **Paper II** integrates an error compensation mechanism. Each client maintains a residual vector  $e_k^i$ , which accumulates the discarded components of past updates. The effective update is computed as

$$u_k^i = g_k^i + e_k^i, \quad \tilde{u}_k^i = S_{B_k^i}(u_k^i),$$

and the residual is updated as

$$e_{k+1}^i = u_k^i - \tilde{u}_k^i.$$

The sparsified update  $\tilde{u}_k^i$  is transmitted to the server, while the residual  $e_{k+1}^i$  ensures that information lost due to sparsification is reintroduced in future rounds. This mechanism is critical for preserving convergence under aggressive compression.

**Federated Update Rule:** At each communication round, the server aggregates the received sparsified updates from participating clients and updates the global model:

$$x_{k+1} = x_k - \gamma \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \tilde{u}_k^i,$$

where  $\gamma$  is the learning rate and  $\mathcal{S}_k$  denotes the set of active clients.

**Algorithmic Summary:** The overall procedure is summarised in Algorithm 1, which highlights the interaction between adaptive sparsification, error compensation, and federated aggregation.

**Theoretical Perspective:** A major contribution of **Paper II** is the theoretical analysis of the proposed adaptive and error-compensated framework. Under standard assumptions on convexity, smoothness,

---

**Algorithm 1** Communication-Adaptive Federated Learning with Error Compensation (FL-CATE)

---

```
1: Initialise global model  $x_0$ , residuals  $e_0^i = 0$ 
2: for each communication round  $k = 0, 1, \dots$  do
3:   for each participating client  $i \in S_k$  in parallel do
4:     Compute stochastic gradient  $g_k^i$ 
5:     Form compensated update  $u_k^i = g_k^i + e_k^i$ 
6:     Select sparsification level  $B_k^i$  via efficiency maximisation
7:     Sparsify update  $\tilde{u}_k^i = S_{B_k^i}(u_k^i)$ 
8:     Update residual  $e_{k+1}^i = u_k^i - \tilde{u}_k^i$ 
9:     Transmit  $\tilde{u}_k^i$  to server
10:  end for
11:  Server aggregates updates and updates model  $x_{k+1}$ 
12: end for
```

---

and bounded stochastic gradient variance, convergence guarantees are established for both convex and non-convex objectives. These results are elaborated in Subsection 4.2.4.

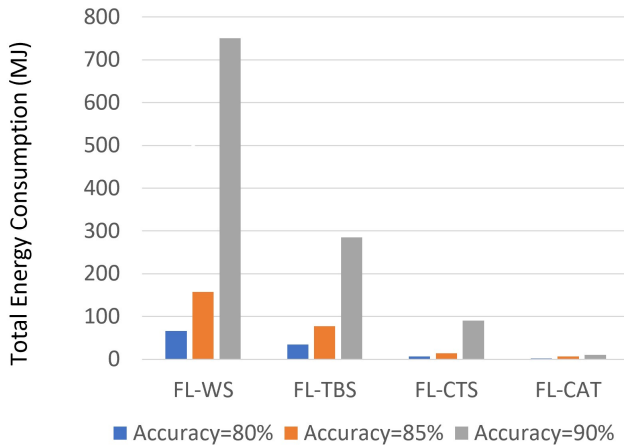
#### 4.2.4 Results

The experimental results reported in **Paper I** and **Paper II** demonstrate that adaptive sparsification leads to substantial communication savings without sacrificing learning performance. Across different experimental settings, the proposed methods consistently outperform fixed-sparsification baselines.

##### **Results (Paper I): Energy-Aware Adaptive Sparsification**

**Paper I** evaluates the proposed FL-CAT algorithm against three baselines: FL-WS (federated learning without sparsification), FL-TBS (threshold-based sparsification), and FL-CTS (constant-K sparsification). Experiments are performed on a network of five IoT edge devices collaboratively training a logistic regression model for MNIST handwritten digit classification, with a cloud server performing aggregation. Each experiment is repeated five times with random data splits across clients, and the results are averaged. For the baselines, the sparsification threshold in FL-TBS and the constant sparsification level in FL-CTS are chosen via grid search (as reported in **Paper I**), whereas FL-CAT selects the sparsification level automatically at each communication round. We report results for the affine communication cost model in terms of (i) the total energy consumed, (ii) the total data transferred, and (iii) the accuracy-energy trade-off.

Fig. 4.2 shows that, under the affine cost model, FL-CAT achieves



**Figure 4.2:** Total energy consumption under the affine communication cost model (**Paper I**).

a substantial reduction in total energy consumption compared to all baselines while reaching the same target accuracies. On average, FL-CAT consumes approximately 97.2%, 94.2%, and 70.4% less energy than FL-WS, FL-TBS, and FL-CTS, respectively. Importantly, these gains are achieved without any manual tuning of sparsification parameters.

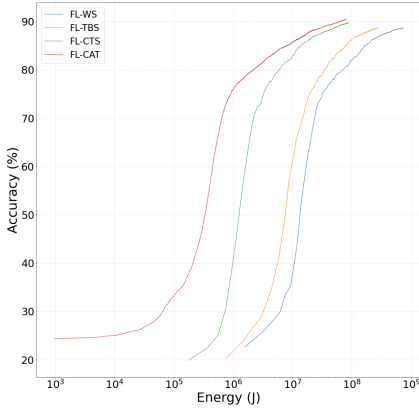
The accuracy–energy trade-off is illustrated in Fig. 4.3a. All methods converge to similar final accuracies close to 90%, but FL-CAT reaches these accuracy levels at a significantly lower energy cost. The advantage of adaptive sparsification is especially pronounced in the early stages of training, up to 80% accuracy.

Fig. 4.3b compares the total amount of communicated data. Under the affine model, FL-CAT reduces data transmission by approximately 95%, 89.5%, and 47% compared to FL-WS, FL-TBS, and FL-CTS, respectively, while achieving a comparable or higher accuracy.

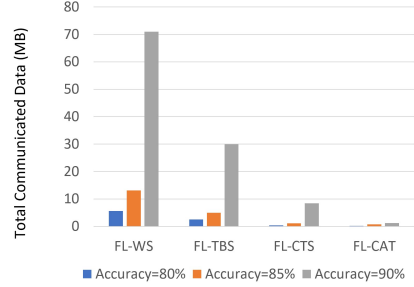
*Packet model (summary).* Similar trends are observed under the packet-based communication model (IEEE 802.15.4). FL-CAT consistently achieves lower energy consumption and reduced data transfer while maintaining comparable accuracy, indicating that the benefits of efficiency-driven adaptive sparsification generalise across different communication cost models.

**Paper II** further demonstrates that incorporating error compensation improves convergence stability, particularly under aggressive compression and heterogeneous cost conditions.

**Theoretical Results (Paper II).** **Paper II** establishes convergence guarantees for the proposed error-compensated and sparsified federated



(a) Accuracy versus energy



(b) Total data transferred

**Figure 4.3:** Accuracy–energy trade-off and total data transfer under the affine communication cost model (**Paper I**).

learning algorithm for both convex and non-convex optimisation problems. Consider a federated learning setup with  $n$  clients, where each client  $i$  minimises a local objective function  $f^i : \mathbb{R}^d \rightarrow \mathbb{R}$ , and the global objective is defined as

$$f(x) := \frac{1}{n} \sum_{i=1}^n f^i(x).$$

Let  $\{x_k\}_{k \in \mathbb{N}}$  denote the global model iterates produced by the algorithm, and let  $x^*$  be an optimal solution when it exists.

The analysis assumes that each local objective function has an  $L$ -Lipschitz continuous gradient and that stochastic gradients are unbiased with bounded variance  $\sigma^2$ . For the convex case, it is further assumed that each  $f^i$  is convex. Under these standard assumptions, and for a suitably chosen constant step size  $\gamma = \bar{\gamma}/T$ , the algorithm achieves the convergence rate

$$\mathbb{E} \left[ f \left( \frac{1}{K} \sum_{k=0}^{K-1} x_k \right) - f(x^*) \right] = \mathcal{O} \left( \frac{1}{K} \right) + \mathcal{O}(\bar{\gamma}),$$

where the residual error depends on the gradient noise variance, data heterogeneity across clients, and the compression quality parameter  $\alpha$ . This result shows that error-compensated sparsification preserves the standard  $\mathcal{O}(1/K)$  convergence rate for convex objectives despite aggressive communication compression.

For non-convex optimisation, which is common in deep learning, **Paper II** establishes convergence to a first-order stationary point under

an additional *data heterogeneity* assumption. Specifically, it is assumed that the local gradients satisfy

$$\|\nabla f^i(x) - \nabla f(x)\|^2 \leq \zeta^2 \quad \text{for all } x \in \mathbb{R}^d,$$

where  $\zeta$  quantifies the level of heterogeneity across clients. Under this assumption and an appropriately chosen step size, the algorithm guarantees that a randomly selected iterate  $\hat{x}_K$  satisfies

$$\mathbb{E}[\|\nabla f(\hat{x}_K)\|^2] = \mathcal{O}\left(\frac{1}{K}\right) + \mathcal{O}(\bar{\gamma}),$$

with constants that depend explicitly on the compression parameter  $\alpha$ , the gradient noise variance  $\sigma^2$ , and the heterogeneity level  $\zeta$ .

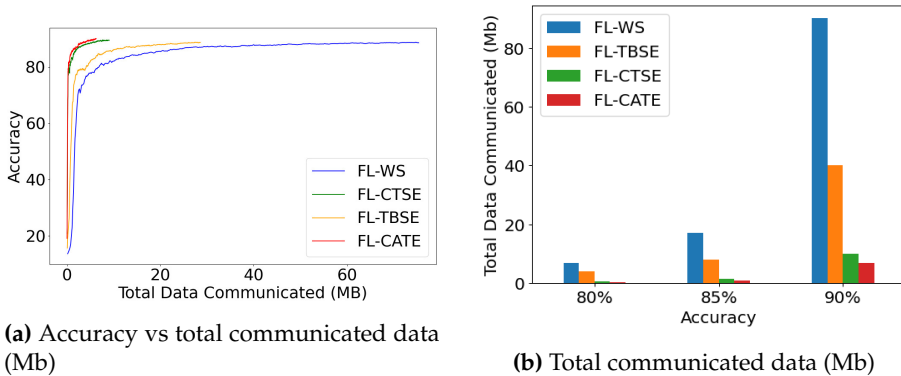
Moreover, by selecting a diminishing step size  $\gamma = \mathcal{O}(1/\sqrt{K})$ , the algorithm achieves the standard non-convex convergence rate

$$\mathbb{E}[\|\nabla f(\hat{x}_K)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

matching the best-known rates for compressed federated learning methods, while not requiring unbiased compression operators. The heterogeneity parameter  $\zeta$  does not need to be known to run the algorithm, but it directly influences the tightness of the convergence bound, capturing the intuitive effect that higher data heterogeneity makes optimisation more challenging.

Overall, the theoretical results show that the proposed error-compensated sparsification framework provides provable convergence guarantees for both convex and non-convex objectives while explicitly characterising the impact of communication compression and data heterogeneity.

**Experimental Results (Paper II): Cost-Adaptive Sparsification with Error Compensation.** Building on the energy-aware adaptive sparsification of **Paper I**, **Paper II** extends the experimental scope in three key ways. First, it evaluates a *general cost-adaptive* framework (FL-CATE) that can tune sparsification with respect to different communication costs, including *transmitted bits*, *energy consumption*, and *communication time* (rather than focusing mainly on energy / data under a specific cost model). Second, it incorporates *error compensation* into both the proposed method and the sparsified baselines, enabling stable learning under stronger compression and allowing a fairer comparison in modern compressed FL settings. Third, it validates robustness across *convex and non-convex* learning problems - *Linear Regression (LR)* and *Multilayer Perceptron (MLP)* on MNIST - and further tests *generalisability and scalability* by training a larger *Convolutional Neural Network (CNN)* on Fashion-MNIST while varying the number of clients (5 to 25) and adding noise to cost measurements.

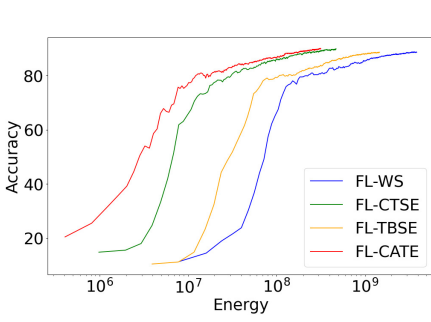


**Figure 4.4:** Minimising communicated bits (**Paper II**, Exp. 1: MNIST + logistic regression, convex).

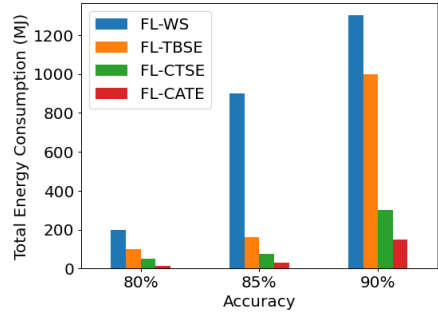
In **Paper I**, FL-CAT had already demonstrated that adaptive sparsification can substantially reduce energy and data transfer while preserving accuracy under the affine model. **Paper II** strengthens this message by showing that the same core principle – *adapting sparsification online based on an explicit cost model* – yields large efficiency gains not only for energy, but also for bits and latency, and that these gains persist for non-convex models and larger-scale CNN training. In addition, by integrating error compensation, **Paper II** shows that the cost savings can be achieved *without sacrificing convergence quality* even under aggressive sparsification.

Fig. 4.4 highlights a core new message of **Paper II**: FL-CATE can directly optimise *bits* as the target cost. As shown in Fig. 4.4a, FL-CATE achieves the same accuracy trajectory as the baselines while communicating substantially fewer bits throughout training. The bar plot in Fig. 4.4b shows that, for convex optimisation, FL-CATE reduces the total communicated data by approximately 92.7%, 83.93%, and 31.5% compared to FL-WS, FL-TBSE, and FL-CTSE, respectively, while reaching the same accuracy targets. This extends the adaptive savings of **Paper I** (energy/data efficiency under specific models) to an explicitly *bits-driven* objective, with error compensation enabled throughout.

Fig. 4.5 connects directly to the main results of **Paper I** by revisiting energy-aware efficiency under a packet-based setting with error compensation. Fig. 4.5a shows that FL-CATE preserves the learning trajectory while reducing energy throughout training. The bar plot in Fig. 4.5b shows that, for convex optimisation, FL-CATE reduces the total energy consumption by approximately 91.3%, 82.7%, and 55% compared to FL-WS, FL-TBSE, and FL-CTSE, respectively. Relative to **Paper I**, the key added value here is that the same adaptive princi-

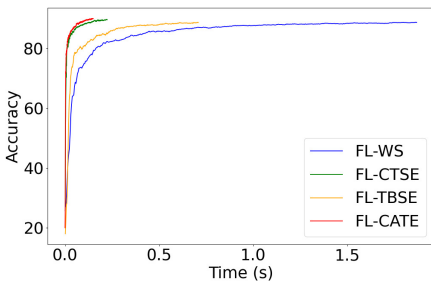


(a) Accuracy vs energy (mJ)

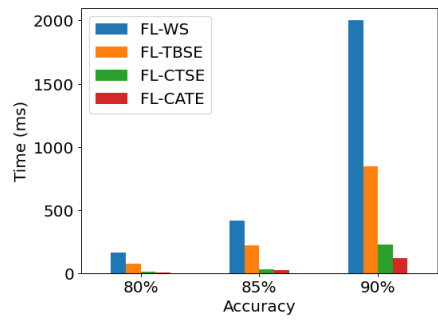


(b) Total energy consumption (mJ)

**Figure 4.5:** Minimising energy consumption (**Paper II**, Exp. 3: MNIST + logistic regression, convex).



(a) Accuracy vs time (s)



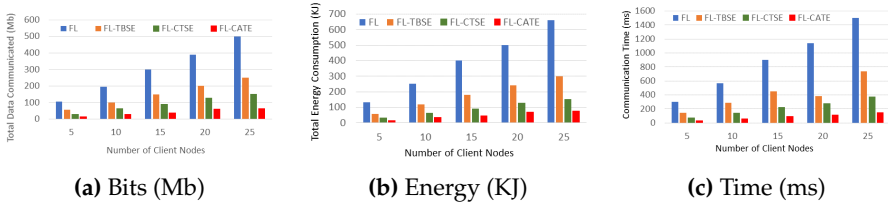
(b) Communication time (ms)

**Figure 4.6:** Minimising communication time (**Paper II**, Exp. 5: MNIST + logistic regression, convex).

ple is validated under a broader experimental matrix and under *error-compensated* sparsified FL.

A major experimental extension beyond **Paper I** is the explicit latency objective. Fig. 4.6 shows that FL-CATE can also tune sparsification to minimise communication time while maintaining comparable accuracy. In particular, Fig. 4.6b shows that, for convex optimisation, FL-CATE reduces the total communication time by approximately 94.2%, 85.85%, and 45.5% compared to FL-WS, FL-TBSE, and FL-CTSE, respectively. These results broaden the relevance of the adaptive sparsification idea from energy efficiency (**Paper I**) to delay-sensitive FL settings.

Finally, Fig. 4.7 summarises experiments that go beyond **Paper I** in both model complexity and system scale. Using a CNN trained on Fashion-MNIST while varying the number of clients from 5 to 25 (and adding noise to cost calculations), FL-CATE consistently outper-



**Figure 4.7:** Scalability and generalisability (**Paper II**, Exp. 7–9: Fashion-MNIST + CNN, varying clients).

forms the baselines for all three cost objectives (bits, energy, and time), as shown in Figs. 4.7a–4.7c. This demonstrates that the proposed cost-adaptive sparsification remains effective for larger models with hundreds of thousands of parameters and under a changing network size.

**Other results and experiments (Paper II).** In addition to the key results highlighted above, **Paper II** reports complementary experiments for (i) non-convex training on MNIST using an MLP (Exp. 2, 4, 6), showing that FL-CATE maintains strong cost reductions even when the optimisation landscape is non-convex and models are larger; and (ii) loss-versus-cost plots for bits, energy, and time, showing that FL-CATE reaches lower loss levels with a substantially lower communication cost. The experiments also include noise in the cost calculations (Signal-to-Noise Ratio (SNR) 30–50) as part of the simulation setup, which tests the robustness of the adaptive tuning to imperfect cost estimation. Across all nine experiments, the central empirical conclusion is consistent: FL-CATE matches or improves the learning quality while reducing the targeted communication cost, and it achieves these gains without the manual hyperparameter tuning of sparsification levels, unlike threshold-based or constant sparsification baselines.

### 4.3 Aim 2: Constraint-Aware Online Learning

This section addresses **Aim 2** of the thesis and answers **SQ2**. The aim is to develop AI-based decision-making methods that operate under explicit and dynamic constraints, and that balance performance optimisation with long-term constraint satisfaction. The section synthesises the contributions of **Paper III** and **Paper V**, which study constrained online learning and reinforcement learning for IoT systems with limited resources.

While the two papers focus on different application domains, they share a common methodological theme: both formulate decision-making

as a sequential process under uncertainty, where constraints cannot be ignored and must be handled explicitly within the learning algorithm.

### 4.3.1 Motivation

Many IoT systems often operate to optimise certain objectives like reliability and throughput under operational constraints, such as limited energy budgets, processing capacity, or bandwidth budgets [84]. Unlike unconstrained optimisation problems, violations of such constraints can have long-term consequences, including device failure, service degradation, or safety risks. At the same time, fully conservative strategies that always satisfy constraints may lead to poor performance, especially during early learning phases when system dynamics are not yet well understood.

This tension motivates learning-based decision-making methods that allow *controlled exploration* while ensuring that constraints are respected in the long run. Traditional optimisation techniques often require complete system models and static assumptions, making them ill-suited for dynamic IoT environments. **Paper III** and **Paper V** address this challenge from complementary perspectives. **Paper III** focuses on online decision-making with explicit constraint budgets using multi-armed bandit models. **Paper V** formulates constrained decision-making problems using reinforcement learning for data stream processing at the network edge.

**Motivation for Paper III:** Decision-making under uncertainty with resource constraints has been widely studied in online learning, particularly through multi-armed bandit (MAB) models such as UCB and Thompson Sampling [85]. Extensions to constrained MABs address fixed-budget settings [50] and safety-aware exploration [86; 87], while multi-objective bandits consider trade-offs across reward dimensions [88; 89]. However, most existing approaches assume static or predictable constraints, limiting their applicability to dynamic IoT and wireless systems [90; 91]. Related work in online convex optimisation addresses time-varying constraints under stronger information assumptions [92–94]. To address this gap, **Paper III** introduces a stochastic bandit formulation with a dynamically shrinking violation budget and proposes a Budgeted UCB algorithm that enables controlled exploration under evolving constraints. The proposed approach provides theoretical guarantees on both regret and constraint violations, thereby bridging the gap between constrained bandit theory and practical decision-making in dynamic IoT environments [4].

**Motivation for Paper V:** Data stream processing in fog and IoT networks has been studied from architectural and system perspectives

[95–98]. Some works analyse general architectures and challenges [95; 96], while others focus on latency reduction [97] or distributed edge processing [98]. However, energy efficiency, data-intensive workloads, and data transfer costs are often neglected. Scheduling and offloading approaches based on heuristics or optimisation have also been proposed [99–101], but they are poorly suited for dynamic fog environments [96]. Reinforcement learning has therefore gained attention [102; 103], yet existing RL-based methods target coarse-grained tasks and do not address unbounded data stream processing [104]. **Paper V** addresses this gap by formulating high-speed stream processing as a dynamic decision-making problem that jointly optimises energy consumption and delay while accounting for data migration costs [96].

### 4.3.2 Setup

#### Constrained Online Decision-Making

In **Paper III**, the system is modelled as a stochastic multi-armed bandit problem with resource constraints. There are two feedback signals per action: a *reward* (e.g. throughput) and a *cost* (e.g. energy consumption). At each round  $t$ , the learner observes a dynamically changing constraint threshold  $C_t$  and must select an arm that balances reward maximisation against constraint satisfaction. At each time step, an agent selects one of several available actions (arms), each yielding a stochastic reward and consuming a certain amount of a limited resource. The objective is to maximise cumulative reward while ensuring that the total constraint violation remains within a predefined budget. The key feature of the formulation in **Paper III** is that the constraint budget is allowed to be *dynamic*. In particular, limited constraint violations are permitted during early learning stages to facilitate exploration, but violations must decay over time to guarantee long-term feasibility.

#### Reinforcement Learning for Edge Data Stream Processing

**Paper V** studies a data stream processing problem in IoT edge environments. At each decision epoch, an edge node must decide how to process incoming data streams under limited computational and energy resources. Possible actions include local processing, offloading to a neighbouring node or cloud server, or deferring processing.

This problem is modelled as a Markov decision process (MDP), where system states capture resource availability and workload conditions, actions correspond to processing decisions, and rewards reflect a trade-off between energy consumption and processing performance.

Constraints arise naturally from limited energy budgets and processing capacity.

### 4.3.3 Approach and Methods

#### Adaptive Budgeted Bandits (**Paper III**)

**Paper III** addresses online decision-making under uncertainty when actions must satisfy *time-varying resource constraints*. The central methodological contribution is the proposed Budgeted UCB algorithm, which extends classical Upper Confidence Bound (UCB) methods to explicitly account for constraint violations while retaining strong exploration–exploitation guarantees.

Unlike standard constrained bandit formulations that enforce constraints only in expectation or via long-term averages, **Paper III** introduces a *decaying violation budget*. This allows the algorithm to tolerate a controlled number of constraint violations during early exploration phases while ensuring that violations vanish asymptotically as learning progresses. This design directly reflects practical IoT settings, where occasional early violations may be acceptable to identify high-performing actions.

The Budgeted UCB algorithm maintains separate upper confidence bounds for rewards and costs. For each arm  $a$ , it tracks the empirical means  $\hat{\mu}_r(a)$  and  $\hat{\mu}_c(a)$  and constructs confidence bounds

$$\text{UCB}_i(a) = \hat{\mu}_i(a) + \sqrt{\frac{2 \ln t}{N(a)}}, \quad i \in \{r, c\},$$

where  $N(a)$  is the number of times arm  $a$  has been played.

To regulate safety, the algorithm defines a linearly decaying violation allowance

$$\delta_t = \delta_0 \left(1 - \frac{t-1}{T_{\text{bud}}}\right),$$

and it monitors the empirical violation rate

$$v_t = \frac{1}{t-1} \sum_{s=1}^{t-1} \mathbf{1}\{c_s > C_s\}.$$

If  $v_t \leq \delta_t$ , the algorithm operates in an *exploration mode*, selecting the arm with the highest reward UCB. Once the violation rate exceeds the remaining budget, the algorithm switches to a *safety mode*, restricting attention to arms whose cost UCB satisfies the current constraint and selecting the best among them. If no arm appears safe, it selects the arm with the smallest cost UCB to minimise further violations.

---

**Algorithm 2** Budgeted UCB with Decaying Violation Budget (**Paper III**)

---

```
1: Initialise  $N(a) = 0, S_r(a) = 0, S_c(a) = 0$  for all arms  $a$ 
2: for  $t = 1$  to  $T$  do
3:   Observe constraint threshold  $C_t$ 
4:   Compute  $\text{UCB}_r(a)$  and  $\text{UCB}_c(a)$  for all arms
5:   Update violation allowance  $\delta_t$ 
6:   Compute empirical violation rate  $v_t$ 
7:   if  $v_t \leq \delta_t$  then
8:      $a_t \leftarrow \arg \max_a \text{UCB}_r(a)$      $\triangleright$  throughput-driven exploration
9:   else
10:     $\mathcal{F}_t \leftarrow \{a : \text{UCB}_c(a) \leq C_t\}$ 
11:    if  $\mathcal{F}_t \neq \emptyset$  then
12:       $a_t \leftarrow \arg \max_{a \in \mathcal{F}_t} \text{UCB}_r(a)$ 
13:    else
14:       $a_t \leftarrow \arg \min_a \text{UCB}_c(a)$ 
15:    end if
16:  end if
17:  Play  $a_t$ , observe  $(r_t, c_t)$ , update statistics
18: end for
```

---

The key methodological novelty of **Paper III** lies in the explicit integration of a *shrinking violation budget* into the bandit decision rule. This mechanism enables efficient early exploration while guaranteeing that both the average regret and the average violation rate converge to zero over time. Compared to virtual-queue or penalty-based approaches, Budgeted UCB enforces safety more directly and avoids persistent constraint breaches.

Overall, **Paper III** provides a principled and lightweight framework for adaptive decision-making under dynamic constraints, making it particularly well suited for large-scale IoT systems, where resource limits fluctuate and centralised control is infeasible.

### Constraint-Aware Reinforcement Learning (**Paper V**)

**Paper V** addresses constrained decision-making using reinforcement learning. Rather than enforcing hard constraints at every decision step, the proposed RL framework learns policies that implicitly balance performance objectives and resource consumption through reward shaping and state augmentation. **Paper V** addresses the problem of dynamic data stream processing and offloading in fog and edge networks, where incoming data patterns, resource availability, and application requirements evolve over time. Unlike traditional heuristic or optimisation-

based approaches that rely on static assumptions and accurate system models, **Paper V** adopts a reinforcement learning (RL) framework that enables an edge node to learn optimal processing and offloading decisions directly from interaction with the environment.

**MDP Formulation for Data Stream Processing:** The decision-making problem is modelled as an MDP, making RL a natural choice. The state captures the current data backlog at the edge node, represented as a vector  $\vec{x} \in \mathbb{N}^T$  describing the amount of data with different remaining deadlines, subject to a finite storage capacity. The action consists of deciding how much data to process locally and how much to offload to a remote server, represented by a pair  $(\vec{p}, \vec{c})$ , with explicit constraints on processing and communication capacities. The state transition dynamics are stochastic due to random data arrivals, reflecting realistic streaming environments.

The reward function is designed to reflect two competing objectives: minimising energy consumption and minimising the amount of data with missed deadlines. These objectives are combined through a scalarisation parameter  $\lambda \in [0, 1]$ , allowing explicit control over objective priorities. This formulation enables the system to express different operating regimes, such as energy-efficient operation or delay-sensitive processing, within a unified framework.

**Average-Reward Reinforcement Learning via R-Learning:** A key methodological choice in **Paper V** is the use of *R-learning* instead of standard Q-learning. While Q-learning optimises discounted cumulative rewards, the objective in **Paper V** is to optimise long-term *average performance*, expressed in terms of average energy consumption and average unprocessed data per time step. R-learning directly targets the average reward criterion, making it better aligned with continuous, infinite-horizon data stream processing tasks in fog networks.

The algorithm maintains a state-action value function  $\hat{q}(\vec{x}, \vec{u})$  and an estimate of the average reward  $\bar{R}$ . At each time step, the agent selects actions using an exploration strategy (e.g.  $\epsilon$ -greedy), observes the immediate reward and next state, and updates both the value estimates and the average reward estimate. This allows the agent to gradually learn an offloading policy that balances energy consumption and deadline violations according to the chosen priority parameter  $\lambda$ .

**Algorithmic Significance:** The proposed RL-based offloading (RLO) approach enables fully adaptive decision-making without requiring prior knowledge of data arrival distributions, channel conditions, or exact energy costs. By optimising the average reward and explicitly parameterising objective priorities, the method provides a flexible and interpretable mechanism for dynamic data stream processing. This makes **Paper V** particularly suitable for real-world fog and IoT environments,

where system characteristics and application requirements change over time and must be handled online without centralised control.

#### 4.3.4 Results

The results in **Paper III** demonstrate that adaptive budgeted bandit algorithms achieve strong performance compared to unconstrained baselines while maintaining long-term constraint satisfaction. Empirical evaluations show that allowing controlled early violations leads to faster learning and improved cumulative reward without compromising feasibility.

**Theoretical Results (Paper III).** **Paper III** analyses the performance of the proposed Budgeted UCB algorithm for stochastic multi-armed bandits with dynamically enforced constraints. Consider a bandit problem with  $K$  arms, where at each round  $t$  the learner selects an arm  $a_t$ , receives a stochastic reward, and incurs a stochastic cost. Let  $\Delta_t$  denote the instantaneous regret with respect to the optimal budget-respecting policy, and define the cumulative regret and cumulative constraint violations as

$$R(T) := \sum_{t=1}^T \Delta_t, \quad V(T) := \sum_{t=1}^T \mathbf{1}\{\mu_c(a_t) > C_t\},$$

where  $\mu_c(a)$  is the expected cost of arm  $a$ , and  $C_t$  is the (possibly time-varying) constraint threshold.

The analysis shows that Budgeted UCB achieves sublinear regret while ensuring that constraint violations vanish asymptotically. With probability at least  $1 - 1/T$ , the algorithm satisfies

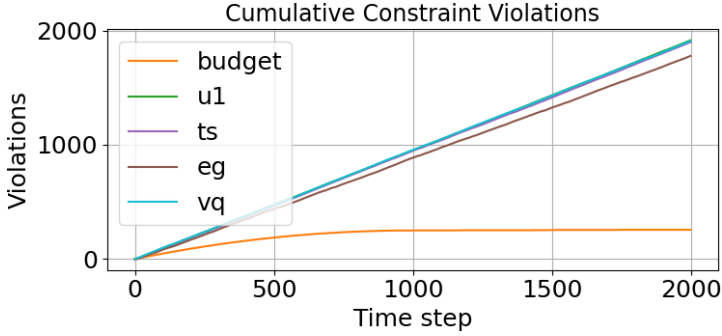
$$R(T) = \mathcal{O}\left(\sqrt{KT \ln T}\right), \quad V(T) = \mathcal{O}(\ln T).$$

The regret bound matches the optimal order of classical UCB algorithms, while the violation bound grows only logarithmically in time.

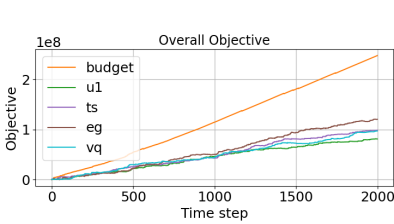
The key idea underlying the analysis is a two-phase behaviour. During an initial exploration phase, the algorithm allows a controlled number of constraint violations through a decaying violation budget. Once this budget is exhausted, the algorithm restricts its actions to arms that are estimated to be feasible. Violations in this safety phase occur only due to statistical estimation errors and can be bounded using concentration inequalities, resulting in at most logarithmic growth.

As a consequence, both the average regret and the average violation rate converge to zero, i.e.

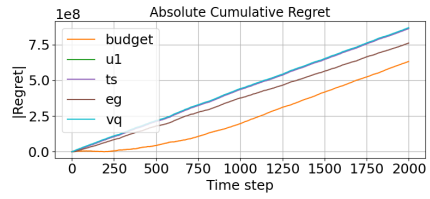
$$\frac{R(T)}{T} \rightarrow 0, \quad \frac{V(T)}{T} \rightarrow 0 \quad \text{as } T \rightarrow \infty.$$



(a) Cumulative constraint violations



(b) Overall objective ( $\Lambda = 10^6$ )



(c) Cumulative absolute throughput regret

**Figure 4.8:** Performance evaluation of Budgeted UCB under randomly varying energy constraints (**Paper III**).

These results demonstrate that Budgeted UCB achieves the optimal exploration-exploitation trade-off of standard UCB while simultaneously enforcing long-term constraint satisfaction, making it well suited for dynamic IoT decision-making scenarios with evolving resource limits.

**Experimental Results (Paper III):** Paper III also includes simulation-based experiments to evaluate the proposed Budgeted UCB algorithm in environments with randomly varying energy constraints. The performance is assessed in terms of cumulative constraint violations, overall objective value, and throughput regret, and is compared against unconstrained baselines and a virtual-queue-based method.

Fig. 4.8a shows the cumulative constraint violations over time. Budgeted UCB confines violations to grow only logarithmically, in accordance with its decaying violation budget. In contrast, the unconstrained baselines rapidly converge to a single high-throughput arm and thereafter violate the energy constraint almost every round. The virtual-queue-based method (vq), which does not enforce a hard safety constraint, continues selecting high-energy arms even when over budget, resulting in substantially higher violation counts.

Fig. 4.8b reports the overall objective, defined as the cumulative

throughput penalised by constraint violations with a large penalty factor  $\Lambda$ . By strictly limiting violations, Budgeted UCB preserves nearly all of its raw throughput, leading to a steadily increasing net objective. In contrast, unconstrained methods incur severe penalties once they identify and repeatedly select energy-intensive arms, causing their overall objective to stagnate or decline. The vq method exhibits an early spike in violations, which significantly degrades its objective value despite high instantaneous throughput.

Finally, Fig. 4.8c illustrates the cumulative absolute throughput regret with respect to the clairvoyant constrained optimum. Budgeted UCB exhibits sublinear regret growth: after an initial exploration phase and occasional adjustments near the violation threshold, it quickly converges to the best feasible arm. The baseline algorithms accumulate substantially larger regret due to aggressive exploration and repeated violations that prevent them from consistently operating at the constrained optimum.

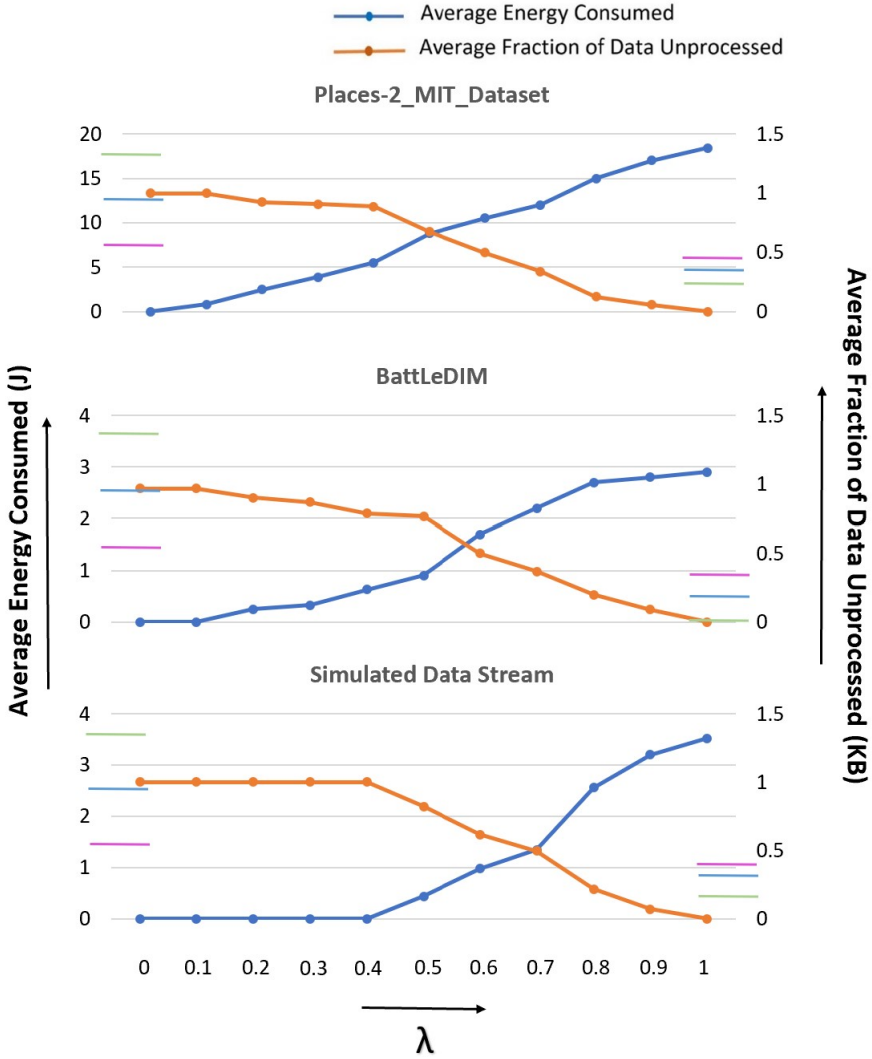
Additional experiments in **Paper III** consider linearly varying energy constraints. The results show that Budgeted UCB continues to maintain logarithmic growth in constraint violations while achieving sublinear regret, demonstrating robustness to smoothly changing feasibility regions. In contrast, baseline methods either violate constraints persistently or suffer a large amount of performance degradation when adapting to the evolving constraint.

Scalability experiments further confirm that the performance of Budgeted UCB scales favourably with the number of arms. Both the regret and violation rates grow in line with the theoretical bounds, indicating that the proposed algorithm remains effective in larger decision spaces without requiring additional tuning or centralised coordination.

Overall, these results demonstrate that Budgeted UCB achieves a favourable balance between throughput and safety in environments with time-varying constraints, validating the theoretical guarantees established in **Paper III**.

**Results (Paper V):** **Paper V** demonstrates that the proposed RL0 algorithm enables the explicit and continuous adjustment of objective priorities through a scalar parameter  $\lambda$ . Fig. 4.9 illustrates how varying  $\lambda$  produces a smooth and interpretable trade-off between the average energy consumption and the fraction of processed data for one simulated and two real data streams. In contrast, baseline methods operate at fixed points and do not allow such trade-off control.

The dual-axis  $\lambda$ -plots in Fig. 4.9 show that system behaviour can be tuned to satisfy application-specific constraints. For example, under an average energy budget of 1.5 J per time step, selecting  $\lambda \leq 0.72$  maximises the amount of processed data while respecting the energy



**Figure 4.9:**  $\lambda$ -plot showing the trade-off between average energy consumption and average unprocessed data as the priority parameter  $\lambda$  varies from 0 to 1. Results are shown for two real data streams and one simulated data stream (adapted from **Paper V**).

constraint. Similarly, enforcing that at least 75% of the input data be processed within its deadline requires  $\lambda \geq 0.8$ . This demonstrates that RLO provides a flexible and intuitive mechanism for priority adaptation, which is particularly important for dynamic data stream processing in IoT and fog environments, where operational requirements evolve over time [96].

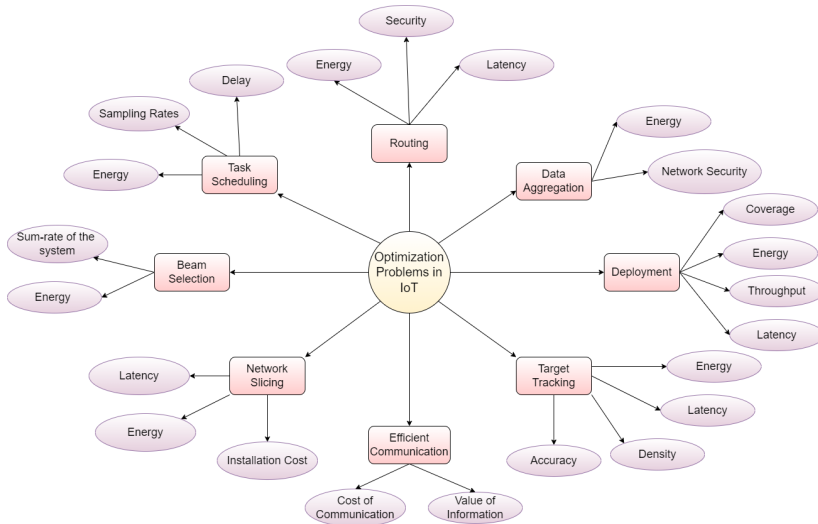
Beyond the illustrative trade-off shown in Fig. 4.9, additional experimental results in **Paper V** show that reinforcement learning-based data stream processing policies consistently outperform static and heuristic baselines across varying workloads and resource conditions. The learned policies adapt online to changes in the input rate and system dynamics, achieving lower energy consumption and improved processing performance over time. These findings further motivate the study of dynamically changing objectives and preferences, which is addressed in **Paper VI**.

Taken together, the results of **Paper III** and **Paper V** illustrate how AI-based decision-making methods can handle explicit and dynamic constraints in IoT systems. Bandit-based methods provide lightweight solutions with theoretical guarantees, while reinforcement learning offers greater modelling flexibility for state-dependent and long-horizon decision problems. These complementary approaches collectively answer **SQ2** and support the overarching research question **RQ**.

#### 4.4 Aim 3: Preference-Adaptive Multi-objective Reinforcement Learning

This section addresses **Aim 3** of the thesis and answers **SQ3**. The aim is to study AI-based decision-making methods that explicitly handle multiple, potentially conflicting objectives in IoT systems and that can adapt to changing objective preferences over time. In highly dynamic environments such as the IoT, reinforcement learning (RL) methods have gained increasing attention as effective alternatives. RL is particularly attractive since routing policies can be learned directly through interaction with the environment. These limitations motivate the development of new RL-based solutions tailored to large-scale and dynamically evolving IoT networks. The section synthesises the contributions of **Paper IV** and **Paper VI**, which focus on multi-objective reinforcement learning (MORL) as a principled framework for such problems.

While **Paper IV** provides a conceptual and methodological foundation for multi-objective and constrained reinforcement learning in the IoT, **Paper VI** demonstrates the practical application of MORL in a distributed IoT routing scenario.



**Figure 4.10:** Objectives in some common optimisation problems in the IoT. Adapted from **Paper IV**.

#### 4.4.1 Motivation

Many decision-making problems in IoT systems involve multiple objectives that cannot be reduced to a single scalar metric without loss of important structure. Typical examples include trade-offs between energy consumption and communication reliability, latency and throughput, or performance and resource usage. Fig. 4.10 shows an overview of the objectives in some common optimisation problems in the IoT, as presented in **Paper IV**. These objectives are often conflicting, and their relative importance may change over time depending on system conditions or application requirements.

Traditional approaches typically combine multiple objectives into a single scalar reward using fixed weights. While this simplifies optimisation, it assumes that objective preferences are known a priori and remain static. In practice, however, IoT systems often operate in dynamic environments where priorities evolve, for example, as battery levels decrease or network conditions degrade. This motivates multi-objective decision-making frameworks that explicitly represent and reason about multiple objectives. **Paper IV** and **Paper VI** adopt multi-objective reinforcement learning to address this challenge, enabling adaptive behaviour without requiring repeated retraining for each new preference setting. **Paper IV** supplies a review of existing optimisation methods for the IoT and provides conceptual frameworks and scopes for the application of MORL in the IoT.

**Paper VI** proposes and applies an MORL algorithm for the rout-

ing scenario. Many RL-based routing protocols have already been proposed for routing [105–116]. However, the majority of existing RL-based approaches, including deep RL methods, either (i) consider single or fixed multiple objectives and require retraining when system preferences between objectives change, or (ii) depend on centralised control architectures, such as Software-Defined Networking (SDN)-based solutions, which suffer from scalability limitations, privacy concerns, and vulnerability to single points of failure [117]. **Paper VI** presents a dynamic and fully distributed multi-objective Q-learning routing algorithm that learns multiple per-preference Q-tables in parallel. It also proposes, with a theoretical near-optimal guarantee, a novel greedy interpolation policy to adapt to unseen preferences.

#### 4.4.2 Setup

##### Multi-objective Reinforcement Learning

In MORL, the reward signal is vector-valued rather than scalar. Let  $\mathbf{r}_t \in \mathbb{R}^M$  denote the reward vector at time step  $t$ , where each component corresponds to a distinct objective. The goal is not to optimise a single cumulative reward, but to learn policies that capture the trade-offs among objectives.

A common way to operationalise MORL is through scalarisation, where a preference vector  $\boldsymbol{\lambda} \in \mathbb{R}^M$  maps vector rewards to scalar values. However, unlike standard RL, MORL frameworks often seek to learn policies that are reusable across multiple preference vectors.

##### Conceptual Foundations (**Paper IV**)

**Paper IV** surveys and formalises multi-objective and constrained reinforcement learning problems in IoT systems. It identifies representative decision-making tasks – such as routing, task offloading, and resource allocation – and discusses how vector-valued rewards and constraints naturally arise in these settings.

The chapter introduces key MORL concepts, including Pareto optimality, policy sets, and preference-aware decision-making, and provides guidance on selecting appropriate solution methods based on system requirements and computational constraints.

##### Routing as a Multi-objective Decision Problem (**Paper VI**)

**Paper VI** focuses on routing in IoT networks, where each node must make forwarding decisions based on local observations and limited information. Routing is formulated as a multi-objective decision-making

problem, with the primary objectives being energy consumption and communication reliability.

The system is modelled as a distributed MDP, where each node acts as a learning agent. Routing decisions influence long-term network behaviour, including node energy depletion and packet delivery success, making the problem well suited for MORL.

#### 4.4.3 Approach and Methods

**Paper IV** adopts a structured literature review and conceptual analysis methodology to examine existing approaches to multi-objective and constrained reinforcement learning in IoT systems, and to identify key modelling choices and open research challenges that inform the algorithmic developments in **Paper VI**.

**Multi-objective Learning Framework in Paper VI:** Each routing decision yields a vector-valued reward capturing both energy consumption and communication reliability. Instead of fixing objective weights during training, the problem is parameterised by a scalar preference variable  $\beta \in [0, 1]$ , which defines a convex scalarisation of the two objectives,

$$r_\beta(s, a) = \beta r^{\text{Energy}}(s, a) + (1 - \beta) r^{\text{PDR}}(s, a).$$

This formulation enables routing decisions to be adapted to different trade-offs between energy efficiency and reliability, reflecting diverse and evolving system requirements.

A key methodological choice in **Paper VI** is to decouple policy learning from preference specification. Rather than learning a single policy for a fixed preference, the proposed approach learns value functions for a finite set of preference values. Once trained, these value functions can be reused to generate routing decisions for different preferences without retraining, enabling flexible and efficient adaptation at decision time.

**Preference Parameterisation and Grid Interpolation:** To support continuous preference adaptation, **Paper VI** introduces a finite preference grid  $\mathcal{B} \subset [0, 1]$ . For each  $\beta \in \mathcal{B}$ , a corresponding optimal  $Q$ -function is learned using reinforcement learning. For an arbitrary preference  $\beta \notin \mathcal{B}$ , the proposed *Grid-Interpolated Policy* (GIP) constructs an interpolated  $Q$ -table from its two neighbouring grid points,

$$Q_\beta^{\text{Int}} := \rho \widehat{Q}_\beta + (1 - \rho) \widehat{Q}_{\bar{\beta}},$$

where  $\underline{\beta} = \max\{b \in \mathcal{B} : b \leq \beta\}$ ,  $\bar{\beta} = \min\{b \in \mathcal{B} : b \geq \beta\}$ , and  $\rho = (\beta - \underline{\beta}) / (\bar{\beta} - \underline{\beta})$ . Routing decisions are then made greedily with respect to  $Q_\beta^{\text{Int}}$ .

This interpolation-based design enables fast, online adaptation to changing preferences without restarting the learning process. It also introduces a principled trade-off between computational cost and approximation accuracy, controlled by the resolution of the preference grid.

**Distributed Learning and Decision-Making:** An important aspect of the proposed approach is its inherently distributed design. Learning and decision-making are performed locally at each network node, based on the locally observed states of its neighbouring nodes, without requiring a centralised controller or global state information. Each node maintains and updates its own  $Q$ -functions for the preference grid, enabling fully decentralised routing decisions.

The shared structure of the preference grid allows nodes to reuse learned value functions across different operating regimes, while preference changes are handled locally through interpolation rather than global coordination. This avoids repeated retraining and significantly reduces communication and synchronisation overhead.

Such a distributed formulation is well suited to large-scale IoT networks, where centralised control is often impractical due to scalability, latency, privacy, and single-point-of-failure concerns. By combining distributed reinforcement learning with preference-aware interpolation, **Paper VI** supports scalable, adaptive, and robust routing in dynamic network environments.

**Dynamic Preference Adaptation:** A central methodological contribution of **Paper VI** is its ability to handle dynamically changing preferences. Preference values may change across episodes or even at every decision step, reflecting evolving system conditions or operational priorities. The proposed distributed MORL algorithm allows each node to adapt its routing behaviour online as preferences change while relying on a shared set of learned value functions.

This capability is particularly important in IoT deployments, where objectives such as energy efficiency and reliability are rarely static. By separating learning from preference selection and embedding adaptation within a distributed framework, the proposed approach enables dynamic, preference-aware decision-making with minimal overhead.

#### 4.4.4 Results

The results in **Paper VI** demonstrate that the proposed MORL-based routing approach effectively adapts to changing objective preferences. Compared to single-objective baselines, the method achieves a better balance between energy efficiency and communication reliability across a wide range of operating conditions. Empirical evaluations show that

routing policies learned through MORL generalise well across different preferences, reducing the need for retraining and enabling rapid adaptation. This leads to an improved network lifetime and more reliable data delivery under dynamic conditions.

**Theoretical Results (Paper VI).** Paper VI establishes formal performance guarantees for the proposed GIP. Let  $Q_\beta$  denote the optimal Q-function corresponding to preference  $\beta$ . The analysis relies on two problem-level constants. First, let  $H$  denote the worst-case expected episode length under  $\beta$ -optimal control,

$$H := \sup_{\beta \in [0,1]} \sup_{s \in \mathcal{S}} \sup_{\pi \in \Pi_\beta^*} \mathbb{E}_s^\pi[K],$$

where  $K$  is the random episode length. Second, define

$$\Gamma := \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |r^{\text{Energy}}(s,a) - r^{\text{PDR}}(s,a)|,$$

which captures the maximum discrepancy between the two objectives.

A key intermediate result shows that the optimal Q-function varies smoothly with the preference parameter. We measure the approximation error in the  $\|\cdot\|_\infty$  norm in order to obtain a *uniform* guarantee over all state–action pairs. Since the policy is executed greedily by selecting  $\arg\max_a Q_\beta(s,a)$  at each state, the performance depends on pointwise differences between action values. Controlling the worst-case deviation  $\max_{s,a} |Q_\beta(s,a) - Q_{\beta'}(s,a)|$  therefore ensures that no individual state–action value is significantly misestimated. Other norms such as  $\ell_2$  or  $\ell_1$  provide average-case control and may hide large local errors in a small subset of states, which is undesirable in safety- and resource-sensitive IoT routing scenarios. The  $\ell_\infty$  norm thus aligns naturally with greedy policy execution and yields a clean, dimension-independent worst-case bound. Moreover, translating an  $\ell_2$  bound into a worst-case statement would introduce dimension-dependent looseness via norm inequalities, whereas the  $\ell_\infty$  formulation directly yields a grid-resolution-based design rule.

**Theorem 1 (Lipschitz continuity of optimal Q-functions)** For all  $\beta, \beta' \in [0,1]$ ,

$$\|Q_\beta - Q_{\beta'}\|_\infty \leq \Gamma(H+1)|\beta - \beta'|.$$

Assuming that the learned Q-tables at each grid point are uniformly  $\varepsilon$ -accurate,

$$\|\widehat{Q}_\beta - Q_\beta\|_\infty \leq \varepsilon \quad \text{for all } \beta \in \mathcal{B},$$

the main approximation guarantee for the GIP is obtained.

**Theorem 2 (Uniform approximation bound for GIP)** For any  $\beta \in [0, 1]$ , the interpolated  $Q$ -table satisfies

$$\|Q_\beta^{Int} - Q_\beta\|_\infty \leq \varepsilon + \Gamma(H + 1)(\bar{\beta} - \beta).$$

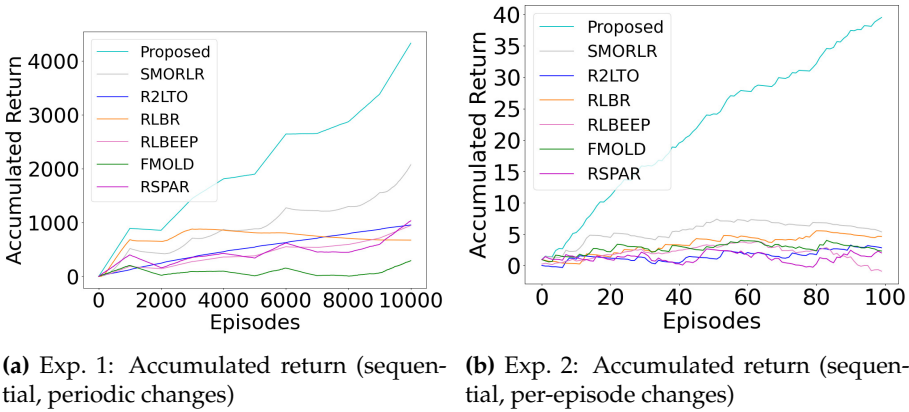
This bound decomposes into a learning error term and an interpolation error term, showing that the GIP is uniformly near-optimal across the entire preference range. Refining the preference grid improves approximation accuracy at the cost of additional computation and storage.

**Experimental Results (Paper VI):** Paper VI evaluates the proposed *distributed dynamic preference Q-learning (distributed DPQ)* routing method in a custom Python simulator (IEEE 802.15.4 abstraction) on a  $10 \times 10$  grid topology ( $N = 100$  nodes). Each episode routes one packet from a randomly chosen source to a fixed sink, with unreliable nodes dropping packets with probability  $p_{\text{drop}}$ . Energy is modelled via per-step and per-hop transmission costs under finite initial node energy. The preference parameter  $\beta \in [0, 1]$  controls the trade-off between packet delivery ratio (PDR) and energy, where a smaller  $\beta$  prioritises reliability and a larger  $\beta$  prioritises energy savings. Four experimental settings are considered by combining (i) *sequential vs simultaneous* exploration–exploitation and (ii) *periodic* preference changes (every 1000 episodes) vs *frequent* preference changes (every episode). The proposed method is compared against RLBEPP [118], R2LTO [119], RLBR [120], FMOLD [121], RSPAR [122], and a static MORL baseline SMORLR (which re-learns from scratch after preference changes).

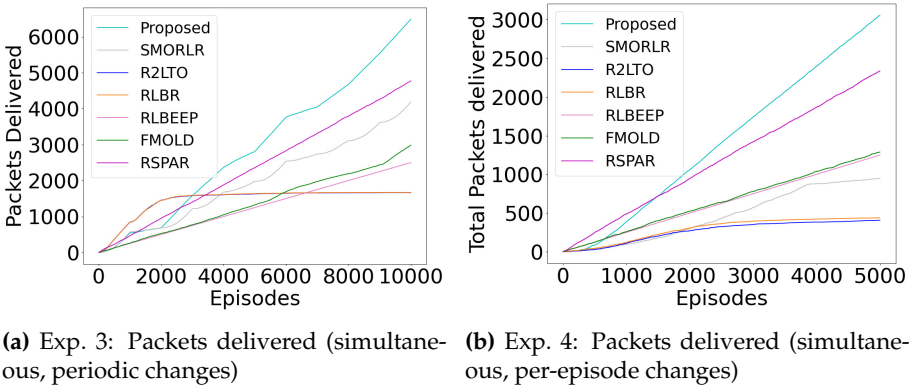
**Key result 1: Fast adaptation and highest overall reward under dynamic preferences.** The main performance indicator in Paper VI is the *overall reward*, which matches the active preference  $\beta_m$ :

$$r^{\text{Overall}}(s, a) = \beta_m r^{\text{Energy}}(s, a) + (1 - \beta_m) r^{\text{PDR}}(s, a). \quad (4.7)$$

Under sequential exploration–exploitation with periodic preference changes (Experiment 1), the proposed method achieves consistently higher episodic rewards and a substantially larger accumulated return than all baselines. This is because the method learns and reuses preference-aware value information, whereas SMORLR must re-learn after each preference update and static baselines cannot adapt. When preferences change every episode (Experiment 2), the proposed method still maintains the strongest accumulated return, indicating robustness under rapidly shifting objectives (Fig. 4.11).



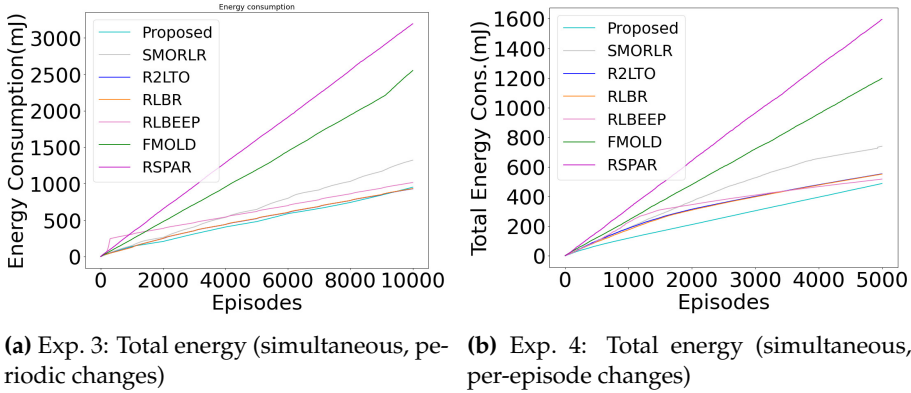
**Figure 4.11:** Overall reward (accumulated return) under sequential exploration–exploitation in **Paper VI**.



**Figure 4.12:** PDR performance (cumulative packets delivered) under simultaneous exploration–exploitation in **Paper VI**.

**Key result 2: Higher reliability when PDR is prioritised, without sacrificing adaptation speed.** **Paper VI** also reports the underlying objectives explicitly. Under simultaneous exploration–exploitation (Experiments 3 and 4), the proposed method delivers more packets once sufficient learning has taken place and maintains higher cumulative packet delivery than all baselines (Fig. 4.12). The gains are most visible in regimes where the preference places higher weight on PDR, confirming that the routing policy can shift toward reliability when required. In contrast, energy-focused RL baselines do not explicitly optimise PDR, while SMORLR adapts too slowly under changing preferences.

**Key result 3: Competitive energy efficiency across preferences, including under frequent changes.** The proposed method remains energy-



**Figure 4.13:** Energy consumption (cumulative/total) under simultaneous exploration–exploitation in **Paper VI**.

**Table 4.1:** Sensitivity results for window length  $W = 200$  (mean  $\pm$  std).

Method	Reward ( $\uparrow$ )	PDR ( $\uparrow$ )	Energy ( $\downarrow$ )
DPQ coarse_2	<b>0.610 <math>\pm</math> 0.374</b>	0.815 $\pm$ 0.298	0.437 $\pm$ 0.579
DPQ fine_11	<b>0.605 <math>\pm</math> 0.374</b>	0.809 $\pm$ 0.285	0.101 $\pm$ 0.017
SMORLR (Static-Q)	0.292 $\pm$ 0.223	0.403 $\pm$ 0.158	0.140 $\pm$ 0.025
R2LTO	0.350 $\pm$ 0.185	0.545 $\pm$ 0.031	0.150 $\pm$ 0.005
RLBR	0.350 $\pm$ 0.185	0.545 $\pm$ 0.031	0.150 $\pm$ 0.005
RLBEED	0.350 $\pm$ 0.185	0.545 $\pm$ 0.031	0.150 $\pm$ 0.005

efficient across preference regimes. Under simultaneous exploration–exploitation, it achieves lower cumulative energy consumption than all baselines while still achieving strong PDR and overall reward. This shows that the improved reliability does not come from excessive energy expenditure, and that preference adaptation does not destabilise energy behaviour (Fig. 4.13).

**Sensitivity analysis: robustness to preference persistence and grid resolution.** To assess robustness to how long a preference remains active, **Paper VI** evaluates performance over testing windows of length  $W \in \{50, 200, 500\}$  episodes and compares two discretisations of the preference space: a coarse grid (DPQ coarse\_2) and a fine grid (DPQ fine\_11). Table 4.1 reports a representative case ( $W = 200$ ). Both DPQ variants achieve the highest reward among all methods, indicating robust performance even when preferences persist for moderate durations. The fine grid also improves energy behaviour while maintaining a similarly high reward, consistent with finer preference discretisation yielding a better approximation of the continuous preference space.

**Summary of the remaining results.** The full **Paper VI** results include additional plots for episodic (not just cumulative) reward, energy, and PDR under both exploration strategies and both preference-change regimes. These additional results reinforce the same conclusion: the proposed distributed DPQ-learning method adapts rapidly after preference changes by reusing learned preference-conditioned value information, outperforms static routing baselines across dynamic regimes, and consistently improves the long-run scalarised objective while maintaining strong PDR and competitive energy consumption. The sensitivity analysis for other window lengths ( $W = 50$  and  $W = 500$ ) shows the same qualitative trend as Table 4.1, confirming that the performance advantage persists under both shorter and longer preference persistence.

**Results Summary (Paper IV):** **Paper IV** complements the results of **Paper VI** by positioning MORL within a broader design space of IoT decision-making problems. **Paper IV** provides a structured synthesis of existing work on multi-objective reinforcement learning (MORL) in IoT systems. The paper surveys representative MORL-based approaches across a range of application domains, including routing, task scheduling, resource allocation, virtual machine placement, and data stream processing. For each class of problems, the analysis focuses on how objectives are modelled, how trade-offs are handled, and how preference information is incorporated into the learning process.

A central result of **Paper IV** is a systematic categorisation of MORL approaches based on their preference-handling mechanisms. Across the surveyed literature, most solutions rely on manually designed reward functions or fixed preference vectors that are selected offline, either through heuristics or hyperparameter tuning. In several cases, separate learning models are trained for individual objectives and combined using externally specified weights. While these approaches demonstrate the feasibility of MORL in IoT settings, they typically assume static objectives and do not support online adaptation when preferences change.

Table 4.2 summarises representative MORL-based IoT solutions, highlighting the problem domain, the adopted MORL methodology, and the considered objectives. From this comparison, **Paper IV** identifies several recurring characteristics of existing MORL solutions. Preference specification is commonly fixed during training, and changes in objective weighting often require retraining or manual reconfiguration. Moreover, many approaches assume centralised control or stable system conditions, which limits their applicability in large-scale and dynamic IoT deployments.

Based on these findings, **Paper IV** highlights the need for MORL frameworks that (i) decouple policy learning from preference speci-

MORL in IoT			
Ref.	Problem / Application	MORL Approach	Objectives
[123]	Workflow scheduling	Manually designed reward	Workflow completion time, Cost of virtual machines
[124]	IoT-based canal control	Reward network learns preference vector and feeds it to DQN	Speed, Safety, Efficiency
[125]	Task scheduling	Preference vector tuned via hyperparameter optimisation	Task execution time, Processing cost, Resource utilisation
[126]	Virtual machine placement	Manually designed preference vector	Reliability, Interference, Power consumption
[127]	Virtual machine placement	Tabular Q-learning with fixed preference vector	Load balancing across CPU, memory, and bandwidth
[128]	Resource allocation for IoV	Tabular Q-learning with manually designed reward	Reliability, Latency
[129]	Cloud resource scheduling	DRL with tuned preference vector	Energy, Quality of service
[130]	UAV-assisted IoT networks	Extended DDPG with fixed preference vector	Data rate, Harvested energy, UAV energy consumption
[131]	Routing	Multiple DQNs combined using a user-defined preference vector	Delay, Network lifetime, Throughput
[5]	Data stream processing and offloading	R-learning with decision-maker-defined preference vector	Energy, Delay

**Table 4.2:** Multi-objective optimisation problems in the IoT, representative MORL approaches, and their objectives (adapted from **Paper IV**).

fication, (ii) support dynamic and online preference adaptation, and (iii) scale to distributed IoT environments. These identified gaps directly motivate the algorithmic developments presented in **Paper VI**, which address preference-adaptive decision-making with formal performance guarantees.

Together, **Paper IV** and **Paper VI** demonstrate that multi-objective reinforcement learning provides both conceptual clarity and practical benefits for AI-based decision-making in distributed IoT systems. Overall, these contributions address **SQ3** by showing how MORL enables adaptive, preference-aware decision-making in dynamic IoT environments, thereby supporting the main research question **RQ**.

## 4.5 Summary

This chapter summarises the main results of the thesis by organising the contributions of the included papers around the three research aims. Each section focused on a specific class of decision-making problems in IoT systems, including federated learning, bandit-based resource allocation, and reinforcement learning for routing and data processing.

Although the problems differ in form, they share common challenges such as uncertainty, limited resources, and competing objectives. The results show how AI-based methods can be designed to balance trade-offs between energy, communication cost, accuracy, latency, and reliability. Both theoretical analysis and simulations were used to support the findings. Together, the results demonstrate that adaptive and learning-based approaches are well-suited for practical IoT scenarios. The next chapter discusses the broader implications of these results, their limitations, and possible directions for future work.



# 5. Concluding Remarks

This chapter concludes the thesis by reflecting on the overall research aims, the problem formulations adopted in the individual papers, and the insights gained from their combined contributions. The chapter is structured as follows. Section 5.1 discusses how the different papers address the research aims through distinct but complementary problem setups and learning frameworks. Section 5.2 synthesises the main findings by explicitly answering the research questions. Section 5.3 outlines the main limitations of the presented work, and Section 5.4 discusses promising directions for future research.

## 5.1 Discussion

The overall aim of this thesis has been to study how learning-based methods can support adaptive decision-making in IoT and wireless systems operating under resource constraints and multiple, potentially conflicting objectives. To address this broad goal, the thesis was structured around three specific research aims. Each aim motivated a different class of problem formulations and learning paradigms, ranging from federated learning and bandits to reinforcement learning and multi-objective optimisation. Rather than relying on a single unified model, the thesis intentionally explores multiple setups, each chosen to match the structure of the underlying system-level challenge.

### 5.1.1 Aim I: Communication-Efficient Federated Learning

The first aim focuses on reducing communication and energy costs in distributed learning systems while maintaining learning performance. This aim is primarily addressed in Papers I and II, which consider federated learning scenarios with energy- and communication-constrained edge devices.

To address this aim, the learning problem is formulated as a distributed optimisation task where model updates must be exchanged repeatedly between clients and a server. The key challenge is that communication itself is costly, especially in IoT settings, where devices are battery-powered and operate over constrained wireless links. Rather

than treating communication constraints as fixed external limitations, the papers explicitly model the communication cost and incorporate it into the learning process.

Both papers adopt adaptive gradient sparsification as the core mechanism. Instead of using static compression levels selected via offline hyperparameter tuning, the sparsification level is adjusted online at each iteration. The problem setup therefore couples learning dynamics with communication cost models, using metrics such as energy consumption, transmitted bits, or latency. Paper I introduces this idea in an energy-focused setting, while Paper II generalises the framework to multiple cost models and incorporates error compensation, enabling stronger theoretical guarantees and improved convergence behaviour.

The chosen setup reflects the nature of the aim: communication efficiency is best handled at the algorithmic level, close to the learning updates themselves. By formulating the problem as an online trade-off between information content and communication cost, the papers demonstrate how resource awareness can be embedded directly into distributed learning algorithms.

### 5.1.2 Aim II: Constraint-Aware Online Learning

The second aim addresses decision-making under explicit constraints that evolve over time. This aim is mainly studied in Papers III and V, which move beyond federated learning and consider online decision-making problems faced by individual IoT nodes.

In Paper III, the problem is formulated as a constrained multi-armed bandit with a dynamic violation budget. The bandit setup is appropriate here because the decision-maker repeatedly selects actions under uncertainty while receiving limited feedback. Constraints are not treated as static limits but as quantities that can be temporarily violated as long as long-term budgets are respected. This formulation allows a principled study of regret and constraint violations simultaneously, which aligns well with the aim of balancing performance and feasibility in resource-limited systems.

Paper V considers a different but related setting: edge data stream processing and offloading. Here, the decision-maker must choose whether to process data locally, offload it, or delay processing, subject to energy, delay, and storage constraints. This problem is naturally sequential and state-dependent, making reinforcement learning a suitable method. Unlike the bandit setting, the system dynamics and long-term consequences of actions play a central role. The setup therefore emphasises learning from interaction over time while still explicitly modelling resource constraints.

Together, these papers illustrate that explicit constraints can be handled at different abstraction levels. Bandits provide a clean framework for studying constraint violations and guarantees, while reinforcement learning enables richer system dynamics and control decisions. The choice of setup in each paper reflects the complexity of the underlying system being modelled.

### 5.1.3 Aim III: Preference-Adaptive Multi-objective Reinforcement Learning

The third aim focuses on multi-objective decision-making in IoT systems, where trade-offs between objectives change over time. This aim is primarily addressed in Paper VI, with conceptual support from Paper IV.

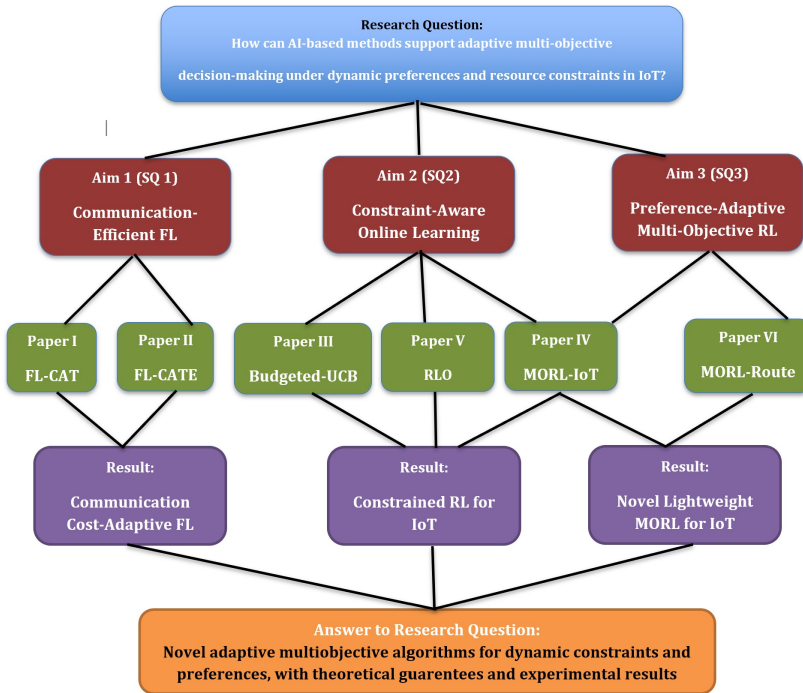
Paper VI studies routing in IoT networks, where objectives such as energy consumption and packet delivery reliability are inherently conflicting. Unlike traditional approaches that rely on fixed scalarisation weights, the problem is formulated to allow preferences to vary dynamically. The learning setup is fully distributed and based on multi-objective Q-learning, enabling nodes to adapt routing decisions in real time without centralised control or retraining.

The formulation reflects the central challenge of this aim: IoT systems often operate under changing application requirements, and learning algorithms must generalise across preference settings. By learning multiple value functions and exploiting structural properties of the value space, the proposed approach supports fast adaptation while retaining theoretical guarantees.

Paper IV complements this contribution by providing a broader perspective on multi-objective and constrained reinforcement learning in the IoT. Rather than focusing on a single algorithm, it frames the challenges, design choices, and open problems associated with applying MORL in resource-constrained networks. This conceptual grounding helps position the technical contributions of the other papers within a unified research landscape.

## 5.2 Conclusion

This section concludes the thesis by explicitly answering the research questions based on the combined insights from all six papers. While some papers provide concrete algorithmic solutions, others contribute by framing the broader methodological landscape in which these solutions operate. Figure 5.1 summarises the overall structure of the thesis



**Figure 5.1:** Conceptual overview of the thesis contributions, showing the connection between the research question, the three research aims, the associated papers, and their resulting contributions to adaptive multi-objective decision-making in the IoT.

and illustrates how the three research aims and corresponding papers collectively answer the main research question.

### 5.2.1 Answering RQ1: How can communication and energy costs in distributed learning systems, such as federated learning, be reduced while preserving convergence and model performance?

The results in Papers I and II show that communication and energy costs in distributed learning can be significantly reduced by making learning algorithms resource-aware. Rather than relying on static compression or heuristic tuning, adaptive mechanisms that optimise the trade-off between information content and communication cost can be integrated directly into the learning process. Theoretical and empirical results demonstrate that such adaptive schemes can achieve equal or better learning performance while substantially lowering communication overhead.

### 5.2.2 Answering RQ2: How can online learning methods handle explicit and time-varying constraints in a principled and theoretically grounded manner?

Papers III and V demonstrate that constraints can be incorporated into learning-based decision-making in a principled manner. In Paper III, constraints are modelled explicitly through dynamic violation budgets in a bandit framework, allowing provable guarantees on both regret and constraint satisfaction. Paper V extends this perspective to sequential decision-making problems, where reinforcement learning is used to balance long-term performance with resource constraints such as energy, delay, and storage.

Paper IV complements these contributions by providing a unifying conceptual framework for constrained and multi-objective reinforcement learning in IoT systems. It clarifies how constraints, objectives, and preferences interact in practical IoT optimisation problems, and situates the algorithmic results of Papers III and V within a broader class of learning-based control problems. Together, these works show that explicit constraints can be treated as first-class components of learning problems across different abstraction levels.

### 5.2.3 Answering RQ3: How can multi-objective reinforcement learning adapt routing decisions in IoT systems under dynamically changing trade-offs between energy consumption and reliability?

Paper VI shows that multi-objective reinforcement learning enables IoT systems to adapt to changing preferences without retraining or centralised coordination. By learning policies that generalise across objective weightings, nodes can adjust their behaviour in real time as application requirements shift while retaining theoretical guarantees.

Paper IV provides the conceptual foundation for this result by highlighting the inherent multi-objective nature of IoT systems and reviewing key MORL principles such as Pareto optimality, scalarisation, and preference adaptation. By framing preference-aware and preference-adaptive learning as a central requirement for future IoT networks, Paper IV strengthens the interpretation and generality of the results obtained in Paper VI.

### 5.2.4 Design Guidelines for AI-Driven IoT Systems Under Resource Constraints

Based on the theoretical analyses and empirical findings across Papers I–VI, we summarise the following practical design guidelines for AI-driven

IoT systems operating under limited communication and energy resources.

## **Guidelines for communication- and energy-efficient federated learning (Papers I and II)**

- 1. Dimension device memory explicitly for compression and error compensation.** Client nodes must allocate memory not only for local models and gradients but also for a compression-related state. In FL-CATE, each client maintains an error-compensation (residual) buffer of the same dimension as the transmitted update. For a model with  $d$  parameters and floating-point precision FPP bits, this results in a minimum memory requirement of approximately  $2d \times \text{FPP}$  bits per client (for the update and residual buffer). On resource-constrained devices, this effectively bounds the maximum feasible model size and favours lightweight architectures or partial model updates.
- 2. Measure and expose per-bit communication cost at the hardware level.** Adaptive sparsification relies on communication cost models that depend on transmitted bits, energy, or latency. These costs are highly hardware- and channel-dependent and should be characterised at deployment time. For energy-aware operation, this includes estimating parameters such as the transmission power, channel gain, bandwidth, and baseline radio overhead. System designers should expose these measurements through a lightweight interface between the radio or operating system and the learning module, enabling compression decisions based on actual hardware conditions rather than nominal specifications.
- 3. Verify and exploit compute–communication asymmetry.** In low-power IoT nodes, radio transmission typically consumes significantly more energy than local floating-point operations. When this regime holds, it is beneficial to spend additional local computation (e.g. sparsification scoring, residual updates) to reduce transmitted bits. Designers should verify this asymmetry by comparing the measured energy per transmitted bit versus per multiply–accumulate operation.
- 4. Limit compression aggressiveness according to compute and memory capacity.** Although FL-CATE can adapt compression levels online, highly aggressive sparsification can increase the local processing time and cause growth in residual buffers. Devices

with limited CPU frequency or memory bandwidth should impose upper bounds on sparsification aggressiveness to ensure stable timing and memory usage. In practice, this can be achieved by restricting the search space of  $B$  or reducing the frequency of compression-level optimisation per communication round.

5. **Match communication cost objectives to dominant hardware constraints.** FL-CATE supports optimisation with respect to different communication objectives, including transmitted bits, energy consumption, and latency. System designers should select the cost objective that reflects the dominant hardware bottleneck. Battery-powered sensors should prioritise energy-aware tuning, delay-sensitive applications should emphasise latency-aware tuning, and bandwidth-constrained deployments should focus on minimising transmitted bits. The same learning algorithm can support all three objectives by switching cost models, facilitating reuse across heterogeneous hardware platforms.

### **Guidelines for budget-aware online decision-making under dynamic constraints (Paper III)**

1. **Keep per-action state small and persistent.** Budgeted UCB requires storing only a play count and cumulative reward and constraint statistics per action, resulting in  $\mathcal{O}(K)$  memory usage, where  $K$  denotes the number of available discrete actions (arms). This makes the algorithm suitable for microcontroller-class devices, but also implies that the action space should be discretised coarsely when memory is limited (e.g. a small number of power levels or transmission modes).
2. **Expose constraint thresholds directly from hardware or control logic.** The algorithm assumes the constraint threshold  $C_t$  is available at each decision round. In practice, this requires battery monitors, power controllers, or network schedulers to provide explicit, real-time constraint values to the decision layer, rather than embedding constraint handling implicitly within the learning algorithm.
3. **Measure constraint cost using hardware-level signals.** Accurate per-action measurement of constraint feedback (e.g. energy used, transmit power, airtime) is essential. These measurements should be obtained from hardware counters or calibrated low-level models at the radio or power-management level, since systematic bias in cost estimation directly translates into constraint violations.

4. **Tune violation budgets and decision timescale conservatively.** The initial violation allowance ( $\delta_0$ ) and decay horizon ( $T_{\text{bud}}$ ) should match how much extra load the device can safely handle early on without reducing its lifetime (e.g. temporary battery drain or heat buildup). The bandit decision interval should align with how quickly constraints change: slowly varying constraints permit longer decision cycles, while fast-varying constraints require more frequent decisions to avoid reacting to outdated signals.

**Guidelines for energy-aware stream processing and offloading at the edge (Paper V)** In the following,  $d(k)$  denotes the amount of data arriving at time step  $k$ ,  $T$  is the maximum allowable processing delay,  $X$  is the local storage capacity,  $P$  and  $C$  are the per-step local processing and transmission limits, respectively,  $f_e$  is the edge CPU frequency,  $B_u$  is the uplink bandwidth, and  $p(k)$  and  $c(k)$  denote the amounts of data processed locally and offloaded at time step  $k$ .

1. **Dimension local storage based on delay window and arrival rate.** Local memory should be provisioned to store all unprocessed data over the full delay horizon  $T$ . Specifically, the required storage capacity must satisfy

$$X \geq \sum_{i=0}^{T-1} d(k-i),$$

which in the worst case reduces to  $X \geq T \cdot d_{\text{max}}$ , where  $d_{\text{max}} = \max_k d(k)$  is the maximum per-step data arrival. In practice, this corresponds to maintaining  $T$  age-indexed buffers, each capable of holding up to  $d_{\text{max}}$  data units.

2. **Enforce processing and transmission limits as hard per-step constraints.** The local processing limit  $P$  and transmission limit  $C$  should be enforced by the OS or firmware as strict per-step caps derived from hardware parameters such as the CPU frequency  $f_e$  and available uplink bandwidth  $B_u$ . These limits define the feasible action space and must accurately reflect the device's actual capabilities.
3. **Use calibrated energy models for computation and communication.** Energy consumption should be modelled using hardware-calibrated parameters, with the local computation energy represented as  $e_{\text{lp}}(k) = A(k)p(k)$  and the offloading energy as  $e_{\text{rp}}(k) = B(k)c(k)$ . Here,  $A(k)$  depends on the CPU frequency and processing density, while  $B(k)$  captures the transmission power, channel

gain, and data-size reduction. Separating computation and communication costs is essential, as they scale differently with workload and channel conditions.

4. **Select control interval relative to arrival rate and deadline.** The decision interval should be chosen such that backlog growth between decisions remains bounded. In particular, the control period should be short relative to both the data arrival rate and the processing deadline  $T$ . High-rate or tight-deadline streams require finer decision intervals, while lower-rate streams allow coarser scheduling with reduced control overhead.

### Guidelines for dynamic multi-objective routing (Paper VI)

1. **Design routing as a multi-objective, preference-aware problem.** Instead of optimising a single static metric (e.g. hop count), maintain value estimates for multiple objectives (energy, reliability, latency) and combine them through a preference parameter that can vary over time. This enables the routing policy to adapt dynamically to changing application priorities (e.g. emergency vs routine operation) without redesigning the algorithm.
2. **Exploit structure in preference space to avoid retraining.** Paper VI establishes the Lipschitz continuity of the optimal  $Q$ -function with respect to the preference parameter. This structural property enables interpolation between previously learned preference points, allowing real-time adaptation without retraining from scratch. In practice, preference interpolation significantly reduces computational overhead when priorities change.
3. **Favour energy-balanced routing over myopic shortest paths.** Including residual energy and congestion in the reward discourages the overuse of specific nodes and distributes traffic more evenly across the network. Compared to shortest-path or static-metric routing, this approach substantially improves the network lifetime while maintaining reliability.
4. **Tune preference-grid resolution according to per-node memory and compute limits.** The preference grid size  $|\mathcal{B}|$  provides a direct accuracy–resource trade-off. Since learning is fully distributed, each node stores  $Q$ -values only for its local state space. Let  $|S_{\text{local}}|$  denote the number of locally represented states and  $|A|$  the number of candidate next-hop actions. Each preference point requires storing  $|S_{\text{local}}| \cdot |A|$   $Q$ -values, resulting in a per-node storage requirement of  $|\mathcal{B}| \cdot |S_{\text{local}}| \cdot |A|$  floating-point values. With 64-bit

precision, this corresponds to approximately  $8|\mathcal{B}||S_{\text{local}}||A|$  bytes per node. System designers can therefore select  $|\mathcal{B}|$  to balance interpolation accuracy against memory and computational constraints in resource-limited IoT devices.

**Overall synthesis:** Across all components of the system, communication should be treated as the primary scarce resource in the targeted IoT regime. Local computation can often be leveraged to reduce or better schedule communication. Adaptive compression, explicit budget-aware learning, and preference-adaptive multi-objective optimisation together provide a principled framework for designing scalable and energy-efficient IoT systems.

### 5.3 Limitations

Despite the contributions of this thesis, several limitations should be acknowledged. These limitations relate to modelling assumptions, algorithmic choices, the evaluation methodology, and the scope, and they also help contextualise the applicability of the results.

First, many of the proposed methods rely on simplified system models. Communication costs, energy consumption, and network dynamics are typically represented using analytical or abstracted models. While these models are standard in the literature and capture key trade-offs, real-world IoT deployments often exhibit additional sources of uncertainty, such as hardware heterogeneity, non-stationary wireless conditions, protocol overheads, and device failures. As a result, the quantitative gains observed in simulation may differ in deployments in operational systems.

Second, several learning algorithms assume sufficient exploration and stationarity over the learning horizon. In practice, IoT environments may change abruptly due to node mobility, traffic bursts, or failures, which can violate these assumptions. While adaptivity is a central theme of this thesis, most algorithms still rely on implicit stability over certain time scales to guarantee convergence or performance bounds.

Third, scalability remains a challenge. Some approaches rely on tabular representations, discrete action spaces, or per-preference learning. While these choices are well suited for constrained IoT devices and enable theoretical analysis, they may limit applicability in high-dimensional settings or large-scale networks. Extending these methods to richer state representations without sacrificing interpretability or guarantees is non-trivial.

Fourth, the evaluation of several contributions is primarily simulation-based. Although simulations allow controlled experimentation and repeatability, they cannot fully capture the complexity of real-world deployments, such as protocol interactions, background traffic, or long-term hardware degradation. Experimental validation on physical testbeds would therefore be an important next step to strengthen the practical relevance of the results.

Finally, each paper addresses a specific aspect of adaptive decision-making in IoT systems, often in isolation. While this modular approach is useful for clarity and depth, real IoT systems typically involve tightly coupled decisions across communication, computation, routing, and learning layers. The lack of fully integrated cross-layer evaluations limits the ability to assess emergent behaviours arising from the interaction of multiple adaptive components.

## 5.4 Future Directions

The limitations identified above point toward several promising directions for future research. These directions extend the contributions of this thesis and align with broader challenges in learning-enabled IoT systems.

### 5.4.1 Beyond Image Data: Streaming and Medical Time Series

While the federated learning experiments in Papers I and II use standard image datasets, the proposed adaptive sparsification methods are agnostic to the underlying data modality. The algorithms operate on model updates and gradient statistics, and therefore can equally compress gradients produced by models trained on time series, tabular data, or textual data.

Importantly, part of this thesis already addresses non-image streaming data. In Paper V, we consider time-sensitive streaming data at the network edge and develop the RLO framework, which learns whether to process locally, offload, or buffer incoming data under energy and latency constraints. The evaluation includes both real and synthetic data streams, demonstrating applicability beyond static image datasets.

Similarly, the formulations in Papers III and IV model rewards and costs abstractly. The reward may represent task quality under any modality, while the cost terms capture energy, bandwidth, or latency. The algorithms therefore do not rely on image-specific assumptions.

Medical time series data, such as ECG, EEG, or ICU monitoring signals, introduce additional challenges. First, physiological signals exhibit strong temporal correlations and potential concept drift, requiring

continuous online adaptation. Second, clinically relevant events are often rare, making false negatives particularly costly. Third, strict latency constraints may require decisions within seconds, limiting batching and delayed communication. Finally, privacy and regulatory requirements further motivate on-device processing and federated learning.

The frameworks developed in this thesis are well suited to such settings. Adaptive sparsification (Papers I and II) can be combined with sequence models (e.g. 1D CNNs or recurrent architectures) to reduce communication overhead while preserving model quality. The RLO framework (Paper V) and the constrained and multi-objective RL formulations (Papers IV and VI) can encode latency and safety requirements directly into reward functions or constraints. Budgeted bandits (Paper III) provide a natural mechanism for adaptive sampling or feature selection under time-varying energy or bandwidth budgets.

Evaluating the proposed methods on real medical time series datasets constitutes an important direction for future work. However, the algorithms themselves do not rely on image-specific properties and extend naturally to streaming and physiological data.

#### 5.4.2 Other Future Work

Another important direction is scalability through function approximation. Incorporating deep learning or other approximation techniques could enable the proposed methods to handle larger state and action spaces, continuous variables, and richer system observations. A key challenge here is to balance scalability with stability, interpretability, and theoretical guarantees, especially in safety- or resource-critical IoT applications.

A second direction concerns cross-layer and joint optimisation. Many of the problems studied in this thesis focus on a single decision layer, such as communication efficiency, routing, or edge processing. Future work could explore unified frameworks that jointly optimise decisions across multiple layers, for example, combining routing, learning, and computation offloading into a single adaptive control problem. Such integration could reveal new trade-offs and coordination challenges that are not visible when layers are studied independently.

A third direction is moving from preference-aware learning to preference learning. In several works, objective weights or preferences are assumed to be provided externally. In practice, these preferences may be implicit, uncertain, or evolving. Learning preferences from data, user behaviour, or system feedback would enable more autonomous and human-aligned IoT systems, particularly in dynamic environments.

Another important avenue is robustness and safety. Future research could explicitly address robustness to distribution shifts, adversarial behaviour, or partial observability. Incorporating safety constraints, risk-sensitive objectives, or worst-case guarantees into learning-based decision-making remains an open and practically relevant challenge for IoT systems deployed in critical infrastructure.

Finally, experimental validation and deployment represent an essential step forward. Implementing and testing the proposed algorithms on real IoT testbeds, edge platforms, or large-scale emulators would provide valuable insights into their practical feasibility, overheads, and long-term behaviour. Such studies could also inform the design of new abstractions and models that better bridge the gap between theory and practice.

Overall, the research directions outlined above suggest that adaptive, learning-based decision-making for IoT systems remains a rich and evolving field. The contributions of this thesis provide a foundation for further work toward scalable, robust, and autonomous IoT networks.



# References

- [1] SHUBHAM VAISHNAV, MARIA EFTHYMIOU, AND SINDRI MAGNÚSSON. **Energy-efficient and adaptive gradient sparsification for federated learning.** In *ICC 2023-IEEE International Conference on Communications*, pages 1256–1261. IEEE, 2023. [ix](#)
- [2] SHUBHAM VAISHNAV, SARIT KHIRIRAT, AND SINDRI MAGNÚSSON. **Communication-adaptive-gradient sparsification for federated learning with error compensation.** *IEEE Internet of Things Journal*, **12**(2):1137–1152, 2025.
- [3] SHUBHAM VAISHNAV, PRAVEEN KUMAR DONTA, AND SINDRI MAGNÚSSON. **Adaptive budgeted multi-armed bandits for IoT with dynamic resource constraints.** In *GLOBE-COM 2025 - 2025 IEEE Global Communications Conference*, pages 4535–4540, 2025.
- [4] SHUBHAM VAISHNAV AND SINDRI MAGNÚSSON. **Multi-objective and constrained reinforcement learning for IoT.** In PRAVEEN KUMAR DONTA, ABHISHEK HAZRA, AND LAURI LOVÉN, editors, *Learning techniques for the Internet of Things*, pages 153–170. Springer Nature Switzerland, Cham, 2024. [xix](#), [46](#), [81](#)
- [5] SHUBHAM VAISHNAV AND SINDRI MAGNÚSSON. **Intelligent processing of data streams on the edge using reinforcement learning.** In *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1265–1270. IEEE, 2023. [100](#)
- [6] SHUBHAM VAISHNAV, PRAVEEN KUMAR DONTA, AND SINDRI MAGNÚSSON. **Dynamic and distributed routing in IoT networks based on multi-objective Q-learning.** *IEEE Internet of Things Journal*, pages 1–1, 2026. [ix](#)
- [7] DIMITRI BERTSEKAS. *Dynamic programming and optimal control: Volume I*, **4**. Athena Scientific, 2012. [24](#)
- [8] KAISA MIETTINEN. *Nonlinear multiobjective optimization*, **12**. Springer Science & Business Media, 1999. [24](#)
- [9] MATTHIAS EHRGOTT. *Multicriteria optimization*. Springer, 2005. [24](#)
- [10] DIMITRI BERTSEKAS AND STEVEN E SHREVE. *Stochastic optimal control: The discrete-time case*, **5**. Athena Scientific, 1996. [24](#)
- [11] EITAN ALTMAN. *Constrained Markov decision processes*. Routledge, 2021. [24](#)
- [12] MARTIN L PUTERMAN. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014. [24](#)
- [13] STEPHEN BOYD AND LIEVEN VANDENBERGHE. *Convex optimization*. Cambridge University Press, 2004. [24](#)
- [14] MATTHIAS EHRGOTT AND STEFAN RUZIKA. **Improved  $\epsilon$ -constraint method for multiobjective programming.** *Journal of Optimization Theory and Applications*, **138**(3):375–396, 2008. [24](#)
- [15] CRISTINA BAZGAN, STEFAN RUZIKA, CLEMENS THIELEN, AND DANIEL VANDERPOOTEN. **The power of the weighted sum scalarization for approximating multiobjective optimization problems.** *Theory of Computing Systems*, **66**(1):395–415, 2022. [24](#)

- [16] KALYANMOY DEB, AMRIT PRATAP, SAMEER AGARWAL, AND TAMT MEYARIVAN. **A fast and elitist multiobjective genetic algorithm: NSGA-II.** *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. 24
- [17] QINGFU ZHANG AND HUI LI. **MOEA/D: A multiobjective evolutionary algorithm based on decomposition.** *IEEE Transactions on Evolutionary Computation*, 11(6):712–731, 2007. 24
- [18] OZAN SENER AND VLADLEN KOLTUN. **Multi-task learning as multi-objective optimization.** *Advances in Neural Information Processing Systems*, 31, 2018. 24
- [19] BO LIU, XINGCHAO LIU, XIAOJIE JIN, PETER STONE, AND QIANG LIU. **Conflict-averse gradient descent for multi-task learning.** *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021. 24
- [20] SIMONE PARISI, MATTEO PIROTTA, AND MARCELLO RESTELLI. **Multi-objective reinforcement learning through continuous Pareto manifold approximation.** *Journal of Artificial Intelligence Research*, 57:187–227, 2016. 24
- [21] SHIE MANNOR AND NAHUM SHIMKIN. **A geometric approach to multi-criterion reinforcement learning.** *Journal of Machine Learning Research*, 5(Apr):325–360, 2004. 24
- [22] JOSHUA ACHIAM, DAVID HELD, AVIV TAMAR, AND PIETER ABBEEL. **Constrained policy optimization.** In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017. 25
- [23] CHEN TESSLER, DANIEL J MANKOWITZ, AND SHIE MANNOR. **Reward constrained policy optimization.** *arXiv preprint arXiv:1805.11074*, 2018. 25
- [24] DONGSHENG DING, KAIQING ZHANG, TAMER BASAR, AND MIHAILO JOVANOVIĆ. **Natural policy gradient primal-dual method for constrained Markov decision processes.** *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020. 25
- [25] YUJUN LIN, SONG HAN, HUIZI MAO, YU WANG, AND WILLIAM J DALLY. **Deep gradient compression: Reducing the communication bandwidth for distributed training.** *arXiv preprint arXiv:1712.01887*, 2017. 25
- [26] SEBASTIAN U STICH, JEAN-BAPTISTE CORDONNIER, AND MARTIN JAGGI. **Sparsified SGD with memory.** *Advances in Neural Information Processing Systems*, 31, 2018. 25, 28, 70
- [27] SAI PRANEETH KARIMIREDDY, QUENTIN REBJOCK, SEBASTIAN STICH, AND MARTIN JAGGI. **Error feedback fixes SignSGD and other gradient compression schemes.** In *International Conference on Machine Learning*, pages 3252–3261. PMLR, 2019. 25
- [28] JAKUB KONEČNÝ, H BRENDAN MCMAHAN, FELIX X YU, PETER RICHTÁRIK, ANANDA THEERTHA SURESH, AND DAVE BACON. **Federated learning: Strategies for improving communication efficiency.** *arXiv preprint arXiv:1610.05492*, 2016. 25, 69
- [29] XIANG LI, KAIXUAN HUANG, WENHAO YANG, SHUSEN WANG, AND ZHIHUA ZHANG. **On the convergence of FedAvg on non-iid data.** *arXiv preprint arXiv:1907.02189*, 2019. 25
- [30] NAOTO KIMURA AND SHAHRAM LATIFI. **A survey on data compression in wireless sensor networks.** In *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, 2, pages 8–13. IEEE, 2005. 26, 28
- [31] MARK HOROWITZ. **1.1 Computing’s energy problem (and what we can do about it).** In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14. IEEE, 2014. 27
- [32] BERT GYSELINCKX, RAFFAELLA BORZI, AND PHILIPPE MATTELAER. **Human++: Emerging technology for body area networks.** In *Wireless technologies*, pages 221–240. CRC Press, 2017.

- [33] SPARSH MITTAL. **A survey of techniques for improving energy efficiency in embedded computing systems.** *International Journal of Computer Aided Engineering and Technology*, 6(4):440–459, 2014. [27](#)
- [34] USMAN RAZA, PARAG KULKARNI, AND MAHESH SOORIYABANDARA. **Low power wide area networks: An overview.** *IEEE Communications Surveys & Tutorials*, 19(2):855–873, 2017. [27](#)
- [35] BRENDAN MCMAHAN, EIDER MOORE, DANIEL RAMAGE, SETH HAMPSON, AND BLAISE AGUERA Y ARCAS. **Communication-efficient learning of deep networks from decentralized data.** In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. [28](#), [47](#), [69](#)
- [36] KEITH BONAWITZ, HUBERT EICHNER, WOLFGANG GRIESKAMP, DZMITRY HUBA, ALEX INGERMAN, VLADIMIR IVANOV, CHLOE KIDDON, JAKUB KONEČNÝ, STEFANO MAZZOCCHI, BRENDAN MCMAHAN, ET AL. **Towards federated learning at scale: System design.** *Proceedings of Machine Learning and Systems*, 1:374–388, 2019.
- [37] PETER KAIROUZ AND H BRENDAN MCMAHAN. **Advances and open problems in federated learning.** *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 2021. [28](#)
- [38] RICHARD S SUTTON AND ANDREW G BARTO. *Reinforcement learning: An introduction*. MIT Press, 2018. [37](#), [40](#), [41](#), [42](#)
- [39] KAN ZHENG, ZHE YANG, KUAN ZHANG, PERIKLIS CHATZIMISIOS, KAN YANG, AND WEI XIANG. **Big data-driven optimization for mobile networks toward 5G.** *IEEE Network*, 30(1):44–51, 2016. [37](#), [40](#)
- [40] AXEL ABELS, DIEDERIK ROIJERS, TOM LENAERTS, ANN NOWÉ, AND DENIS STECKELMACHER. **Dynamic weights in multi-objective deep reinforcement learning.** In *International Conference on Machine Learning*, pages 11–20. PMLR, 2019. [38](#)
- [41] RUNZHE YANG, XINGYUAN SUN, AND KARTHIK NARASIMHAN. **A generalized algorithm for multi-objective reinforcement learning and policy adaptation.** *Advances in Neural Information Processing Systems*, 32, 2019. [38](#)
- [42] YINING LU, ZILONG WANG, SHIYANG LI, XIN LIU, CHANGLONG YU, QINGYU YIN, ZHAN SHI, ZIXUAN ZHANG, AND MENG JIANG. **Learning to optimize multi-objective alignment through dynamic reward weighting.** *arXiv preprint arXiv:2509.11452*, 2025. [38](#)
- [43] MOHAMMAD MIRZANEJAD, MORTEZA EBRAHIMI, PETER VAMPLEW, AND HADI VEISI. **An online scalarization multi-objective reinforcement learning algorithm: TOPSIS Q-learning.** *The Knowledge Engineering Review*, 37:e7, 2022. [38](#)
- [44] JOHANNES DORNHEIM. **GTLO: A generalized and non-linear multi-objective deep reinforcement learning approach.** *arXiv preprint arXiv:2204.04988*, 2022. [38](#)
- [45] JUNLIN LU, PATRICK MANNION, AND KARL MASON. **Inferring preferences from demonstrations in multi-objective reinforcement learning.** *Neural Computing and Applications*, 36(36):22845–22865, 2024. [38](#)
- [46] QIAN LIN, ZONGKAI LIU, DANYING MO, AND CHAO YU. **An offline adaptation framework for constrained multi-objective reinforcement learning.** *Advances in Neural Information Processing Systems*, 37:140292–140319, 2024. [38](#)
- [47] MD ZAHANGIR ALAM, SURYAIA RAHMAN, MD ASIF BIN KHALED, ASHRAFUL ISLAM, AND ABBAS JAMALIPOUR. **A graph-assisted digital-twin-driven multiagent shared offloading for Internet of Vehicles.** *IEEE Internet of Things Journal*, 12(11):17349–17363, 2025. [38](#)
- [48] ZESONG FEI, BIN LI, SHAOSHI YANG, CHENGWEN XING, HONGBIN CHEN, AND LAJOS HANZO. **A survey of multi-objective optimization in wireless sensor networks: Metrics, algorithms, and open problems.** *IEEE Communications Surveys & Tutorials*, 19(1):550–586, 2016. [39](#), [44](#), [51](#)

- [49] SÉBASTIEN BUBECK AND NICOLO CESA-BIANCHI. **Regret analysis of stochastic and non-stochastic multi-armed bandit problems.** *arXiv preprint arXiv:1204.5721*, 2012. 40
- [50] ASHWINKUMAR BADANIDIYURU, ROBERT KLEINBERG, AND ALEKSANDRS SLIVKINS. **Bandits with knapsacks.** *Journal of the ACM (JACM)*, 65(3):1–55, 2018. 40, 81
- [51] YUXI LI. **Deep reinforcement learning: An overview.** *arXiv preprint arXiv:1701.07274*, 2017. 43
- [52] RONGPENG LI, ZHIFENG ZHAO, QI SUN, CHENYANG YANG, XIANFU CHEN, MINJIAN ZHAO, HONGGANG ZHANG, ET AL. **Deep reinforcement learning for resource management in network slicing.** *IEEE Access*, 6:74429–74441, 2018. 43
- [53] MINGZHE CHEN, URSULA CHALLITA, WALID SAAD, CHANGCHUAN YIN, AND MÉROUANE DEBBAH. **Artificial neural networks-based machine learning for wireless networks: A tutorial.** *IEEE Communications Surveys & Tutorials*, 21(4):3039–3071, 2019. 43
- [54] KALYANMOY DEB, KARTHIK SINDHYA, AND JUSSI HAKANEN. **Multi-objective optimization.** In *Decision sciences*, pages 161–200. CRC Press, 2016. 46
- [55] DIEDERIK M ROIJERS, PETER VAMPLEW, SHIMON WHITESON, AND RICHARD DAZELEY. **A survey of multi-objective sequential decision-making.** *Journal of Artificial Intelligence Research*, 48:67–113, 2013. 46, 52
- [56] WEI YANG BRYAN LIM, NGUYEN CONG LUONG, DINH THAI HOANG, YUTAO JIAO, YING-CHANG LIANG, QIANG YANG, DUSIT NIYATO, AND CHUNYAN MIAO. **Federated learning in mobile edge networks: A comprehensive survey.** *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020. 47
- [57] TAKAYUKI NISHIO AND RYO YONETANI. **Client selection for federated learning with heterogeneous resources in mobile edge.** In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2019. 48
- [58] JACOB DEVLIN, MING-WEI CHANG, KENTON LEE, AND KRISTINA TOUTANOVA. **BERT: Pre-training of deep bidirectional transformers for language understanding.** *arXiv preprint arXiv:1810.04805*, 2018. 48
- [59] KEN PEFFERS, TUURE TUUNANEN, MARCUS A ROTHENBERGER, AND SAMIR CHATTERJEE. **A design science research methodology for information systems research.** *Journal of Management Information Systems*, 24(3):45–77, 2007. 55, 56
- [60] SHIRLEY GREGOR AND ALAN R HEVNER. **Positioning and presenting design science research for maximum impact.** *MIS Quarterly*, pages 337–355, 2013. 55
- [61] ROEL WIERINGA. *Design science methodology for information systems and software engineering.* Springer, 2014. 55
- [62] BERNARD WARNER. **The sciences of the artificial.** *Journal of the Operational Research Society*, 20(4):509–510, 1969. 56
- [63] ALAN R HEVNER, SALVATORE T MARCH, JINSOO PARK, AND SUDHA RAM. **Design science in information systems research.** *MIS Quarterly*, pages 75–105, 2004. 56
- [64] YANN LECUN. **The MNIST database of handwritten digits.** <http://yann.lecun.com/exdb/mnist/>, 1998. 60
- [65] HAN XIAO, KASHIF RASUL, AND ROLAND VOLLGRAF. **Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms.** *arXiv preprint arXiv:1708.07747*, 2017. 60
- [66] JAMES R NORRIS. *Markov chains*. Number 2. Cambridge University Press, 1998. 61

- [67] BOLEI ZHOU, AGATA L., ADITYA K., AUDE O., AND ANTONIO T. **Places: A 10 million image DB for scene recognition.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 61
- [68] STELIOS G VRACHIMIS, DEMETRIOS G ELIADES, RICCARDO TAORMINA, AVI OSTFELD, ZORAN KAPELAN, SHUMING LIU, MARIOS KYRIAKOU, PAVLOS PAVLOU, MENGNING QIU, AND MARIOS M POLYCARPOU. **BattLeDIM: Battle of the leakage detection and isolation methods.** In *Proc., 2nd Int. CCWI/WDSA Joint Conf*, 2020. 61
- [69] XAVIER CARON, RACHELLE BOSUA, SEAN B MAYNARD, AND ATIF AHMAD. **The Internet of Things (IoT) and its impact on individual privacy: An Australian perspective.** *Computer Law & Security Review*, 32(1):4–15, 2016. 63
- [70] CATHY O’NEIL. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishers, New York, 2016. 64
- [71] ARTHUR L SAMUEL. **Some studies in machine learning using the game of checkers. II—Recent progress.** *IBM Journal of Research and Development*, 11(6):601–617, 1967. 64
- [72] NORBERT WIENER. *The human use of human beings: Cybernetics and society.* Number 320. Da Capo Press, 1988. 64
- [73] VINCENT C MÜLLER. **Ethics of artificial intelligence and robotics.** 2020. 65
- [74] JOSEPH WEIZENBAUM. **On the impact of the computer on society: How does one insult a machine?** *Science*, 176(4035):609–614, 1972. 66
- [75] RONALD E ANDERSON. **ACM code of ethics and professional conduct.** *Communications of the ACM*, 35(5):94–99, 1992. 66
- [76] TIAN LI, ANIT K. SAHU, MANZIL ZAHEER, MAZIAR SANJABI, AMEET TALWALKAR, AND VIRGINIA SMITH. **Federated optimization in heterogeneous networks.** *Proceedings of Machine Learning and Systems*, 2:429–450, 2020. 69
- [77] TAO SUN, DONGSHENG LI, AND BAO WANG. **Decentralized federated averaging.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 70
- [78] SHIQI LI, QI QI, JINGYU WANG, HAIFENG SUN, YUJIAN LI, AND F RICHARD YU. **GGS: General gradient sparsification for federated learning in edge computing.** In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2020. 70
- [79] HAIJIAN SUN, XIANG MA, AND ROSE QINGYANG HU. **Adaptive federated learning with gradient compression in uplink NOMA.** *IEEE Transactions on Vehicular Technology*, 69(12):16325–16329, 2020. 70
- [80] SARIT KHIRIRAT, SINDRI MAGNÚSSON, ARDA AYTEKIN, AND MIKAEL JOHANSSON. **A flexible framework for communication-efficient machine learning.** In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, pages 8101–8109, 2021. 70
- [81] FRANK SEIDE, HAO FU, JASHA DROPPA, GANG LI, AND DONG YU. **1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs.** In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 70
- [82] HANLIN TANG, SHAODUO GAN, CE ZHANG, TONG ZHANG, AND JI LIU. **Communication compression for decentralized training.** *Advances in Neural Information Processing Systems*, 31, 2018. 70
- [83] PETER RICHTÁRIK, IGOR SOKOLOV, AND ILYAS FATKHULLIN. **EF21: A new, simpler, theoretically better, and practically faster error feedback.** *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021. 70

- [84] JIANGFENG XIAN, JUNLING MA, XIAOJUN MEI, HUAFENG WU, NASIR SAEED, DEZHI HAN, MARIO DONATO MARINO, AND KUAN-CHING LI. **Robust coarse-to-fine 3D-target-localization algorithm for underwater-IoT-based networks: Design and performance evaluation under uncertain multi-parameters.** *IEEE Internet of Things Journal*, 2025. 81
- [85] PETER AUER, NICOLO CESA-BIANCHI, AND PAUL FISCHER. **Finite-time analysis of the multiarmed bandit problem.** *Machine Learning*, 47(2-3):235–256, 2002. 81
- [86] YANAN SUI, ATHANASIOS GOTOVOS, JOEL BURDICK, AND ANDREAS KRAUSE. **Safe exploration for optimization with Gaussian processes.** In *International Conference on Machine Learning*, pages 997–1005. PMLR, 2015. 81
- [87] AHMADREZA MORADIPARI, CHRISTOS THRAMPOULIDIS, AND MAHNOOSH ALIZADEH. **Stage-wise conservative linear bandits.** *Advances in Neural Information Processing Systems*, 33:11191–11201, 2020. 81
- [88] MADALINA M DRUGAN AND ANN NOWÉ. **Designing multi-objective multi-armed bandits algorithms: A study.** In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013. 81
- [89] SHANGSHANG WANG, SIMENG BIAN, XIN LIU, AND ZIYU SHAO. **Neural constrained combinatorial bandits.** *IEEE Transactions on Networking*, 2025. 81
- [90] TIANYI CHEN AND GEORGIOS B GIANNAKIS. **Bandit convex optimization for scalable and dynamic IoT management.** *IEEE Internet of Things Journal*, 6(1):1276–1286, 2018. 81
- [91] QINGSONG LIU AND ZHIXUAN FANG. **Learning to schedule tasks with deadline and throughput constraints.** In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pages 1–10, 2023. 81
- [92] MICHAEL J NEELY AND HAO YU. **Online convex optimization with time-varying constraints.** *arXiv preprint arXiv:1702.04783*, 2017. 81
- [93] XUANYU CAO AND KJ RAY LIU. **On the time-varying constraints and bandit feedback of online convex optimization.** In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2018.
- [94] XUANYU CAO AND KJ RAY LIU. **Online convex optimization with time-varying constraints and bandit feedback.** *IEEE Transactions on Automatic Control*, 64(7):2665–2680, 2018. 81
- [95] RUDYAR CORTÉS, XAVIER B., OLIVIER M., AND PIERRE S. **Stream processing of health-care sensor data: Studying traces to identify challenges from big data perspective.** *Procedia Computer Science*, 52:1004–1009, 2015. 82
- [96] SHUSEN YANG. **IoT stream processing and analytics in the fog.** *IEEE Communications Magazine*, 55(8):21–27, 2017. 82, 90
- [97] GAYASHAN AMARASINGHE, MARCOS D. DE ASSUNCAO, AARON H., AND SHANIKA K. **A data stream processing optimisation framework for edge computing applications.** In *2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC)*, pages 91–98. IEEE, 2018. 82
- [98] RUSTEM DAUTOV AND SALVATORE DISTEFANO. **Stream processing on clustered edge devices.** *IEEE Transactions on Cloud Computing*, 2020. 82
- [99] MENG-HSI CHEN, BEN LIANG, AND MIN DONG. **Joint offloading and resource allocation for computation and communication in mobile cloud with computing access point.** In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE, 2017. 82

- [100] GUIFU MA, HAORAN LI, XIAOWEI WANG, XIAOLONG CHEN, YOUANG BIAN, MANJIANG HU, XUEPENG WANG, AND JIN ZHANG. **Mobility-aware task splitting and computation resource allocation for distributed multi-access edge computing enabled vehicular network.** In *2021 International Conference on Mechanical, Aerospace and Automotive Engineering*, pages 164–170, 2021.
- [101] ZHIQING TANG, XIAOJIE ZHOU, FUMING ZHANG, WEIJIA J, AND WEI Z. **Migration modeling & learning algorithms for containers in fog computing.** *IEEE Transactions on Services Computing*, 12(5):712–725, 2018. 82
- [102] K JIANG, HUAN ZHOU, DAWEI LI, XUXUN LIU, AND SHOUZHI XU. **A q-learning based method for energy-efficient computation offloading in edge computing.** In *2020 29th International Conference on Computer Communications and Networks*, pages 1–7. IEEE, 2020. 82
- [103] HUAN ZHOU, KAI JIANG, XUXUN LIU, XIUHUA LI, AND VICTOR CM LEUNG. **Deep reinforcement learning for energy-efficient offloading in edge computing.** *IEEE Internet of Things Journal*, 9(2):1517–1530, 2021. 82
- [104] DIANNE SV MEDEIROS, HELIO N CUNHA N, MARTIN A LOPEZ, LUIZ CLAUDIO S MAGALHÃES, NATALIA C FERNANDES, ALEX B VIEIRA, EDELBERTO F SILVA, AND DIOGO M F MATTOS. **A survey on data analysis on large-scale wireless networks: Online stream processing, and challenges.** *Journal of Internet Services and Applications*, 11(1):1–48, 2020. 82
- [105] BRIAN KIM, JUSTIN H KONG, TERRENCE J MOORE, AND FIKADU T DAGEFU. **Deep reinforcement learning based routing for heterogeneous multi-hop wireless networks.** *arXiv preprint arXiv:2508.14884*, 2025. 92
- [106] NEHA SHARMA, VENKATA SAAI PRANEETH THOTA, YUVARAJ TANKALA, SHRADDHA TRIPATHI, AND OM JEE PANDEY. **OptRISQL: Toward performance improvement of time-varying IoT networks using Q-learning.** *IEEE Transactions on Network and Service Management*, 21(3):3008–3020, 2024.
- [107] YIJIE WANG, ZHIYUAN QU, ZHONGLIANG ZHAO, XIANBIN CAO, YANG LIU, AND TONY QS QUEK. **EMOR: Energy-efficient mixture opportunistic routing based on reinforcement learning for lunar surface ad-hoc networks.** *IEEE Transactions on Communications*, 2024.
- [108] SIJIN YANG, LEI ZHUANG, JIANHUI ZHANG, JULONG LAN, AND BINGKUI LI. **A multipolicy deep reinforcement learning approach for multiobjective joint routing and scheduling in deterministic networks.** *IEEE Internet of Things Journal*, 11(10):17402–17418, 2024.
- [109] QIANG HE, YU WANG, XINGWEI WANG, WEIQIANG XU, FULIANG LI, KAIQI YANG, AND LIANBO MA. **Routing optimization with deep reinforcement learning in knowledge defined networking.** *IEEE Transactions on Mobile Computing*, 23(2):1444–1455, 2024.
- [110] MENGQIN WANG, YANLING WEI, XUELIANG HUANG, AND SHAN GAO. **An end-to-end deep reinforcement learning framework for electric vehicle routing problem.** *IEEE Internet of Things Journal*, 11(20):33671–33682, 2024.
- [111] ISHITA CHAKRABORTY, PRODIPTO DAS, AND BUDDHADEB PRADHAN. **An intelligent routing for Internet of Things mesh networks.** *Transactions on Emerging Telecommunications Technologies*, 34(11):e4628, 2023.
- [112] D PRABHU, R ALAGESWARAN, AND S MIRUNA JOE AMALI. **Multiple agent based reinforcement learning for energy efficient routing in WSN.** *Wireless Networks*, 29(4):1787–1797, 2023.
- [113] LIU YANG, YIFEI WEI, F. RICHARD YU, AND ZHU HAN. **Joint routing and scheduling optimization in time-sensitive networks using graph-convolutional-network-based deep reinforcement learning.** *IEEE Internet of Things Journal*, 9(23):23981–23994, 2022.

- [114] CHENYI LIU, MINGWEI XU, YUAN YANG, AND NAN GENG. **DRL-OR: Deep reinforcement learning-based online routing for multi-type service requirements.** In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021.
- [115] HOSSAM FARAG AND CEDOMIR STEFANOVIĆ. **Congestion-aware routing in dynamic IoT networks: A reinforcement learning approach.** In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2021.
- [116] XUANCHENG GUO, HUI LIN, ZHIYANG LI, AND MIN PENG. **Deep-reinforcement-learning-based QoS-aware secure routing for SDN-IoT.** *IEEE Internet of Things Journal*, 7(7):6242–6251, 2020. 92
- [117] MOHAMED SAID FRIKHA, SONIA METTALI GAMMAR, ABDELKADER LAHMADI, AND LAURENT ANDREY. **Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey.** *Computer Communications*, 178:98–113, 2021. 92
- [118] ALI FORGHANI ELAH ABADI, SEYYED AMIR ASGHARI, MOHAMMADREZA BINESH MARVASTI, GOLNOUSH ABAEI, MORTEZA NABAVI, AND YVON SAVARIA. **RLBEEP: Reinforcement-learning-based energy efficient control and routing protocol for wireless sensor networks.** *IEEE Access*, 10:44123–44135, 2022. 96
- [119] SALAH EDDINE BOUZID, YOUSSEF SERRESTOU, KOSAI RAOOF, AND MOHAMED NAZIH OMRI. **Efficient routing protocol for wireless sensor network based on reinforcement learning.** In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–5. IEEE, 2020. 96
- [120] WENJING GUO, CAIRONG YAN, AND TING LU. **Optimizing the lifetime of wireless sensor networks via reinforcement-learning-based routing.** *International Journal of Distributed Sensor Networks*, 15(2):1550147719833541, 2019. 96
- [121] MAHMOOD R. MINHAS, SATHISH GOPALAKRISHNAN, AND VICTOR C.M. LEUNG. **Multi-objective routing for simultaneously optimizing system lifetime and source-to-sink delay in wireless sensor networks.** In *IEEE International Conference on Distributed Computing Systems Workshops*, pages 123–129, 2009. 96
- [122] JUAN COTA-RUIZ, PABLO RIVAS-PEREA, ERNESTO SIFUENTES, AND RAFAEL GONZALEZ-LANDAETA. **A recursive shortest path routing algorithm with application for wireless sensor network localization.** *IEEE Sensors Journal*, 16(11):4631–4637, 2016. 96
- [123] YUANDOU WANG, HANG LIU, WANBO ZHENG, YUNNI XIA, YAWEN LI, PENG CHEN, KUNYIN GUO, AND HONG XIE. **Multi-objective workflow scheduling with deep-Q-network-based multi-agent reinforcement learning.** *IEEE Access*, 7:39974–39982, 2019. 100
- [124] TAO REN, JIANWEI NIU, JIAHE CUI, ZHENCHAO OUYANG, AND XUEFENG LIU. **An application of multi-objective reinforcement learning for efficient model-free control of canals deployed with IoT networks.** *Journal of Network and Computer Applications*, 182:103049, 2021. 100
- [125] BOONHATAI KRUEKAEW AND WARANGKHANA KIMPAN. **Multi-objective task scheduling optimization for load balancing in cloud computing environment using hybrid artificial bee colony algorithm with reinforcement learning.** *IEEE Access*, 10:17803–17818, 2022. 100
- [126] LUCA CAVIGLIONE, MAURO GAGGERO, MASSIMO PAOLUCCI, AND ROBERTO RONCO. **Deep reinforcement learning for multi-objective placement of virtual machines in cloud datacenters.** *Soft Computing*, 25(19):12569–12588, 2021. 100
- [127] AREZOO GHASEMI AND ABOLFAZL TOROGHI HAGHIGHAT. **A multi-objective load balancing algorithm for virtual machine placement in cloud data centers based on machine learning.** *Computing*, 102:2049–2072, 2020. 100

- [128] YAPING CUI, LIJUAN DU, HONGGANG WANG, DAPENG WU, AND RUYAN WANG. **Reinforcement learning for joint optimization of communication and computation in vehicular networks.** *IEEE Transactions on Vehicular Technology*, **70**(12):13062–13072, 2021. [100](#)
- [129] ZHIPING PENG, JIANPENG LIN, DELONG CUI, QIRUI LI, AND JIEGUANG HE. **A multi-objective trade-off framework for cloud resource scheduling based on the deep Q-network algorithm.** *Cluster Computing*, **23**:2753–2767, 2020. [100](#)
- [130] YU YU, JIE TANG, JIAYI HUANG, XIUYIN ZHANG, DANIEL KA CHUN SO, AND KAI-KIT WONG. **Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm.** *IEEE Transactions on Communications*, **69**(9):6361–6374, 2021. [100](#)
- [131] GAGANDEEP KAUR, PRASENJIT CHANAK, AND MAHUA BHATTACHARYA. **Energy-efficient intelligent routing scheme for IoT-enabled WSNs.** *IEEE Internet of Things Journal*, **8**(14):11440–11449, 2021. [100](#)