PhD-FSTM-2023-042

The Faculty of Science, Technology and Medicine

DISSERTATION

Presented on 12/05/2023 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITE DU LUXEMBOURG
EN INFORMATIQUE

by

Arsham Mostaani
Born on 20 May 1987 in Esfahan, Iran

INDIRECT TASK-ORIENTED COMMUNICATION DESIGN FOR CONTROL AND DECISION MAKING IN MULTI-AGENT SYSTEMS

Dissertation defense committee

Dr. Björn Ottersten, Dissertation Advisor
Professor and Director of SnT, University of Luxembourg
Dr. Bhavani Shankar, Chairman
Professor and Head of SPARC, SnT, University of Luxembourg
Dr. Symeon Chatzinotas, Vice Chairman
Professor and Head of SIGCOM, SnT, University of Luxembourg
Dr. Jakob Hoydis, Member
Principal Research Sceintist, Nvidia, France
Dr. Samson Lasaulce, Member
CNRS Director of Research, Université de Lorraine, France

Indirect Task-oriented Communication Design for Control and Decision Making in Multi-Agent Systems

Arsham Mostaani

Abstract

As 5G is rolling out, we commence witnessing a surge of data-hungry applications in various domains such as IoT, industry 4.0, and autonomous vehicles. In contrast to the previous generations of cellular networks, 5G will serve many services in which a (subsystem of) (an intelligent) machine is the receiving end of the communications. As soon as the receiving end of communications is no longer human, the ultimate goal of data transmission deviates from the traditional communication and data transmission systems. Under these circumstances, the communication is carried out only to address information deficiency at the receiving end. In particular, the receiving machine has a deficiency of information when the computations that it intends to carry out require further data than what is available to it. Communications are, thus, carried out to deliver the relevant/useful data required to perform the desired computations at the receiving end. This scenario demands a fresh approach for the design of the data transmission schemes since only the data that helps improve the computations of the receiving end are required to be transmitted.

Many of the services provided by cellular networks are traditionally designed to serve their primary users - humans. In contrast to humans, machines do not appreciate the extra descent quality of the received data/communications. The shift taking place in the number of non-human users, now, asks for revolutionary designs at all subsystems of a complete communication pipeline, where the relevance of data is taken into account when designing the specific subsystem. The relevance/usefulness of data can for instance change the way channel coding schemes behave by allowing the channel coder to know which data is more important to be protected. The relevance/usefulness of data can help the data compression schemes perform much more effectively, by discarding the part of data that will be useless in the computations of the receiving end. The relevance/usefulness can also help redesign the power/user scheduling schemes by giving priority to the users who are sharing more useful/relevant data.

The importance of massive research effort required to address these challenges becomes even more pronounced when we notice that by 2030, thirty billion machines will be served by communication networks. Task-oriented communication is an emerging field often overlapping with control theory, estimation theory, communication theory and machine learning, whose mission is to perform this fresh design at all layers of communication systems.

The focus of this thesis is on the task-oriented design of data compression/quantization methods. In particular, we limit ourselves to the design of data quantization algorithms for the control tasks and thus the task-oriented design of quantization for estimation tasks is out of the scope of this thesis. A very wide range of control tasks is classified under the control of multi-agent systems, where the current thesis finds its context. In multi-agent systems that operate under partial observability, inter-agent communications can prove as an essential tool to improve the overall performance of the system. We study different data compression schemes for communications between agents under different topologies of communication networks between agents. We also introduce schemes that have different capacities to scale with the number of agents in the system.

In particular, in chapter 3, we perform an indirect design of the communications in a multi-agent system (MAS) in which agents cooperate to maximize the averaged sum of discounted one-stage rewards of a collaborative task. Due to the bit-budgeted communications between the agents, each agent should efficiently represent its local observation and communicate an abstracted version of the observations to improve the collaborative task performance. We first show that this problem can be approximated as a form of data-quantization problem which we call task-oriented data compression (TODC). We then introduce the state-aggregation for information compression algorithm (SAIC) to solve the formulated TODC problem. It is shown that SAIC is able to achieve near-optimal performance in terms of the achieved sum of discounted rewards. The proposed algorithm is applied to a geometric consensus problem and its performance is compared with several benchmarks. Numerical experiments confirm the promise of this indirect design approach for task-oriented multi-agent communications.

Subsequently, in chapter 4, we consider a task-effective quantization problem that arises when multiple agents are controlled via a centralized controller (CC). While agents have to communicate their observations to the CC for decision-making, the bit-budgeted communications of agent-CC links may limit the overall performance of the system which is

measured by the system's average sum of stage costs/rewards. As a result, each agent should compress/quantize its observation such that the average sum of stage costs/rewards of the control task is minimally impacted. We address the problem of maximizing the average sum of stage rewards by proposing two different Action-Based State Aggregation (ABSA) algorithms that carry out the indirect and joint design of control and communication policies in the multi-agent system. While the applicability of ABSA-1 is limited to single-agent systems, it provides an analytical framework that acts as a stepping stone to the design of ABSA-2. ABSA-2 carries out the joint design of control and communication for a multi-agent system. We evaluate the algorithms - with average return as the performance metric - using numerical experiments performed to solve a multi-agent geometric consensus problem. The numerical results are concluded by introducing a new metric that measures the effectiveness of communications in a multi-agent system.

In our last technical chapter 5, we present a novel approach for designing scalable task-oriented quantization and communications in cooperative multi-agent systems (MAS). The proposed approach utilizes a task-oriented communication framework to enable efficient communication of observations between agents while optimizing the average return performance of the MAS, a parameter that quantifies the fulfilment of MAS's task. Our approach uses the concept of the value of information to design quantization schemes that scale with the number of agents in the system. The designed quantization scheme enables agents to communicate task-relevant observations while minimizing the number of bits to be communicated. Computing the value of information, however, does not scale with the increasing number of agents in the MAS. We observe that one can reduce the computational cost of obtaining the value of information by exploiting insights gained from studying a similar two-agent system - instead of the original N-agent system. We then quantize agents' observations such that their more valuable observations are communicated more precisely. We show analytically that under a wide range of problems, the proposed scheme is applicable. Our numerical results show that the proposed approach achieves significant improvements in reducing the computational complexity of the centralized training phase for the design of inter-agent communications in MAS problems while maintaining the average return performance of the system.

Acknowledgements

It is with immense gratitude that I express my acknowledgments for the completion of my doctoral thesis. The journey of earning a doctoral degree has been both intellectually and emotionally demanding, and I could not have accomplished it without the unwavering support of my loved ones.

First and foremost, I would like to express my heartfelt appreciation to my parents and my family for their constant love, guidance, and encouragement. Their unwavering support throughout my life has provided me with the opportunity and the right mindset for growth. Their endless sacrifices and tireless efforts have made it possible for me to pursue higher education, and I am forever grateful for their support.

I am also deeply grateful to my wife, who has been my rock throughout the challenging days of my doctoral journey. Her unwavering support, understanding, and love have helped me to navigate through the uncertainties and difficulties that come with a PhD degree. I am grateful for her patience, understanding, and unconditional love.

I would like to extend my sincere appreciation to my supervisors, whose trust in me and my work gave me enough freedom to pursue my research desires. Their guidance, mentorship, and constructive feedback have been invaluable throughout my doctoral journey. I am grateful for the opportunities that they provided me with to learn, grow, and develop my skills. The breadth and depth of knowledge possessed by my supervisors were like a guiding light in the darkness of my research journey. Their expertise, wisdom, and ability to ask thought-provoking questions helped me to refine my ideas, sharpen my analytical skills, and stay focused on my research goals. I am grateful for the opportunity to learn from them, and their contributions have been instrumental in shaping my academic and intellectual development.

Finally, I would like to express my special thanks to Dr. Shahbazpanahi, who played

a pivotal role in the first year of my PhD as well is in the journey of self-discovery. His unwavering support, insightful guidance, and constructive feedback have helped me to find sufficient resources within myself to continue with the rough journey of a PhD.

In conclusion, I would like to thank everyone who has contributed to my academic and personal growth, and whose support has made it possible for me to achieve my academic goals. My gratitude extends to everyone who has provided me with their time, resources, and wisdom throughout my doctoral journey.

Preface

This PhD thesis has been carried out from Dec. 2019 to October 2022, at the SIGCOM research group, SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, under the supervision of Prof. Bjorn Ottersten, Prof. Symeon Chatzinotas and Dr. Thang X. Vu. The time-to-time evaluation of the PhD thesis was duly performed by the CET members constituting the supervisors and co-supervisors.

Contents

This PhD thesis entitled "Indirect Task-Oriented Communication Design for Control and Decision Making in Multi-Agent Systems" is divided into six chapters. In Chapter 1, the background, motivation, a non-formal problem statement and an overview of the thesis structure is provided. Chapter 2 surveys the potential theories that can be used to solve task-oriented communication problems. It also introduces some of the applications of task-oriented communications at a high level. Next, the design of distributed task-oriented communications and distributed control policies for a multi-agent system is carried out at 3. Chapter 4 provides the design of task-oriented communications at a slightly different setting where the control policy is designed for a centralized setting. Chapter 5 Introduces a scalable approach to designing task-oriented communications for multi-agent systems comprised of a higher amount of agents. Finally, Chapter 6 provides further detailed application scenarios for task-oriented communications, for which market demand is already detected. This section, eventually, concludes with some remarks and future work.

Support of the Thesis

This work is supported by European Research Council (ERC) advanced grant 2022 (Grant agreement ID: 742648).

Declaration

Except where acknowledged in a customary manner, the material presented in this thesis is, to the best of the authors' knowledge, original and has not been submitted in whole or part for a degree in any university.

Publication List

Journals

[I] A. Mostaani, T. X. Vu, S. Chatzinotas and B. Ottersten, "Task-Oriented Data Compression for Multi-Agent Communications Over Bit-Budgeted Channels," in IEEE Open Journal of the Communications Society, vol. 3, pp. 1867-1886, 2022, doi: 10.1109/OJCOMS.2022.3213213.

[II] A. Mostaani, T. X. Vu, S. K. Sharma, V. -D. Nguyen, Q. Liao and S. Chatzinotas, "Task-Oriented Communication Design in Cyber-Physical Systems: A Survey on Theory and Applications," in IEEE Access, vol. 10, pp. 133842-133868, 2022, doi: 10.1109/ACCESS.2022.3231039.

[III] A. Mostaani, T. H. Vu, S. Chatzinotas and B. Ottersten, "Task-Effective Compression of Observations for the Centralized Control of a Multi-agent System Over Bit-Budgeted Channels," *IEEE IoT Journal* - under revision.

[IV] A. Mostaani, T. H. Vu, S. Chatzinotas and B. Ottersten, "On Understanding the Value of Observations: Task-Oriented Communication Design at Scale," *IEEE Transactions on Communications* - submitted.

Conference/Workshop Papers

[V] A. Mostaani, O. Simeone, S. Chatzinotas and B. Ottersten, "Learning-based Physical Layer Communications for Multiagent Collaboration," 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, 2019, pp. 1-6, doi: 10.1109/PIMRC.2019.8904190.

[VI] A. Mostaani, T. X. Vu, S. Chatzinotas and B. Ottersten, "State Aggregation for Multiagent Communication over Rate-Limited Channels," *2020 IEEE Global Communications Conference (GLOBECOM),* Taipei, Taiwan, 2020, pp. 1-7, doi: 10.1109/GLOBECOM42002.2020.9322138.

[VII] A. Mostaani, T. H. Vu, S. Chatzinotas and B. Ottersten, "Centralized Control of a Multi-Agent System Via Distributed and Bit-Budgeted Communications," *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, Glasgow, UK, 2023, pp. 24-29. (presented but not published yet)

[VIII] A. Mostaani, T. H. Vu, H. Habibi, S. Chatzinotas and B. Ottersten, "Learning multi-agent communications: a scalable approach" *2023 IEEE Global Communications Conference (GLOBECOM),* (Submitted).

Contents

List of Figures

List of Tables

Abbreviations

5G	Fifth Generation of Mobile Communications
6G	Sixth Generation of Mobile Communications
ABSA	Action Based State Aggregation
ADCs	Analog to Digital Converters
CC	Central Controller
CMF	Conditional Mass Function
CPS	Cyber-Physical System(s)
CIC	Conventional Information Compression
DL	Deep Learning
DACs	Digital to Analog Converters
Dec-POMDP	Decentralized Partially Observable Markov Decision Process
DRL	Deep Reinforcement Learning
ESAIC	Extended State Aggregation for Information Compression
GDPR	General Data Protection Regulation
IoT	Internet of Things
i.i.d.	Independent and identically distributed
IIoT	Industrial Internet of Things
JCCD	Joint Communications and Control Design
KPI	Key Performance Indicator
LBIC	Learning Based Information Compression
MAP	Maximum a Posteriori
MEC	Multi-access edge computing
MAP	Markov Decision Process
MMSE	Minimum Mean Square Error
MARL	Multi-Agent Reinforcement Learning

MAS	Multi-Agent System
MMDP/ MAMDP	Multi-Agent Markov Decision Process
POMDP	Partially Observable Markov Decision Process
RL	Reinforcement Learning
r.v.	Random variable
SAIC	State Aggregation for Information Compression
SON	Self-Organizing Networks
TRL	Technology Readiness Level
TODC	Task-Oriented Data Compression
UAV	Unmanned Aerial Vehicle
URLLC	Ultra Reliable Low Latency Communications
V2I	Vehicle to Infrastructure
V2V	Vehicle to Vehicle
V2X	Vehicle to everything

Notations

$\mathbf{x}(t)$	A generic random variable generated at time t
$\mathsf{x}(t)$	Realization of $\mathbf{x}(t)$
$\mathcal{X}$	Alphabet of $\mathbf{x}$(t)
$\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
$\mathbb{P}(\mathcal{X})$	Power set of $\mathcal{X}$
$p_{\mathbf{x}}(\mathsf{x}(t))$	Shorthand for $\Pr(\mathbf{x}(t) = \mathsf{x}(t))$
$H(\mathbf{x}(t))$	Information entropy of $\mathbf{x}(t)$ (bits)
$I(\mathbf{x}(t), \mathbf{y}(t))$	Mutual information between the r.v. $\mathbf{x}(t)$ and the r.v. $\mathbf{y}(t)$ (bits)
$\mathcal{X}_{-\mathbf{x}}$	$\mathcal{X} - \{\mathbf{x}\}$
$\mathbb{E}_{p(\mathsf{x})}\{\mathbf{x}\}$	Expectation of the random variable X over the probability distribution $p(\mathsf{x})$
$\delta(\cdot)$	Dirac delta function
$\mathsf{tr}(t)$	Realization of the system's trajectory at time t

Chapter 1

Task-Oriented Communication Design: Background and Motivation

The emergence of the Internet of Things (IoT) has led to an explosion in the amount of digitized data being generated and distributed. Characterized by the integration of digital devices with sensors, networks, and software, IoT creates a network of interconnected devices that can communicate with each other. These devices sometimes happen to have the agency to take actions/make decisions turning them to agents and the whole IoT network into a cyber-physical system (CPS)[1]. The data generated by these devices/agents can be used to improve the efficiency and effectiveness of decision-making in different sectors of various industries, including manufacturing, healthcare, transportation and, agriculture. When an industry can benefit from the wealth of data generated by millions of smart devices and agents, the industry is, indeed, transformed into a huge CPS that can make more informed decisions in its different subsystems given online streams of data.

Data is at the heart of a CPS. The ability to collect, communicate data and to compute in real-time enables intelligent data-driven decision-making and control. The data generated by different subsystems of CPSs can provide insights into everything from gaining awareness in autonomous driving systems/agents to informed agricultural actions made by smart

[1] A cyber-physical system encompasses broader systems that integrate physical processes with computing and communication capabilities to achieve specific objectives. IoT can be seen as a subset of CPS, as it focuses specifically on the interconnectedness of devices via the internet.

agents who are aware of the environmental conditions via the transmission of data inside an agricultural CPS. With this data, businesses/industries can improve their processes, reduce costs, and create new products and services.

However, managing/communicating/analysing the massive amounts of data generated by IoT and CPS is a major challenge. The volume and velocity of data generated by these systems can overwhelm traditional communication, data processing, and storage techniques. Moreover, processing this data with sufficient agility while maintaining the performance of the CPS/IoT in their computational task is a critical issue. To overcome these challenges, new data processing and communication technologies must be developed.

1.1 Background

With computing and learning power becoming more pervasive than ever, the need for more data becomes pronounced too. In some scenarios, the computing device has direct access to the data it requires for its computations, in other scenarios the data is accessible only through a communication channel [1–5]. Nevertheless, the problem of communicating data to address the data deficiency at a computing centre is fundamentally different from communicating data for connecting people. As soon as the receiving end of communications is no longer a human, the ultimate goal of data transmission deviates from the traditional communication and data transmission systems. Under these circumstances, the communications are carried out to deliver the relevant/useful data required to perform the desired computations at the receiving end.

In traditional task-agnostic communication systems, the transmitter has never had the agency to decide on the usefulness of the transmitted data. Task-agnostic communication systems are designed by solving what Shanon refers to as the technical communication problem [6]. Given the limitations of the communication channel, the ultimate goal of solving the technical problem is to transmit the original data over a noisy channel with the least possible expected error at the receiving end. Thus, the performance of a communication system could be measured using different distortion metrics such as mean squared error. These distortion metrics allow us to measure the difference between the original data at the transmitter, and the regenerated data at the receiver while there is no particular stress on the importance of

one bit (sequence) over another [7]. Treating bit (sequence)s equally, however, proves to be highly sub-optimal in some cases - this is due to the fact that some bit (sequence)s might convey more valuable information [8].

In modern communications systems, there will be an increasing need for understanding the (semantic/task) value of a bit (sequence) towards the computations that are performed at the receiving end [8–10]. The relevance/value of data can for instance change the way channel coding schemes behave by allowing the channel coder to know which data is more important to be protected [4, 11]. The relevance/value of data can help the data compression schemes perform much more effectively, (by reducing the granularity of)/discarding the part of data that will be (less useful)/useless in the computations of the receiving end [5, 9]. The value can also help redesign the power/user scheduling schemes by giving priority to the users who are sharing more valuable data [12, 13].

Under this modern setting, the transmitter will have the agency to decide on the value/ relevance of the communicated data. The challenge, however, will be how to quantify the value/ relevance of the data in a universal fashion. Note that, in different computational tasks a certain bit (sequence) can have a non-constant value/relevance. Therefore, proposing an *indirect* measure of the value of data that is applicable across different tasks (and not just for a single specific task) is of the essence. The next challenge is how to incorporate the value of data in designing a quantization, data compression, error correction code or a user scheduling algorithm. Task-oriented communication is an emerging field often overlapping with control theory, estimation theory, communication theory and, machine learning, whose mission is to address similar questions leading to fresh designs at all layers of communication systems.

1.2 Motivation

Many of the data services provided by cellular networks are traditionally designed to serve the network's primary users - humans. However, humans are no longer the only users of cellular networks. It is projected that by 2030, approximately 30 billion IoT devices will be connected to cellular networks [14]. The recent explosion in IoT together with other data-driven use cases and their reliance on huge datasets collected by edge devices "have raised

legitimate concerns that the increasing data traffic might soon overwhelm the capacity of current networks despite ongoing efforts to increase their capacity and efficiency" [15]. In contrast to humans, machines do not appreciate the extra descent quality of the received data/communications. Nevertheless, communication networks are primarily designed with the ultimate goal of satisfying their main users - humans so far. The shift taking place in the number of non-human users, now, asks for revolutionary designs at all subsystems of a communication pipeline, where the relevance of data is taken into account in designing the specific subsystem [16].

When in it comes to the introduction of new cellular network generations, authors of [17] believe that some services are first provided by (an odd-numbered) generation mainly for business purposes/users. The generation to follow, then, scales the pre-existing technologies to make them suitable and affordable for mass usage among all consumers. While 5G is the first to offer new technologies for tactile internet [18] and massive machine-type communications, 6G is expected to broaden the application of these technologies by making them available and affordable to all consumers. Task-oriented communications can significantly reduce the cost of tactile services while improving their latency.

Task-oriented communication systems, go beyond the naive forwarding of data samples for processing at the receiver end and attempt to realize the importance of the communicated data for the processing task at the receiver. This approach will help (i) put less burden on the limited radio resources of the network [15], (ii) reduce the size of data being transmitted leading to improved delays for delay-sensitive applications (iii) decrease the complexity of processes to be performed at the receiver via a joint communication and computation design approach [19], (iv) enabling wider use of narrow-band technology standards such as Cat-NB1 (NB-IoT) and LoraWan that support only low throughput applications requiring low power and long-range [20] and, (v) reducing the storage size required at the receiving end to save the data.

1.3 Research problem statement and objectives

According to Shannon and Weaver, communication problems can be divided into three levels [6]: (i) technical problem: given channel and network constraints, how accurately can

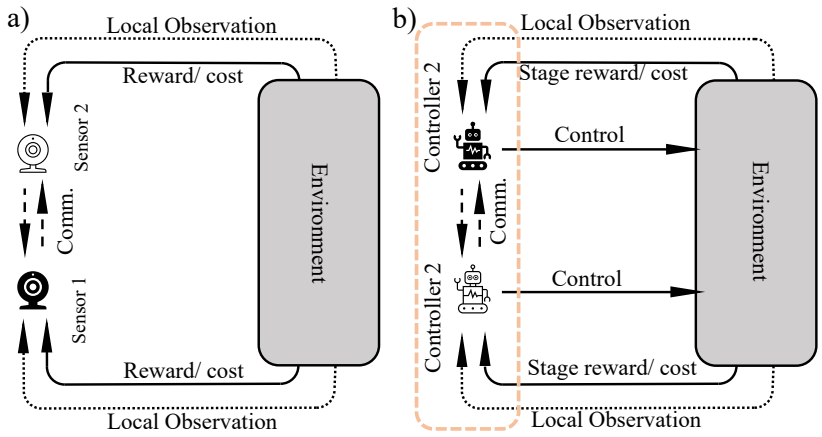


Figure 1.1: Task-effective communications for a) an estimation vs. b) a control task .

the communication symbols/bits be transmitted? (ii) semantic problem: given channel and network constraints, how accurately the communication symbols can deliver the desired meaning? (iii) effectiveness problem: given channel and network constraints, how accurately the communication symbols can help to fulfil the desired task? While the traditional communication design addresses the technical problem, recently, the semantic problem [1, 16, 21–23] as well as the effectiveness problem [4, 9, 11, 24–29] have attracted extensive research interest.

In contrast to Shannon's technical-level communication framework, semantic communication can enhance performance by exploiting prior knowledge between source and destination [30, 31]. The semantic-based designs, however, are not necessarily task-effective [32]. One can design transmitters which compress the data with the least possible compromise on the semantic meaning being transmitted [1, 21] while the transmission can be task-unaware [33]. In contrast to semantic level and technical level communication design, the performance of a task-effective communication system is ultimately measured in terms of the average return/cost linked to the task [11]. In the (task-)effectiveness problem, we are not concerned only about the communication of meaning but also about how the message exchange is helping the receiving end to improve its performance in the expected cost/reward of an estimation task [26, 27, 29, 31, 34] or a control task [4, 9, 11, 13, 25, 27, 35].

There are fundamental differences between the design of task-effective communications for an estimation vs. a control task - Fig. 1.1. (i) In the latter, each agent can produce a control signal that directly affects the next observations of the agent. Thus, in control tasks the source of information - local observations of the agent - is often a stochastic process with memory - e.g. linear or Markov decision processes - [4, 9, 11]. In the estimation tasks, however,

the source of information is often assumed to be an i.i.d. stochastic process [26, 29, 34]. (ii) In the control tasks, a control signal often has a long-lasting effect on the state of the system more than for a single stage/time step e.g., a control action can result in lower expected rewards in the short run but higher expected rewards in the long run. This makes the control tasks intrinsically sensitive to the time horizon for which the control policies are designed. Estimation tasks, specifically when the observation process is i.i.d., can be solved in a single stage/ time step - since there is no influence from the solution of one stage/ time step to another i.e., each time step can be solved separately [34, 36]. (iii) The cost function for estimation tasks is often in the form of a difference/distortion function while in the control tasks, it can take on many other forms.

In this thesis, we focus on the effectiveness problem for the control tasks. In particular, we investigate the distributed communication and control design of a multiagent system (MAS) with the ultimate goal of maximizing the expected summation of per-stage rewards also known as the expected return. By nature, this is a joint communication and control design problem in nature. To better understand the joint design nature of the problem, note that, according to [37], when communications incur no cost on the objective of the system, the optimal communication strategy is to transmit all the data available at the transmitter to the receiving end - no priority is given to a certain bit (sequence). The controllers (of the agents in the MAS) are, subsequently, designed to maximize the expected return under full observability. Whenever we deviate from transmitting all the data available at the transmitter to the receiver, which is the basis of the task-effectiveness problem, we have to foresee what implications this might have on the design of the controller at the receiving end as well as on the expected return of the MAS. This leaves us with the design of a controller at the receiving end too; a controller that can achieve (near) optimal expected return performance while having access to only a part of the observations made by the transmitting end(s).

As mentioned earlier, task-oriented design of communications can provide us with fresh and revolutionary results at different subsystems of a communication pipeline 1.1. The focus of this thesis, however, is on task-oriented data compression schemes for multi-agent systems. In particular, multiple agents select control actions and communicate in the MAS to accomplish a collaborative task with or without the help of a central controller (CC). Accordingly, one potential topology for the communication network of the agents - thoroughly studied in chapters 3 and 5 - is the full mesh topology. Under this topology, a decentralized joint design

of communications and control is carried out between every pair of agents. In this scenario, the design of communications and control are carried out in a decentralized fashion. The communication network topology of the MAS that is studied in this thesis is star topology - 4. In the star topology, the hub node is the central controller and the peripheral nodes are the agents - Fig. 5.1. Under this scenario, the control policy is designed in a centralized fashion while the data compression carried out on the observations of the peripheral nodes is carried out distributively.

1.4 Overview of the thesis structure

The focus of this thesis is on the task-oriented design of data compression/quantization methods. In particular, we limit ourselves to the design of data quantization algorithms for the control tasks and thus the task-oriented design of quantization for estimation tasks is out of the scope of this thesis. While usually, task-oriented communications is a more general term, within the context of this thesis, whenever referring to the algorithms developed here, task-oriented communications and task-oriented quantization are used interchangeably.

A very wide range of control tasks are classified under the control of multi-agent systems, where the current thesis finds its context. In multi-agent systems that operate under partial observability, inter-agent communications can prove as an essential tool to improve the overall performance of the system. We study different data compression schemes for communications between agents under different topologies of communication networks between agents. We also introduce schemes that have different capacities to scale with the number of agents in the system.

Fig. 1.2, provides an outline of this manuscript. By distinguishing the different features of the problems solved in each of the technical chapters, this figure directs readers to the problems of their interest. Further details about the arrangement of the thesis and the specific mission of each chapter are provided as follows. In chapter 3, we perform an indirect design of the communications in a multi-agent system (MAS) in which agents cooperate to maximize the averaged sum of discounted one-stage rewards of a collaborative task. Due to the bit-budgeted communications between the agents, each agent should efficiently represent its local observation and communicate an abstracted version of the observations to improve

the collaborative task performance. We first show that this problem can be approximated as a form of data-quantization problem which we call task-oriented data compression (TODC). We then introduce the state-aggregation for information compression algorithm (SAIC) to solve the formulated TODC problem. It is shown that SAIC is able to achieve near-optimal performance in terms of the achieved sum of discounted rewards. The proposed algorithm is applied to a geometric consensus problem and its performance is compared with several benchmarks. Numerical experiments confirm the promise of this indirect design approach for task-oriented multi-agent communications.

Subsequently, in chapter 4, we consider a task-effective quantization problem that arises when multiple agents are controlled via a centralized controller (CC). While agents have to communicate their observations to the CC for decision-making, the bit-budgeted communications of agent-CC links may limit the overall performance of the system which is measured by the system's average sum of stage costs/rewards. As a result, each agent should compress/quantize its observation such that the average sum of stage costs/rewards of the control task is minimally impacted. We address the problem of maximizing the average sum of stage rewards by proposing two different Action-Based State Aggregation (ABSA) algorithms that carry out the indirect and joint design of control and communication policies in the multi-agent system. While the applicability of ABSA-1 is limited to single-agent systems, it provides an analytical framework that acts as a stepping stone to the design of ABSA-2. ABSA-2 carries out the joint design of control and communication for a multi-agent system.

We evaluate the algorithms - with average return as the performance metric - using numerical experiments performed to solve a multi-agent geometric consensus problem. The numerical results are concluded by introducing a new metric that measures the effectiveness of communications in a multi-agent system.

In our last technical chapter 5, we present a novel approach for designing scalable task-oriented quantization and communications in cooperative multi-agent systems (MAS). The proposed approach utilizes a task-oriented communication framework to enable efficient communication of observations between agents while optimizing the average return performance of the MAS, a parameter that quantifies the fulfilment of MAS's task. Our approach uses the concept of the value of information to design quantization schemes that scale with the number of agents in the system. The designed quantization scheme enables agents to communicate

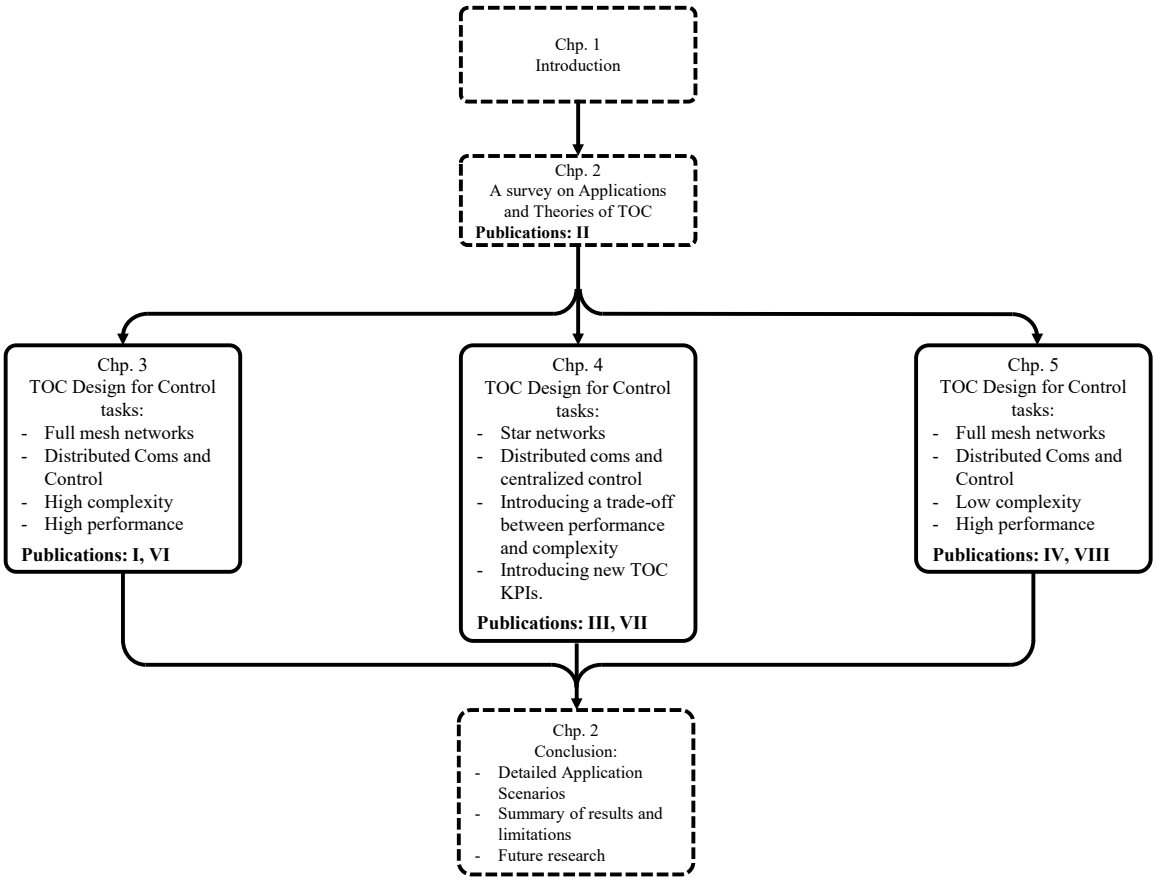


Figure 1.2: Visual arrangement of the thesis is provided in this figure. Each chapter of the thesis is also mapped to the corresponding publication(s) listed in the preface.

task-relevant observations while minimizing the number of bits to be communicated. Computing the value of information, however, does not scale with the increasing number of agents in the MAS. We observe that one can reduce the computational cost of obtaining the value of information by exploiting insights gained from studying a similar two-agent system - instead of the original N-agent system. We then quantize agents' observations such that their more valuable observations are communicated more precisely. We show analytically that under a wide range of problems, the proposed scheme is applicable. Our numerical results show that the proposed approach achieves significant improvements in reducing the computational complexity of the centralized training phase for the design of inter-agent communications in MAS problems while maintaining the average return performance of the system.

Chapter 2

Task-Oriented Communication Design in Cyber-Physical Systems: A Survey on Theory and Applications

2.1 Introduction

Traditionally, communications system design has been guided by task-agnostic principles, which aim at efficiently transmitting as many correct bits as possible through a noisy channel and under some constraints. The design approaches have been largely based on information and coding theories, where the former sets the upper bounds on the system capacity, whereas the latter focuses on achievable techniques to approach the bounds with infinitesimal error probabilities. Despite the abstraction level of these theories, they have been successfully extended to an impressive number of communication network topologies. In this direction, digital communications has made extraordinary leaps in terms of performance, allowing robust information transfer for multi-user systems even in the face of adverse channel conditions.

However, in the era of cyber-physical systems, the effectiveness of communications is not dictated simply by the throughput performance indicators (e.g., bit rate, latency, jitter, fairness etc.), but most importantly by the efficient completion of the task in hand, e.g., remotely

controlling a robot, automating a production line, collaboratively sensing/communicating through a drone swarm etc. It should be noted that according to projections, by 2023, half of the worldwide network connections will be among machines rather than humans [38]. Machines and its components (e.g., sensors, processors, actuators) - in contrast to humans - operate based on objective quantifiable processes, which in theory can be modelled and used as side information during the communication design process. Moreover, coordination through communication messages will be imperative in terms of achieving the completion of complex tasks with the help of multiple agents, be it integrated modules such as robots, drones, vehicles or individual components thereof. In this future cyber-physical world, it becomes critical to investigate a new paradigm for designing communication strategies for multi-agent cyber-physical systems, which can be adapted or tailored on a case-by-case basis by analyzing jointly the nature of the targeted collaborative task objective and the constraints of the underlying communication infrastructure.

Looking back to conventional communication systems, the design of the vast majority is currently based on the source-channel separation principle, which suggests that the source information can be compressed independently of the communication channel and subsequently suitable redundancy should be added to combat the adversities of the channel itself. However, it should be noted that this principle only applies under strict conditions [39–41]. More importantly, the source compression is based on the statistical properties of the input distributions, which do not reveal the importance/value of each sample with respect to the task completion. At the same time, current communication infrastructure largely depends on the concept of layering, which is meant to create abstractions which lead to simplified system design. However, the same abstractions create rigid interfaces that prevent the higher layers (i.e., applications/tasks) from directly affecting/adapting the lower layers of communication system designs.

Incorporating this view of the problem in the design process could be a daunting task, since it requires a combination of principles from information, communication, control theories and computer science theory. In fact, all of the aforementioned scientific communities have already realised the value of this new paradigm and have made initial steps to address it from their own point of view. Nevertheless, a common framework for task-oriented communication design is still lacking, mainly due to the following challenges:

- Divergence of *model assumptions* e.g., communication model (channel, network topology), statistical system models (partially observable Markov decision process, independent identically distributed processes), local versus global rewards/utility functions
- Divergence of *performance metrics*, e.g., mutual information, discounted empirical error/risk, error/cost functions within time horizons
- Divergence of *mathematical tools*, e.g., rate-distortion theory, strong-weak coordination theories, dynamic programming, successive approximation, stochastic optimization and reinforcement learning.

2.1.1 Semantic and Task-Oriented Communications

The design of communication systems mainly involves three different levels of problems, namely, technical, semantic and effectiveness [41, 42]. Out of these, technical problems are related to the accuracy of information transmission (which may be a finite set of discrete symbols, one or many continuous functions of time and/or space coordinates) from a transmitter to a receiver. On the other hand, semantic problems are associated with how precisely and accurately the transmitted symbols can communicate the desired meaning and involve the comparison of the interpreted meaning at the receiver with the intended meaning by the transmitter based on contents, requirements and semantics. In other words, semantic communications deals with the transfer of a concept or information content from a source to the destination without going into the details of how the message is being communicated to the receiver [43, 44]. In contrast to the Shannon's framework based technical-level communications, semantic communications can provide performance enhancement due to the fact that it can exploit the prior knowledge between source and destination in the design process [45–47]. However, the semantic design does not consider the implications of the usefulness of the information for the task on designing the communications [46]. The third category of problems (i.e. effectiveness) focuses on how effectively the received information can help to accomplish the desired task/performance metric [41, 48]. This design paradigm is recently defined as a goal-oriented approach in the literature [43], which is termed as task-oriented design in this thesis[1]. Compared with the conventional Shannon-based technical framework, new

[1]We prefer the term task-oriented, because the term "goal" is often associated with humans and sounds too ambitious for cyber-physical systems comprised of machines.

paradigms of semantic and task-oriented communications are expected to create a paradigm shift in future communication networks in terms of enhancing effectiveness and reliability without the need of additional resources such as energy and bandwidth.

Task-oriented communication enables the involved entities/agents to achieve a common goal/task and its design should focus on achieving the joint objective under task-oriented constraints and specifications by utilizing the provided resources (radio spectrum, computation, energy, etc.) and suppressing the information that is not relevant to the achievement of the goal. The effectiveness of a communication design can be achieved by defining a clear goal, therefore leading to a task-oriented design. This communication framework aims to effectively fulfill the predefined goal/objective by transmitting only the information relevant to a particular goal rather than the all raw information that would be communicated in the Shannon's framework based approach. In such a task-oriented design, the performance of the system can be evaluated in terms of the degree by which a particular goal is fulfilled while utilizing the available amount of resources. In contrast to semantic communications, task-oriented design also utilizes the resources and entities (computation, actuation and control devices, and network nodes) usually dealt at the technical level with the objective of enhancing the effectiveness of the predefined goal [43]. As compared to the existing works focused on semantic communications [45,46,49,50], task-oriented communications in this chapter will focus on the design of cyber-physical systems, which aims to enhance the task effectiveness without going into the details of semantics.

Most importantly, the fact that a communication message has new semantic information does not necessarily mean that it will be useful for the task. Take the tracking example covered by [46]. Consider some tracking information observed at the transmitter side (by the sensor) that is not visible by the receiver side (actuator). Since they are new to the receiver, this tracking information is said to have semantic value and hence worthy of communication. In a task-oriented way of thinking, however, the designer would also take into account that how this new tracking information will make any difference in the actuator's decision. If the new tracking information will not change the decision of the actuator, it has no value to be communicated - from the task-oriented point of view, see [51](Sec. 4) for further readings on the differences between the task-oriented and semantic-based design of communications through the lens of graph theory. In addition, compared with the existing layered-based designs (i.e., technical, semantic and effectiveness), the task-oriented design

framework in this chapter does not consider the layered approach in [42] but envisioned focusing on task effectiveness-based design without explicitly semantic modelling. In this regard, this chapter envisages holistic policies for multi-agent cyber-physical systems to enable the joint design of communication strategies for the underlying resource-constrained B5G/6G network infrastructures and suitable action policies towards maximizing the task-oriented reward. Consequently, other information-related semantics (e.g., Age/Value of Information) should not explicitly affect the task-oriented design framework, but they could potentially be derived as a byproduct of the information distillation policy for each inter-agent connection.

Task-oriented communication systems, go beyond the naive forwarding of data samples for processing at the receiver end and attempt to realize the importance of the communicated data for the processing/computing task at the receiver. This approach will help (i) put less burden on the limited radio resources of the network [15], (ii) reduce the size of data being transmitted leading to improved delays for delay-sensitive applications (iii) decrease the complexity of processes to be performed at the receiver via a joint communication and computation design approach [19], (iv) enabling wider use of narrow-band technology standards such as Cat-NB1 (NB-IoT) and LoraWan that support only low throughput applications requiring low power and long-range [20] and, (v) reducing the storage size required at the receiving end to save the data

2.1.2 Technological Enablers for Task-Oriented Communications Design

The novel design methodology of task-oriented communications design (TOCD) framework departs from the conventional layered-based design and demands suitable technological enablers. In order to meet the overall system task effectiveness of TOCD, it is essential for the communications networks to be highly specialized and adapted to distinct requirements of different tasks. We envisage that these task-specific requirements can be fulfilled by the recently developed technologies such as software defined radio (SDR) and software defined networking (SDN) [52], open radio access network (O-RAN) [53], 5G new radio (5GNR) numerologies [54]. Softwarization SDR/SDN enables highly flexible PHY and NET layer configurations dedicated to different tasks via software updates which can be executed anywhere and anytime. On the system level, O-RAN creates not only a common interface for infrastructures from various vendors but also flexible functionality splits for different appli-

cation needs. In parallel, it allows drawing data from various communication blocks and utilizing them to understand and optimize the performance of current configurations. Therefore, the deployment of SDR/SDN under the O-RAN environment is expected to achieve high performance and cost-efficient network configurations. On the waveform level, 5GNR numerologies allow customized radio resource blocks to meet diverse task-specific requirements without changing the transmission protocols. Furthermore, private 5G networks [55] can deliver ultra-low latency and high bandwidth connections and enhanced security level for Industrial 4.0 applications serving very large number of network elements, e.g., smart manufacturing and autonomous vehicles. A private network dedicated to a specific cyber-physical system can be fully customized to its needs based on the TOCD framework.

Despite of the rise of private networks, a large part of the communication systems will be still operating over shared infrastructure. In this context, a crucial feature of the networks to enable efficient adoption of the TOCD is the orchestration capability to harmonize the diverse, and sometimes conflicting, task-specific requirements from different applications. Such challenges create a multi-objective (cooperative and/or competitive) game on the infrastructure resource allocation (computation, storage and communication bandwidth). With the recent advances in virtualization technology, network function virtualization (NFV) [56] and network slicing [57] play a key role for coexistence and infrastructure sharing to provide distinct task-oriented network configuration profiles. In cyper-physical systems, the tasks' requirements can significantly vary on both network resources (computation, storage) and communication resources (power, bandwidth). For example, a conventional eMBB application requires a large bandwidth with tolerable latency, while a smart manufacturing application requires a moderate data rate and extremely high reliable and low latency connections. In such cases, RAN slicing [58] has full potential to accommodate these communications challenges. Nevertheless, we believe that the diverse requirements of the task-oriented framework can be efficiently tackled by careful tuning of the aforementioned technological enablers without implementing additional overarching communication layers.

2.1.3 Contributions

Despite the promising studies (e.g. [43, 51]) and an urgent need for formalizing a general framework, there is still a large gap in the existing literature since task-oriented communi-

cations has not been systematically reviewed. In this survey, we make a first step towards highlighting the importance of the new paradigm and surveying various approaches from the wider scientific community, which can help towards a common understanding. Our contributions are summarized as follows:

- We conduct an extensive literature review from a theoretical perspective, classifying the contributions across three major communities, i.e., information/communication theory, control theory and computer science.
- We formulate a common conceptual task-oriented communication design (TOCD) framework to clarify and justify the selected terminology, assumptions and definitions, which will then form the basis for the general problem description for the task-oriented communications design.
- To validate the framework, we focus on specific use cases, where we have collected the major literature which can be studied under the prism of this new paradigm. Properly addressed, the implications of these topics can be far-reaching across a range of real-world applications, such as industrial internet of things, multi-UAV systems, tactile internet, autonomous vehicles, distributed learning systems, internet of skills, smart manufacturing plants and 5G and beyond self-organizing networks.
- Finally, we envision a number of critical open research topics within the proposed TOCD framework and suggest potential approaches for tackling them.

2.1.4 Organization

This survey chapter is structured as follows. In Section 2.2, we review the relevant literature from a theoretical perspective, classifying the contributions across information/communication, control theories and computer science. In Section 2.3, we introduce a common conceptual framework along with the corresponding assumptions to clearly delimit the problems that are addressed within this survey. In Section 2.4, we first focus on concrete application areas by specifying the general framework to match the underlying system model and then structure the relevant literature. Section 2.5 offers a list of open research topics and challenges pertinent to task-oriented communications system design.

Table 1. Table of notations

Symbol	Meaning
$\mathbf{x}(t)$	A generic random variable generated at time t
$\langle \mathbf{x}, \mathbf{y} \rangle$	Product of two vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathsf{x}(t)$	Realization of $\mathbf{x}(t)$
$\mathcal{X}$	Alphabet of $\mathbf{x}$(t)
$\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
$p_{\mathbf{x}}\big(\mathsf{x}(t)\big)$	Shorthand for $\Pr\big(\mathbf{x}(t) = \mathsf{x}(t)\big)$
$H\big(\mathbf{x}(t)\big)$	Information entropy of $\mathbf{x}(t)$ (bits)
$I\big(\mathbf{x}(t); \mathbf{y}(t)\big)$	Mutual information of $\mathbf{x}(t)$ and $\mathbf{y}(t)$
$\mathbb{E}_{p(\mathsf{x})}\{\mathbf{x}\}$	Expectation of the random variable X over the probability distribution $p(\mathsf{x})$

2.1.5 Notations

The notations used throughout the chapter are listed in Table 1. In general, bold font is used for matrices or scalars which are random and their realizations follow simple font.

2.2 Theoretical Concepts for Task-Oriented Communications Design

This section focuses on theoretical problems and insights which can have implications or applications on task-oriented communication design. Although the borders are often blurry, we classify contributions across three main axes: information/communication theory, computer science and control theory. Fig. 2.1 presents a summary of the classification, main theories and selected references.

2.2.1 Information/Communication Theory

Relevant results in Information Theory can be traced all the way back to 1971 [59], where the state of the system as well as the observations of agents are modelled as information sources with memory and approached through the lens of rate distortion theory. The paper introduces a quantitative measure to capture the amount of information that is stored in average in the memory of the information source. It is then shown that when source coding is designed subject to any constraint on the distortion, the best achievable rate (corresponding to that distortion constraint) is lower than the achievable rate for a similar source which is

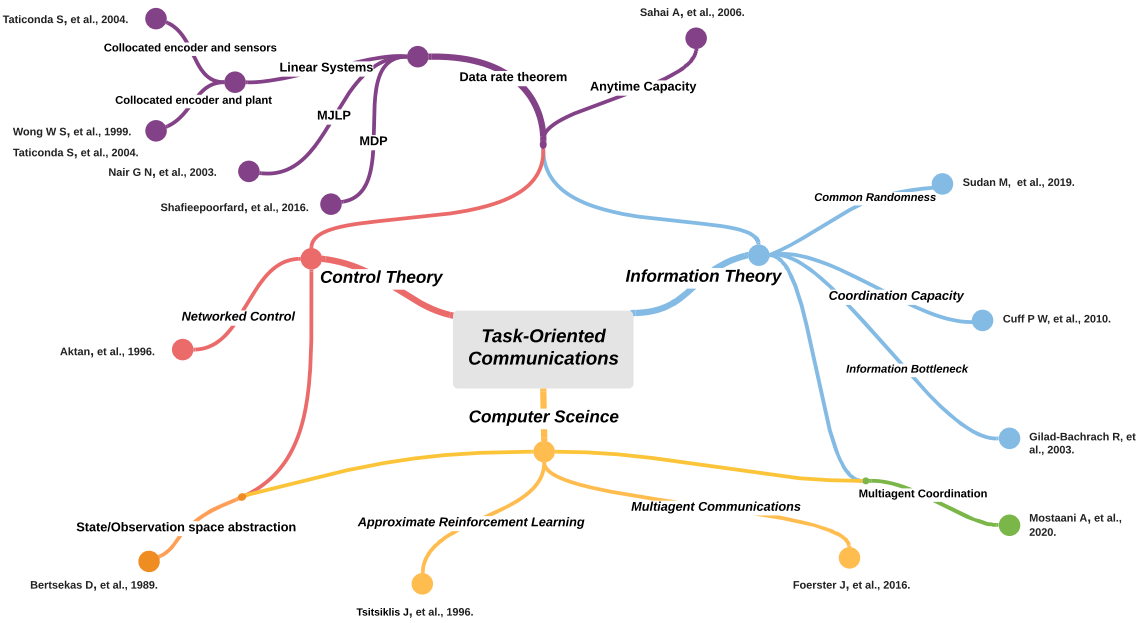


Figure 2.1: Main theories and selected references.

memoryless. The difference between the achievable rate for a source with memory and a similar source without memory is quantified to be less than or equal to the measure of the memory of the information source. The results of the paper are general enough to apply to the sources with finite memory of any size, e.g., Markov sources with memory size of $L = 1$. These results are fundamentally different. Due to the wealth of literature, this section focuses mainly on recent results and surveys, which can be used by the readers to trace back other useful references.

The concept of common randomness [60] is recurrent through the information-theoretic approaches relevant to our framework. Even though its applications are much wider, we focus here on the problem of generating shared uniformly distributed bits across a network of agents, using initial correlated observations as side information, complemented with interactive communication. Common or shared randomness can facilitate the execution of numerous distributed tasks, such as secret key generation, distributed computation, channel coding over arbitrary channels, synchronization, consensus, leader election etc.

Coordination Capacity: In the information theory community, the probing of the system through different agents is often modelled as a joint distribution of actions that has to be achieved across the agents given constraints on the communication rates. In this context, the

work in [61] introduced the concepts of empirical and strong coordination. In this work, the concept of common randomness plays an important role and it is defined as a source of random samples which is available at multiple nodes even if there is no communication among them. Empirical coordination is tightly connected with distortion theory, since the objective is to achieve a desired joint probability distribution in a set of nodes driven by the communicated random samples generated in another set of nodes. The authors begin with toy examples of two or three nodes by examining various topologies such as the cascade and the broadcast channel. Strong coordination extends the paradigm to temporal sequences of samples. The authors in [62] introduce another variation, termed imperfect empirical coordination, aiming to bound the total variation between the joint type of actions and the desired distribution.

Anytime Capacity: Another notable contribution in [63] focused on the intersection of information and control theory, by studying the concept of anytime capacity. In this case, the aim of communication is specifically targeted on stabilizing an unstable linear process (e.g., a plant control loop). While previous works have focused on erasure channels [64], the work in hand addresses noisy channels. The focus is on a small-scale scalar system model with a single observer who communicated over a noisy channel with a single controller. The controller can send both control signals to the system and feedback to the observer with a one step delay. The main result is the "equivalence" between stabilization over noisy feedback channels and reliable communication over noisy channels with feedback. A key point in the model is that the decoder of the controller has to provide increasingly reliable estimates for all received past messages, as there is no side information about the message timing as required by the system. The reliability of the messages should increase sufficiently and rapidly over time to assure the stability of the system.

Information Bottleneck: The information bottleneck [65] is an interesting construct with hidden implications towards task-oriented communication design. Let us consider the following formulation of three random variables:

$$\begin{aligned} \max_{\mathbf{t}\in\mathcal{T}} \quad & I(\mathbf{x};\mathbf{t}) \\ \mathtt{s.t.} \quad & I(\mathbf{t};\mathbf{y}) < R. \end{aligned} \tag{2.1}$$

The aim of the information bottleneck is to compress the information in $\mathbf{y}$ into $\mathbf{t}$ following the rate constraint R, such that $\mathbf{t}$ can provide the most useful information about $\mathbf{x}$ in the mutual

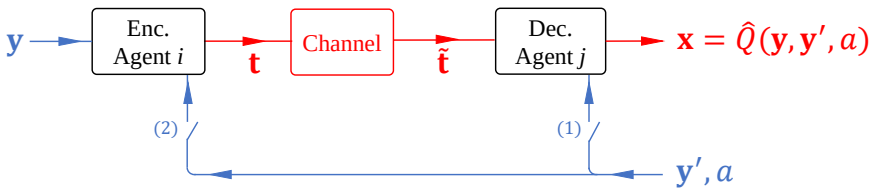


Figure 2.2: Distributed task based source coding.

information sense. In this context, it has been recognized that the information bottleneck problem provides a method to extract the information in $\mathbf{y}$ which is most relevant for estimating or approximating $\mathbf{x}$ [66]. Now, let us consider a toy example of sequential decision making for two agents, where one receives an observation $\mathbf{y}$ that has to be, at least partly, communicated to the other one in order to maximize the expected cumulative reward. The random variable $\mathbf{x}$ can be seen as the expected cumulative reward of the system given action a and the current state of the system, where the state of the two agents is jointly defined by $\langle \mathbf{y}, \mathbf{y}' \rangle$. The parameter $\mathbf{t}$ is the communication message that the first agent is about to transmit based on its observation $\mathbf{y}$ to facilitate the control decision a by the agent j. Accordingly, by solving a (conditional) information bottleneck problem, at the side of encoder, we can optimize the communication of observations $\mathbf{y}$ of agent i, by compressing them to $\mathbf{t}$ while ensuring that the compressed communication message $\mathbf{t}$ has yet the maximum possible information about the conditional expected return $\mathbf{x}$. Fig. 2.2 illustrates the setting of this problem. When switch (1) and (2) are both off the problem reduces to a rate-distortion problem with memory at encoder of the agent i. The work done in [67] was first to show the connection of the information bottleneck problem and rate-distortion with logarithmic loss. To stay consistent with the framework that will be proposed in section 2.3, we assume the switch (1) to be always on. Accordingly, conditioned on the switch (2) being on, the coding in the encoder of agent j should be done such that it maximizes the conditional mutual information $I(\mathbf{x}; \mathbf{t}|\mathbf{y}')$. A solution to the conditional information bottleneck problem was provided by [68].

Multi-agent coordination under imperfect observations: The contribution in [69] and [70] lies in the intersection of information and communication theory. The focus is on multi-agent systems that have to coordinate to maximize their long-term utility functions based on system observations impaired by an i.i.d. process. This particular example focuses on distributed power control and the system state contains the channel gains, which are partially

and imperfectly known to each agent. The authors formalized the optimal problem and provided an achievable solution based on sequential best-response dynamics.

2.2.2 Control Theory

After the advent of the networked control systems [71, 72], the pioneering works proposed in [73, 74] was successful in developing a cohesive data rate theorem. The paper studies the stability of a control system under a rate-limited but reliable communication channel. It is assumed that the rate of communication is time invariable, the state process is a discrete random variable, and the disturbances that the system is subject to are bounded. They show that the necessary conditions on the rate of the communication channel is independent of the algorithms used for encoding and decoding the communication messages. It has also been shown that these necessary conditions on the channel rate are independent from the information patterns of the networked control system.

There has been a plethora of subsequent works afterwards to extend this setting, to more realistic scenarios [75–79]. The authors of [75] extended this framework to cases where the support of system disturbances is unbounded. Martins et al. accommodated time-varying rates for the communication channel in the framework, however, their results were limited to first order linear systems with bounded disturbances [76]. The work in [77] widened the applicability of the data rate theorem to finite dimensional linear systems with unbounded disturbance and time varying rate for the communication channel. In contrast to the previous works [75–77], Liu et al. in [80] considered the joint effect of all the two parameters of latency and data rate by finding a region of stability that indicates the necessary and sufficient values of data rate as well as latency. Huang also considered the joint effect of all the three parameters of latency, reliability and data rate in the control system [81]. It was shown by Kostina that for a fully observable linear system, a lower bound for the rate-cost function can be computed even when the system disturbances are not generated by a Gaussian process [5]. The results provided by their research can even be generalized to partially observed linear systems, when the observation noise is Gaussian. The rate-cost function is the minimum required bit rate that can guarantee the system state to be upper-bounded by a certain value.

Interested readers can find more details about the solutions proposed for the control of a

linear system over a communication network in the following surveys and books [82]. While most of the works discussed above consider the state process to be generated by a linear system with added Gaussian system/measurement noise, fewer works have targeted the state processes which are generated by Markov jump linear processes (MJLP) [83–86] or Markov Decision Process [87–89].

State aggregation for dynamic programming has been studied for long among the communities of control theory and operations research [88, 90–93]. Successive convex approximation (SCA) is one of the main tools leveraged to form a trade-off between the accuracy of the solution of the dynamic programming problem and its computational complexity [90, 94]. Later, adaptive algorithms for state aggregation were proposed which could recompose the aggregation of states during the iterations [88]. The major driving force for researchers, in the past, to work on this problem has been to enhance the computational efficiency of the algorithms. However, the algorithms that they have made available, can now be used to efficiently represent the state of environment while minimizing the degradation of the objective function. This area is also well explored by the community of computer science and is addressed in the next subsection.

Shafieepoorfard and Raginsky have studied the problem of controlling a Markov Decision Process while the agent is subject to observation constraint [95]. In particular, the observations of the agent here are considered to have a limited mutual information with the state of the system. Most of the literature in control theory society either treats a given medium for observations as a given constraint or considers transmissions of the sensory system as an additional cost [96]. The work in [95] together with [97] can be considered among the very first few papers which solve the problem of finding a stochastic control function in conjunction with the control problem. The problem of joint design of the observation and control policy is studied in this paper in its very general form as the policies are not considered as deterministic but as stochastic functions, where the transitions of the system are also considered to be stochastic. The problem of one-shot control-communication policy optimization under mutual information constraint is first shown by authors to be a form of rate-distortion problem. Armed with this analytical result, authors consequently use the Bellman equations to reformulated the core problem as a one-shot control policy optimization under mutual information constraint. This particular way that the infinite horizon control problem under observation constraint is formulated by [95] was first introduced by Sims in his seminal

work [98] which won him a Nobel prize.

In [98], Sims explained the limited correlation between the behavior of economic agents and the information they have access to, observing the information through a rate limited channel. One application for this novel way of viewing/modeling economic agents is to solve the permanent income problem where there is limited information about the labour as well as the wealth, which is a dynamic programming problem with mutual information constraint on the state information of the system. In [9] and [35] the concept of information constrained dynamic programming is brought into the context of multi-agent coordination. Agents are limited to local observations but are allowed to communicate through rate-limited channels. Therein, the authors have developed a state aggregation algorithm which enables each agent to compress its generated communication message while maintaining their performance in the collaborative task. Similarly, in [4], authors introduced task-based joint source channel coding to solve the problem of multiagent coordination over noisy communication channels. To understand how these works are relevant to our framework, it should be noticed that, in fact, the information constraint on the observation of the agents in the aforementioned works can be translated as the limitation of the communication channel between the observer of the environment and the controller.

2.2.3 Computer Science

Function approximation has played an essential role in the reinforcement learning (RL) literature to overcome the limitations of the Q-learning method. The authors of [99, 100] have built the foundation of function approximation RL. Therein, the convergence of function approximations which are linearly combined from basis functions over a state space is established. This result opens up a wide range of applications of RL as it only requires a compact representation of the cost-to-go function, in which there are fewer number of parameters than states. In [101], a neural network based on reinforcement learning, namely neural fitted Q (NFQ), was proposed. NFQ comprises of a multi-layer perceptron which is able to store and reuse the transition experience. It is shown therein that NFQ can effectively train a model-free Q-value and achieve the control policy after few communications rounds. The authors of [102] considered similar fitted Q-iteration applied to continuous state and continuous actions batch reinforcement learning. The goal is to achieve a good policy gen-

eralized from a sufficient number of generated trajectories. The authors have developed first finite-time bound for value-function based algorithms applied to continuous state and action problems. The authors of [103] extended the temporal-difference learning proposed in [99] to stochastic control settings via convergence analysis of several variations of Q-learning under function approximation. Therein, the condition for the approximate methods to converge with probability one is identified. The advantages of approximate RL have been successfully demonstrated in real-world environments in Atari game in [104,105]. Therein, the deep Q-network was able to learn the policies directly from the pixels and the game score and outperformed all existing learning models.

A lossless compression scheme has been proposed by [106–108] for the collaborative tasks where the observations of the distributed decision makers are generated by a (Decentralized) partially observable Markov decision process (Dec-)POMDP. The authors suggest an optimal clustering scheme to partition the history of observations of an agent such that no loss in the team objective occurs. While the main intention of the paper is to reduce the complexity of distributed computations done at each individual decision maker, the proposed algorithm can have substantial applications in data compression when a group of collaborative decision makers intend to communicate through rate-limited channels. Similarly, the work in [9] also proposed k-medians clustering to find a proper abstraction of agents' observations before they communicate to other agents through a communication channel. It is shown that the problem of agents' observation abstraction in a multi-agent setting is, in fact, a generalized version of rate-distortion problem.

A general multi-agent framework has been proposed in [109] for mixed cooperative and competitive multi-agent environments. Moving from the common assumption in multi-agent studies, which assumes observations and policies of all agents are available at every agent, this work proposed to learn the approximation of other agents' policies via the maximization of the log probability of the agent. In order to maintain the robustness of the of proposed policy, the agents employ policy ensembles that are trained for a wide range of policies. Nevertheless, the developed framework allows to evaluate the impact of the communication among the agents on the learning policy by limiting the observed information from other agents.

Efficient communication strategies have been studied in [110] and [111] to save the com-

munications in distributed learning systems. Therein, the authors have proposed a so-called LASG (Lazily aggregated stochastic gradients), which adaptively determines when to communicate. The main idea of LASG is based on the new communication rule that determines the informative content of the gradients after each round based on the difference between the fresh and staled gradients. By properly implementing the rules, both downlink and uplink loads can be reduced, since only nodes with certain informative update communicate. It was shown that the proposed LASG therein achieves similar convergence as the conventional stochastic gradient descent (SGD) and significantly reduces the communications load.

The authors of [112] proposed a framework for states abstraction via aggregating original states into abstract states to reduce state and action spaces. Based on the unified state abstraction framework in [113], the authors therein proposed and proved the existence of four types of approximate abstractions that guarantee the gap to the reward using the true state value function. The developed framework therein was tested in five different problems which shows significant reduction in abstract state space and suboptimal value function. In [114], a representation learning scheme was proposed for hierarchical reinforcement learning in which a high-level controller learns to communicate the goal representation to the lower-level policy that is trained to achieve the goals. It was shown that a good choice of representation learning policy leads to a bounded suboptimality. Although developed for the single-agent system, these frameworks are useful to reduce the communication message in multi-agent networks.

2.3 Task-Oriented Communication Framework and Scope

As surveyed in the preceding section, the task-oriented communication design has been investigated through many different viewpoints. There can be found many direct/indirect task-oriented design schemes of the communications system in the literature of computer science, information/communication theory and control theory (e.g. [115,116]). While the theoretical tools introduced in Section II were mostly focused on indirect schemes, the succeeding section will explore direct schemes. In contrast to indirect schemes, the direct schemes aim at guaranteeing or improving the performance of the cyber-physical system at a particular task by designing a task-tailored communication strategy. Since direct schemes are specifically designed for a particular task, they can hardly be generalized to other application scenarios. On the other hand, direct schemes can take the advantage of the available side knowledge

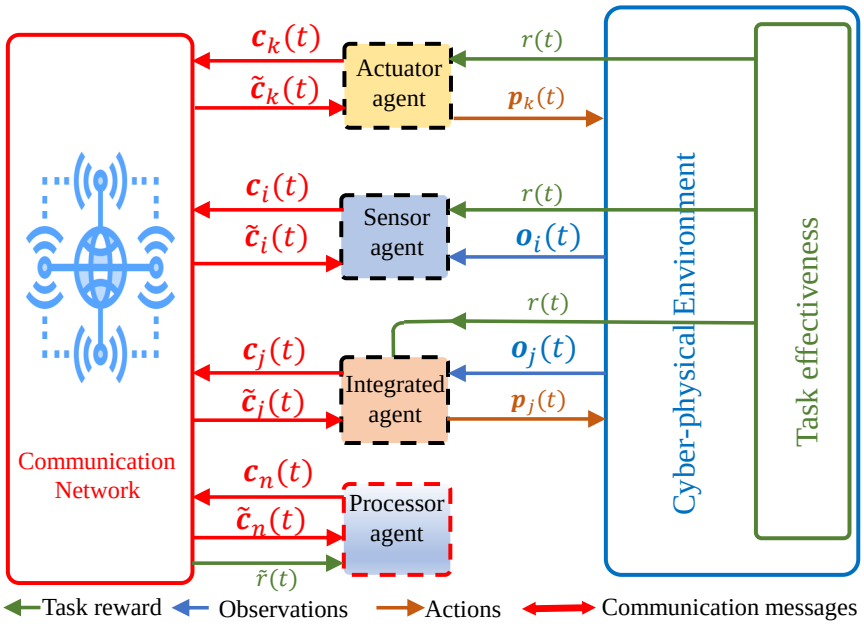


Figure 2.3: Proposed task-oriented communication design framework for cyber-physical systems. There are four types of agents with different levels of interaction with the environment.

about the particular task which are meant to facilitate.

Therefore, we believe that having a unified problem framework would allow us to find commonalities among various task-oriented communication problems, use the available methods in the literature to solve a wider range of task-oriented communication problems and distinguish the differences between them. In this section, we formulate a framework, which is sufficiently generic to capture various examples of task-oriented communication design, as will be shown later in Section 2.4. Furthermore, the proposed framework will establish a common terminology, clarify the underlying assumptions and delimit the targeted problem set.

2.3.1 TOCD Overview and Agent Types

The proposed TOCD framework shown in Fig. 2.3 targets a proactive design approach to enhance the task effectiveness of cyber-physical systems, which are captured via three major components: *i*) the environment module, *ii*) the multi-agent module, and *iii*) the communication module. The environment is a core component that defines a set of parameters determining the environment state, which is controlled by the agents' actions and is in turn translated into task effectiveness levels. The multi-agent module includes a number of agents

with different levels of interaction with the environment module in terms of observations and probing actions. Finally, the communication module dictates the communication capabilities (network topology, medium access type, etc.) and constraints (rate, power, energy, interference, codeword length etc.) which dictate the nature of the inter-agent information exchange. The TOCD aims at optimizing the multi-agent module by jointly designing its communication strategies and action policies, using as input 1) the task effectiveness values from the environment module; 2) the capabilities of the multi-agent module; and 3) the constraints of the communication module.

The TOCD framework classifies the agents into four main types: sensors, actuators, processors and integrated agents, as depicted in Fig. 2.3. The sensor agents directly receive the task effectiveness signal and observe the environment states through its sensory measurements, which will then be sent to other agents via the communication networks. We note that the actions of sensor agents do not directly change the environment states. Examples of sensor agents include the sensors in the sensor networks or separate sensory modules in industrial plants. On the other hand, actuator agents can only have access on the environment states via communication messages with other agents, but their actions directly change the state of the environment. The integrated agents are the most complete and contain the features of both sensor and actuator agents. Therefore, they fully interact with the environment including direct observation of the system states and influence the environment via their actions. The fourth type of agent, processors, helps other agents with heavy computational or consensus tasks. Therefore, they do not directly observe or influence the environment state. In fact, the processor agent receives feedback on the current task effectiveness via the other agents through the communication network. It is worth noting that although sensor and processor agents do not directly send probing signals (actions) to the environment, their actions still have impacts on the system states by influencing other agents' actions.

2.3.2 Joint Communications and Actions Policies in TOCD

Consider a multi-agent system with K agents $\mathcal{K} := \{1, \ldots, K\}$. A generic agent k at any time slot t can observe the system state through the local observation signal $\mathbf{o}_k(t) = \boldsymbol{s}_k \in \mathcal{S}_k$, where $\mathcal{S}_k$ is the set of the local system states. Then, the global system state at time step t can be represented by $\boldsymbol{s} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K] \in \mathcal{S}$, where $\mathcal{S} := \bigcup_{k \in \mathcal{K}} \mathcal{S}_k$. An agent k at any time step

t can execute an action ${}_k(t)$, which can affect the overall state of the system. Let us denote $\boldsymbol{p} := [{}_1(t), {}_2(t), \ldots, {}_K(t)] \in \mathcal{P}$ as the system action(s). Furthermore, let $\boldsymbol{s}(\cdot)$ and $\boldsymbol{p}(\cdot)$ denote the sequence of the system states and the actions over time, respectively. The overall objective is to complete an abstract task whose performance modelled by a task effectiveness function $\mathrm{T}((\cdot), \boldsymbol{p}(\cdot)) \in [0, 1]$. In order to make a decision on , we need to extract the information $\boldsymbol{c} = [\boldsymbol{c}_1(t), \boldsymbol{c}_2(t), \ldots, \boldsymbol{c}_K(t)] \in \mathcal{C}$ contained in $\boldsymbol{s}$ that is relevant to the task, where $\boldsymbol{c}_k(t)$ is the communication message sent by agent k at time step t. The communication network between the agents is characterized via the communication operator $h : \mathcal{C} \to \widetilde{\mathcal{C}} : \boldsymbol{c} \mapsto \tilde{\boldsymbol{c}}$, where $\tilde{\boldsymbol{c}} = [\tilde{\boldsymbol{c}}_1(t), \tilde{\boldsymbol{c}}_2(t), \ldots, \tilde{\boldsymbol{c}}_K(t)]$ and $\tilde{\boldsymbol{c}}_k(t)$ is the message received by agent k at time step t. We note that while $\mathcal{C}$ stands for the alphabet from which the agents can select their communication messages, the set $\tilde{\mathcal{C}}$ is the set of complex numbers as the received signal can be a complex number. The task-oriented communication and action problem is defined as follows.

Definition 1. *The TOCD aims at jointly optimizing the* ***communication policy*** *and* ***action policy*** *to maximize the task effectiveness* $\mathrm{T}(\boldsymbol{s}(\cdot), \boldsymbol{p}(\cdot))$*, defined as below.*

- ***Communication policy*** $\pi^{(C)} : \mathcal{S} \to \mathcal{C} : \boldsymbol{s} \mapsto \boldsymbol{c}$*. Since the extracted information from* $\boldsymbol{s} \in \mathcal{S}$ *needs to be transferred to the decision maker(s) via the communication channel(s)* $h : \mathcal{C} \to \tilde{\mathcal{C}}$*, the communication policy* $\pi^{(C)}$ *may include both information distillation and source/channel coding policies.*

- ***Action policy*** $\pi^{(P)} : \mathcal{S} \times \widetilde{\mathcal{C}} \to \mathcal{P} : \boldsymbol{s} \times \tilde{\boldsymbol{c}} \mapsto \boldsymbol{p}$ *that decides the action* $\boldsymbol{p}$*, that can be either a global action in the coordinator or the distributed actions in the local agents, based on the observed system states* $\boldsymbol{s}$ *and the communicated information* $\tilde{\boldsymbol{c}}$*.*

Note that the above-defined problem can be subjected to additional constraints of the communications operator h. Compared against the conventional optimization problem that searches for an optimal decision policy by directly communicating states $\pi^* : \mathcal{S} \times \mathcal{S} \to \mathcal{P}$, the TOCD has twofold advantages: 1) *reduce the communication cost* because the communication policy $\pi^{(C)}$ guarantees $H(\boldsymbol{c}) \leq H(\boldsymbol{s})$, i.e., the entropy of the sent information is smaller than the raw states, and 2) *reduce the complexity to optimize the decision process*, because with $H(\tilde{\boldsymbol{c}}) \leq H(\boldsymbol{c}) \leq H(\boldsymbol{s})$ the input space to learn the action policy $\pi^{(P)}$ can be reduced.

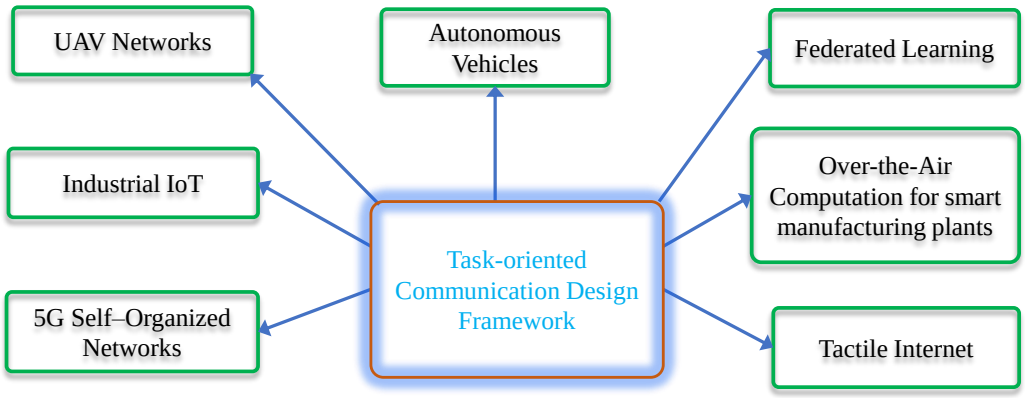


Figure 2.4: Application areas of the proposed TOCD framework.

The proposed TOCD does not assume a specific model for the state evolution of the system, although any task-oriented design of the inter-agent communication requires a consistent and unique objective function to be in place. The objective function here is the basis upon which we can measure the performance of the agents' collaborative probing of the system and it is assumed to be shown as a side information based on the targeted task. In some example scenarios, the objective function can be the accomplishment of an industrial task with enough precision in an industrial IoT framework, e.g., not missing or mistaking the target within a time horizon in a UAV object tracking framework, or having a minimal sum of errors throughout a limited time horizon in the central server of a distributed training system. Nevertheless, the proposed TOCD sets a list of fundamental assumptions which justify the need of cooperation among the agents, as follows:

- The framework consists of at least $K \geq 2$ agents.
- There is a single, common, and consistent single-variate objective function, which is the effectiveness of the task at hand and it should not vary through the time horizon for which the problem is solved. The study for competing or non-aligned objectives for different agents is not considered by our framework.
- There is at least one agent with strict local observation, i.e., $H(\boldsymbol{s}) > H(\boldsymbol{s}_k) \geq I(\boldsymbol{s}_k; \boldsymbol{s}) > 0$ for some agent k, otherwise there will be no need for communication among agents, where $I(.;.)$ denotes the mutual information operator.

- The actions selected by agents affect both the obtained reward as well as the state process. In other words, we are less interested in the scenarios where state and action processes are independent - these scenarios usually arise in distributed estimation problems which form another rapidly growing literature [26, 117–122]. [2]

- We assume that the local/global response signal of the system is available to all agents at no cost. In case that agents have local response signals, we assume that the global response signal can be represented as a function of the local responses.

In the next section, we will show how the proposed TOCD framework can be applied in various application domains.

Remark 2. *We do not assume that the objective function is known in its analytic form by the task-oriented communication designer. Potential side information about the equations governing the evolution of the system state and the objective function can be exploited to design task-oriented communication policies using optimization/dynamic programming techniques. Nevertheless, even if we only have access to sampled data points of the objective function, one could resort to machine learning techniques for data-aided communication policy design. Machine learning techniques can also be promising to give rather generic solutions to the problem of task-oriented communication design, since they can generalize over tasks and systems that are not analytically tractable.*

2.4 Applications

In this section, we demonstrate how the proposed framework captures the most popular applications of task-oriented communication networks, covering seven application areas as depicted in Fig. 2.4. Furthermore, in order to understand how the literature employs theories and tools, we provide Table 2.2 which maps the application scenarios to the major task-oriented design and learning techniques.

[2] Note that in these scenarios, the state-process is usually considered to be momoryless source of information which technically differentiates between the methods that are useful distributed estimation problems and task-based communication problems that are the scope of our work.

Table 2.2: Applications classified by tools/techniques

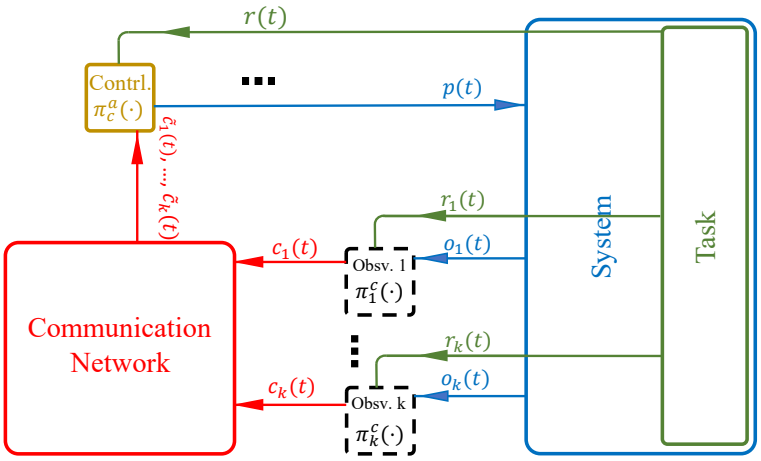
	Data-rate Theorem & Networked Control	State Abstraction/ Approximate RL	Information Bottleneck	Multi-agent Communications & Coordination	MDP and its Extensions	Domain-Specific (non generic) Methods
Industrial IoT	[80, 81] [123–128]	-	-	[129]	[81, 129]	-
Distributed Learning	-	-	[130]	-	-	[131–137]
Self-Organized Networks	-	[138–140] [141–143]	-	[138]	-	-
Over the air Computations	-	-	-	-	-	[144–146]
Tactile Internet	[147–149]	[150]	-	-	[151]	[152, 153]
UAV/ Autonomous Vehicle Networks	-	[154–166] [167, 168] [25, 169]	[170, 171]	[172–175] [168, 176] [25, 164]	[168]	-

2.4.1 Industrial IoTs

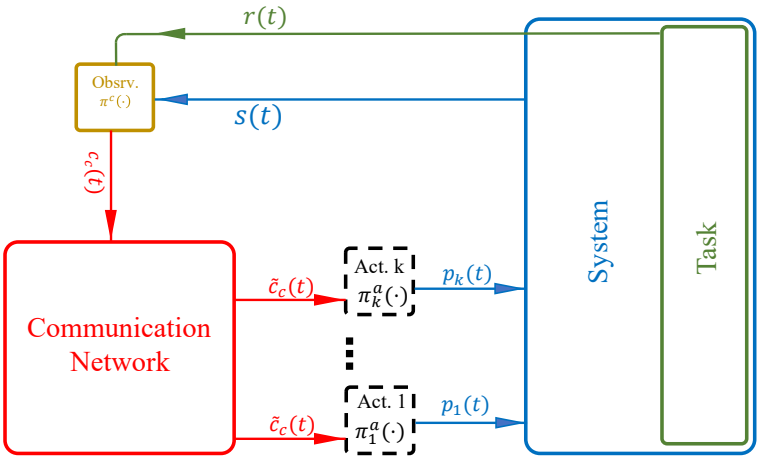
The industrial Internet of Things (IIoT) is the generic framework that exploits the abundance of available data being generated by sensors and other devices to improve the efficiency, reliability and accuracy of an industrial manufacturing process. The availability of data generated by various devices and sensors is playing a key role here which allows each manufacturing process to be performed while having access to a much more sophisticated view of the current state of the system. Meanwhile, not every observation made at any part of the system is useful for all the stages of the manufacturing process. Given possible limitations in the rate of communications in rural areas, or the limited processing power of actuators and controllers in the manufacturing process, extracting useful information becomes of the essence. Consider only the primary activities of a manufacturing value chain, i.e. inbound logistics, operations and outbound logistics, this huge system is comprised of many thousands of tiny elements which can generate (many) megabytes of data per second. In such a huge and complex system, communication and processing power are indeed bottlenecks of the system. Task-oriented communications would facilitate and automate the process of extracting the useful data generated by any element of the system for any controller/actuator of the system. The block diagram of the task-oriented communications for the industrial IoTs is depicted in Fig. 2.5. In this figure, the system response $r(t)$ is the negative of the stage cost function received after the system is probed by the plant i through the signal $p_i(t)$.

In [124] and [125], an IIoT system is studied, where a number of plants (e.g. chemical plants or robots) are controlled by a single controlling unit through slow/fast fading channels. The aim is to reduce the communication latency to meet the ultra low-latency requirements of industry. The authors propose a coding-free communication scheme for the central controller with the plants to optimize a shared cost function among all the plants under a power constraint at the central controller. A power control algorithm is proposed to solve the problem. The works [127] and [128] showed that ML-based techniques can greatly improve the system performance in terms of the spectrum sensing and sum throughout for the complex IIoT networks. In [124], the global stage cost function is not locally observable by each plant. Similarly, authors in [177] also consider a coordination problem where the system response signal is partially observable by the agents. The authors of [123] considered the implications of the limited information rate of a channel on the system stability, where the plants are collo-

cated with sensors and their uplink channels are lossy. The effect of time-variant information rate is considered jointly with the potentially unbounded system disturbances. In particular, the authors have analytically measured the effect of information rate of the channel on the estimation error of the system state. This allows to acquire a precise requirement on the information rate of the plant's uplink channel depending on the magnitude of the unstable mode of the system. For multidimensional linear systems, the necessary information rate of the channel is acquired based on the summation of unstable eigenvalues of the open loop. In [123], however, the joint effect of communication parameters (i.e. latency, reliability and data-rate) is not considered.



(a) Collocated plant and sensors



(b) Collocated plant and Actuators

Figure 2.5: An industrial internet of things problem illustrated using the Task-Based Communication framework.

The authors in [129] developed a multi-agent RL cooperative caching framework which allows edge servers cooperatively learn the optimal caching decision. Each edge server acts as an agent to individually perform the action to predict the location and content request of IIoT devices by applying the K-order Markov chain and long short term memory (LSTM). The proposed approach is capable of both improving the cache hit ratio and reducing content access delay. The trade-off between the reliability of a communication message and its latency is studied in [81] for an industrial IoT application. To study this trade-off, the impact of the age of information that is received by the plant(s) from the central controller is investigated. This is done by obtaining the value function of every possible code length through value iteration algorithm. To be able to run a value iteration algorithm, the interaction of the central controller with the whole system, including the plants, is modelled by a semi-Markov decision process. Therein, the length of error correction codes is designed considering the current state of the system rather than considering only the channel state information (CSI) of the communication network. In that sense, the error correction block of the agents are co-designed with their control policy block.

In [80] the system to be controlled is linear and the uplink channel of the plant(s) is considered to be noise-free. Whereas, the downlink channel of the plants is assumed to be an AWGN feedback channel addressing the needs of low-mobility industrial IoT applications. The paper obtains a region of stability that indicates the necessary and sufficient values of communication parameters (i) information rate as well as (ii) length of error correction code blocks such that the stability of the system is insured. The paper assumes an ideal quantization and control policy to be followed across the network. Accordingly, the quantization as well as the control policy consider the history of all the communication and feedback signals exchanged between the plant and the controller. The paper also computes the average lower and upper bound of the task's cost function where the bounds are obtained as functions of communication parameters, information rate and length of error correction code blocks. The authors of [126] solved a problem very similar to that of [81], where the difference is that in the former, the uplink channel of the plant(s) is considered to be noisy.

Although the separability of the estimator of the system's state from the controller is studied in a work such as [178], the separability of communication and control designs remain largely unknown. The authors in [9] have studied the separability of control design and source coding under mild conditions. Necessary data rate to guarantee a bounded cost function is

obtained by [74, 75, 179], where communication channels are considered to be noise-free.

2.4.2 UAV Communications Networks

Unmanned areal vehicles (UAV) plays an important role in the development of 5G and beyond systems due to their flexible and low-cost deployments. UAVs can serve as a complementary application of the existing infrastructure to stand-alone service in remote areas or emergency scenarios.

Compared to the single-UAV system, the main challenge of multi-UAV system is how to efficiently coordinate the UAVs' operations under limited communication resources. However, most of the works on multi-UAV rely on existing communications design and only focus on the UAVs' action policy [154–158, 180–183]. This conventional communications, which is designed for per-link performances (bit-rate maximization or packet-error rate minimization), in general is not optimal for the joint task in hand. By using the proposed TOC framework, each UAV can jointly optimize its action policy and communication message to be exchanged with the centralized controller (or neighboring UAVs). In the K-UAV systems, each UAV observes the environment via its location (local state) $\boldsymbol{x}_k \in \mathcal{X}$ and received message $\tilde{\mathbf{c}}_k \in \tilde{\mathcal{C}}$ from the centralized controller (or its neighbors), from which the UAV jointly determine its communication message $\boldsymbol{c}_k \in \mathcal{C}$ and next movement (action) $\boldsymbol{a}_k \in \mathcal{A}$. The performance of the collaborative task in UAV systems is then determined via a general utility function $\Phi(\boldsymbol{x}, \boldsymbol{a})$, where $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K]$ and $\boldsymbol{a} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$ are the UAVs' locations and movements, respectively. Under the TOC design, each UAV k optimizes the joint communication and action policy $\pi_k^{(c,a)} : \mathcal{X} \times \tilde{\mathcal{C}} \to \mathcal{C}_k \times \mathcal{A}$ that maps the local observation $\boldsymbol{x}_k$ and received messages $\tilde{\boldsymbol{c}}_k$ to a tuple of the local encoded message $\boldsymbol{c}_k$ and local movement $\boldsymbol{a}_k$. The block diagram of TOC framework for UAV systems is depicted in Fig. 2.6. In the following, we review the most relevant works on UAV communications systems, although most of them either assume perfect or design the communications separately from action policies.

In [155], a reinforcement learning-based sense-and-send framework was proposed for UAV networks. Therein, a BS communicates with multiple UAVs which sense data and then send it back to the BS. The objective is to maximize the sensed data sent back to the BS. To simplify the model, the authors adopt probabilistic sensing model and orthogonal sub-carriers are assumes, hence there is no interference. The channel gain is modelled as either line-of-

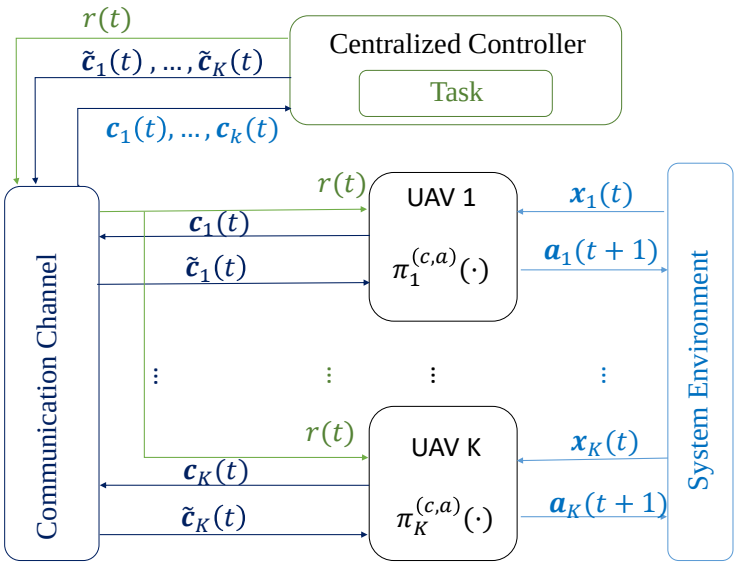


Figure 2.6: The proposed task-oriented communications framework applied to multi-UAV and Autonomous Vehicles networks.

sight (LoS) and non-LoS, depending on distance between the BS and the UAV. By modeling the system states as MDP, with three steps transmission protocol: beacon, sensing, and data transmission, a refined action space is proposed to accelerate the learning performance. The authors of [156] proposed a Deep Q-network in multi-UAV added communications systems in which multiple UAVs serve ground users. The objective is to optimize the movements of the UAVs to maximize the system sum-rate, subject to a constraint on the minimum number of served users. Dueling DQN which uses neural network and dueling update is employed as the learning solution. In [159], a deep reinforcement transfer learning was developed that allows UAVs to "share" and "transfer" learning knowledge, which can reduce the learning time and improve learning quality significantly. In [157], the authors proposed a simultaneous target assignment and path planning based on multi-agent deep deterministic policy gradient (MADDPG) for multi-UAV system. The target is to find the shortest path for all the UAVs while avoiding collision of the task assignment. Full knowledge is assumed to be available to all UAVs during the training. The authors of [158] proposed a deep RL (DRL)-based algorithm for multi-UAV networks to optimize the 3D-deployment of the UAVs. The target is to maximize the quality of experience, defined as a weighted sum of scores of rate and delay. Q-learning and deep Q-learning is used to model the problem. However, similar to the previous work, the communication among the UAVs is not addressed. Very recently, RL-based approaches have been widely developed for the trajectory design [160, 161, 165], computation-

offloading scheduling [162,163], online motion planning [164], and energy minimization [166].

In [180], the authors studied the interference management of the UAV-aided cellular network, in which the UAVs and the users share the same frequency bandwidth and communicate with the BSs. Similar to [167], the objective in [180] is to maximize the average user's rate via the UAV's trajectory optimization. The problem is modeled as a non-cooperative game theory problem with full information. Thus, the communication among the UAVs is not addressed. The authors of [182] considerd the UAV-aided sense-and-send application in which the UAVs, after sensing the target, compete to access the sub-channel to send the sensed data to a mobile device. By taking into account all possible UAVs' actions, the state transition probability is derived, from which a DRL method is proposed assuming centralized architecture. In [183], the authors studied the UAV networks for maximizing the coverage of an area of interest, dividing into clusters. Each UAV provides service to one cluster. The goal is to design the UAVs' navigation policy to maximize the average coverage, as well as fairness among the clusters. Although modeling the problem as POMDP, it assumes unconstrained information exchange among the UAVs for free. The authors of [172] investigated coordinated flight problem in multi-UAV systems in which the UAVs exchange their local location in order to execute given missions. Therein, an adaptive binary coding scheme is proposed based on adaptive zooming that adjusts the quantization level according the moving average of input binary signal. In [181], the authors proposed an RL-based framework for UAV-aided network slicing. Assuming mobile edge computing (MEC) UAV, each UAV serves a number of tasks in an area of interest. The UAV can choose to perform the jobs in its region on its own, or offload to a neighboring UAV. Assuming the job arrival follows the Switched Batch Bernoulli Process, the transition probability matrix is derived taking into account the states of the region and queues at the UAVs. Although the communication among the UAVs is considered, the communication rate is fixed and it does not actively interact with the learning process.

One of the most important research topics in UAV study is UAV swarms, a (large) group of UAVs which collaboratively perform some task. The major challenge in UAV swarms compared to UAV-assisted communication networks lies in the lack of centralized control due to its very large and highly dynamic topology. Unlike UAV-aided communications, where the interaction among the UAVs is aided by a centralized node, e.g., base station, the communication in UAV swarm is usually done via mesh networks. Due to the large topology,

each UAV can obtain only partial observation of the environment and needs to cooperate with other UAVs to improve the learning process. An efficient way to leverage the collaboration in UAV swarms is to employ an interaction graph, which maintains a set of neighbor nodes for each UAV. This principle is considered in [173] and [174], which study the impact of the communication among the UAVs to the distributed reinforcement learning task. In particular, they demonstrate via a rendezvous problem that allowing more exchanged information among neighbor UAVs can significantly accelerate the learning process and results in higher rewards. The authors of [175] analyzed the communication impacts on the multi-robot multi-target tracking problem, where each flying robot can only communicate with its neighbours within its transmission range. Therein, two learning algorithms are proposed to achieve agreement among the robots under limited communication time. The work in [168] provided an extensive survey on the use-cases of machine learning for inter-robot communication design - including robot communications under rate-limit and time constraints to name a few. The timeliness of the data in UAV networks is also studied in [184, 185]. These scenarios all fall under the umbrella of TOCD for autonomous robots/UAVs/Vehicles.

2.4.3 Autonomous Vehicles

Autonomous driving is dependent on the efficient processing of data gathered from various sensors including radar, camera, and light detection and ranging (LiDAR), and involves a complex design process to have a dependable and flexible real-time system. One important use case of autonomous driving is cooperative automated driving, in which one crucial challenge is to ensure the safety gap (< 5 ms) between the vehicles, thus requiring stringent communication requirements in terms of latency and reliability [186]. The underlying entities should be fully coordinated with the help of suitable communication mechanisms, such as mmWave, cellular and visible light communications, in order to ensure the full dependability of autonomous driving systems. Furthermore, in order to guarantee the reliability of information transmission via redundancy, multiple communications links could be utilized in parallel. In this regard, the proposed TOC framework can be applied to autonomous driving systems in a similar manner as in Fig. 2.6. The main differences compared with UAV systems are larger dimensions of both action and observation spaces and more stringent requirements of task effectiveness, which in consequence requires more powerful communication channels.

The transformation from manual control to fully automated driving in autonomous vehicles demands for the efficient management of control authority between the automation and human driver by avoiding human-machine conflicts. In this regard, haptic sharing control could be one promising approach to dynamically adapt the control authority between the human driver, and to suggest suitable actions while exploiting the environmental perception and the driver's state [187]. A haptic sharing control architecture proposed in [187] comprises two hierarchical levels, namely tactile and operation levels, which are responsible for taking driving decisions and to provide helpful actions to track the planned trajectory of the vehicle. This approach incorporates the tactile variables such as driving activities of human driver and the control authority into the planning algorithm so that the automation can better resemble the driver's strategy for planning the vehicle trajectories.

An important design aspect in cooperative automated driving system is the transformation of the single vehicle perception/control in self-driving vehilces to multi-vehicle perception/control, as a vehicle's perception field is dependent on the local coverage of sensors embedded in that vehicle. This requires the need of cooperative perception and manuevering [188], in which TI can play an important role to enable the reliable and fast transmission of haptic information related to the driving trajectories along with sensor information via the underlying vehicular communication network. Furthermore, in cooperative adaptive cruise control (CACC) or platooning applications, which comprise several cars autonomously following the leaders, there arise stringent communication requirements in terms of reliability and update frequency to enhance the safety and traffic flow efficiency [189]. This demands for the design of novel communication strategies for synchronized communications and dynamic adaptation of transmit power, and task-based communication could be promising in addressing these issues.

Another important design aspect for autonomous driving is to design a human-computer interaction (HCI) system with the human-in-the-loop in a co-adaptive manner, as the complete removal of human involvement might be impractical because of various involved uncertainties including human behavior, environmental variations and user requirements. In this regard, authors in [190] used a customizable traffic simulator, which utilizes Artificial Intelligence (AI) to predict the traffic quality at the intersection and can be used as a feedback to the human driver's decision under uncertainties. It has been demonstrated that the proposed cooperative AI-enabled decision making platform can increase the safety and average traffic

delay as compared to the individual automated and human-operated traffic systems.

Furthermore, in highly automated driving systems, it is important to enhance the trust towards automation process, and one promising approach in this direction could be to enhance the situational awareness by displaying the spatial information of close traffic objects via a vibrotactile display [191]. Since the levels of trust in automation vary dynamically depending on the knowledge about the surrounding traffic status, the display of spatial information of nearby vehicles in a vibro-tactile display captured via a haptic stimulus can be significantly useful in designing an automated driving system. Another crucial aspect in automated driving is to ensure reliable interactions between human drivers and automated driving systems so that possible collisions due to the divergence of actions taken by human drivers and automated driving system can be avoided. The existing research works related to such interactions mainly follow the experimental approaches, which are usually expensive and time consuming. This has led to the need of efficient models for future automated driving technologies, which can predict and interpret the human driver's interaction with the automated driving system [170].

In order to adapt the task-based design of autonomous vehicles in dynamic driving situations, it is crucial to gather 3D information about the road and surrounding vehicles accurately in real-time with the help of vehicular to infrastructure (V2I) or vehicle to vehicle (V2V) communications. However, most of the existing methods to capture the 3D road perception focus on a single task/aspect even if that particular aspect is not so important, thus leading to larger delay for autonomous vehicles in completing all the required tasks [192]. Although vision-based methods benefit from the use of deep learning, they suffer from the loss of 3D information. In this regard, multi-task deep learning [169] could be promising due to its potential to improve the performance of the individual tasks and to enhance the overall efficiency of the network. Pedestrian detection and estimation of time to cross the street are important issues to be addressed in the design of autonomous vehicle systems. The DL-enabled task-based design should consider a detection model, a classification model and prediction model, which deal with the localization and recognition of the pedestrians, distinguishing the pedestrian actions and estimation of pedestrian actions, respectively [169]. Also, a loss function considering the learnable weight of each task can be utilized to train the underlying deep neural network in order to enhance the performance of individual tasks and to balance the loss of each task.

Furthermore, heterogeneous service requests coming from the autonomous vehicle users comprise multiple tasks which are dependent on the availability of the limited resources. The task-based design with the success ratio of task execution as the target performance metric can be carried out at the autonomous vehicle cloud, which is connected with the underlying V2I and V2X architectures [193]. Although cluster head of V2V architecture, which connect with the autonomous vehicle cloud, can allocate tasks to the vehicles for execution and can predict the task execution time, it may not provide accurate prediction of task completion time due to the limited storage and computing resources. Also, in the V2I architecture comprising vehicles and infrastructure nodes, mobility of vehicles poses challenges in designing efficient task-allocation strategies. In this regard, the efficient allocation of tasks while considering the communication node's stability and computing capability with the objective of minimizing the task completion time required to guarantee the smooth execution of requested services and to enhance the task execution success ratio is an interesting research problem [193]. It should be noted that in contrast to most existing works focused on offloading tasks in edge computing environments, the focus of this paper is on designing a task-oriented communication network with the objective of maximizing the task effectiveness metrics, i.e., the performance metrics which can maximize the task-oriented reward.

In cooperative automated driving applications, vehicles need to communicate not only with other vehicles but also other road-users including bicycles, motorcycles, pedestrians and road-side IoT units over the underlying 5G-V2X or short-range communications networks [176]. The main crucial aspects to be considered during task-based design of these applications include how effectively vehicles coordinate with other vehicles, pedestrians and road-side units, how inter-vehicular cooperation can be utilized for better situational awareness of road-side conditions and how effective is the designed task-oriented protocol, see e.g. this work [25], in terms of giving collision warning at the intersections, defining mechanisms for overtaking, merging traffic and platooning in the highways, and designing policy rules for governing road traffic as well as ethical and trustworthy interactions with the central unit/cloud server.

2.4.4 Distributed Learning Systems

Distributed learning systems have emerged due to the growing size of training data sets. Careful design of the communication workload and privacy-preservation of the clients/edge

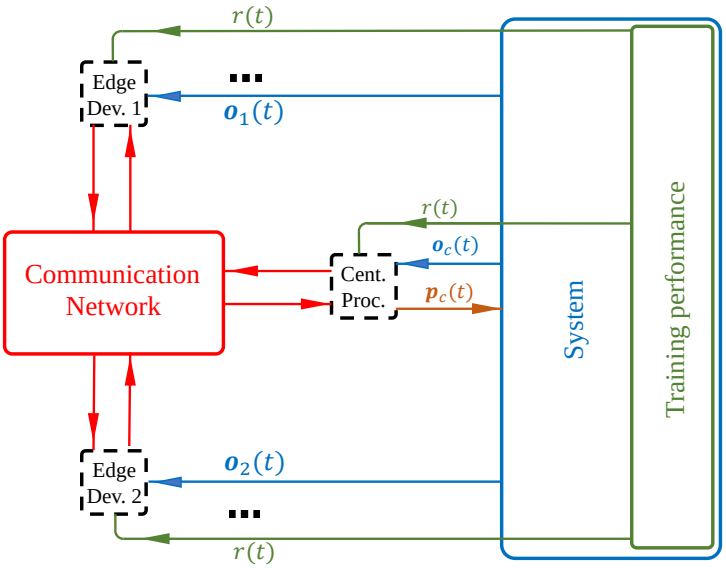


Figure 2.7: A centralized federated learning system illustrated using the task-based communication framework.

devices are sought to be the enabling means in these systems [194, 195]. Lately, federated learning has received much attention as an alternative setting: a parameterized global model is trained under the coordination of a parameter server with a loose federation of participating edge devices/clients [194–196]. Because of the major role that communication plays in all variants of distributed learning systems, it is one of the active areas of research, where direct task-oriented communication design has proven significantly efficient. This subsection details, with examples from the literature, why/how task-based communication design is helpful for distributed learning systems.

In federated learning systems, as illustrated in Fig. 2.7, the global model is trained using a number of different data points $\{X_l(t)\}_{l=1}^{n'}$ that are spread over different edge devices at time t. Clients are allowed to send communication messages $\mathbf{c}(t)$ to the parameter server as to facilitate the convergence of the model parameters $\mathbf{p}(t)$. The training data set $\{X_{l,i}(t)\}_{l=1}^{n'_i}$ available at client i within time step t, together with the latest available model parameters $\mathbf{p}(t-1)$ are interpreted here as the local observations of that agent/edge device in our universal framework i.e., $\mathbf{o}_i(t) = \langle \{X_{l,i}(t)\}_{l=1}^{n'_i}, \mathbf{p}(t-1) \rangle$. We also define the state of the system to be the collection of all data points available at all edge devices, that is, the state is jointly observable by all clients. This definition of state is technically correct since (state-probation pairs $\mathbf{s}(t), \mathbf{p}(t)$ will still be jointly sufficient statistics for the stage cost $r(\mathbf{s}(t), \mathbf{p}(t))$ and next stage state $\mathbf{s}(t+1)$. Note that the stage cost $r(\mathbf{s}(t), \mathbf{p}(t))$ here captures the expected loss

corresponding to our training model caused at all data points $\{X_l(t)\}_{l=1}^{n'}$.

In connection with our universal framework, the task of a distributed learning system is to optimize a parameterized model by solving

$$\underset{\pi^{(C)},\pi^{(P)}}{\operatorname{argmin}} \quad \mathbb{E}_{\mathbf{s}}\{\mathcal{J}(\mathbf{p}(t),\hat{\mathbf{s}}(t))\} \tag{2.2}$$

where $\mathcal{J}(\mathbf{p}(t),\hat{\mathbf{s}}(t)) = \sum_{t=t_0}^{t_{MAX}} r(\hat{\mathbf{s}}(t),\mathbf{p}(t))$ captures the sum of losses corresponding to our training model caused at the data points $\{X_l(t)\}_{l=1}^{n'}$, at all times, and $\pi^{(C)},\pi^{(P)}$ stand for the communications and action policies defined in Section 2.3.

A naive strategy to carry out the communications between an edge device and the parameter server is to communicate all the local data of each edge device. The communication strategy, however, can be much more efficient if each edge device computes an update to the current global model maintained, and only communicates the update [197]. In standard federated learning, this is done by applying SGD distributively over the local data set $\{X_{l,i}(t)\}_{l=1}^{n_i}$ available at each edge device i and communicating the average gradient of the loss function at each node i and iteration t, to the parameter server

$$\mathbf{c}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} \nabla_{\mathbf{p}}\mathcal{J}(\mathbf{p}(t),\mathbf{o}_{i,j}(t)). \tag{2.3}$$

In fact, the large size of data-sets and privacy of edge-devices as well as possible changes in data-sets are the main reasons that we are not willing to communicate all the state information to the parameter server. In this sense, even the very first variants of federated learning [194–196] introduced some form of task-based communication, where communication messages $\mathbf{c}_i(t)$ of the clients to the central controller, are of much smaller size than the observations of clients and yet the task can be accomplished with no compromise on the performance. That is, the size of observations of each client is thousands of times larger than the size of gradient updates being communicated.

However, these techniques applied, the communication between the edge devices and the parameter server is yet seen to be a major bottleneck. The authors in [131], together with similar works [132–137], introduced schemes to reduce the size of communication messages beyond the standard federated learning. These methods are shown to enhance the speed of convergence as well as to overcome communication bandwidth constraints.

The authors of [131] introduced a particular coding scheme that heuristically identifies and sends the important gradients as to reduce the size of communication messages. Following the proposed scheme, the authors in [131] reported a significant task-oriented compression ratio while virtually no loss is seen in optimizing the cost function. This scheme first finds a threshold for the magnitude of the gradient vectors, above which the computed gradient of a node will be considered for communication to the master node. The gradient vectors with lower size than the threshold, however, will remain in the node and will be accumulated with the rest of gradients that will/already have computed by the same node and have not be qualified for transmission. While in [132–134], the communication channels are considered to be rate-limited but error-free, works done in [135–137] (partially) consider the effects of the physical layer features of the communication links on the problem of federated learning [130].

Another useful way of adopting TOCD for the purpose of distributed learning systems is to consider the value/importance of data-sets available at each client to optimize the resource management in the communication network [198]. As an instance, the authors in [199] and [200] optimize user scheduling by incorporating the importance/value of the clients' data-set for the estimation taking place at the server. In particular, the work in [199] introduced an indicator to capture the importance of data-set of each client, according to which scheduling of client-server communications is optimized. One unique aspect of the proposed metric, is that the scheme also considers the quality of the communication channel between a client and server to design the importance indicator.

One way to look at TOCD for federated learning is that a form of compression of the input data is carried out by TOCD such that we can still obtain sufficient statistics about their corresponding labels. While recent research results testify the applicability of information bottleneck with the same purpose on deep neural networks [201, 202], very few research is done to harness the potential of information bottlenecks within the distributed learning systems [130].

2.4.5 Over-The-Air Computation in Smart Manufacturing Plants

In IoT networks, massive amounts of data are generated, collected, and leveraged to help complete a predefined task. For example, in smart manufacturing plants, wireless data needs to be collected from thousands of sensors. However, we are not interested in the value of each

individual data source, instead, we aim to obtain the "fusion" of the information contained in all data sources, e.g., computing sums or arithmetic averages. On one hand, transferring raw measurements from a large amount of different data sources to the same data collector is not spectrum efficient, especially when the measurements can be encoded to small data packets. On the other hand, the computation of massive data in an edge device as data collector with limited computation capacity can be also challenging.

Therefore, the technique called over-the-air computation (AirComp) has been developed to enable communication- and computation-efficient data fusion of the sensing data from large amount of the concurrent sensor transmissions. It takes into account the underlying task, such as computing a function, directly at the physical layer, by exploiting the superposition property of the wireless channel. In other words, it allows an efficient target function computation over the "air".

The AirComp is defined as follows. Consider K wireless sensors, each having a measurement signal $s_k \in \mathbb{R}$, $k \in \mathcal{K} := \{1, 2, \ldots, K\}$ to send. On the receiver side, we expect to derive the function of the measurements of the form

$$f(s_1, \ldots, s_k) = F\left(\sum_{k=1}^{K} f_k(s_k)\right). \tag{2.4}$$

Given the multiple access communication channel $h : \mathbb{C}^K \to \mathbb{C}$, AirComp aims at finding a set of pre-processing functions $\zeta_k : \mathbb{R} \to \mathbb{C}$ and a post-processing function $\psi : \mathbb{C} \to \mathbb{R}$ such that

$$f(s_1, \ldots, s_k) = \psi\Big(h(\zeta_1(s_1) \ldots, \zeta_K(s_K))\Big). \tag{2.5}$$

With the pre-processing of the measurement signal, as well as the post-processing of the received channel output, we can directly obtain the desired computation of the target function in (2.4) by effectively integrating the communication and the computation policies. Moreover, in this way the receiver's computational task $f(\cdot)$ defined in (2.4) of processing K signals is decomposed into $K + 1$ small tasks $\{\zeta_1(\cdot), \ldots, \zeta_K(\cdot), \psi(\cdot)\}$ that can be distributed among sensors and the receiver, with each of them processing only one signal.

Without loss of generality, AirComp also falls into the class of task-oriented communications system design. The task is to compute $f(s_1, \ldots, s_k)$ at the receiver (e.g., a center unit) side. The problem defined above can be aligned with the task-oriented communication

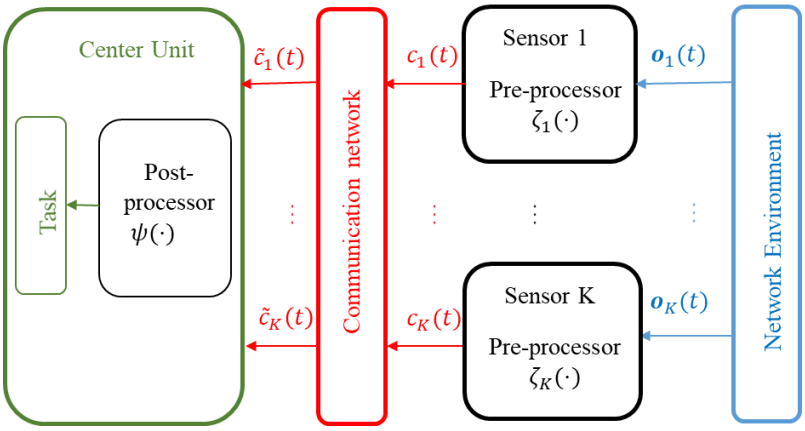


Figure 2.8: Over-the-air computation in smart manufacturing plants aligned with the task-oriented communications system design

framework defined in Section 2.3 by simply integrating the computing policy into the communication policy, i.e., the encoding (pre-processing) and decoding (post-processing) policies included in the communication policy already provide the direct action to compute the desired target function as shown in Fig. 2.8.

The authors of [144] derived theoretical bounds on the mean squared error for a certain AirComp function (sum of the signals) computation in a fast-fading scenario with CSI available at the transmitter. In [145], the AirComp problem is tackled without the explicit knowledge of channel information but under the assumption of slow fading. Then, in [146] a AirComp scheme was proposed for distribution approximation of a larger class of functions than the previous works with theoretically proven bound over fast-fading channels that can deal with correlated fading and requires no CSI.

A popular application of AirComp computational techniques is the computation of distributed gradient descent to solve the empirical risk minimization problem for machine learning models. One example is the training of neural networks in a central mode but with distributed data reported by a large number of local agents such as sensors in the smart manufacturing plant. In [136], the authors proposed to use AirComp computation over wireless channels to help efficiently compute distributed stochastic gradient descent in the federated learning paradigm. The author in [203] extended the idea to channels with fading channel information at either the transmitter or receiver side. In [146], the authors showed

the application of a proposed AirComp scheme to the regressor and classifiers in vertical federated learning.

2.4.6 5G and Beyond Self-Organizing Networks

Many works proposed reinforcement learning-based solutions for various self-organizing network (SON) use cases, as well as the coordination between the SON functions [141–143]. However, in these works, the selection of the network state information and the optimization of the learning function were considered as two independent processes. Expert knowledge is exploited to select the features and use them as the network state information, which may cause either insufficient information, thus poor optimization performance, or too much redundant information, thus high complexity of the learning model.

The task-oriented communications system design can be leveraged to jointly optimize the information exaction and control optimization problems. Let us take the SON function mobility load balancing (MLB) as an example. MLB is a function where cells suffering congestion can transfer load to other cells which have spare resources, by adjusting their handover (HO) control parameters (for details of HO parameters refer to [204]). Many works have proposed to solve the multi-agent MLB problem with the following centralized model and manually selected observations by using deep reinforcement learning [138–140]. Given a set of cell sites (hereafter referred to as cells) $\mathcal{K} = \{1, 2, \ldots, K\}$, each cell can configure its HO control parameters $\boldsymbol{p}_k \in \mathbb{R}^N$ and obtain a partial observation $\boldsymbol{o}_k \in \mathbb{R}^M$ of the global network state $\boldsymbol{s} \in \mathcal{S}$, where the entropy yields $H(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K) = H(\boldsymbol{s})$ and $H(\boldsymbol{o}_k) < H(\boldsymbol{s})$ for $k \in \mathcal{K}$. Let $\boldsymbol{p} := [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K]$ and $\boldsymbol{o} := [\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K]$ denotes the collections of HO control parameters and observations in the multi-cell network system respectively. The objective is to minimize the global cost function $\mathcal{J}(\boldsymbol{s}, \boldsymbol{p})$ with the information contained in the observations $\boldsymbol{o}$. It is obvious that to learn a centralized model that takes all actions and observations $[\boldsymbol{p}; \boldsymbol{o}] \in \mathbb{R}^{(NM)^K}$ into account will lead to an extremely high model complexity.

To reduce the computational complexity, a task-oriented communication design can be considered as shown in Fig. 2.9. Such design enables distributed execution of the joint training of the local communication and control polices $\pi_k^{(c,a)}$, $k \in \mathcal{K}$. At time slot t, each cell observes a partial information $\boldsymbol{o}_k(t)$ from the environment, with the policy $\pi_k^{(c,a)}(t)$, it derives an encoded message $\boldsymbol{c}_k(t) \in \mathcal{C}_k$ which extracts the information of the local observation, and

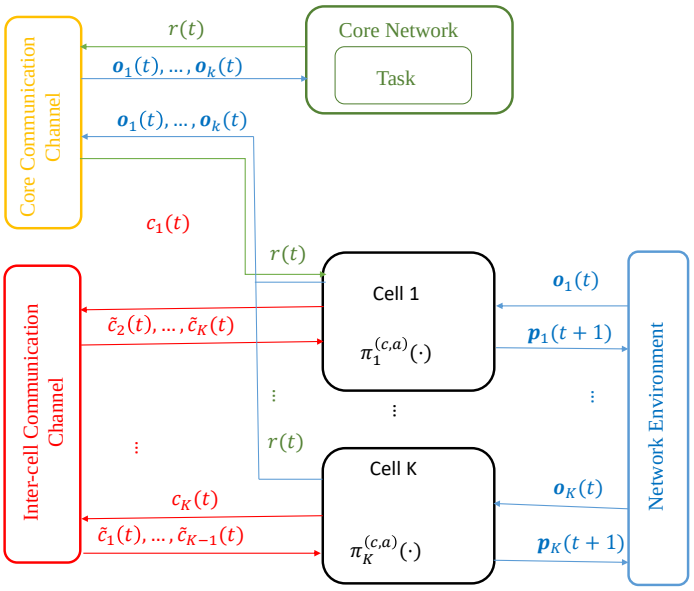


Figure 2.9: Task-oriented communication design for mobility load balancing problem in 5G and beyond SON

sends it to other cells via inter-cell communication channel (e.g., the X2 interface in 5G networks). After receiving the messages from the other cells $\{\tilde{\boldsymbol{c}}_l : l \in \mathcal{K} \setminus \{k\}\}$ (note that with lossless channel we may have $\tilde{\boldsymbol{c}}_k = \boldsymbol{c}_k$), the local policy $\pi_k^{(c,a)}(t)$ also decides the new configuration of the HO parameters $\boldsymbol{p}_k(t+1)$. On the other hand, to evaluate the global performance and compute the reward for all cells, each cell sends its observation to the core network (we assume lossless channel between the cells and the core network since they usually communicate with wired connection), and the core network who defines the task computes the common reward $r(t)$ based on the received observations, and sends it to the cells as the feedback of the current state $\boldsymbol{s}(t)$ and joint actions $\boldsymbol{p}(t)$. Each cell learns jointly a policy $\pi_k^{(c,a)}(t) : \mathbb{R}^M \times \Pi_{l \in \mathcal{K} \setminus \{k\}} \mathcal{C}_l \to \mathcal{C}_k \times \mathbb{R}^N$ that maps the local observation $\boldsymbol{o}_k$ and received messages $\{\boldsymbol{c}_l : l \in \mathcal{K} \setminus \{k\}\}$ to a tuple of the local encoded message $\boldsymbol{c}_k$ and local control parameters $\boldsymbol{p}_k$. In this way, the model in each cell has an input space with the cardinality $|R|^M \prod_{l \in \mathcal{K} \setminus \{k\}} |\mathcal{C}_l|$, which is dramatically smaller than the input space cardinality $|R|^{(NM)^K}$ of the aforementioned centralized training based on the full observations and actions $(\boldsymbol{o}, \boldsymbol{p})$.

2.4.7 Tactile Internet

To enhance the degree of immersion of the user in distant communications, it is known that the communication of haptic information can play a crucial role. In the well-known scenario of teleoperation/telepresence, a human user interacts with a remote environment through: (i) a human system interface, (ii) a communication link, and (iii) the teleoperator. Such interactions involve both communications messages and action policies and should be carefully designed. Fig. 2.10 presents how the proposed task-oriented communication framework can be used to model Tactile Internet scenario.

When teleoperation is performed over a communication channel with potential delays, noise, and uncertainties, it can be shown that achieving good performance in the teleoperation task can be formulated as a task-oriented communication problem [18]. Haptic information can be divided in two different classes. The first class, called kinesthetic information, includes data related to muscle activation and movements. The second class, called tactile data, refers to the perception of pressure, texture, and temperature [205].

Communication of the first category of haptic data, kinesthetic data, involves the movement of an actuator of the teleoperator in such a way that a particular task is done with the best possible performance. Some examples for the tasks that require communication of kinesthetic data can be medical teleoperations or playing a musical instrument remotely. In the both of the mentioned examples, achieving a good level of performance in the task is not equivalent to reducing the distortion between a reference (desired) action signal and the controlled action process. While a wrong movement of the teleoperator parallel to the surface of a piano keys generates no undesired musical note, an error in the controlled action process along the axis vertical to the surface of the piano can generate a wrong musical note. Let us recall that even wrong movements vertical to surface of the piano, depending on the location of the actuator and the magnitude of actuation error, may or may not cause any cost in task. Accordingly, the cost function of the task cannot be simply characterized by a mean squared error of the action process. Instead, a task-oriented cost function should be considered to achieve (close to optimal) performance in the task. Moreover, the communication of kinesthetic data is of one particular sensitivity. Due to the stability requirements of the control loop on the teleoperator end, signal processing algorithms that are used at both sending and receiving ends should not cause large algorithmic delays [147, 148].

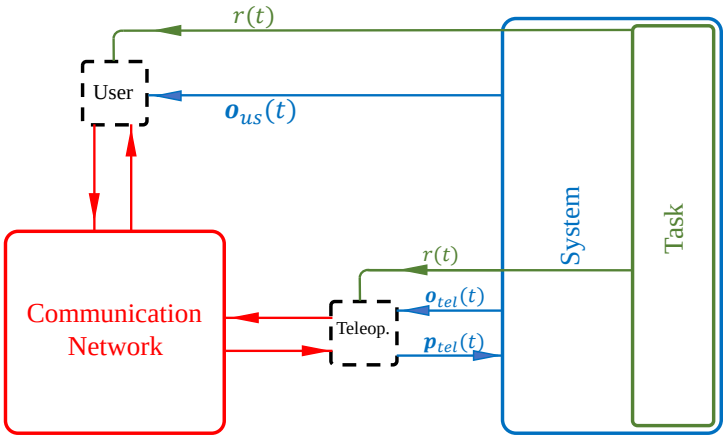


Figure 2.10: TOCD framework applied to the Tactile Internet.

On the other hand, task-oriented communication of tactile information, the second category of haptic information, can also contribute to the improvement of the system performance. In this case, the cost function of the task is imposed by the perceptive abilities of an average human. Although there might be a difference between the communicated tactile signal and the signal received at the other end of the communication channel, if the difference between the two signal is beyond the perception of an average human, no cost should be associated to this communication error [152, 153]. The authors in [153] utilize the deadband principle to reduce the rate of communication of haptic information. The deadband principle is understood based on the perceptive abilities of an average human. According to this principle, the haptic information is transmitted only if the difference between the last communicated haptic data and the current available haptic data is perceptible by the human operator. While it is obvious that widening the deadband will result in more distortion, this may not affect the performance of task, which can be measured by the level of preciseness the task is performed by the user [148]. Accordingly, finding the relationship between the achieved compression rate of the haptic information and the performance of the task is considered to be of substantial value [148].

The above-mentioned relationship is studied in [5], under several assumptions: (i) the uplink channel is noise-free, (ii) the cost function is considered to be a regularized quadratic function of the error in state of the system, (iii) the downlink channel noise is an additive white Gaussian noise, and (iv) and the system state is generated by a stochastic linear model.

2.5 Challenges and Open Problems

In this section, we envision challenges in developing efficient task-oriented communications solutions in future cyber-physical systems. After that, we present various potential research problems based on the task-oriented design framework. We enumerate the existing challenges in two different categories, fundamental challenges of the framework and application-specific challenges.

2.5.1 Challenges of the Framework

Indirect Design

One advantageous aspect of task-agnostic communication design has been its capability to perform fairly well across all possible tasks - simply because of the fact that the specifics of the task were never taken into account and the communication systems were supposed to transmit all potentially useful/useless to the receiving end. As attested by [206], "a unified framework to support various tasks is still missing in multi-user semantic communications.". While the mission of the task-oriented communication design is to tailor effective ways of communication tailored to each specific task, we want to avoid designing all communication layers for every single task. The indirect design of task-oriented communications allows for addressing all/ a wide range of tasks using a unified framework. Unlike earlier task-oriented quantization techniques that tailor a quantization scheme to certain applications [29], this work proposes an *indirect* design for its task-oriented quantization scheme - SAIC, ABSA and, ESAIC. The *indirect* design is carried out in a fashion that it never benefits from any explicit domain knowledge about any specific task e.g., geometric consensus problems. Accordingly, the *indirect* design of the algorithms allows them to be effectively applied beyond the geometric consensus problems and to a much wider range of tasks where the conditions of the algorithms are met.

This thesis attempts to take the first steps towards designing an *indirect* task-effective data compression theory. While the data compression algorithm proposed by this thesis is designed in an *indirect*[3] fashion i.e., not for a specific task, we demonstrate its applicability in

[3]By using the word *indirect* here we are not referring to the concept of indirect access to the source of information [207] - this usage of the word falls in the nomenclature of source coding and information theory. In fact, we are referring to the concept being introduced by the control theory nomenclature in which an indirect

a specific task: a geometric consensus problem under finite observability [208] in all chapters 3, 4 and, 5. While being a well-defined mathematical object, the value $V(\cdot)$ of observations - defined in chapters 3 and 5, is an indirect metric of measuring the value of observation information across different control tasks defined on MDPs. Also in chapter 4, we develop novel indirect KPIs to measure the effectiveness of a task-oriented data quantization scheme across different tasks.

Scalability

Although not a surprising challenge, the scalability of the envisioned methodology is of paramount importance, especially in cases of large networks of agents with limited capabilities. A substantial research effort is put to address the issue of scalability in multi-agent systems [209–212, 212, 213]. The overall complexity of MARL, however, increases with the addition of each agent to the system. Even with the use of attention mechanisms for the centralized training authors have not been able to go beyond linearly increasing the complexity of the centralized training with respect to the number of agents [214]. The issue of scalability of the design of communications for a multi-agent system is properly studied and addressed in 5. To the best of our knowledge, this chapter is the first to reduce the complexity of the centralized training phase from exponential time complexity - $\mathcal{O}(|\mathcal{P}|^K \times \prod_{k\in\mathcal{K}}|\mathcal{S}_k|$ - to constant time complexity - $\mathcal{O}(1)$ - with respect to the number of agents.

Reward Signals

The current framework assumes the availability of the common reward for all agents at no cost. In practice, the reward has to be sent to the agents via (wireless) communication channels. Therefore, the distribution of rewards among the agents must be taken into account when designing the communication strategy in such cases. Another challenge in designing a task-oriented communications system is how to properly take into consideration different reward functions for different agents. In an extreme case, the agents can be competitive rather than collaborative [215]. In these cases, the communications should be adapted to

design is generic enough to be used for unmodelled system dynamics and not a certain dynamic [115]. Thus the schemes - such as SAIC - which enjoy an indirect design can be applied to all/a wider range of tasks. In contrast to indirect schemes, "the direct schemes aim at guaranteeing or improving the performance of the cyber-physical system at a particular task by designing a task-tailored communication strategy" [24].

specify individual target, which in turn affects both action and communication policies of the agents.

One other issue related to the reward signal is that sometimes visits to the desired state-action pairs - that hold the large rewards - are very unlikely to happen during the training phase. This is essentially due to the very large state-action space of the underlying MDP. To solve this problem, a very large number of training episodes are required - making the RL algorithms extremely slow in solving the problem. To solve this issue, reward shaping is proposed in the literature of RL [216]. Reward shaping is the practice to design a second reward signal $r_2(\cdot) : \mathcal{S} \times \mathcal{P} \times \mathcal{S} \to \mathbb{R}$ using which the agent is trained. This reward signal together with other main attributes of the environment create a virtual emulator that the agent(s) is/are trained on. While the ultimate goal is still to maximise the average discounted sum of $r(\cdot)$ rewards, the second reward signal is used to improve the training of the agent(s). The aim of designing this second reward signal is to guide the agent(s) towards learning a better (or optimal) policy in a faster way. While the above-mentioned challenges about the reward signal are still the area of active research [217,218], we have provided some preliminary results on the issue of reward shaping for multi-agent systems in 3.5.3.

Training

Another issue is how the training is performed. It is well known that centralized training can provide the optimal solution for the multi-agent system. However, the centralized training architecture is not always available, especially in dynamic multi-agent systems. This requires efficient distributed training design, in which the communications among the agents during the training phase is crucial and needs to be properly designed. So far, the communication is optimized based on the codebook and resource allocation philosophy. While this method is efficient for point-to-point and small networks, it is difficult to generalize to massive networks or take latency and privacy aspects into account. Further, how to configure the learning rate if each agent when training a multi-agent system is another open problem. While usually, a symmetric learning rate is considered for multi-agent reinforcement learning systems [4, 9], there are particular examples where asymmetric learning rates are proven to outperform [219]. Very little is known about how changes in the learning rates would affect the performance of the multi-agent system. Optimizing the learning rates for different agents in the MAS has

been out of the scope of this thesis.

Intrinsic Features of the Task

In task-oriented approaches, we are aiming to maximize system performance towards the task-related KPIs - given all the constraints on the communication resources, e.g., rate and latency. One of the unique aspects of the task-oriented communication problems that were absent in the task-agnostic "technical problem[4]", is the remarkable impact that the characteristics of the task can have on the analytical studies. Oftentimes, the existing works on the task-oriented communication design, cannot predict/guarantee how their prescribed solutions would perform in general at every possible task. That is to say, before we apply a certain task-oriented quantization scheme to a specific task, our analysis can hardly how effective the algorithm is in reducing the minimum information rate required for the communications across users/agents [9,11]. There are some intrinsic features in all tasks that would change the extent to which a part of the original communication message can be discarded (see e.g., [46]) and how they describe the "semantics of information" metric to be a "context-dependent" metric that maps qualitative information attributes to their "application dependent value". It is, however, not perfectly clear, yet, what these intrinsic task features are and how they impact the extent to which we can improve the effectiveness of using communication resources. Finding these features and how they influence the extent to which communication resources can be used effectively are some of the unique theoretical challenges faced in this framework. These challenges are not addressed in this thesis and will stay open for future research.

System Memory

While many of the existing solutions for task-oriented communication design are relying on the memorylessness of the source of information [118, 119], this assumption is violated in almost every control task, where the current state of the system is determined based on its previous state(s) and the latest control decision made. Violating this assumption makes it hard (in general) to use classical tools offered by information theory that rely on asymptotic equipartition property property of the information processes.

[4] As mentioned by Shannon in his landmark work [220]

Temporal Dynamics

In complex topologies, the temporal dynamics of the communication network are an important challenge. The physical mobility of certain agents is a prime factor for the temporal dynamics and in combination with the excessive environment state space can lead to volatile behavior. In this context, communication diversity and connectivity prediction tools, directly dimensioned by their impact on the overall task effectiveness, should be able to produce stable/robust communication and action policies.

2.5.2 Application-Specific Challenges

Industrial Internet-of-Things Networks

The agents in IIoT systems differ from other applications in two main features: light computational capability and limited energy budget. Consequently, energy-efficient communications is expected to be the first design criteria for IIoT applications. Meanwhile, IIoT systems usually demand timely decision-making actions and consequently stringent latency, which contradicts to the energy-efficiency target. This trade-off between energy efficiency and reliability asks for new design method for IIoT networks. Ultra-reliable low latency communications (URLLC) is a promising solution for designing the communications in multi-agent IIoT systems. However, the current URLLC design does not take into account the requirements for the learning process. Therefore, the URLLC property should be jointly considered with the action policies considering the energy and computation capacities of IIoT nodes.

UAV Communications Networks

UAV communications plays an essential role in the future communications networks as not only a stand-alone system in dedicated areas but also complementary parts of the cellular networks [221]. More specifically, multiple UAVs can cooperate to provide communications in isolated areas for rescuing or sensing purposes, or they can aid the macro BSs to enhance handover or provide ultra-reliable communications to ground users [222]. In either cases, reliable communications among the UAVs is key to improve the overall performance of the UAV networks. One challenge in UAV communications networks is how to jointly design the trajectory of every UAV to reduce collision risk and improve the system energy efficiency [223].

However, since the UAVs usually operate in dynamic environment, conventional methods may not be applied due to the lack of proper system modeling. In fact, the UAVs' optimal trajectories are difficult, and sometimes unable to obtain as they depend on the movements and actions of all the UAVs. Therefore, one efficient way is that the UAV optimizes its trajectory and action policy while listening to the others, which requires efficient communications among the UAVs. Efficient communications should cover both how and what to communicate. On one hand, the UAVs should have to be well coordinated in accessing the common channel to avoid interference. On the other hand, each UAV has to determine what message to communicate to other UAVs in order to maximize the UAVs' collaboration. As a result, this asks for a novel design paradigm that optimizes the communications based on specific semantics.

Federated Learning

Federated learning (FL) is an emerging distributed ML framework that allow a large number of edge nodes collaboratively train a shared learning model [224]. FL is capable of addressing many challenges in implementing ML over networks [225, 226]. One of they key challenges in FL is non independently and identically distribute (non-i.i.d.) data among devices and resource constraints [227–229]. Because the update at the edge nodes are based on their locally available data, the contribution to the aggregated system parameters varies from one edge node to another. As a result, always-transmit policy is no longer the optimal transmission in FL. In fact, a node which does not have sufficient data usually generates bias gradient parameters. Sending these parameters to the server does not improve, and sometimes harms the learning process. This asks for a context-aware transmission policy to tackle this issue. It is shown in [230, 231] that a proper transmission policy can improve the FL in terms of both energy efficiency and convergence performance. Another issue in FL is the different privacy levels among the edge nodes [232, 233], which requires unequal-protection transmission and coding designs. Optimizing the source-channel coding to satisfy the privacy requirements and improve the learning convergence is usually difficult and requires novel system design perspectives.

Mobile Edge Computing

MEC will be a key component in the 5GB architecture to implement the intelligence on the network edge. By being equipped with both computational and storage capabilities, MEC nodes are able to determine to perform requested tasks locally or to offload them to the cloud server. Most of the existing works consider single MEC node that can optimally make task offloading decision. When applying to multiple cooperative MEC nodes, such methods, however, no longer render optimal policies. This is because in multiple MEC agents context, one node's action can have affects on other nodes. Furthermore, as the MEC nodes usually share the same communication medium, e.g., channel bandwidth, it requires proper transmission design to efficiently mitigate interference among the MEC agents. The major challenge is how to jointly design the MEC's local action policy with communications policy to balance the exploration-exploitation tradeoff. The authors of [234] have shown the potential of such joint design for IIoT systems, in which edge-device acts as a machine-type agent (MTA). The MTAs collaboratively learn optimal policy for channel access and task offloading in multi sub-carrier D2D environment. At every time slot, each MTA determines to compute the task locally or to offload the task to the MEC server. By modeling the state space including offloading decisions, channel access status and computation task, the authors proposed to use multiagent deep deterministic policy gradients algorithm on actor-critic network at each MTA. then the MTAs exchange their locally trained model to generalize the global model. The benefit of multiagent MEC system presented in [234] is based on an assumption that all the edge nodes can perfectly communicate to each other during the training phase. In many practical cases, such assumption rarely occurs due to imperfect communications among the edge nodes and the highly dynamic network topology. This asks for novel distributed designs of action and communication policies.

5G and Beyond Self-Organizing Networks

Although it has been almost a decade since the concept of SONs was introduced to the next generation mobile networks (NGMN) and 3GPP standards [235], the existing SON solutions have not met the high expectation of the operators to achieve a fully self-aware cognitive network with automated configuration, monitoring, troubleshooting, and optimization. This is because, with the emergence of new wireless devices and applications, it is expected that a

large amount of measurements and signaling overhead will be generated in future networks, while partial and inaccurate network knowledge, together with the increasing complexity of envisioned wireless networks, pose one of the biggest challenges for SON – maintaining global network information at the level of autonomous network elements is simply illusive in large-scale and highly dynamic wireless networks. Another big challenge is the network-wide optimization of strongly interdependent network elements, with the goal of improving the efficiency of total algorithmic machinery on the network level. Thus, to deliver SON solution for the 5G and beyond networks, we need to answer the following two questions:

1. What information should be communicated among network elements to enable cooperation?

2. How to let the network elements achieve consensus with a limited number of probes to improve performance on the global network level?

Task-oriented communications system design can be leveraged to solve the above-mentioned two problems. Instead of considering the selection of the network state information and the optimization of the learning function as two independent processes, we can jointly optimize the information exaction and control optimization problems to improve both data efficiency and computational efficiency for large-scale highly interdependent network systems.

Autonomous Vehicles and Cooperative Automated Driving

As highlighted previously in Sec. IV-E, the main tasks to be considered in the design of autonomous vehicles include pedestrian detection, gathering of 3D information about surrounding environment, estimation of time to cross the street, action recognition, prediction, lane change decision and interaction of driver's interaction with the autonomous driving system. Furthermore, for cooperative driving systems, the important design tasks to be considered include cooperative communications among vehicles, infrastructure nodes and road-side units to avoid the collisions, cooperative platooning decisions under mobility and uncertainties and capturing situational awareness of the road-side information and required adaptation in dynamic situations are important design tasks to be considered. Furthermore, the positioning accuracy of autonomous vehicles may be impacted by issues like latency and packet loss in the underlying V2X communications networks; and in order to reduce the position errors

and possible collisions, it is crucial to design suitable cooperative driving and merging strategies [236]. Another key issue to be addressed is to design robust and reliable cooperative sensing in order to effectively localize the surrounding and road-side objects detected by the onboard sensors and neighboring vehicles based on the available limited data-set [237]. Furthermore, DRL-enabled design of motion planning for autonomous vehicles, comprising trajectory planning, control and strategic decisions, is another promising area for future research, in which the main tasks to be designed include the environmental modeling, generating model abstractions, realization of underlying neural networks, and modeling of states, actions and rewards [238]. Furthermore, from the practical implementation and business perspective, it is interesting to investigate a task-based design in order to enable the cooperation among the key players of autonomous vehicles such as car makers, telecommunication industries and policy makers by balancing their individual interests.

Advanced RL Techniques for Task-oriented Communications

Several advanced versions of RL including Inverse IRL, Safe RL and Multiagent-RL (MARL) seem promising to enable task-oriented communications design in the considered application domains, i.e., AV systems, multi-UAV networks and multi-sensory systems including TI. Among these, IRL determines a reward function to be optimized for a learning agent given either the set of measurements of the agent's behavior or sensory inputs over time in various situations, and benefits from better inherent transferability of the reward function in new dynamic environmental situations as compared to the learned policy in the standard RL, which might need to be discarded in the changed situations [239]. The main research issues associated with IRL along with some recommendations are included below: (i) existing works have mostly considered small-state spaces, which may not fully capture practical autonomous scenarios. To address this, one promising way would be to exploit the deeper version of IRL method by utilizing deep Q-network and other deep learning architectures [240]; (ii) learning a reward function is ambiguous as several reward functions may correspond to the same policy, resulting in the need of suitable accuracy metric to have the fair comparison of the agent's behavior generated from the inferred reward function with the true expert's behavior, and (iii) due to the involved iterative process comprising a constrained search over a space of reward functions, the complexity of the solution to the problem may grow disproportionately along with the problem size and number of required iterations, demanding for the need of

low-complexity methods.

On the other hand, safe RL deals with finding suitable learning policies for the problems/applications (i.e., robotic systems, AVs), where it is essential to respect safety constraints and/or to guarantee reasonable system performance. As compared to the standard RL whose main objective is the long-term return maximization based on the real-valued reward, safe RL aims to consider both the long-term reward maximization and safety involved in the underlying agents/system as the first objective does not necessarily avoid the risk/negative outcomes due to inherent uncertainty of the underlying environment [241]. The main issues in employing safe RL along with some future recommendations are included below: (i) the employed aggressive exploration policy in model-free RL techniques to construct an accurate model might lead to high-risk situations, and also construction of a reasonably accurate model while capturing the underlying dynamics in a safe manner could be problematic in a model-based RL. To address these, some promising approaches could be learning dynamics from the demonstrations and employing policy and relational RL methods with bootstrapping [241]; (ii) most existing solutions are designed for finite MDPs, however, learning in realistic environments needs to deal with the continuous state and action spaces, thus leading to the need of devising suitable safe RL approaches, which can handle continuous actions and state space; (iii) selection of a risk metric might put limitations in using a particular RL algorithm, limiting the applicability of a safe RL algorithm to a particular application domain. This leads to the need of investigating generalized risk metrics and safe RL algorithms, which can be applied across different application areas.

Similarly, MARL algorithms deal with the modeling of the multiple agents in the system, enabling coordination via information exchange among the agents and exploit the intention or hidden information of other agents' behaviors to have effective communications/cooperation. MARL techniques can find significant importance in various applications (i.e., AVs, cooperative cruise control, multi-UAVs) as they need to deal with the multiple learning agents and collaboration is essential among various learning agents in order to fully utilize the exploration space. Existing MARL techniques can be grouped under the frameworks of Markov/stochastic games and extensive-form games, and the former framework can be employed in three settings, namely cooperative (i.e., based on a common reward function or team-average reward), competitive (i.e., zero-sum Markov game) and mixed settings (general-sum game setting) [242]. The main issues associated with MARL along with some recommendations are highlighted

below: (i) MARL algorithms may need to consider multi-dimensional goals, which can be often unclear, and they may fail to converge to the stationary Nash Equilibrium of general-sum Markov games. One approach to analyze the convergence behavior of MARL techniques is to utilize the concept of cyclic equilibrium; (ii) the stationarity assumption behind the convergence of single-agent RL methods becomes no longer valid for MARL methods, demanding for new mathematical tools for MARL analysis; (iii) joint state-action space needs to be taken into account by each agent, however, the dimension of this joint space can increase exponentially with the increase in the number of agents, leading to the issue of combinatorial MARL problem. One approach to tackle this scalability issue it to utilize deep neural networks to design MARL algorithms [243].

2.6 Conclusion

Task-oriented communication has been considered to be a new paradigm for designing communications strategies for multi-agent cyber-physical systems. In this article, we have presented a comprehensive review and classification of the theoretical works across a wide range of research communities. We have then proposed a general conceptual framework for designing a task-oriented system and adapted it for the targeted use cases. Furthermore, we have provided a survey of relevant contributions in eight major application areas. Finally, we have discussed challenges and open issues in the task-oriented communications design.

Chapter 3

Task-Oriented Data Compression for Multi-Agent Communications Over Bit-Budgeted Channels

3.1 Introduction

The design of traditional communication systems has often been carried out according to task-agnostic principles. Information and coding theories drive the major analytical and design techniques, where the former sets the upper bounds on the system capacity, and the latter focuses on techniques for approaching the bounds with infinitesimal error probabilities. Accordingly, digital communications have made astonishing strides in terms of performance, enabling robust information transmission even under adverse channel conditions. However, in the era of cyber-physical systems, the effectiveness of communications is not solely dictated by the traditional performance indicators (e.g., bit rate, latency, jitter, fairness etc.), but most importantly by the efficient completion of the task in hand, e.g., remotely controlling a robot, automating a production line or collaboratively sensing/communicating through a drone swarm.

Machine-to-machine communications occur since the received signals can help the receiving end to make more informed decisions or more precise estimates/computations. In this context, the reliability of the communications is not essential beyond serving the specific needs of the control/estimation/computational task that the receiving end machine is trying

to accomplish. This calls for a fresh look into the design of communication systems that have been engineered with reliability as one of their ultimate goals. The emerging literature on semantic communications as well as goal/task-oriented communications is trying to take the first steps towards the above-mentioned goal, i.e., incorporating the semantics as well as the goal/usefulness of the message exchange into the design of communication systems [24, 43, 244]. By jointly analyzing the features of the collaborative task and the constraints on the underlying communication infrastructure, the communication strategies can be adapted or tailored such that they will be specifically effective for the task.

This chapter attempts to take the first steps towards designing an *indirect* task-effective data compression theory. While the data compression algorithm proposed by this chapter is designed in an *indirect*[1] fashion i.e., not for a specific task, we demonstrate its applicability in a specific task: a geometric consensus problem under finite observability [208]. As attested by [206], "a unified framework to support various tasks is still missing in multi-user semantic communications.". Unlike earlier task-oriented quantization techniques that tailor a quantization scheme to certain application [29], this work proposes an *indirect* design for its task-oriented quantization scheme - SAIC. The *indirect* design is carried out in a fashion that the it never benefits from any explicit domain knowledge about any specific task e.g., geometric consensus problems. Accordingly, the *indirect* design of the algorithms allows them to be applied beyond the geometric consensus problems and to a much wider range of tasks. The framework can be applied where a major communication bottleneck is in place between multiple cooperative decision makers. This bottleneck can occur due to a multitude of reasons (i) the energy lifetime of the communicating agents e.g., in the case of UAV/LEO satellite communications, that forces agents to communicate with low-energy high-range communication protocols [245, 246] (ii) the limitations imposed by the environment on the communication channel e.g., in space/underwater missions or (iii) limited communication resources of the network through which agents communicate. For more on the applications of TODC see [24, 247].

[1] By using the word *indirect* here we are not referring to the concept of indirect access to the source of information [207] - this usage of the word falls in the nomenclature of source coding and information theory. In fact, we are referring to the concept being introduced by the control theory nomenclature in which an indirect design is generic enough to be used for an unmodelled system dynamics and not a certain dynamic [115]. Thus the schemes - such as SAIC - which enjoy from an indirect design can be applied to all/a wider range of tasks. In contrast to indirect schemes, "the direct schemes aim at guaranteeing or improving the performance of the cyber-physical system at a particular task by designing a task-tailored communication strategy" [24].

3.1.1 Task-Oriented Data Compression

In particular, we consider a cooperative scenario where our goal is to optimize the expected return of a multi-agent system that is run on top of an underlying Markov decision process. The system's return is an unknown function of joint observations and control actions of all agents. The system's expected return can be controlled or optimized by selecting the proper joint controls actions at all agents. The partial observability of each agent together with their limitation to merely select local actions necessitates the presence of inter-agent communications to improve the coordination across the multi-agent system. We assume a full mesh communication network between all agents and that all the communication channels in the network are bit-budgeted but error-free. That is, the communication channels are all error-free fixed-rate bit pipes [248] and not variable rate bit-pipes [75] - the fixed rate of communications is constant across all inter-agent communication channels. Under these circumstances, rate-limited communication channels between agents drive the need for task-oriented data compression i.e., the usefulness of each message exchange should be incorporated into the design of the data compression strategy. The communicated messages between agents are useful only when they positively affect the decision-making of the receiving agents towards improving the system's expected return.

The problem we address would be a classic multi-agent Markov decision process (MAMDP) [249] if, each agent's communication message could include all the information inside the agent's observations. We assume, however, that the communication message of each agent is sent over a bit-budgeted communication channel i.e., per each channel use each agent will be able to reliably communicate a bit sequence with a length less than the entropy rate of the observation process. With this information constraint in place, it becomes imperative to carry out the communications at each agent such that they lead to the optimal expected return performance of the MAS. Each agent has to jointly select its control and quantized message at each time step with the aim of optimizing the expected return.

Due to the bit-budgeted communications between the agents, it is necessary for agents to compactly represent their observations in communication messages. As we ultimately measure the performance of the MAS in terms of the expected return, the loss of information caused by the compact representation of the agents' observations needs to be managed in such a way that it minimally affects the obtained return [5, 11]. As such, in this form of compression

scheme which we call task-oriented data compression, *the goal of abstraction is different from conventional compression schemes* whose ultimate aim is to reduce the distortion between the original signal and the decoded/reconstructed signal [250] - see [26, 29], where a similar task-based notion is introduced and a comparison of it with our work in Table 3.1.

3.1.2 Literature Review

As we study the joint communication and control design of a MAS, the topic of this chapter falls under the general category of multi-agent communications [37]. In contrast to many other cooperative multi-agent systems [251], the full state and action information are not available here to each agent. Accordingly, agents are required to carry out communication to overcome these barriers [37]. Earlier works used to address the coordination of multiple agents through a noise-free communication channel, where the agents follow an engineered communication strategy [252–256]. Later the impact of stochastic delays in multi-agent communication was considered on the multi-agent coordination [255], while [256] considers event-triggered local communications. Deep reinforcement learning with communication of the gradients of the agents' objective function was proposed in [3] to learn the communication among multiple agents. In contrast to the above-mentioned works, the presence of noise in the inter-agent communication channel was first studied by [4] where exact reinforcement learning was used to design the inter-agent communications. Later, the authors of [11] proposed a deep reinforcement learning approach to address a similar problem. Papers [4,11,26,29,35] and [13] have contributed to the rapidly emerging literature on task-oriented communications [24]. Noteworthy are also some novel metrics that are introduced in [257] to measure the *positive signaling* and *positive listening* amongst agents which learn how to communicate [3, 4, 13].

The current chapter can also be seen as designing a state aggregation algorithm. In this chapter, state aggregation enables each agent to compactly represent its observations through communication messages while maintaining their performance in the collaborative task. Classical state aggregation algorithms, however, have been used to reduce the complexity of the dynamic programming problems over MDPs [88, 258–260] as well as Partially Observable MDPs [89]. One similar work is [261], which studies a task-based quantization problem. In contrast to our work, the assumption there is that the parameter to be quantized is only measurable and cannot be controlled. In our problem, agents' observations stem from a generative process with memory, an MDP. Similarly, in [262], the authors have in-

troduced a gated mechanism so that reinforcement learning-aided agents reduce the rate of their communication by removing messages which are not beneficial for the team. However, their proposed approach mostly relies on numerical experiments. In contrast, this chapter relies on analytical studies to design a multi-agent communication policy which efficiently coordinates agents over a bit-budgeted channel - the benefits of our analytical approach are briefly explained in the contributions section 3.1.3. State aggregation algorithms are often developed for single-agent scenarios and are used to reduce the complexity of MDPs. To the best of our knowledge, we are the first to design a TODC algorithm using state-aggregation schemes. In particular, we use state-aggregation to design a data compression scheme to compactly represent the observation process of each agent in a multi-agent system.

Conventionally, the communication system design is disjoint from the distributed decision-making design [3, 252–255, 263]. The current work can also be interpreted as a demonstration of the potential of the joint design of the data compression/quantization and control policies. Determining the existence of a quantizer operating at a certain bit-budget to achieve a given figure of expected return is known to be an "intriguing open problem" [5] - even for single agent scenarios. Here we set a non-closed form upper bound on the expected-return performance of the multi-agent system given a quantization data rate/ the finite size of the discrete alphabet of the quantizer. We show how this joint quantization and control design problem is connected to minimizing an absolute error distortion measure via Theorem 1. A similar interpretation of the TODC problem can also be seen in [34]. While relevant, their setup is different from our work as they consider two distortion criteria for the rate-distortion problem.

We will show in section 3.2.2, that, in fact, the decentralized problem we target can be translated as the joint constrained design of the control policies as well as the observation function of a Dec-POMDP to maximize the expected return. While in classic Dec-POMDP problems the observation function is considered to be a fixed function [106], by a constrained design of the observation function, our problem setting offers more flexibility in designing a multi-agent system. The design of the observation function helps to filter the non-useful observation information of each agent while meeting the problem's constraint i.e., the communication bit-budget. The mathematical framework being used here is neither a classic MDP as we have the issue of partial observability, nor is a partially observable MDP (POMDP) [264] as the action vector is not jointly selected at a single entity. Our problem setting is differen-

tiated from Dec-POMDPs due to the fact that in Dec-POMDPs the partial observability is accepted as is, where as in our problem setting we design the lens through which the agents acquire a partial observation/perception of the environment.

Nevertheless, a similar class of problems - often referred to as task-oriented, goal-oriented or efficient communication approaches, has recently received significant attention from the communication society, see e.g., the extensive surveys on similar problems in [24,43,244]. Table 3.1 positions the current work against some of the recent research that is closely related. To date, there is no work in the literature that we are aware of, which provides an analytical approach to the design of task-based communications for the coordination of multiple cooperative agents.

3.1.3 Contributions

The contributions of this chapter are as follows:

Firstly, we develop a general cooperative multi-agent framework in which agents interact over an underlying MDP environment. Unlike the existing works which assume perfect communication links [3, 13, 263, 265], we assume the practical bit-budgeted communications between the agents. We formulate a multi-agent cooperative problem where agents interact over an underlying MDP and can communicate over a bit-budgeted channel. Our goal is to derive the optimal control and communication strategies to maximize the expected return. We will show in section 3.2.2, that an underlying difference in our setting from the Dec-POMDP is that here we carry out a constrained design of each agent's perception function - which is also referred to as the observation function in the literature of the Dec-POMDP [266]. The constraints of this design are dictated by the bit-budget of the inter-agent communication channels.

Secondly, Theorem 1, in section 3.3, derives the interconnection between the joint control and communication/quantization problem and a generalized version of the data quantization problem: TODC problem. In fact, the TODC problem distils all the relevant features of the control task and takes them into account in a novel non-conventional communication design problem. This is the underlying reason behind the effectiveness of the designed communications and is one the contributions in this work differentiating it from existing works in [3–5, 11, 26, 29, 69, 257]. Our analytical studies show that how the value function - the

function that estimates the expected return of the system given the current observation - can be considered as a proper indirect measure of the usefulness of the data to be compressed. Thus, Theorem 1, shows how the usefulness of the (observation) data can be incorporated into the design of the TODC policy.

Thirdly, we propose a novel algorithm - SAIC - as a multi-agent state-aggregation algorithm which designs indirect task-effective communication strategies via solving (an approximated version of) the TODC problem. As a result, the performance of SAIC in terms of the system's expected return is on par with the jointly optimal strategies. To the best of our knowledge, this is the first use of state-aggregation algorithms for data-compression applications (in multi-agent systems) according to which our work differs from the classic state-aggregation literature [88, 258–260] as well as the recent advancements in multi-agent communication literature [3, 257].

Moreover, we extend the existing results in the single-agent state-aggregation literature [259] on the gap between the optimal control and the state-aggregated control schemes, where the former has access to the true state of the environment and the latter has access to an aggregated state of the environment - to reduce the computational complexity. We quantify the same gap for a multi-agent system - Theorem 8. In our work, however, the gap is due to the bit-budget that is introduced on the inter-agent communication channels, whereas in classic state-aggregation literature the gap was a consequence of the constraints on the computational complexity. In addition to that, our theoretical results show that if our proposed method, SAIC, is applied the expected return of the multi-agent communication system - with the bit-budget in place - can stay in close proximity to the optimal expected return that is obtained under jointly optimal strategies.

Last but not least, numerical experiments are carried out on a geometric consensus problem to compare the performance of SAIC with several other benchmark schemes in terms of the optimality of the expected return, for a multi-agent scenario [2]. It is shown that when communication bit-budgets are in place, SAIC is of significant advantage over the benchmarks. In particular, we observe a very tight gap between the performance of SAIC and the optimal control strategy where only the latter runs over perfect communication channels and the former runs over bit-budgeted channels.

[2]Due to the complexity related issues explained in section 3.5 & 3.6, the numerical results are limited to two-agent and three-agent scenarios.

Table 3.1: Comparison between our work and the related prior art

Paper	Information source with memory	Joint coms and control	Distributed	Source/Channel coding Quantization	Implicit/Explicit coms	Analytical/ Data-driven
[26, 29]	×	×	×	Quantization	N/A	Data-driven
[69]	×	✓	✓	N/A	Implicit	Analytical
[5]	✓ (Linear)	✓	×	Quantization	Explicit	Analytical
[3, 257]	N/A	✓	✓	N/A	Explicit	Data-driven
[4, 11]	✓(Markov)	✓	✓	Channel Coding	Explicit	Data-driven
Our work	✓(Markov)	✓	✓	Quantization	Explicit	Analytical

Table 3.2: Table of notations

Symbol	Meaning
$\mathbf{x}(t)$	A generic random variable generated at time t
$\mathsf{x}(t)$	Realization of $\mathbf{x}(t)$
$\mathcal{X}$	Alphabet of $\mathbf{x}$(t)
$\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
$p_{\mathbf{x}}(\mathsf{x}(t))$	Shorthand for $\Pr(\mathbf{x}(t) = \mathsf{x}(t))$
$H(\mathbf{x}(t))$	Information entropy of $\mathbf{x}(t)$ (bits)
$I(\mathbf{x}(t); \mathbf{y}(t))$	Mutual information of $\mathbf{x}(t)$ and $\mathbf{y}(t)$
$\mathbb{E}_{p(\mathsf{x})}\{\mathbf{x}\}$	Expectation of the random variable X over the probability distribution $p(\mathsf{x})$
$\delta(\cdot)$	Dirac delta function
$\mathsf{tr}(t)$	Realization of the system's trajectory at time t

3.1.4 Organization

Section II describes the system model for a cooperative multi-agent task with rate-constrained inter-agent communications. Section III Proposes a scheme for the joint design of communication and control policies that takes the value of information into account to perform data compression. We also provide analytical results on how distant the result of this algorithm can be from the optimal centralized solution. The numerical results and discussions are provided in section IV. Finally, section V concludes the chapter.

3.1.5 Notation

For the reader's convenience, a summary of the notation that we follow in this chapter is given in Table 3.2. Bold font is used for matrices or scalars which are random and their realizations follow simple font.

3.2 System Model

In the multi-agent system, comprised of n agents, at any time step t each agent $i \in \mathcal{N}$ makes a local observation $\mathsf{o}_i(t) \in \Omega$ on environment while the true state of the environment

$$\mathsf{s}(t) = \langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t) \rangle \tag{3.1}$$

is a member of $\mathcal{S} = \Omega^n$. The alphabets Ω and $\mathcal{S}$ define observation space and state space, respectively. The particular observation structure of agents' observations, is referred to as collective observations in the literature [37]. Under collective observability, individual observation of an agent provides it with partial information about the current state of the environment, however, having knowledge of the collective observations acquired by all of the agents is sufficient to realize the true state of environment - eq. (3.1). The columns of the state vector are orthogonal to each other. Note that even in the case of collective observability, for agent i to be able to observe the true state of environment at all times, it needs to have access to the observations of the other agents $j \in \mathcal{N} - \{i\} \triangleq \mathcal{N}_{-i}$ through communications at all times.

The true state of the environment $\mathsf{s}(t)$ is controlled by the joint actions $\mathsf{m}(t) = \langle \mathsf{m}_1(t), ..., \mathsf{m}_n(t) \rangle \in \mathcal{M}^n$ of the agents, where each agent i can only choose its local action $\mathsf{m}_i(t) \in \mathcal{M}$. The environment runs on discrete time steps $t = 1, 2, ..., M$, where at each time step, each agent i selects its domain level action $\mathsf{m}_i(t)$ upon having an observation $\mathsf{o}_i(t)$ of the environment. Dynamics of the environment are governed by a conditional probability mass function (CMF)

$$T(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t)) = p(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t)) \tag{3.2}$$

which is unknown to the agents. $T(\cdot) : \Omega^{2n} \times \mathcal{M}^n \to [0, 1]$ determines the future state of the environment $\mathsf{s}(t+1)$ given its current state $\mathsf{s}(t)$ and the joint actions $\mathsf{m}(t)$. We recall that each agent i's domain level action $\mathsf{m}_i(t)$ can, for instance, be in the form of a movement or acceleration in a particular direction or any other type of action depending on the domain of the cooperative task.

A deterministic reward function $r(\cdot) : \Omega^n \times \mathcal{M}^n \to \mathbb{R}$ indicates the reward of all agents at time step t, where the arguments of the reward function are the joint observations $\mathsf{s}(t)$ and the domain-level joint actions $\mathsf{m}(t)$ of all agents. We assume that the underlying environment over which agents interact can be defined in terms of an MDP [3] determined by the tuple $\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T(\cdot)\}$, where Ω and $\mathcal{M}$ are discrete alphabets, $r(\cdot)$ is a function, $T(\cdot)$ is defined in (3.2) and the scalar $\gamma \in [0, 1]$ is the discount factor. The focus of this chapter is

[3] As defined in the literature [10], the underlying MDP' is the horizon-T' MDP defined by a hypothetical single agent that takes joint actions $\mathbf{m}(t) \in \mathcal{M}^n$ and observes the nominal state $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \ldots, \mathbf{o}_n(t) \rangle$ that has the same transition model $T(\cdot)$ and reward model $R(\cdot)$ as the environment experienced by our multi-agent system.

on scenarios in which the agents are unaware of the state transition probability function $T(\cdot)$ and of the closed form of the function $r(\cdot)$. However we assume that, further to the literature of reinforcement learning [267], a realization of the function $r(\mathsf{s}(t), \mathsf{m}(t))$ will be accessible for all agents at some time steps. Since the tuple $\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T(\cdot)\}$ is an MDP and the state process $\mathbf{s}(t)$ is jointly observable by agents, the system model of this cooperative multi-agent setting, under perfect communications, is also referred to as a multi-agent MDP (MAMDP or MMDP) in the literature of multi-agent decision making [249, 268, 269].

In what follows two problems regarding the above-mentioned setup is detailed i.e., centralized and decentralized control problems. The main intention of this chapter is to address decentralized control which also incorporates inter-agent communications for a system of multiple agents. The centralized control problem, however, is also formalized in subsection 3.2.1 as the optimal expected return obtained for the centralized problem can serve as a lower-bound/(upper-bound) for the decentralized scheme. Moreover, the simpler nature and mathematical notations used for the centralized problem, allow the reader to have a smoother transition to the decentralized problem which is of a more complex nature.

3.2.1 Centralized Control

We consider a scenario in which a central controller has instant access to the observations $\mathsf{o}_1(t), ..., \mathsf{o}_n(t)$ of both agents through a free (with no cost on the objective function) and reliable communication channel. From the central controller's point of view, the environment is the same as the underlying MDP that governs the system $\Big\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T(\cdot)\Big\}$. The goal of the centralized controller is to maximize the expected sum of discounted rewards (3.3). The expectation is computed over the joint PMF of the whole system trajectory $\mathbf{s}(1), \mathbf{m}(1), ..., \mathbf{s}(M), \mathbf{m}(M)$ from time $t = 1$ to $t = M$, where this joint probability mass function (PMF) is generated if agents follow policy $\pi(\cdot)$, eq. (3.4), for their action selections at all times and the initial state $\mathbf{s}(1) \in \mathcal{S}$ is randomly selected by the initial distribution $\mathbf{s}(1) \sim \alpha_{\mathbf{s}}$. For the sake of having a more compact notation to refer to the system trajectory, hereafter, we represent the realization of a system trajectory at time t by $\mathsf{tr}(t)$ which corresponds to the tuple $\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t) \rangle$ and the realization of the whole system trajectory by $\{\mathsf{tr}(t)\}_{t=1}^{t=M}$. Accordingly, the problem boils down to a single agent problem which can be denoted by

$$\max_{\pi(\cdot)} \qquad \mathbb{E}_{p_\pi\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)}\Big\{\sum_{t=1}^{M} \gamma^{t-1} r(\mathbf{s}(t), \mathbf{m}(t))\Big\} \tag{3.3}$$

where the policy π can be expressed as a CMF

$$\pi\Big(\mathsf{m}(t)\Big|\mathsf{s}(t)\Big) = p\Big(\mathsf{m}(t)\Big|\mathsf{s}(t)\Big), \tag{3.4}$$

and $p_\pi(\mathsf{s}(t+1)|\mathsf{s}(t))$ is the probability of transitioning from $\mathsf{s}(t)$ to $\mathsf{s}(t+1)$ when the joint action policy $\pi(\cdot)$ is executed by the central controller. Similarly, $p_\pi(\{\mathsf{tr}(t)\}_{t=1}^{t=M})$ is the joint PMF of $\mathsf{tr}(1), \mathsf{tr}(2), ..., \mathsf{tr}(M)$ when the joint action policy $\pi(\cdot)$ is followed by the central controller.

On one hand, problem (3.3) can be solved using single-agent Q-learning [267] and the solution $\pi^*(\cdot)$ obtained by Q-learning is guaranteed to be the optimal control policy, given some non-restricting conditions [270]. On the other hand, the use-cases of the centralized approach are limited to the applications in which there is a permanent communication link with an unlimited bit-budget between the agents and the controller. Whereas these conditions are not met in many remote applications, where there is no communication infrastructure to connect the agents to the central controller.

Given sufficient training time, and channels with the sufficient rate of communication between the agents and the central controller, the centralized algorithm provides us with a performance upper bound in maximizing the objective function (3.3). Perfect communication between the central controller and distributed agents, however, may not exist due to the resource limitations of the telecommunication/communication network. Thus, the aim of this chapter is to introduce decentralized approaches which are run over practical bit-budgeted communication channels, yet show comparable performance levels. In the distributed scenario, the agents do not communicate with a central controller, but the bit-budgeted communications are performed for inter-agent message exchange. The centralized problem can be presented by an MDP and be solved efficiently by a single agent reinforcement learning algorithm. As explained in the section 3.1.3, the decentralized problem is a more complicated/general form of Dec-POMDP, where we know that a Dec-POMDP is already much more complex than an MDP to solve [266] - to see further insights about the significance and the applications of the decentralized problem see e.g., [24].

3.2.2 Problem Statement

Here we consider a scenario in which the same objective function explained in Eq. (3.3) needs to be maximized by the multi-agent system in a decentralized fashion, Fig. 3.1. Namely, agents with partial observability can only select their own actions. To prevail over the limita-

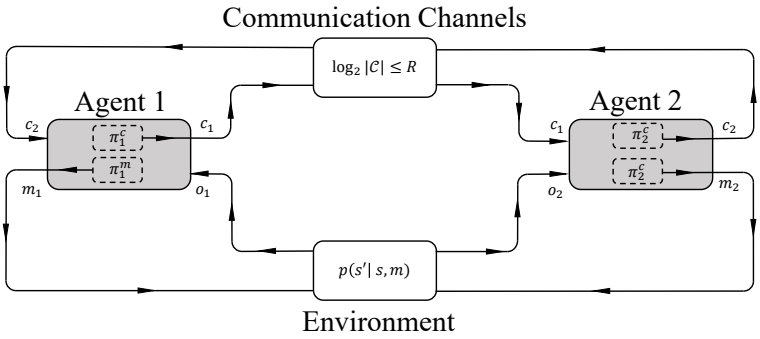


Figure 3.1: An illustration of the decentralized cooperative two-agent system with rate-limited inter-agent communications.

tions imposed by the local observability, agents are allowed to have direct (explicit) communications, and not indirect (implicit) communications [69, 271]. However, the communication is done through a bit-budgeted but reliable channel. The bit-budget of the channel is R-bits per time step. Equivalently, each agent i at every time step t produces and transmits a single digit communication message $\mathbf{c}_i(t) \in \mathcal{C}$ such that

$$log_2|\mathcal{C}| \leq R, \tag{3.5}$$

i.e., the size of the code-books $\mathcal{C}$ for all agents is the same and is less than 2^R. The communication message $\mathbf{c}_i(t)$ produced by agent i is broadcast and received every agent $j \in \mathcal{N}_{-i}$. It should be noted that the design of the channel coding is beyond the scope of this thesis and the main focus is on the compression of agents' observations. In particular we consider R to be time-invariant and to follow:

$$R < \min\{H(\mathbf{o}_1(t)), ..., H(\mathbf{o}_n(t))\}. \tag{3.6}$$

The above-mentioned information constraint which will be in place throughout this thesis together with the observation structure assumed in eq. (3.1) are of the aspects that distinguish our work from many of the related works in the literature of multi-agent communications [4, 11]. Now let the function $\mathbf{g}(t')$ denote the system's *return*:

$$\mathbf{g}(t') = \sum_{t=t'}^{M} \gamma^{t-1} r(\mathbf{s}(t), \mathbf{m}(t)). \tag{3.7}$$

Note that $\mathbf{g}(t')$ is a random variable and a function of t' as well as the trajectory $\{\mathsf{tr}(t)\}_{t=t'}^{t=M}$. Due to the lack of space, here we drop a part of the arguments of this function. In contrast to the centralized problem, the goal of the decentralized problem is to jointly design the communication/quantization as well as $\pi_i^c(\cdot)$ control policies $\pi_i^m(\cdot)$ for each agent $i \in \mathcal{N}$ to maximize the average return of the system. The control policy $\pi_i^m : \mathcal{M} \times \mathcal{C}^{n-1} \times \Omega \to [0, 1]$

of each agent i is defined as CMF

$$\pi_i^m\Big(\mathsf{m}_i(t)\Big|\mathsf{o}_i(t),\mathsf{c}_{-i}(t)\Big) = \\ \Pr\Big(\mathbf{m}_i(t)=\mathsf{m}_i(t)\Big|\mathbf{o}_i(t)=\mathsf{o}_i(t),\mathbf{c}_{-i}(t)=\mathsf{c}_{-i}(t)\Big), \tag{3.8}$$

in which, $\mathsf{c}_{-i}(t) \in \mathcal{C}^{n-1}$ is a vector that includes all communication messages $\mathsf{c}_j(t),\ \forall j \in \mathcal{N}_{-i}$. The communication policy $\pi_i^c : \Omega \times \mathcal{C}^{n-1} \to \mathcal{C}$ of each agent i is a deterministic data quantization (many to one) function:

$$\mathsf{c}_i(t) = \pi_i^c\Big(\mathsf{o}_i(t),\mathsf{c}_{-i}(t)\Big), \tag{3.9}$$

which has a discrete domain $\Omega \times \mathcal{C}$, making the quantizer a discrete quantizer. The joint control policy π^m is a tuple made of n elements with its i-th element being $\pi_i^m(\cdot)$. Similarly, The joint communication policy π^c is another tuple with its i-th element being $\pi_i^c(\cdot)$.

According to the above definitions, the decentralized joint control and communication design problem is formalized as

$$\begin{aligned} &\max_{\pi_i^m,\pi_i^c} && \mathbb{E}_{p_{\pi^m,\pi^c}\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)}\Big\{\mathbf{g}(1)\Big\}, \quad i\in\mathcal{N} \\ &\text{s.t.} && log_2|\mathcal{C}| \leq R, \end{aligned} \tag{3.10}$$

where the expectation is taken over $p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})$ which is the joint PMF of $\mathsf{tr}(1),\mathsf{tr}(2),$ $...,\mathsf{tr}(M)$ when each agent $i \in \mathcal{N}$ follows the action policy $\pi_i^m(\cdot)$ and the communication policy $\pi_i^c(\cdot)$ and the initial state $\mathbf{s}(1) \in \mathcal{S}$ is randomly selected by the initial distribution $\mathbf{s}(1) \sim \alpha_{\mathbf{s}}$. Given communication policy $\pi_i^c(\cdot),\ \forall i \in \mathcal{N}$, we now define the perception function $h_i(\cdot) : \mathcal{S} \to \mathcal{C}^{n-1} \times \Omega$ of agent i which is the lens through which agent i perceives the state $\mathbf{s}(t)$ of the environment.

$$h_i(\mathbf{s}(t)) = \\ \langle \pi_1^c(\mathbf{o}_1(t)), \pi_2^c(\mathbf{o}_2(t)), ..., \mathbf{o}_i(t), \pi_{i+1}^c(\mathbf{o}_{i+1}(t)), ..., \pi_n^c(\mathbf{o}_n(t))\rangle \tag{3.11}$$

Agent i's perception of the environment is characterized by the communication policy $\pi_j^c(\cdot)$ of each agent $j \in \mathcal{N}_{-i}$. Accordingly, agent i uses its sensory signal $\mathbf{o}_i(t)$ together with the received communication signals $\mathbf{c}_{-i}(t)$ to acquire its perception of the environment. While the perception function defined here plays a role very similar to the observation function

in Dec-POMDPs [106], the main difference is that here we design communication policies such that they directly affect the perception of agents from the environment. In contrast, in the case of Dec-POMDPs, the observation function is given. Communication policies $\pi_j^c(\cdot), \forall j \in \mathcal{N}_{-i}$ partially define the perception function of agent i.

To make the problem more concrete, further to (3.8) and (3.9), here we assume the presence of instantaneous and synchronous communications between agents, contrasting with the delayed [4, 272] and sequential communication models. Fig. 3.2 demonstrates this communication model during a single time-step. As such, each agent i at any time step t prior to the selection of its action $\mathsf{m}_i(t)$ receives the communication vector $\mathsf{c}_{-i}(t)$ that encodes the observations of each agent $j \in \mathcal{N}_{-i}$ at time t.

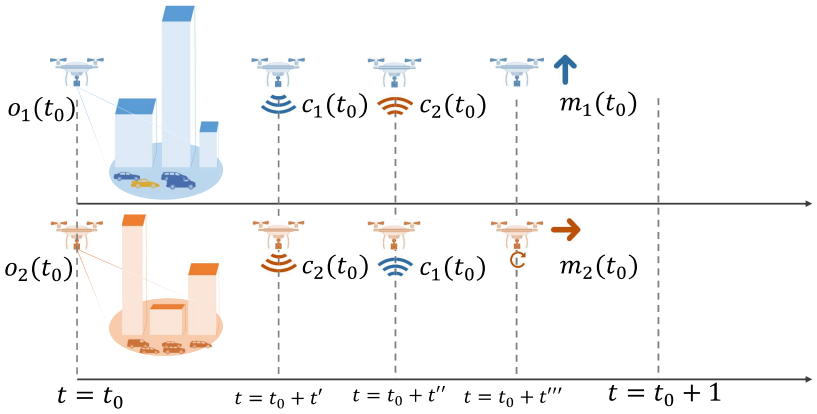


Figure 3.2: Ordering of observation, communications and action selection for synchronous and instantaneous communication model in a multi-UAV object tracking example, with $0 < t' < t'' < t''' < 1$. At time $t = t_0$ both agents (UAVs) make local observations on the environment. At time $t = t_0 + t'$ both agents select a communication signal to be generated. At time $t = t_0 + t''$ agents receive a communication signal from the other agent. At time $t = t_0 + t'''$ agents select a domain level action, here it can be the movement of UAVs or rotation of their cameras etc.

In a general approach, the selection of communication action $\mathsf{c}_i(t)$ at agent i could be conditioned on both $\mathsf{o}_i(t)$ and $\mathsf{c}_{-i}(t)$. Since we assume instantaneous and synchronous inter-agent communications, here we are focused on communication policies of type $\pi_i^c(\mathsf{o}_i(t))$, where communication actions of each agent at each time are selected only based on its observation at that time. For clear reasons, it is not possible to adopt a synchronous and instantaneous inter-agent communication model and yet take the communication message $\mathsf{c}_{-i}(t)$ into account when selecting the communication $\mathsf{c}_i(t)$ at agent i. Here we assume that the communication resources are split evenly amongst the agents, by considering the bit-budget of all

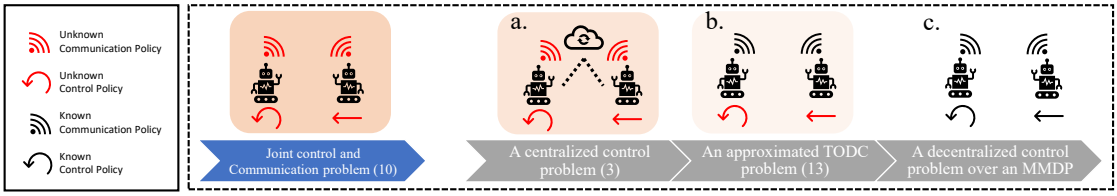


Figure 3.3: Here we show how we approached solving the joint control and communication problem for a distributed multi-agent system in a sequence of steps. According to the legend, one can understand that at the end of each step what are the known and unknown policies. a. This step solves the problem (3.3) for a centralized multi-agent system where the objective is to design one centralized control strategy. b. This step solves the problem (3.13) for a distributed multi-agent system where the objective is to design the communication policies of all agents. c. this step solves the problem for a distributed multi-agent system where the objective is to design the control policies of all agents.

communication channels to be equal to R. As such, each agent $i \in \mathcal{N}$ encodes its observation $\mathsf{o}_i(t)$ to $\mathsf{c}_i(t)$ using a code-book $\mathcal{C}$ of the same length $|\mathcal{C}|$ - with the constraint (5.2) in place.

3.3 State Aggregation for Information Compression (SAIC) in multi-agent Coordination Tasks

The main result of this section - provided by Theorem 3 - is to show that finding the quantization policy in the joint control and quantization problem (3.10) can be approximated by a TODC problem. The goal of this problem is to quantize the observations of all agents according to how valuable these observations are within any specific task. The value of observations should be measured by the value function $V^*(\cdot)$ - eq. (3.25). Lemma 4 approximates the TODC to a k-median clustering of the of observations according to their values, while lemma 5 computes the value function of each agent's observation. The concluding remarks of this section study the convergence and the optimality of the decentralized control policies.

Fig. 3.3 is brought to demonstrate the chronological order according to which a joint communication and quantization is solved by SAIC. Our proposed scheme, SAIC, breaks down the joint communication and quantization problem to smaller problems that are feasible to solve. In this section, however, the subsections are organized according to the logical order that these smaller problems are encountered: (A) in section 3.3.1 , we address the communication design of multi-agent communications by transforming the primary joint control and quantization problem (3.10) to a novel problem (3.12) called TODC - step "b" of the Fig.

3.3. (B) Since solving the TODC problem relies on the knowledge of the value function $V^*(\cdot)$, it is necessary to obtain the value function $V^*(\cdot)$ prior to solving the TODC problem. In section 3.3.2, the optimal value function $V^*(\cdot)$ is obtained via a centralized training phase - step "a" of the Fig. 3.3. Given the knowledge of the value function $V^*(\cdot)$, the TODC problem incorporates the features of the specific control task in the communication design problem. Accordingly, we can separately solve the communication problem with very little compromise on the optimality of the system's expected return. (C) As the final step, in section 3.3.3, decentralized training phase is carried out to distributively design the control policy of each agent given the communication/quantization policy obtained via solving the TODC problem. Decentralized training is shown in step "c" of the Fig. 3.3. Since we follow standard methods to carry out the centralized training - steps "a" of the Fig. 3.3 - we will be mainly focused on deriving and solving the TODC problem and providing guarantees on the performance of the MAS in the decentralized training phase - steps "b" and "c" of the Fig. 3.3 respectively. Fig. 3.4 illustrates how SAIC performs data compression while it maintains the performance of the multi-agent system in its task.

3.3.1 Task-Oriented Data Compression Problem

The main result of this section is provided by Theorem 3. This theorem departs from the joint communication/quantization and control problem and arrives at the task-oriented data compression problem (3.12).

Theorem 3. *The design of the communication policy in problem (3.10) can be approximated as a generalized data quantization problem*

$$\begin{aligned} &\underset{\pi_i^c(\cdot)}{min} && \mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(1)))}\Big\{|V^*(\mathbf{s}(1)) - V^*(h_i(\mathbf{s}(1)))|\Big\} \\ &s.t. && log_2|\mathcal{C}| \leq R, \end{aligned} \tag{3.12}$$

in which the measure of distortion is the absolute difference of the value functions $V^*(\mathbf{s}(t))$ and $V^*(h_i(\mathbf{s}(1)))$ with the source of information $\mathbf{s}(t) \in \Omega^n$ being a Markovian stochastic process. The function $V^*(h_i(\mathbf{s}(1)))$ measures the optimal value of the perceived state $h_i(\mathbf{s}(1))$ from agent i's perspective.

Proof. Appendix 3.7. ■

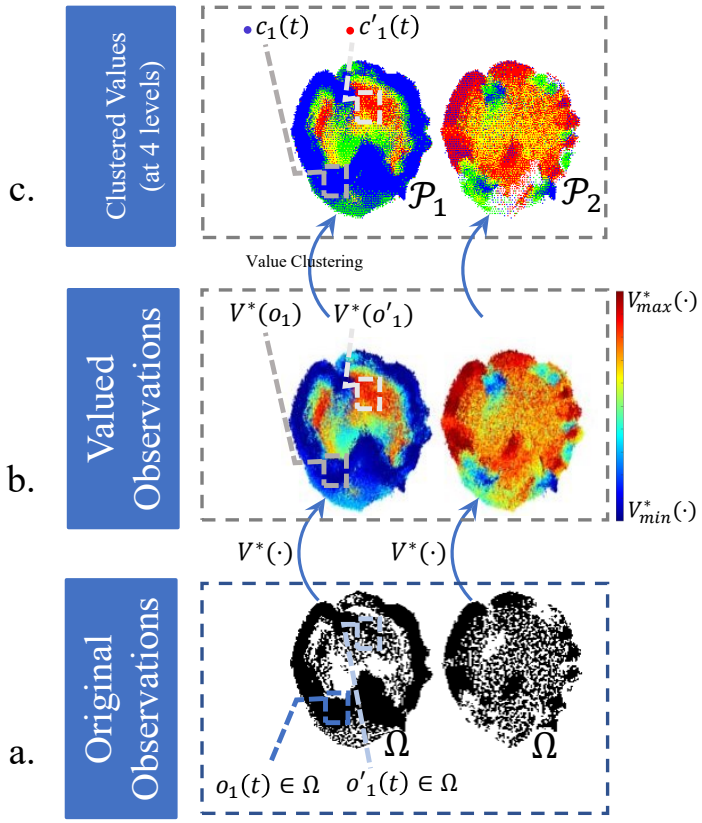


Figure 3.4: The subplots of this figure illustrate how in SAIC we transform a high-dimensional (σ-dimensions) and high-precision observation space into aggregated one-dimensional low-precision/digitized communication message space. This figure is plotted for a scenario where $R = 2$ (bits per channel use) and thus, observation values are clustered at $2^R = 4$ different levels. a. A 2D demonstration of the original high-dimensional and high-precision observation space of agents is shown here in black and white. b. After carrying out the centralized training phase we will obtain the value function $V^*(\cdot)$ - which acts as indirect measure of the usefulness of observation data to be communicated. Now by applying the value function $V^*(\cdot)$ at every point of the original observation space we get valued observations - a one-dimensional high-precision space as the output space of the value function $V^*(\cdot)$. c. By clustering the observation points according to their corresponding values for each agent i we would get a one-dimension and low-precision/digitized communication message space. The quantization illustrated in this diagram is using only 4 levels of quantization that are represented by 4 colours. All the points in the observation space of the agent i which are represented by the same colour, in subplot c, will be represented by a unique communication message - i.e., the accuracy of the original data is reduced and hence requires fewer communication bits to be transmitted. Accordingly, agent 1, after observing $\mathsf{o}_1(t)$ transmits the communication message $\mathsf{c}_1(t)$ which is a compressed version of $\mathsf{o}_1(t)$ while it maintains the performance of the multi-agent team in maximizing their expected return.

In Appendix 3.7.3, we provide more details on how to obtain the value $V^*(h_i(\mathbf{s}(1)))$ of the perceived state from the agent i's point of via Lemma 12. This value function allows us to indirectly quantify the usefulness of agent i's observation. With this interpretation in mind, in the TODC problem (3.12), unlike conventional quantization problems, we are not minimizing the absolute difference between the original signal $\mathbf{s}(1)$ and its quantized version $h_i(\mathbf{s}(1))$. Instead, we are minimizing the distance between how useful/valuable the original signal $\mathbf{s}(1)$ is and how useful the quantized version of the signal $h_i(\mathbf{s}(1))$ are for the task at hand. This is in-line with what many believe as the mission of the goal-oriented/task-oriented communications. Let us recall that the value function here is an *indirect* measure of usefulness, as it can be obtained for any task that can be expressed via Markov Decision Processes - making it a measure of usefulness that is applicable to a plethora of scenarios [24, 247].

The significance of the result obtained by Theorem 3 is multi-fold: (i) Multi-dimensional observations will be transformed to one-dimensional output space of the value functions, reducing the complexity of the clustering algorithm, (ii) It can be shown that the observation points will be linearly separable when being clustered according to the problem (3.12), (iii) It is widely accepted that the mission of goal oriented communications is to incorporate the usefulness/value of the data for the task when designing the task-effective communications. The result of Theorem 3, in which the design of the quantizer relies on the value/usefulness of observations resonates well with this purpose of goal-oriented communications. (iv) It is known that the value of observations starts to grow as we get closer to the ultimate target of the task in hand. With this interpretation of "target" in mind, the finding of Theorem 3 is in line with the adaptive quantization schemes, which stretch the quantization intervals when the observations are far from the target and sharpen the quantization when the observations are closer to the target [75, 273]. This interpretation is also confirmed by our numerical results in section 3.5, Fig. 3.8.

To solve a quantization problem as (3.12) using non-variational techniques, it is customary to approximate/convert a quantization problem by/to a clustering problem [274, 275]. Lemma 4 approximates the quantization problem (3.12) by a clustering problem.

Lemma 4. *The quantization problem (3.12) can be approximated by a clustering problem*

$$\min_{P_i} \quad \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_i(t)\in\mathcal{P}_{i,k}} \left| V^*(\mathsf{o}_i(t)) - \mu_k^{'} \right|, \tag{3.13}$$

where μ_k' is the centroid of the k-th cluster $\mathcal{P}_{i,k}$ and $\mathcal{P}_i = \{\mathcal{P}_{i,1}, \ldots, \mathcal{P}_{i,|\mathcal{C}|}\}$ is a partition of the observation space Ω. Similar to any other quantization function, the quantizer $\pi_i^c(\cdot)$, can be uniquely described by the partition $\mathcal{P}_i$ together with $\mathcal{C}$.

Proof. Appendix 3.8 provides proof and discussions. ■

The problem (3.13), can be solved via k-median clustering. In order to that, one can first perform the k-median clustering on the observation values by solving

$$\min_{V_i} \quad \sum_{k=1}^{2^B} \sum_{V^*(\mathsf{o}_i(t))\in\mathcal{V}_{i,k}} \left| V^*(\mathsf{o}_i(t)) - \mu_k^{''} \right|,$$

where $\mathcal{V}_i$ is the set of all observation values of agent i and $\{\mathcal{V}_{i,1}, ..., \mathcal{V}_{i,|\mathcal{C}|}\}$ is its partition. Afterwards, as shown in Figure 3.4, the observation points should be clustered according to the clustering of their corresponding values. That is, any two distinct observation points $\mathsf{o}_i', \mathsf{o}_i'' \in \Omega$ are clustered together in $\mathcal{P}_{i,j}$ if and only if their values $V^*(\mathsf{o}_i'), V^*(\mathsf{o}_i'') \in \mathcal{V}_{i,j}$ are in the same cluster $\mathcal{V}_{i,j}$.

Theorem 3 together with lemma 4 allows us to find a communication/quantization policy $\pi_i^c(\cdot)$ by clustering the input space Ω of the communication policy according to the values $V^*(\mathsf{o}_i(t))$ of the input points. The performance guarantees for the obtained communication/quantization policy will be shown in section 3.4. One can obtain $V^*(\mathsf{o}_i(t))$ via solving the centralized problem (3.3) by Q-learning. The subsection 3.3.2, details a centralized training approach for obtaining the value observations $V^*(\mathsf{o}_i(t))$.

3.3.2 Centralized Training Phase

While solving the TODC problem can provide us with a task-effective design of quantization policy, to solve (3.12) we need to know the value of observations according to the optimal centralized control policy. By solving the centralized problem (3.3), the value of joint observations and actions $Q^*(\mathsf{s}(t), \mathsf{m}(t))$ can be obtained. Let us recall that the centralized training

phase will only yield an optimal policy if the environment is jointly observable - as described by the joint observability condition.

Joint Observability Condition:

$$\mathsf{s}(t) = \langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t) \rangle. \tag{3.14}$$

Accordingly, following the lemma 5 we can compute the value of each agent's observations $V^*(\mathsf{o}_i(1))$. But before lemma 5, let us first give an intuitive/philosophical meaning of the centralized training and distributed execution. We know that in task-oriented communication design, our goal is to take into account the usefulness/value of the data for the task in hand. Thus we need to first be able to measure the usefulness/value of the data to be transmitted. The centralized training phase is needed to come up with a precise measure of usefulness for the specific task in hand. We have already shown in Theorem 3, that this measure of usefulness is nothing but the value observations $V^*(\mathsf{o}_i(1))$ - yet the exact function values can be known only after the centralized training phase. During the centralized training phase, we assume perfect communication between all agents and a central controller - this is a common practice in the literature of multi-agent communications and coordination [3, 276]. Whereas, in the decentralized training - step "c" of the Fig. 3.3 - as well as in the execution phase, we assume bit-budgeted communications. That is, all the results reported for SAIC in section 3.5 are obtained via bit-budgeted communications.

Lemma 5. *One can compute the* $V^*(\mathsf{o}_i(1))$ *following*

$$V^*(\mathsf{o}_i(t)) = \sum_{\mathsf{o}_{-i}(t) \in \Omega^{n-1}} \max_{m} \; Q^*(\mathsf{s}(t), \mathsf{m}(t)) p(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t)).$$

Proof. Appendix 3.9. ■

Based on (3.15), $V^*(\mathsf{o}_i(1))$ can be computed both analytically (if transition probabilities of environment are available) and numerically. As detailed in Algorithm 1, SAIC first solves a centralized control problem to compute the value $V^*(\mathsf{o})$ for all $\mathsf{o} \in \Omega$ - this is equivalent to the step "a" of the Fig. 3.3 and subplot (b) of the Fig. 3.4. Afterwards, SAIC solves the approximated TODC problem (3.12) by converting it to a k-median clustering (3.13), leading

to an observation aggregation/quantization function for each agent i determined by $\pi_i^c(\cdot)$ - this is equivalent to the step "b" of the Fig. 3.3 and the subplot (c) of the Fig. 3.4. By following this aggregation function, the observations $\mathsf{o}_i(t) \in \Omega$ will be aggregated/quantized such that the performance of the multi-agent system in terms of the objective function it attains is optimized. As SAIC uses a deterministic mapping of observation o_i to produce the communication message c_i, SAIC is guaranteed to have positive signalling [257].

3.3.3 Obtaining Decentralized Control Policies via a Decentralized Training Phase

Upon the availability of the $\pi_i^c(\cdot),\ \forall i \in \mathcal{N}$, which was obtained by solving problem (3.13), we need to find control policies for all agents corresponding to the communication policies $\pi_i^c(\cdot),\ i \in \mathcal{N}$. That is, we now solve the problem (3.10) by plugging the exact communication policy $\pi_i^c(\cdot)\ \forall i \in \mathcal{N}$ into it. Within this training phase - referred to as the decentralized training phase - control Q-tables $Q_i^m(\cdot)\ \forall i \in \mathcal{N}$ are obtained - step "c" of the Fig. 3.3. This training phase, as well as the execution phase of the algorithm, can both be carried out distributively, while agents communicate over bit-budgeted channels using the communication policies obtained before in section 3.3.1. The following remarks are brought to characterize the performance of SAIC, in the decentralized training phase.

We now first define the concept of lumpablity, according to which we will then set a condition - Lumpability Condition - for the correctness of remarks 3 and 4.

Definition 6. ***Lumpability of an MDP:*** *Let $\alpha_{\mathbf{s}}$ be the probability distribution of the initial state of an MDP at the initial step. The MDP is called (strongly) lumpable with respect to the perception function $h_i(\cdot)$ if the transitions between all the perceived states $h_i(\mathbf{s}(t))$ - which are perceived through the lens of $h_i(\cdot)$ - follow Markov rule for every probability distribution $\alpha_{\mathbf{s}}$ of the initial state of the original MDP [260].*

Lumpability Condition: Let the environment as perceived from the perspective of agent i within the decentralized training phase be called an aggregated MDP denoted by $\left\{\Omega \times \mathcal{C}^{n-1}, \mathcal{M}, r(\cdot), \gamma, T'(\cdot)\right\}$, whereas the state space of the aggregated MDP $\Omega \times \mathcal{C}^{n-1}$ is an image of Ω^n under the perception function $h_i(\cdot)$. Now given the definition 6, assuming the

lumpability of the underlying MDP $\left\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T(\cdot)\right\}$ with respect to $h_i(\cdot)$ is equivalent to the assumption that the aggregated $\left\{\Omega \times \mathcal{C}^{n-1}, \mathcal{M}, r(\cdot), \gamma, T(\cdot)\right\}$ is an MDP under every possible $\alpha_{\mathbf{s}}$. This assumption is in place for the correctness of remarks 3 and 4.

Algorithm 1 State Aggregation for Information Compression (SAIC)

Input: γ, α, c
Initialize all-zero table $N_i^m(\mathsf{o}_i(t), \mathsf{c}_{-i}(t), \mathsf{m}_i(t))$, for $i \in \mathcal{N}$
    and Q-table $Q_i^m(\cdot) \leftarrow Q_i^{m,(k-1)}(\cdot)$, for $i \in \mathcal{N}$
    and all-zero Q-table $Q(\mathsf{o}_i(t), \mathsf{o}_j(t), \mathsf{m}_i(t), \mathsf{m}_j(t))$.
Obtain $\pi^*(\cdot)$ and $Q^*(\cdot)$ by solving (3.3) using Q-learning [267].
Compute $V^*(\mathsf{o}_i(t))$ following eq. (3.15), for $\forall \mathsf{o}_i(t) \in \Omega$.
Solve problem (3.13) by applying k-median clustering to obtain $\pi_i^c(\cdot)$, for $i \in \mathcal{N}$.
for *each episode* $k = 1 : K$ **do**
    Randomly initialize local observation $\mathsf{o}_i(t = 1)$, for $i \in \mathcal{N}$ **for** $t_k = 1 : M$ **do**
Select $\mathsf{c}_i(t)$ following $\pi_i^c(\cdot)$, for $i \in \mathcal{N}$
Obtain message $\mathsf{c}_{-i}(t)$, for $i \in \mathcal{N}$
Update $Q_i^m(\mathsf{o}_i(t-1), \mathsf{c}_{-i}(t-1), \mathsf{m}_i(t-1))$, for $i \in \mathcal{N}$
Select $\mathsf{m}_i(t) \in \mathcal{M}$ following UCB, for $i \in \mathcal{N}$
Increment $N_i^m(\mathsf{o}_i(t), \mathsf{c}_{-i}(t), \mathsf{m}_i(t))$, for $i \in \mathcal{N}$
Obtain reward $r(\mathsf{s}(t), \mathsf{m}(t))$, for $i \in \mathcal{N}$
Make a local observation $\mathsf{o}_i(t)$, for $i \in \mathcal{N}$
$t_k = t_k + 1$
end
Compute $\sum_{t=1}^{M} \gamma^t r_t$ for the lth episode
end
Output: $Q_i^m(\cdot)$,
    and $\pi_i^m(\mathsf{m}_i(t)|\mathsf{o}_i(t), \mathsf{c}_{-i}(t))$ by following greedy policy for $i \in \mathcal{N}$

Remark 1: The optimal policy $\pi^*(\cdot)$ is achievable by the centralized training phase. Assuming the joint observability condition to hold, the environment is fully observable for the central controller while the central controller possesses the ability to jointly select the actions for all agents. The problem will thus reduce to a single agent Q-learning applied on an MDP with asymptotic convergence to the optimal policy $\pi^*(\cdot)$.

Remark 2: During the decentralized training phase, each agent, instead of viewing the environment as the original underlying MDP denoted by $\left\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T'(\cdot)\right\}$, views an aggregated form of the original MDP denoted by $\left\{\Omega \times \mathcal{C}^{n-1}, \mathcal{M}, r(\cdot), \gamma, T'(\cdot)\right\}$. The aggregated MDP experienced by agent i will be an MDP itself, if the joint observability and lumpability conditions hold.

Remark 3: The MAS, during the decentralized training phase, will be composed of n different MDPs with identical state space $\Omega \times \mathcal{C}^{n-1}$, action space $\mathcal{M}$ and reward signal. The

resulting multi-agent environment will be, according to the definition, a multi-agent MDP (MMDP) [269].

Remark 4: Within the distributed training phase, distributed Q-learning is applied to a deterministic MMDP [4], which leads to an asymptotically optimal control policy [249] [5]. For this remark to be true lumpability and joint observability conditions must hold.

Note that the control policy $\pi_i^{m,SAIC}(\cdot)$ that is obtained within the distributed training phase of SAIC is optimal for the given communication policy $\pi^{c,SAIC}(\cdot)$, that was obtained within the centralized training phase. Therefore, $\pi_i^{m,SAIC}(\cdot)$ is not necessarily an optimal solution to the problem (3.10). In Theorem section 3.4, however, we set an upper-bound on the possible loss on the expected return of the system due to the joint selection of $\pi_i^{m,SAIC}(\cdot)$ and $\pi^{c,SAIC}(\cdot)$.

3.4 Characterizing the error bound of SAIC

As discussed in section 3.3, SAIC uses two approximations to solve the original joint quantization and control problem. It was not, however, explained that how these approximation would impact the performance of SAIC in terms of the system's average return. By extending the results of [259] to a multi-agent scenario, we characterize the performance gap of SAIC proposed in section 3.3. Instead of measuring the difference between the average return obtained by SAIC with that of the jointly optimal policies for the problem (3.10), in Theorem 8, we measure the performance gap between the average return attained by SAIC with that of the centralized controller - whereas the latter has had access to perfect communications and as well as full observability of the environment. The measured gap is, indeed, larger than the performance gap between SAIC and a hypothetical jointly optimal solution to (3.10), as in the case of the central controller there is no communication/observation limitation in place. The performance gap between SAIC and the centralized solution provided by Theorem 8 is proposed in terms of the discount factor λ of the task and a positive scalar ϵ. Definition 7 details the notion of ϵ-cost uniform. Lemma 9 is proposed to compute the value of ϵ for SAIC.

[4] The definition of MMDP in [269] is identical to the definition of cooperative MAMDP used in [249].

[5] This training phase can result in an asymptotically optimal control policy of all agents for non-deterministic MMDPs. This, however, will require n additional centralized training phases prior to the decentralized training phase, where n is the number of agents.

Definition 7. *Given a positive number ϵ a subset $\mathcal{P}_{i,k} \subset \Omega$ is said to be ϵ-cost-uniform with respect to the policy $\pi(\cdot)$ if the following conditions hold for two arbitrary observations $\mathsf{o}', \mathsf{o}'' \in \mathcal{P}_{i,k}$:*

$$c_1 : \qquad \mathcal{M}_\pi(\mathsf{o}') = \mathcal{M}_\pi(\mathsf{o}'') \tag{3.15}$$

$$c_2 : \qquad \textit{For any } \mathsf{m} \in \mathcal{M}_\pi(\mathsf{o}') : |Q^\pi(\mathsf{o}', \mathsf{m}) - Q^\pi(\mathsf{o}'', \mathsf{m})| < \epsilon, \tag{3.16}$$

where $\mathcal{M}_\pi(\mathsf{o}') = \{\mathsf{m} \in \mathcal{M} : \pi(\mathsf{m}|\mathsf{o}') > 0\}$.

Theorem 8. *Consider a multi-agent system in which agents are subject to local observability and local action selection. If agents are allowed to communicate through communication channels with a bit-budget R-bits at each time step, the maximum achievable expected return of the multi-agent system following SAIC algorithm will be in a small neighbourhood of the same MAS if it was controlled with a centralized unit under perfect communications:*

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} - \mathbb{E}_{p_{\pi_i^m, \pi_i^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} < \frac{2\,\epsilon}{(1-\gamma)^2}, \tag{3.17}$$

where γ is the discount factor and ϵ should be computed according to lemma 9, conditioned on the lumpability of the original MDP - lumpability condition.

Proof. Appendix 3.10. ■

In Theorem 8, we will show that the error gap between

Lemma 9. *Given the partition $\mathcal{P}_i = \{\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,2^R}\}$ that is obtained by solving eq. (3.38) during the centralized training phase, all subsets $\mathcal{P}_{i,k}$ for $k \in \{1, 2, ..., 2^R\}$ are ϵ-cost-uniform with respect to the optimal joint policy $\pi^*(\cdot)$ where ϵ can be obtained by the following*

$$\epsilon/2 = \max_{k, o_i} \left| V^*(\mathsf{o}_i(t)) - \mu_k^{'} \right|. \tag{3.18}$$

Proof. Following definition 7 and eq. (3.13) the proof is straightforward. ■

3.5 Performance Evaluation

In this section, we evaluate our proposed schemes via numerical results for a particular geometric consensus problem with finite observability called the rendezvous problem. Geometric consensus problems arise in numerous emerging applications such as UAV/vehicle platooning - making them a meaningful application area for the framework proposed by this chapter/thesis [208]. The numerical results achieved by SAIC will prove the suitability of the proposed framework as a potential enabling technology for vehicle/UAV platooning under limited communications.

The rendezvous problem, which is a sub-category of the geometric consensus, has been previously investigated in the literature [265,277], whereas in our case the inter-agent communication channel is set to have a limited bit-budget. The rendezvous problem is of particular interest to us, also because it allows us to consider a cooperative MAS comprising of multiple agents that are required to communicate for their coordination task. In particular, as detailed in subsection 3.5.1, if the communication between agents is not efficient, at any time step t each agent i will only have access to its local observation $\mathsf{o}_i(t)$, which is its own location in the case of rendezvous problem. This mere information is insufficient for an agent to attain the larger reward C_2, but is sufficient to attain the smaller reward C_1. Accordingly, compared with cases in which no communication between agents is present, in the set up of the rendezvous problem, efficient communication policies can increase the attained objective function of the MAS up to six-folds, as will be seen in Fig. 4. The system operates in discrete time, with agents taking actions and communicating in each time step $t = 1, 2, \ldots$. We consider a variety of grid worlds with different size values N and different locations for the goal-point ω^T. We compare the proposed SAIC and LBIC with (i) the centralized Q-learning scheme and (ii) the Conventional Information Compression (CIC) scheme which is explained in subsection 3.5.2. Changing the reward function can also build new scenarios. For example, a reward function that encourages the agents to come together as close as possible but not collide with each other can emulate a vehicle platooning scenario. While useful, it is outside the scope of our work to investigate the response of the multi-agent system to different rewarding schemes. Note that, according to Theorem 3, regardless of the definition of the reward function, the geometric consensus problem (or in general the joint quantization and control problem) can be solved by SAIC if the necessary lumpability and joint observability

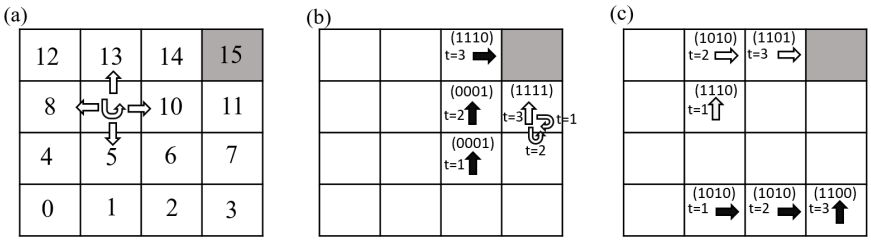


Figure 3.5: The rendezvous problem when $n = 2$, $N = 4$ and $\omega^T = 15$: (a) illustration of the observation space, Ω, i.e., the location on the grid, and the environment action space $\mathcal{M}$, denoted by arrows, and of the goal state ω^T, marked with gray background; (b) demonstration of a sampled episode, where arrows show the environment actions taken by the agents (empty arrows: actions of agent 1, solid arrows: actions of agent 2) and the $B = 4$ bits represent the message sent by each agent. A larger reward $C_2 > C_1$ is given to both agents when they enter the goal point at the same time, as in the example; (c) in contrast, C_1 is the reward accrued by agents when only one agent enters the goal position [4].

conditions are met, and centralized training phase is feasible. As the number of agents n increases, the Q-learning for the centralized training phase becomes increasingly demanding in terms of computational complexity; this is where SAIC's bottleneck lies.

3.5.1 Rendezvous Problem

As illustrated in Fig. 3.5, in a rendezvous problem, multiple agents operate on an $N \times N$ grid world and aim at arriving at the same time at the goal point on the grid. Each agent $i \in \mathcal{N}$ at any time step t can only observe its own location $\mathsf{o}_i(t) \in \Omega$ on the grid, where the observation space is $\Omega = \{0, 1, ..., n^2 - 1\}$. Each episode terminates as soon as an agent or more visit the goal point which is denoted as $\omega^T \in \Omega$. That is, at any time step t that the observation of each agent $i \in \mathcal{N}$ is a member of Ω^T, the episode will be terminated - so the time horizon M is non-deterministic. The subset $\mathcal{S}^T \subset \mathcal{S}$ also defines all state realizations where one or more agents are in the goal location i.e.,

$$\mathcal{S}^T = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \exists i \in \mathcal{N} : \mathsf{o}_i(t) \in \omega^T\}.$$

We also define the subset $\mathcal{S}^T_{n'} \subset \mathcal{S}^T$ that includes all the terminal states where only n' number of agents have arrived at the goal location i.e.,

$$\mathcal{S}^T_{n'} = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \forall i \in \mathcal{N}' : \mathsf{o}_i(t) \in \omega^T\},$$

where $\mathcal{N}' \subseteq \mathcal{N}$ is a subset of all agents with size $|\mathcal{N}'| = n'$. Following the same definition for $\mathcal{S}^T_{n'}$, the subset $\mathcal{S}^T_n$ is equivalent to the set of all terminal states where all agents are at

the goal location. At time $t = 1$, the initial position of all agents is randomly and uniformly selected amongst the non-goal states, i.e., for each agent $i \in \mathcal{N}$ the initial position of the agent is $\mathsf{o}_i(1) \in \Omega - \{\omega^T\}$.

At any time step $t = 1, 2, ...$ each agent i observes its position, or environment state, and acquires information about the position of the other agents by receiving a communication message vector $\mathsf{c}_{-i}(t)$ sent by the other agents $j \in \mathcal{N}_{-i}$ at the time step t. Based on this information, agent i selects its environment action $\mathsf{m}_i(t)$ from the set $\mathcal{M} = \{\text{Right}, \text{Left}, \text{Up}, \text{Down}, \text{Stop}\}$, where an action $\mathsf{m}_i(t) \in \mathcal{M}$ represent the horizontal/vertical move of agent i on the grid at time step t. For instance, if an agent i is on a grid-world as depicted on Fig. 3.5 (a), and observes $\mathsf{o}_i(t) = 4$ and selects "Up" as its action, the agent's observation at the next time step will be $\mathsf{o}_i(t+1) = 8$. If the position to which the agent should be moved is outside the grid, the environment is assumed to keep the agent in its current position. We assume that all these deterministic state transitions are captured by $T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t))$, which can determine the observations of agents in the next time step $t+1$ following

$$\langle \mathsf{o}_1(t+1), ..., \mathsf{o}_n(t+1) \rangle = T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)).$$

Accordingly, given observations $\langle \mathsf{o}_i(t+1), ..., \mathsf{o}_n(t+1) \rangle$ and actions $\langle \mathsf{m}_1(t+1), ..., \mathsf{m}_n(t+1) \rangle$, all agents receive a single team reward

$$r(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) = \begin{cases} C_1, & \text{if } P_1 \\ C_2, & \text{if } P_2, \\ 0, & \text{otherwise,} \end{cases} \tag{3.19}$$

where $C_1 < C_2$ and the propositions P_1 and P_2 are defined as $P_1 : T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) \in \mathcal{S}^T - \mathcal{S}_n^T$ and $P_2 : T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) \in \mathcal{S}_n^T$. When only a subset $\mathcal{N}'$, $|\mathcal{N}| = n' < n$ of agent arrives at the target point ω^T, the episode will be terminated with the smaller reward C_1 being obtained, while the larger reward C_2 is attained only when all agents visit the goal point at the same time. Note that this reward signal encourages coordination between agents which in turn can benefit from inter-agent communications.

Furthermore, at each time step t agents choose a communication message to send to the other agent by selecting a communication action $\mathsf{c}_i(t) \in \mathcal{C} = \{0, 1\}^R$ of R bits, where R (bits per channel use / per time step) is the fixed bit-budget of all inter-agent communication

channels. The goal of the MAS is to maximize the average return by solving the problem (3.10).

3.5.2 Conventional Information Compression In multi-agent Coordination Tasks

As a baseline, we consider a conventional scheme that selects communications and actions separately. For communication, each agent i sends its observation $\mathsf{o}_i(t)$ to the other agents by following policy $\pi_i^c(\cdot)$. According to this policy the agent's observation $\mathsf{o}_i(t)$ will be mapped to a binary bit sequence $\mathsf{c}_i(t)$, using an injective (and not necessarily surjective) mapping $f_1 : \Omega \to \{0,1\}^R$. Consequently, the communication policy π_i^c becomes deterministic and follows

$$\pi_i^c(\mathsf{c}_i(t+1)|\mathsf{o}_i(t)) = \delta\Big(\mathsf{c}_i(t+1) - f_1(\mathsf{o}_i(t))\Big). \tag{3.20}$$

Agent i obtains an estimate $\mathsf{c}_j(t)$ of the observation of all agents $j \in \mathcal{N}_{-i}$ by having access to a quantized version of $\mathsf{o}_j(t)$. This estimate is used to define the environment state-action value function $Q_j^m(\mathsf{o}_i(t), \mathsf{c}_{-i}(t), \mathsf{m}_i(t))$. This function is updated using Q-learning and the UCB policy in a manner similar to Algorithm 1, with no communication policy to be learned.

This communication strategy is proven to be optimal [37], if the inter-agent communication does not impose any cost on the cooperative objective function, the communication channel is noise-free and the bit-budget of communication channels are larger than the entropy rate of the observation process $R \geq H(\mathbf{o}_i)$. Under these conditions, and when the dynamics of the environment are deterministic, each agent i can distributively learn the optimal policy $\pi_i^m(\cdot)$, using value iteration or its model-free variants e.g., Q-learning [249]. While this communication policy is optimal only with a channel bit-budget $R \geq H(\mathbf{o}_j)$, in this thesis, we are focused on the scenarios with $R \leq H(\mathbf{o}_j)$. Therefore, due to the bit-budget of the communication channel, a form of TODC is required.

Note that compression before a converged action policy is not possible, since all observations are a priori equally likely. Thus, we first train the CIC on a communication channel with unlimited capacity. Afterwards, when a probability distribution for observations is obtained, by applying Lloyd's algorithm [274], we define an equivalence relation on the observation space Ω with 2^R numbers of equivalence classes $\mathcal{Q}_1, ..., \mathcal{Q}_{2^R}$. According to the defined equivalence relation by Lloyd's algorithm, we can uniquely define the mapping $f_1 : \Omega \to \{0,1\}^R$ that

maps each agent i's observation $\mathsf{o}_i(t)$ to a communication message $\mathsf{c}_i(t)$. The inverse $f_1^{-1}(\cdot)$ of the quantization mapping that maps agent j's quantized observation $\mathsf{c}_j(t)$ into a estimated observation is not an injective mapping anymore. That is, by receiving the communication message $\mathsf{c}_j(t) \in \mathcal{Q}_k \subset \mathcal{C}$ agent i can not retrieve $\mathsf{o}_j(t)$ but understands the observation of agent j has been a member of $\mathcal{Q}_k$. Note that CIC algorithm has a limitation, as it requires the first round of training to be done over communication channels with unlimited capacity.

3.5.3 Results

To perform our numerical experiments, rewards of the rendezvous problem are selected as $C_1 = 1$ and $C_2 = 10$, while the discount factor is $\gamma = 0.9$. A constant learning rate $\alpha = 0.07$ is applied, and the UCB exploration rate $c = 12.5$. In any figure that the performance of each scheme is reported in terms of the averaged discounted cumulative rewards, the attained rewards throughout training iterations are smoothed using a moving average filter of memory equal to 10% of the experiment iterations. We will use the terms "value of the collaborative objective function", "value of the objective function" and "average return" interchangeably throughout this section. Regardless of the grid-world's size and goal location, the grids are numbered row-wise starting from the left-bottom as shown in Fig. 3.5-a. Apart from Fig. 3.7 that illustrates the result related to a rendezvous problem for a three-agent system, other figures have been obtained when experimenting in a two-agent environment. Fig. 3.6 illustrates the performance of the proposed SAIC as well as six other benchmark schemes

- Centralized Q-learning under perfect communications.
- Learning based information compression (LBIC) is a different indirect scheme to design task-oriented communications which performs the joint design of communication and control policies through reinforcement learning following an algorithm similar to the one proposed in [4].
- CIC, see the details of CIC in subsection 3.5.2.
- Heuristic non-communicative (HNC) algorithm is a direct heuristic scheme which exploits the domain knowledge of its designer about the rendezvous task - making it not applicable to any other task rather than the rendezvous problem. The domain knowledge is utilized to design a control policy where no communication is present. In HNC,

agents approach the goal point and wait next to it for a large enough number of time-steps to make sure the other agent has also arrived there. Only after that, they will get into the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task.

- Heuristic optimal communication (HOC) algorithm is a direct heuristic scheme which exploits the domain knowledge of its designer about the rendezvous task - making it not applicable to any other task rather than the rendezvous problem. The domain knowledge is utilized to design jointly optimal communication and control policies. In HNC, agents approach the goal point and wait next to it until they hear from the other agent it also has arrived there. Only after that, they will get into the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task.

- Hybrid scheme uses the abstract representation of agents' observations according to SAIC with $R = 2$ bits and feeds these latent observations to a centralized controller. The central controller learns the joint action selection of both agents using Q-learning.

It is imperative to recall that, not all the schemes evaluated by Fig. 3.6 are benefit from indirect designs - making them not sufficiently general to be applied to all other multi-agent communication problems with rate-limited inter-agent channels. Regardless of their effectiveness, SAIC, LBIC, CIC and Hybrid are indirect schemes potentially applicable to any other task-oriented compression problem. Whereas, HNC and HOC are tailor-made for the rendezvous problem. In other words, the knowledge that we have about the rendezvous task is already embedded in HNC and HOC to enable the most effective communication/control strategies. HNC and HOC, however, allow us to understand how effective other indirect approaches are even when no knowledge about the specific rendezvous task is embedded in them.

The performance is measured in terms of the expected sum of discounted rewards in a rendezvous problem. The grid-world is considered to be of size $N = 8$ and its goal location to be $\omega^T = 22$. The bit-budget of the channel between the two agents is $R = 2$ bits per time step. Since centralized Q-learning is not affected by the limitation on the channel's bit-budget, it achieves optimal performance after sufficient training, 160k iterations. The CIC, due to the insufficient bit-budget of the communication channel, never achieves the

optimal solution. The LBIC, however, is seen to outperform the CIC, although it is trained and executed fully distributedly. While enjoying a fast convergence, it is observed that the SAIC can achieve optimal performance by less than 1% gap, whereas the performance gap for the LBIC and CIC are much more pronounced ranging from 20% to 30%. The yellow curve showing the performance of the CIC with no communication between agents would show us the best performance of distributed reinforcement learning that can be achieved if no communication between agents is in place without having any domain knowledge - that is present in the HOC and HNC. In fact, the better performance of any scheme compared with the yellow curve, is the sign that the scheme is either benefiting from some effective communication between agents or from some domain knowledge. Note that, when inter-agent communication is unavailable, i.e., $R = 0$ bit per time step, there would be no difference in the performance of the CIC, SAIC or LBIC as all of them use the same algorithm to find out the action policy $\pi_i^m(\cdot)$. We also recall the fact that both the CIC and SAIC require a separate training phase which is not captured by Fig. 5. SAIC requires a centralized training phase - to perform the computations demonstrated in line 5 of the algorithm 1 - and CIC a distributed training phase with unlimited capacity of inter-agent communication channels. The performance of these two algorithms in Fig. 5 is plotted after the first phase of training.

Similar to Fig. 3.6, the performance of SAIC is illustrated in Fig. 3.7, this time in a $n = 3$ three-agent system. In this case, the grid-world is considered to be of size $N = 3$ and its goal location to be $\omega^T = 9$. The bit-budget of the inter-agent communication channels is set to be $R = 1$ bits per time step. The shaded area around the curve corresponding to SAIC, shows the standard deviation of SAIC in the training as well as the execution phases - at any given training episode k the width of the shaded curve is equal to the standard deviation of SAIC's return from the training episode k to the episode $k - 1000$. This figure illustrates the very robust performance of SAIC in a three-agent scenario. For this particular experiment we used decaying epsilon greedy policies with the starting value of $\epsilon = 1$ and the ending value of $\epsilon = 0.03$. To overcome the issue of credit assignment in multi-agent systems - see e.g., [276] to get familiar with the concept, here we used a different reward function via which we trained the agents. Accordingly, given observations $\langle \mathsf{o}_i(t+1), ..., \mathsf{o}_n(t+1) \rangle$ and actions $\langle \mathsf{m}_1(t+1), ..., \mathsf{m}_n(t+1) \rangle$, all agents receive a single team reward

$$r(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) = \begin{cases} C_2^{n'-1}, & \text{if } P_3, \\ 0, & \text{otherwise,} \end{cases} \tag{3.21}$$

where the proposition P_3 is defined as $P_3 : T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) \in \mathcal{S}_n^{T'}$. When a subset $\mathcal{N}'$, $|\mathcal{N}| = n' \leq n$ of agent arrives at the target point ω^T, the episode will be terminated with the reward $C_2^{n'-1}$ being obtained, while the largest reward C_2^{n-1} is attained only when all agents visit the goal point at the same time. Note that this reward signal encourages coordination between agents which in turn can benefit from inter-agent communications.

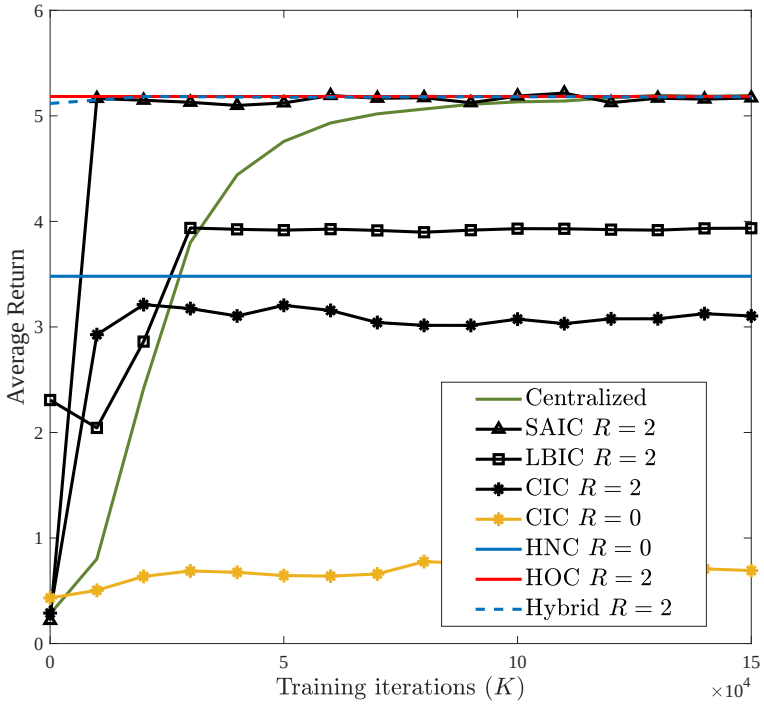


Figure 3.6: A comparison between all seven schemes in terms of the achievable objective function with the bit-budget of $R = 2$ bits per channel use/time steps and number of training iterations/episodes $K = 200k$.

To explain the underlying reasons for the remarkable performance of the SAIC, Fig. 3.8 is provided so that equivalence classes $\{\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,2^R}\}$ computed by the SAIC can be seen - all the locations of the grid shaded with the same colour belongs to the same ϵ-cost-uniform equivalence class. The SAIC is extremely efficient in performing state aggregation such that the loss of observation information barely incurs any loss on the achievable sum of discounted rewards - also depicted in Fig. 5. The Fig. 3.8-(a), illustrates the state aggregation adopted by the SAIC, for which the average return is illustrated in Fig. 4. It is illustrated in Fig. 3.8-(a) that how the SAIC performs observation compression with ratio $R_c = 3 : 1$, while it leads to nearly no performance loss for the collaborative task of the MAS. Here the definition of compression ratio follows $R_c = \lceil H(\mathbf{o}_i(t)) \rceil / \lceil H(\mathbf{c}_i(t)) \rceil$. It was observed in 3.8 that the observation clusters identified by SAIC have not been linearly separable under

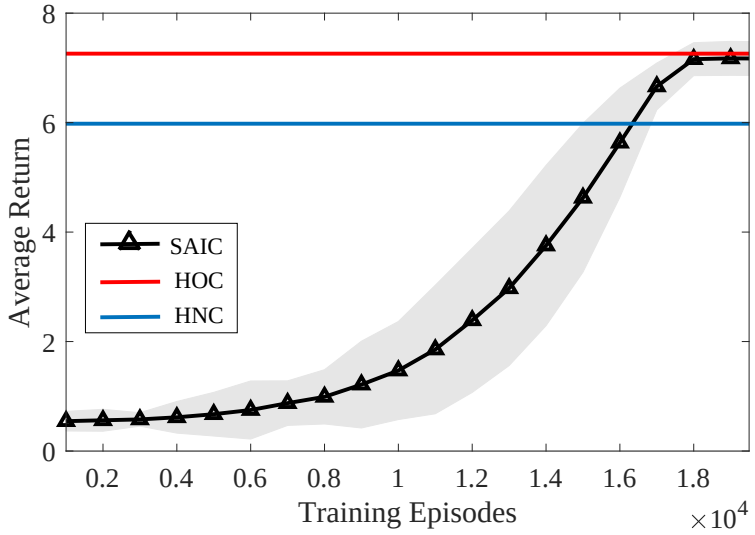


Figure 3.7: A comparison between SAIC, HOC and HNC within a three-agent system in terms of the system's average return with the bit-budget of $R = 1$ bit per time steps and number of training iterations/episodes $K = 20k$. The shaded area around SAIC's curve shows the standard deviation of SAIC in its performance.

their original representation. In contrast, when clustered according to their values, as seen in Fig. 3.9, observation points become linearly separable. Fig. Fig. 3.9, allows us to see how precise the approximation of $V_{\pi^{m*},\pi^c}(\mathbf{o}_i(1), \mathbf{c}_{-i}(1))$ by the value function $V^*(\mathsf{o}_i(t), \mathsf{c}_{-i}(t))$ is - suggested by lemma 12. The figure illustrates the values for both $V_{\pi^{m*},\pi^c}(\mathbf{o}_i(1), \mathbf{c}_{-i}(1))$ and $V^*(\mathsf{o}_i(t), \mathsf{o}_{-i}(t))$, where $\mathsf{o}_i(t) = 21$ and $\mathsf{o}_{-i}(t)$ can take on possible values in Ω. For instance the values 7.2 mentioned on the right down corner of the grid demonstrates the value of $V^*(\mathsf{o}_i(t), \mathsf{o}_j(t))$ when $\mathsf{o}_i(t) = 20$ and $\mathsf{o}_j(t) = 7$. This figure also allows finding the value of ϵ for all ϵ-cost-uniform groups.

We also investigate the impact of channel bit-budget R on the value of average return achieved by the LBIC, SAIC and CIC, in Fig. 3.10. In this figure, the normalized value of average return achieved for any scheme at any given R is shown. As per (3.22), the average return for the scheme of interest is computed by $\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}$, where $\pi_i^m(\cdot)$ and $\pi_i^c(\cdot)$ are obtained by the scheme of interest after solving (3.10) with a given value of R. The average return is then normalized by dividing it to the average return $\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}$ that is obtained by the optimal centralized policy $\pi^*(\cdot)$. The policy policy $\pi^*(\cdot)$ is the optimal solution to (3.3) under no communications constraint.

$$\frac{\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}}{\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}}. \tag{3.22}$$

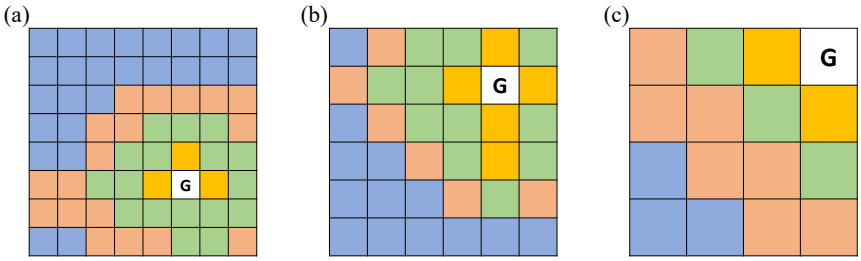


Figure 3.8: State aggregation for multi-agent communication in a two-agent rendezvous problem with grid-worlds of varied sizes and goal locations. The observation space is aggregated to four equivalence classes, $R = 2$ bits, and the number of training episodes has been $K = 1500k$, $K = 1000k$ and $K = 500k$ for figures (a) and (b) and (c) respectively. Locations with similar colours represent all the agents' observations which are grouped into the same equivalence class. The data compression ratio R_c has been seen to be 6:2, 5:2 and 4:2 in subplots a), b) and c) respectively. It is also observed that the observation clusters identified by SAIC have not been linearly separable under their original representation. In contrast, when clustered according to their values, observation points become linearly separable - see also Fig. 3.9 .

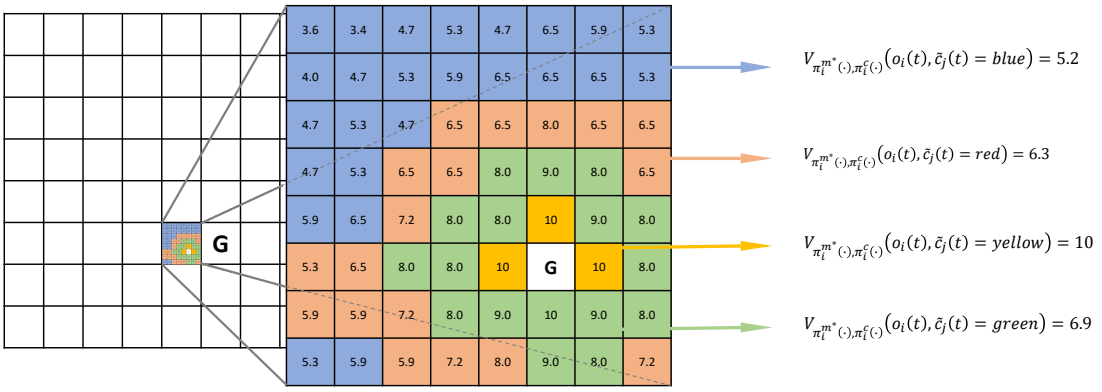


Figure 3.9: Left grid-world shows the observation space Ω, amongst which one particular observation is chosen $\mathsf{o}_i(t) = 20$. While agent i makes this observation, agent j can potentially be at any other 64 locations of the greed. The value function $V^*(\mathsf{o}_i(t) = 20, \mathsf{o}_j(t))$ for all $\mathsf{o}_j(t) \in \Omega$ is depicted in the right grid-world, e.g. a number at location 22, shows the value function $V^*(\mathsf{o}_i(t) = 20, \mathsf{o}_j(t) = 22) = 10$. You can also see the values of $V_{\pi^{m*},\pi^c}(\mathbf{o}_i(t), \mathbf{c}_j(t))$ for $\mathsf{o}_i(t) = 20$ and all possible $\mathsf{c}_j(t) \in \mathcal{C}$ with $R = 2$ bits.

Accordingly, when the normalized objective function of a particular scheme is seen to be close to the value 1, it implies that the scheme has been able to compress the observation information with almost zero loss with respect to the achieved objective function. On one hand, it is demonstrated that the SAIC achieves the optimal performance while running with 2 bits of inter-agent communications, while it takes the CIC at least $R = 4$ bits to get to achieve a sub-optimal value of the objective function. The LBIC, on the other hand, provides more than 10% performance gain in very low rates of communication $R \in \{1, 2, 3\}$ bits per time step, compared with CIC and 20% performance gain compared with SAIC at $R = 1$

bits per time step.

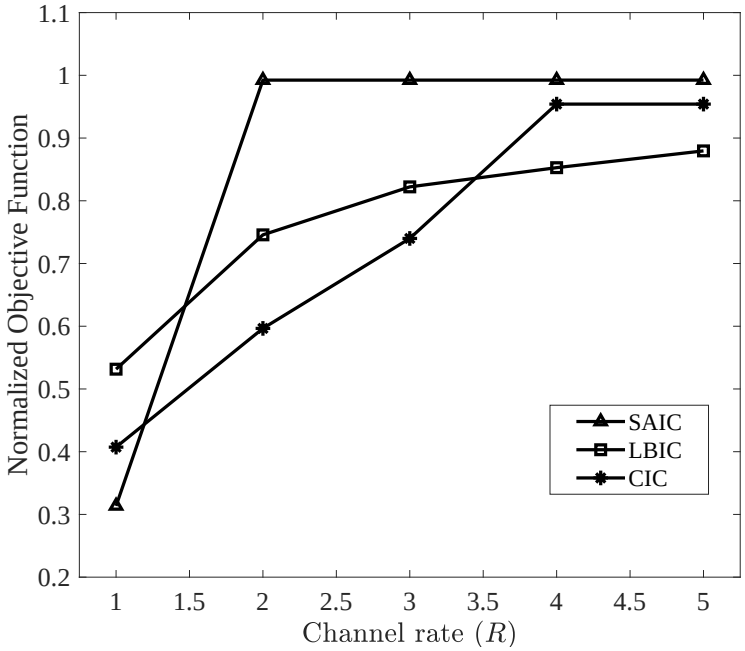


Figure 3.10: A performance comparison between several multi-agent communication and control schemes under different achievable bit rates. All experiments are performed where $N = 8$ and $\omega^T = 21$, similar to the grid-world of Fig. 3.8 -a. The number of training episodes/iterations for any scheme at any given channel bit-budget R has been $K = 200K$.

Fig. 3.11, studies the normalized objective functions attained by the LBIC, SAIC and CIC under different compression ratios R_c. A whopping 40% performance gain is acquired by the SAIC, in comparison to the CIC, at high compression ratio $R_c = 3 : 1$. This is equivalent to 66% of saving in the bit-budget with no performance drop with respect to the collaborative objective function. The SAIC, however, underperforms the LBIC and CIC at very high compression ratio of $R_c = 6 : 1$. This is due to the fact that the condition mentioned in remark 2 is not met at this high rate of compression. Moreover, the CIC scheme is seen not to achieve the optimal performance even at the compression rate of $R_c = 6 : 5$ which is due to the fact that by exceeding the compression ratio $R_c = 1 : 1$ each agent i may lose some information about the observation $\mathsf{o}_j(t)$ of the other agent which can be helpful in taking the optimal action decision.

As demonstrated through a range of numerical experiments, the weakness of conventional schemes for compression of agents' observations is that they may lose/keep information regardless of how useful they can be towards achieving the optimal objective function. In contrast, the task-based compression schemes SAIC and LBIC, for communication bit-budgets

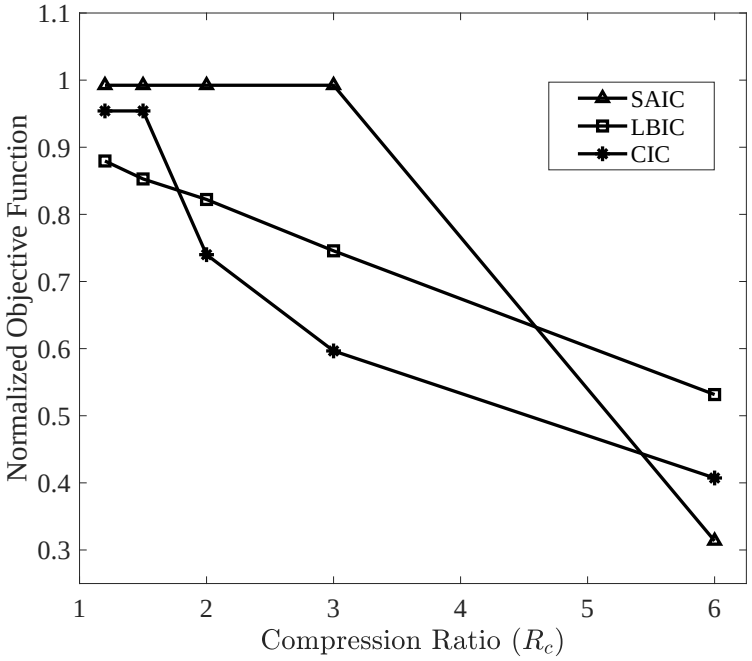


Figure 3.11: A performance comparison between several multi-agent communication and control schemes under different rates of data compression. All experiments are performed where $N = 8$ and $\omega^T = 21$. The number of training episodes/iterations for any scheme at any given bit-budget R has been $K = 200K$.

(very) lower than the entropy of the observation process, manage to compress the observation information not to minimize the distortion but to maximize the achievable value of the objective function. Even though the numerical example provided in section IV, evaluates the performance of SAIC in a problem with a very low communication bit-budget, our theoretical results are applicable in scenarios with higher communication rates, as long as the processing unit that is deployed to solve the problem (3.3) is of sufficient computational resources to solve the problem in the desire time window.

3.6 Conclusion

We have investigated the distributed joint design of communications and control for an MAS under bit-budgeted communications with the ultimate goal of maximizing the system's expected return. Since we consider a limited bit-budget for the multi-agent communication channels, task-based compression of agents' observations has been of the essence. Our proposed scheme, SAIC, which derives and solves the TODC problem can be differentiated from the conventional data quantization algorithms in the sense that it does not aim at achieving minimum possible distortion between the original signal and its reconstructed version - given

a bit-budget for inter-agent communications. In contrast, SAIC aims at achieving the minimum possible distortion between the (learned) usefulness/value of the original observation signal and the learned usefulness/value of the the reconstructed observation signal - given a bit-budget for inter-agent communications. We have demonstrated the outstanding performance of SAIC compared with the conventional data compression algorithms, by up to a remarkable 40% improvement in the achieved objective function, when being imposed with tight constraints on the communication bit-budget.

To maximize the system's expected return, we could show analytically, how one can disentangle the TODC from the control problem - given the possibility of a centralized training phase. Our analytical studies confirm that despite the separation of the TODC and control problems, we can ensure very little compromise on the MAS's average return - compared with the jointly optimal control and quantization. Since the computational complexity of Q-learning in the centralized training phase is order of $|\Omega^n \times \mathcal{M}^n|$ time complexity [278], the addition of one single agent will multiply the complexity of the centralized training by $|\Omega \times \mathcal{M}|$. Thus, the complexity of the centralized training phase becomes a hurdle for the scalability of SAIC to a high number of agents. Accordingly, improving the scalability of the algorithm as well as extending the results for non-symmetric variable bit-budgets can be useful avenues to improve the applicability of the proposed schemes.

3.7 Proof of Theorem 3

To prove this theorem we first introduce a definition in subsection 3.7.1, together with two lemmas and their proofs in subsections 3.7.2 and 3.7.3. Lastly, we complete the proof of Theorem 3, in subsection 3.7.4 leveraging the above-mentioned.

3.7.1 Task-based information compression problem: a definition

Definition 10. *[Task-based information compression (TBIC) problem] Let the higher order function Π^{m^*} be a map from the vector space $\mathcal{K}^c$ of all possible joint communication policies $\pi^c = \langle \pi_1^c(\cdot), ..., \pi_n^c(\cdot) \rangle$ to the vector space $\mathcal{K}^m$ of optimal corresponding joint control policies $\pi^m = \langle \pi_1^{m^*}(\cdot), ..., \pi_n^{m^*}(\cdot) \rangle$. Upon the availability of Π^{m^*}, by plugging it into the problem (3.10),*

we will have a new problem

$$\begin{aligned} &\max_{\pi_i^c} && \mathbb{E}_{p_{\Pi^{m^*},\pi^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\Big\{\mathbf{g}(1)\Big\}, \quad i \in \mathcal{N} \\ &s.t. && log_2|\mathcal{C}| \leq R, \end{aligned} \tag{3.23}$$

where we maximize the system's return only with respect to the joint communication policies π^c. The joint optimal control policies $\langle \pi_1^{m^*}(\cdot), ..., \pi_n^{m^*}(\cdot)\rangle$ are automatically computed by the mapping $\Pi^{m^*}(\pi_1^c(\cdot), ..., \pi_n^c(\cdot))$. The problem is called here as the TBIC problem.

3.7.2 Reformulating the objective function: a lemma

Lemma 11. *The objective function of the decentralized problem (3.10) can be expressed as*

$$\begin{aligned} &\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t'}^{t=M})}\{\mathbf{g}(t')\} = \\ &\mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(t')))}\Big\{\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=2}^{t=M}|h_i(\mathbf{s}(t')))}\{\mathbf{g}(t')|h_i(\mathbf{s}(t'))\}\Big\} = \\ &\mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(t')))}\Big\{V_{\pi^m,\pi^c}(h_i(\mathbf{s}(t')))\Big\}, \end{aligned} \tag{3.24}$$

for all $i \in \mathcal{N}$, where $V_{\pi^m,\pi^c}(h_i(\mathbf{s}(t')))$ is the solution to the Bellman equation corresponding to the joint control and communication policies π^m, π^c.

Proof. Considering the definition of the value function, given in (3.25), the proof is immediately concluded when applying Adam's law on the expectation of the value function

$$V_{\pi^m,\pi^c}(h_i(\mathbf{s}(t'))) = \mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t'+1}^{t=M})}\{\mathbf{g}(t')|h_i(\mathbf{s}(t'))\}. \tag{3.25}$$

■

3.7.3 Value of the perceived state of environment: a lemma

Lemma 12. *Using the knowledge of the solution $\pi^*(\cdot)$ to the centralized problem, we can find the optimal value of a perceived state $V^*(h_i(\mathsf{s}(t)))$ in terms of the value of the underlying state $V^*(\mathsf{s}(t))$ by*

$$V^*(h_i(\mathsf{s}(t))) = \sum_{\mathsf{o}_1(t)\in\Omega} ... \sum_{\mathsf{o}_n(t)\in\Omega} V^*(\mathsf{s}(t))\ p(\mathsf{o}_{-i}(t)|\mathsf{c}_{-i}(t)). \tag{3.26}$$

Proof.

$$V^*(h_i(\mathsf{s}(t')))) = \tag{3.27}$$
$$\mathbb{E}_{p(\{tr\}_{t'}^{M}|h_i(\mathsf{s}(t')))}\Big\{\sum\nolimits_{t=t'}^{M}\gamma^{t-1}r(\mathbf{s}(t),\mathbf{m}(t))|h_i(\mathsf{s}(t'))\Big\} =$$
$$\mathbb{E}_{p(\{\mathsf{tr}\}_{t'}^{M}|h_i(\mathsf{s}(t')))}\Big\{\mathbf{g}(t^{'})|h_i(\mathsf{s}(t'))\Big\} = \tag{3.28}$$
$$\sum\nolimits_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})p\Big(\{\mathsf{tr}\}_{t'}^{M}|h_i(\mathsf{s}(t'))\Big),$$

where the conditional probability $p\Big(\{\mathsf{tr}\}_{t'}^{M}|h_i(\mathsf{s}(t'))\Big)$ can be extended following the law of total probabilities

$$V^*(h_i(\mathsf{s}(t'))) = \sum\nolimits_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})\Bigg[\sum_{\mathsf{o}_1(t)\in\Omega}\dots\sum_{\mathsf{o}_n(t)\in\Omega}$$
$$p\Big(\{\mathsf{tr}\}_{t'}^{M}|\mathsf{o}_i(t^{'}),\mathsf{o}_{-i}(t^{'}),\mathsf{c}_{-i}(t^{'})\Big)p(\mathsf{o}_{-i}(t')|\mathsf{c}_{-i}(t'))\Bigg], \tag{3.29}$$

where $\mathsf{o}_{-i}(t')$ is the observation vector of all agents $i \in \mathcal{N}_{-i}$. In eq. (3.29) $\mathsf{o}_i(t^{'}),\mathsf{o}_{-i}(t^{'})$ are sufficient statistics and can be replaced by $\mathsf{s}(t')$ and the second summation can be shifted to have

$$V^*(h_i(\mathsf{s}(t'))) =$$
$$\sum_{\mathsf{o}_1(t)\in\Omega}\dots\sum_{\mathsf{o}_n(t)\in\Omega}\sum_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})p\Big(\{\mathsf{tr}\}_{t'}^{M}|\mathsf{s}(t'))\Big)p(\mathsf{o}_{-i}(t)|\mathsf{c}_{-i}(t)), \tag{3.30}$$

where $\sum_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})p\Big(\{\mathsf{tr}\}_{t'}^{M}|\mathsf{s}(t'))\Big)$ can be replaced with $V^*(\mathsf{s}(t))$, concluding the proof. ■

3.7.4 Proof of Theorem 3

Proof. Further to the result of lemma 11 and eq. (3.24), the original problem (3.10) can be expressed by

$$\begin{aligned}\max_{\pi_i^m(\cdot),\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathsf{s}(1)))}\Big\{V_{\pi^m,\pi^c}(h_i(\mathsf{s}(1)))\Big\},\\ \text{s.t.} \quad & log_2|\mathcal{C}|\leq R,\end{aligned} \tag{3.31}$$

for $i \in \mathcal{N}$. Now by following definition 10 and plugging $\Pi^{m^*}(\cdot)$ into the problem (3.31) we obtain the TBIC problem

$$\begin{aligned} \max_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\Pi^{m^*}(\pi^c),\pi^c}(h_i(\mathbf{s}(1)))}\Big\{V_{\Pi^{m^*}(\pi^c),\pi^c}(h_i(\mathbf{s}(1)))\Big\}, \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R, \;\; i \in \mathcal{N}. \end{aligned} \tag{3.32}$$

We continue by following lemma 12, to be able to substitute $V_{\Pi^{m^*}(\pi^c),\pi^c}(h_i(\mathbf{s}(1)))$ with its approximator $V^*(h_i(\mathbf{s}(1)))$. This brings us to the approximated TBIC problem

$$\begin{aligned} \max_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^*,\pi^c}(h_i(\mathbf{s}(1)))}\Big\{V^*(h_i(\mathbf{s}(1)))\Big\} \;\; i \in \mathcal{N} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R. \end{aligned} \tag{3.33}$$

Note that the optimizers of the problem (3.33) and (3.34) are identical since the additional term $\mathbb{E}\{V^*(\mathbf{s}(t))\}$ is independent from the communication policy $\pi_i^c(\cdot)$. Furthermore, the problem (3.34) is now expressed as a form of data quantization problem with mean absolute difference of the value functions $V^*(\mathbf{s}(t))$ and $V^*(h_i(\mathbf{s}(1)))$ as the measure of distortion. This interpretation of problem (3.34) can be better understood later by seeing the eq. (3.35).

$$\begin{aligned} \min_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(1)))}\Big\{V^*(\mathbf{s}(1)) - V^*(h_i(\mathbf{s}(1)))\Big\} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R, \end{aligned} \tag{3.34}$$

and since $V^*(\mathbf{s}(1))$ is always larger than $V^*(h_i(\mathbf{s}(1)))$, the problem above can also be written as

$$\begin{aligned} \min_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(1)))}\Big\{|V^*(\mathbf{s}(1)) - V^*(h_i(\mathbf{s}(1)))|\Big\} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R, \end{aligned} \tag{3.35}$$

concluding the proof of Theorem 3. ■

3.8 Proof of Lemma 4

Proof. The term $\mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(1)))}\Big\{V^*(\mathbf{s}(1)) - V^*(h_i(\mathbf{s}(1)))\Big\}$ can be estimated by computing it over the empirical distribution of $\mathbf{s}(1)$. Note that the empirical joint distribution of $h_i(\mathbf{s}(1))$ can be obtained by following the communication policy $\pi_i^c(\cdot)$ on the empirical distribution of

$\mathbf{s}(1)$. Therefore, the problem (3.34) can be rewritten as

$$\begin{aligned} \min_{\pi_i^c(\cdot)} \quad & \sum_{\mathsf{o}_i(1)\in\Omega} \dots \sum_{\mathsf{o}_n(1)\in\Omega} \Big| V^*(\mathsf{s}(t)) - V^*(h_i(\mathsf{s}(t))) \Big|, \ \forall i \in \mathcal{N} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R. \end{aligned} \tag{3.36}$$

Quantization levels are disjoint sets $\mathcal{P}_{i,k} \subset \Omega$, where their union $\cup_{k=1}^{2^R} \mathcal{P}_{i,k}$ will cover the entire Ω. Each quantization level is represented by only one communication message $\mathsf{c}_j(t) = \mathsf{c}_k \in \mathcal{C}$. Further to lemma 12, the value of $V^*(h_i(\mathbf{s}(t)))$ can be computed by empirical mean (3.26).

The quantization problem (3.36) becomes a k-median clustering problem

$$\min_{P_i} \quad \sum_{j\in N_{-i}}^{o_j(t)\in\Omega} \sum_{k=1}^{2^R} \sum_{\mathsf{o}_i(t)\in\mathcal{P}_{i,k}} \Big| V^*(\mathsf{o}_i(t), \mathsf{o}_j(t)) - \mu_k \Big|, \tag{3.37}$$

where $\mathcal{P}_i = \{\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,2^R}\}$ is a partition of Ω, and the first summation $\sum_{j\in N_{-i}}^{o_j(t)\in\Omega}$ is a concatenation of $n-1$ summations each one acting over $\mathsf{o}_j(t) \in \Omega$ where $j \in \mathcal{N}_{-i}$.

By taking the mean of $V^*(\mathsf{s}(t))$ over the empirical distribution of $\mathsf{o}_j(t)$, $\forall j \in \mathcal{N}_i$, we can also marginalize out $\mathsf{o}_j(t)$, $\forall j \in \mathcal{N}_i$. Again, it does not change the solution of the problem and we will have

$$\min_{P_i} \quad \sum_{k=1}^{2^R} \sum_{\mathsf{o}_i(t)\in\mathcal{P}_{i,k}} \Big| V^*(\mathsf{o}_i(t)) - \mu_k^{'} \Big|, \tag{3.38}$$

in which $\mu_k^{'} = \sum_{\mathsf{o}_j(t)\in\mathcal{P}_{i,k}} \mu_k$ will approximate $V^*(\mathsf{c}_i(t))$. ■

To gain more insight about the meaning of this task-based information compression, it is useful to take a look at the conventional quantization problem which is adapted to our problem setting in eq. (3.39), where $\mathbf{c}_j = \pi_j^c(\mathbf{o}_j(1))$. In fact, the compression scheme applied in the CIC, explained in subsection (3.5.2), is obtained by solving the following problem

$$\min_{\pi_i^c(\cdot)} \sum_{\mathsf{o}_i(1)\in\Omega} \Big| \mathsf{o}_i(t) - \mathsf{c}_i(t) \Big|^2, \quad \text{s.t. } log_2|\mathcal{C}| \leq R, \tag{3.39}$$

which can be solved optimally by the Lloyd's algorithm [274].

3.9 Proof of Lemma 5

Proof. Further to the law of iterated expectations, $V^*(\mathsf{o}_i(t'))$ can be expressed as

$$V^*(\mathsf{o}_i(t')) = \mathbb{E}_{p(\mathbf{o}_{-i}(t'))}\Bigg\{\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t'+1}^{t=M}|\mathsf{o}_i(t'),\mathbf{o}_{-i}(t'))}\Bigg\{ \mathbf{g}(t')|\mathbf{o}_i(t') = \mathsf{o}_i(t'), \mathbf{o}_{-i}(t')\Bigg\}\Bigg\} = \sum_{\mathsf{o}_{-i}(t')\in\Omega^{n-1}} p(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t'))\mathbb{E}_{\pi^*}\Bigg\{\mathbf{g}(t')|\mathsf{o}_i(t'), \mathsf{o}_{-i}(t')\Bigg\} \tag{3.40}$$

where the expectation of the last term is the optimal value of the state $\mathsf{s}(t') = \langle \mathsf{o}_i(t'), \mathsf{o}_{-i}(t')\rangle$ of the underlying MDP

$$V^*(\mathsf{s}(t')) = \mathbb{E}_{\pi^*}\Bigg\{\mathbf{g}(t')|\mathsf{o}_i(t'), \mathsf{o}_{-i}(t')\Bigg\}. \tag{3.41}$$

Following Bellman optimality equation $V^*(\mathsf{s}(t'))$ can be obtained by centralized Q-learning following

$$\begin{aligned} V^*(\mathsf{s}(t')) &= \max_{m\in M^n} Q^*(\mathsf{s}(t'), \mathsf{m}(t')) \\ &= \mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t'+1}^{t=M}|\mathsf{o}_i(t'),\mathsf{o}_{-i}(t'))}\Bigg\{\mathbf{g}(t')|\mathsf{o}_i(t'), \mathsf{o}_{-i}(t')\Bigg\}. \end{aligned} \tag{3.42}$$

Using (3.40) and (3.42) we can simply compute $V^*(\mathsf{o}_i(t'))$ by

$$V^*(\mathsf{o}_i(t)) = \sum_{\mathsf{o}_{-i}(t)\in\Omega^{n-1}} \max_m \; Q^*(\mathsf{s}(t), \mathsf{m}(t))p(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t)). \tag{3.43}$$

■

3.10 Proof of Theorem 8

Proof. Without loss of generality, we have written the proof of this theorem for a two agent scenario to improve the readability. Given the proof for the two-agent system, the extension to a multi-agent system is straightforward. According to the [259](Lemma 1), optimal state values of the aggregated MDPs (the environment as is seen by one agent during the decentralized training phase of SAIC) are in a small neighbourhood of the optimal values

corresponding to the optimal solution to the original underlying MDP:

$$\begin{aligned} &\forall \mathsf{o}_j \in \Omega \text{ and } \text{ and } \forall i \in \{1,2\}, j \neq i : \\ &|V^*(\mathsf{o}_i, \mathsf{o}_j) - V_i^m(\mathsf{o}_i, c_j^{(k)})| < \frac{2\epsilon}{(1-\gamma)^2}, \end{aligned} \tag{3.44}$$

where $V_i^m(\cdot)$ is the value function corresponding to $\pi_i^{m,SAIC}(\cdot)$. The communication signal $c_j^{(k)} \in \mathcal{C}$ is agent j's communicated message and at the same time is the k-th element of the communication space $\mathcal{C} = \{c^{(1)}, c^{(2)}, ..., c^{|\mathcal{C}|}\}$ i.e., $c_j^{(k)} = c^{(k)}$. Following the eq. (3.24), one can write the expected return of the system under centralized scheme as :

$$\begin{aligned} &\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0))\Big\} = \\ &\sum_{\mathbf{o}_j \in \Omega} \sum_{\mathbf{o}_i \in \Omega} V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)) p_{\mathbf{o}_i, \mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)), \end{aligned} \tag{3.45}$$

where the second expectation is taken over the joint probability distribution $p_{\pi^*}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0))$ of $\mathbf{o}_i$ and $\mathbf{o}_j$ when following the action policy $\pi^*(\cdot)$. This equation can be extended for multi-agent case only by taking a summation over each agent's observation space on the left-hand side. Similarly, following the eq. (3.24), one can write the expected return of the system that is run by SAIC as:

$$\begin{aligned} &\mathbb{E}_{p_{\pi^m, \pi^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^m(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0))\Big\} = \\ &\sum_{k=1}^{|\mathcal{C}|} \sum_{\mathbf{o}_i \in \Omega} V^m(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)) p_{\mathbf{o}_i, \mathbf{c}_j}(\mathsf{o}_i(t_0), \mathsf{c}_j^{(k)}(t_0)). \end{aligned} \tag{3.46}$$

We can rewrite the joint probability $p_{\mathbf{o}_i, \mathbf{c}_j}(\mathsf{o}_i(t_0), \mathsf{c}_j^{(k)}(t_0))$ as

$$p_{\mathbf{o}_i, \mathbf{c}_j}(\mathsf{o}_i(t_0), \mathsf{c}_j^{(k)}(t_0)) = \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} p_{\mathbf{o}_i, \mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)), \tag{3.47}$$

where the subset $\mathcal{P}_{i,k} \subset \Omega$ stands for the set of all observation realizations o_j that are represented by $\mathsf{c}_j^{(k)}(t_0)$ according to the policy $\pi_i^{c,SAIC}(\cdot)$. Given eq. (3.47), one can express

eq. (3.46) - the expected return of the MAS under SAIC - also as

$$\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^m(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0))\Big\} = \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0)\in\mathcal{P}_{i,k}} \sum_{\mathbf{o}_i\in\Omega} V^m(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)). \tag{3.48}$$

In order for eq. (3.45) to have the arrangement of its summations similar to eq. (3.48), it is sufficient to break its left-hand summation to two parts

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0))\Big\} = \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0)\in\mathcal{P}_{i,k}} \sum_{\mathbf{o}_i\in\Omega} V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)), \tag{3.49}$$

Further to equations (3.49)-(3.48), the difference between the achievable expected return of the centralized scheme and SAIC can be explained by

$$\begin{aligned}\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} - \mathbb{E}_{p_{\pi_i^m,\pi_i^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0)\in\mathcal{P}_{i,k}} \sum_{\mathbf{o}_i\in\Omega} V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)) - \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0)\in\mathcal{P}_{i,k}} \sum_{\mathbf{o}_i\in\Omega} V^m(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)).\end{aligned} \tag{3.50}$$

We now proceed by factorizing the joint probability $p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0))$ which yields

$$\begin{aligned}\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} - \mathbb{E}_{p_{\pi_i^m,\pi_i^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0)\in\mathcal{P}_{i,k}} \sum_{\mathbf{o}_i\in\Omega} p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)) [V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)) \\ - V^m(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0))]\end{aligned} \tag{3.51}$$

Since $[V^*(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)) - V^m(\mathbf{o}_i(t_0), \mathbf{c}_j(t_0))]$ is upper-bounded by a constant term $\frac{2\epsilon}{(1-\gamma)^2}$, its weighted sum is also upper bounded by the same term $\frac{2\epsilon}{(1-\gamma)^2}$. Thus we conclude the proof of Theorem 8. We are unsure if the suggested bound is tight. The results obtained in the performance evaluation indicates a large difference between the bound offered above and the performance bound between SAIC and the optimal centralized control. ■

Chapter 4

Task-Effective Compression of Observations for the Centralized Control of a Multi-agent System Over Bit-Budgeted Channels

4.1 Introduction

As 5G is rolling out, a wave of new applications such as the internet of things (IoT), industrial internet of things (IIoT) and autonomous vehicles is emerging. It is projected that by 2030, approximately 30 billion IoT devices will be connected [14]. With the proliferation of non-human types of connected devices, the focus of the communications design is shifting from traditional performance metrics, e.g., bit error rate and latency of communications to the semantic and task-oriented performance metrics such as meaning/semantic error rate [1,21] and the timeliness of information [31]. To evaluate how efficiently the network resources are being utilized, one could traditionally measure the sum rate of a network whereas in the era of the cyber-physical systems, given the resource constraints of the network, we want to understand how effectively one can conduct a (number of) task(s) in the desired way [16,24]. We are witnessing a paradigm shift in communication systems where the targeted performance metrics of the traditional systems are no longer valid. This imposes new grand challenges in designing the communications towards the eventual task-effectiveness [24]. This line of research is also driven partly due to the success of new machine learning technologies/ algorithms under

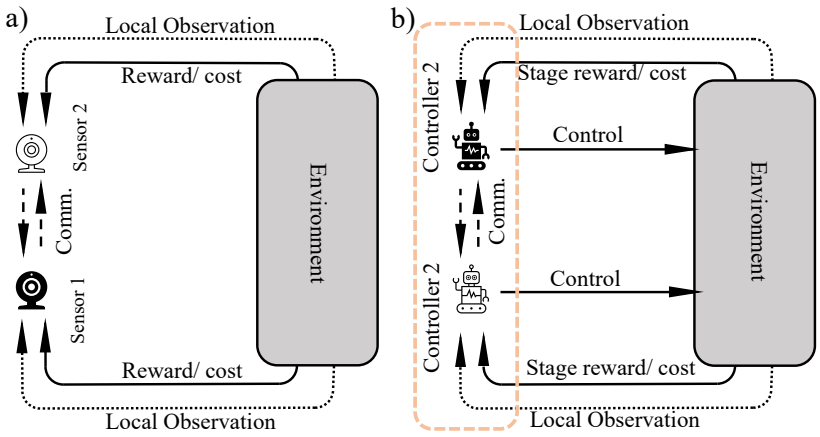


Figure 4.1: Task-effective communications for a) an estimation vs. b) a control task - the orange dashed box is detailed in Fig. 4.2 and Fig. 4.3.

the title of "emergent communications" in multi-agent systems [3]. Transfer of these new technologies/ideas to communication engineering is anticipated to have a disruptive effect in multiple domains of the design of communication systems.

According to Shannon and Weaver, communication problems can be divided into three levels [6]: (i) technical problem: given channel and network constraints, how accurately can the communication symbols/bits be transmitted? (ii) semantic problem: given channel and network constraints, how accurately the communication symbols can deliver the desired meaning? (iii) effectiveness problem: given channel and network constraints, how accurately the communication symbols can help to fulfil the desired task? While the traditional communication design addresses the technical problem, recently, the semantic problem [1, 16, 21–23] as well as the effectiveness problem [4, 9, 11, 24–29] have attracted extensive research interest.

In contrast to Shannon's technical-level communication framework, semantic communication can enhance performance by exploiting prior knowledge between source and destination [30, 31]. The semantic-based designs, however, are not necessarily task-effective [32]. One can design transmitters which compress the data with the least possible compromise on the semantic meaning being transmitted [1, 21] while the transmission can be task-unaware [33]. In contrast to semantic level and technical level communication design, the performance of a task-effective communication system is ultimately measured in terms of the average return/cost linked to the task [11]. In the (task-)effectiveness problem, we are not concerned only about the communication of meaning but also about how the message exchange is helping the receiving end to improve its performance in the expected cost/reward of an estimation

task [26, 27, 29, 31, 34] or a control task [4, 9, 11, 13, 25, 27, 35].

There are fundamental differences between the design of task-effective communications for an estimation vs. a control task - Fig. 4.1. (i) In the latter, each agent can produce a control signal that directly affects the next observations of the agent. Thus, in control tasks the source of information - local observations of the agent - is often a stochastic process with memory - e.g. linear or Markov decision processes - [4,9,11]. In the estimation tasks, however, the source of information is often assumed to be an i.i.d. stochastic process [26, 29, 34]. (ii) In the control tasks, a control signal often has a long-lasting effect on the state of the system more than for a single stage/time step e.g., a control action can result in lower expected rewards in the short run but higher expected rewards in the long run. This makes the control tasks intrinsically sensitive to the time horizon for which the control policies are designed. Estimation tasks, specifically when the observation process is i.i.d., can be solved in a single stage/ time step - since there is no influence from the solution of one stage/ time step to another i.e., each time step can be solved separately [34, 36]. (iii) The cost function for estimation tasks is often in the form of a difference/distortion function while in the control tasks it can take on many other forms.

Throughout this thesis, we focus on the effectiveness problem for the control tasks. In particular, we investigate the distributed communication design of a multiagent system (MAS) with the ultimate goal of maximizing the expected summation of per-stage rewards also known as the expected return. Multiple agents select control actions and communicate in the MAS to accomplish a collaborative task with the help of a central controller (CC) - i.e. the communication network topology of the MAS is a star topology with the hub node being the central controller and the peripheral nodes being the agents - Fig. 4.2. The considered system architecture can find applications in several domains such as Internet of Things, emerging cyber-physical systems, real-time interactive systems, vehicle-to-infrastructure communication [279] and collaborative perception [280].

4.1.1 Related works: Task-effective communications for control tasks

Authors in [4, 9, 11, 13, 25, 27, 35] consider task-effective communication design under different settings. While [25], utilizes the task-effective communication design for the specific problem of the design of application-tailored protocols over perfect communication channels, the

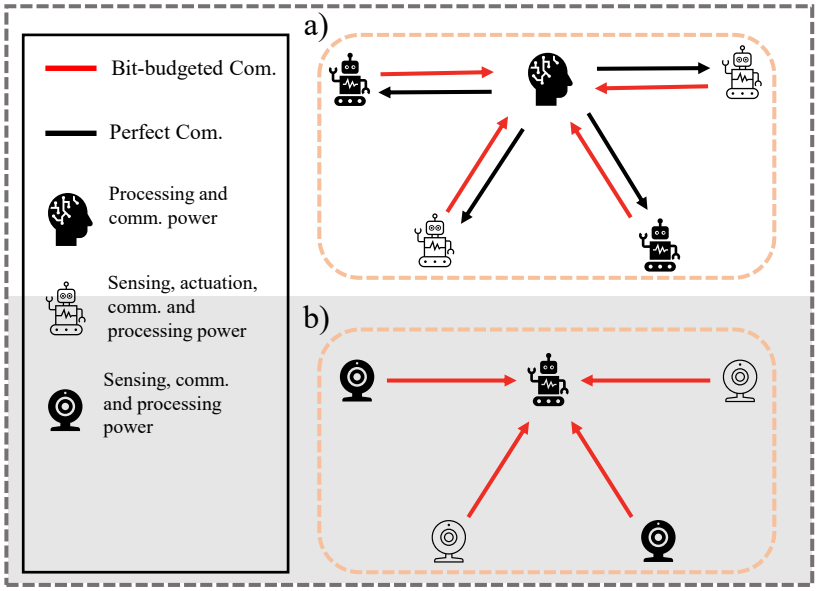


Figure 4.2: Communication topology and its applicable scenarios a) Centralized control of an MAS with collocated actuators and sensors, b) Distributed sensing with a single controller collocated with a single actuator. The orange dashed box is detailing the same box in Fig. 4.1 and Fig. 4.3 .

communication channel is considered to be imperfect in [4, 9, 11, 13, 27, 35]. Authors in [27] provide algorithmic contributions to the design of task-effective joint source channel coding for single agent systems. Task-effective joint source and channel coding for MAS is targeted by [4, 11, 27], whereas [9, 35] are focused on task-effective data compression and quantization. Similar to the current chapter, a star topology for the inter-agent communication is considered in [11, 25] whereas [25] assumes perfect communications between the hub node and the peripherals and [11] assumes imperfect communication channels at the down-link of the peripheral nodes. In contrast to all the above-mentioned work, this chapter is - to the best of our knowledge - the first to study the star topology with the uplink (agent to hub) channel be imperfect (bit-budgeted) - Fig. 4.2. Accordingly, each agent observes the environment and communicates an abstract version of its local observation to the CC via imperfect (bit-budgeted) communication channels - red links in Fig. 4.2. Subsequently, CC produces control actions that are communicated to the agents via perfect communication channels - black links in Fig. 4.2. The control actions are selected by the CC such that they maximize the average return of the collaborative task, where the return is a performance metric linked to the accomplishment of the task.

4.1.2 Contributions

In our earlier work [9], we have developed a generic framework to solve task-oriented communication problems - for a multi-agent system (MAS) with full mesh connectivity. The current work can be considered as an adoption of that framework to a new problem setting for the design of task-effective communications where agents follow a star network topology for their connectivity. In this direction, the current work transcends the applicability of the proposed framework beyond the specific problem that was solved in [9] and provides further insights into how the framework can be used in wider terms and under a wider range of settings. In particular the contributions of this work are listed below.

- Firstly, we consider a novel problem setting in which an MAS is controlled via a central controller who has access to agents' local observations only through bit-budgeted distributed communications. This problem setting can be used in collaboration perception systems as well as vehicle-to-infrastructure communications, which cannot been addressed by the problem settings investigated in the prior similar art.

- Secondly, our analytical studies establish the relationship between the considered joint communication and control design problem and conventional data quantization problems. In particular, lemma 13 shows how the problem approached in this chapter is a generalized version of the conventional data quantization. This formulation is useful as it helps to find an exact solution to the problem under stronger conditions via ABSA-1 and under milder conditions via ABSA-2.

- Moreover, our analytical studies help us to craft an indirect [1] task-effective data quantization algorithm - ABSA-2. Designing a task-effective data quantization for ABSA-2 can equivalently be translated as an indirect approach to feature selection for an arbitrary deep Q-network. Relying on the analysis carried out for ABSA-1, ABSA-2 designs distributed and bit-budgeted communications between the agents and CC. ABSA-2 is seen to approach optimal performance by increasing the memory of the CC. In fact, increasing the memory of CC leads to higher computational complexity. Therefore,

[1] By an indirect algorithm here we mean an approach that is not dependent on our knowledge from a particular task. Indirect approaches are applicable to any/(wide range of) tasks. In contrast to indirect schemes, we have direct schemes that are specifically designed for a niche application [29]. As defined by [24]: "the direct schemes aim at guaranteeing or improving the performance of the cyber-physical system at a particular task by designing a task-tailored communication strategy".

ABSA-2 is said to strike a trade-off between computational complexity and task efficiency.

- Numerical experiments are carried out on a geometric consensus task to evaluate the performance of the proposed schemes in terms of the optimality of the MAS's expected return in the task. ABSA-1 and ABSA-2 are compared with several other benchmark schemes introduced by [9], in a multi-agent[2] scenario with local observability and bit-budgeted communications.

- Finally, we will introduce a new metric, called task relevant information, for the measurement of effectiveness in task-oriented communication policies that - in comparison with the existing metrics such as positive listening and positive signalling - better explains the behaviour of a variety of task-effective communication schemes. The proposed metric is capable of measuring the effectiveness of a task-oriented communication/compression policy without the need of testing a jointly designed control policy and testing the jointly designed policies in the desired task.

4.1.3 Technical approach

Our goal is to perform an efficient representation of the agents' local observations to ensure meeting the bit-budget of the communication links while minimizing the effect of quantization on the average return of the task. To achieve this, we first need to design task-effective data quantization policies for all agents. In task-effective data quantization, one needs to take into account the properties of the average return function and the optimal control policies associated with the task [28]. In addition to the design of the quantization policies for all agents, we also need the control policy of the CC to be capable of carrying out near-optimal decision-making despite its mere access to the quantized messages - resulting in a joint control and data compression problem. We formulate the joint control and data compression problem as a generalized form of data compression: task-oriented data compression (TODC). Following this novel problem formulation, we propose two indirect action-based state aggregation algorithms (ABSA): (i) ABSA-1 provides analytical proof for a task-effective quantization i.e, with optimal performance in terms of the expected return. In this direction, ABSA-1 relaxes

[2]Due to the complexity related issues explained in section 4.4, the numerical results are limited to two-agent and three-agent scenarios.

the assumption of the lumpability of the underlying MDP, according to which [9][condition. 6], the performance guarantees of the proposed method were established. Since ABSA-1 is only applicable when the system is composed of one agent and the CC we also propose ABSA-2. Following the analytical results of ABSA-1, given the help of MAP estimation to relax the aforementioned limitation of ABSA-1, and benefiting from a DQN controller at the CC; ABSA-2 will be introduced as a more general approach. (ii) ABSA-2 solves an approximated version of the TODC problem and carries out the quantization for any number of agents communicating with the CC. Thanks to a deep Q-network controller utilized at the CC, ABSA-2 can solve more complex problems where the controller benefits from a larger memory. Thus, ABSA-2 allows trading complexity for communication efficiency and vice versa. Finally, we will evaluate the performance of the proposed schemes in the specific task: a geometric consensus problem under finite observability [208].

4.1.4 Organization

The rest of this chapter is organized as follows. Section II describes the MAS and states the joint control and communication problem. Section III proposes two action-based state aggregation algorithms. Section IV shows the performance of the proposed algorithms in a geometric consensus problem. Finally, Section V concludes the chapter. For the reader's convenience, a summary of the notation that we follow in this chapter is given in Table 4.1. Bold font is used for matrices or scalars which are random and their realizations follow simple font.

Table 4.1: Table of notations

Symbol	Meaning
$\mathbf{x}(t)$	A generic random variable generated at time t
$\mathsf{x}(t)$	Realization of $\mathbf{x}(t)$
$\mathcal{X}$	Alphabet of $\mathbf{x}$(t)
$\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
$p_{\mathbf{x}}(\mathsf{x}(t))$	Shorthand for $\Pr(\mathbf{x}(t)=\mathsf{x}(t))$
$H(\mathbf{x}(t))$	Information entropy of $\mathbf{x}(t)$ (bits)
$\mathcal{X}_{-\mathbf{x}}$	$\mathcal{X}-\{\mathbf{x}\}$
$\mathbb{E}_{p(\mathsf{x})}\{\mathbf{x}\}$	Expectation of the random variable X over the probability distribution $p(\mathsf{x})$
$\mathsf{tr}(t)$	Realization of the system's trajectory at time t

4.2 System model and problem statement

The problem setting we introduce here can be used to analyse both scenarios illustrated in Fig. 4.2. Nevertheless, to use our language consistently, we focus on the scenario (a) of that figure throughout the manuscript. In particular, when we use the term "agent" we refer to an object which certainly has all the following hardware capabilities: sensing, actuation, communication and data processing. A MAS, however, may not be comprised of mere agents, but of a combination of agents and perhaps other objects that has at least the hardware capabilities for communication and data processing power. The central controller here is supposed to have the hardware capability to process relatively larger data as well as the capability of communications. The interactions inside the MAS and outside the MAS with the environment are illustrated in Fig. 4.3.

4.2.1 System model

We consider a MAS in which multiple agents $i \in \mathcal{N} = \{1, 2, ..., N\}$ collaboratively solve a task with the aid of a CC. Following a centralized action policy, CC provides the agents with their actions via a perfect communication channel while it receives the observations of agents through an imperfect communication channel [3]. The considered setting is similar to conventional centralized control of MASs [9,276], except for the fact that the communications from the agents to the CC are transmitted over a bit-budgeted communication channel. The agent-hub communications are considered to be instantaneous and synchronous [9]. This is in contrast with the delayed [4, 272] and sequential/iterative communication models [281–283]. We note that there is no direct inter-agent communication in the considered system - communications occur only between agents and the central controller. The system runs on discrete time steps t. The observation of each agent i at time step t is shown by $\mathbf{o}_i(t) \in \Omega$ and the state $\mathbf{s}(t) \in \mathcal{S}$ of the system is defined by the joint observations $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \ldots, \mathbf{o}_N(t) \rangle$[4] . The control action of each agent i at time t is shown by $\mathbf{m}_i(t) \in \mathcal{M}$, and the action vector $\mathbf{m}(t) \in \mathcal{M}^N$ of the system is defined by the joint actions $\mathbf{m}(t) \triangleq \langle \mathbf{m}_1(t), ..., \mathbf{m}_N(t) \rangle$. The observation space Ω, state-space $\mathcal{S}$, and action space $\mathcal{M}$ are all discrete sets. The environment is governed by an underlying[5] Markov Decision Process that is described by the

[3] In this work we follow a common assumption used in the networked control literature [74] according to which the bit-budget only limits the uplink communications of the agents and not their downlink. Accordingly, the agents select their control actions as is dictated to them by the central controller.

[4] According to this definition, at any given time t the observations of any two agent $i, j \in \mathcal{N}$ are linearly independent in the Euclidean space. The same conditions are true for the control actions of arbitrary agents.

[5] As defined in the literature [10], the underlying MDP' is the horizon-T' MDP defined by a hypothetical single agent that takes joint actions $\mathbf{m}(t) \in \mathcal{M}^N$ and observes the nominal state $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \ldots, \mathbf{o}_N(t) \rangle$ that

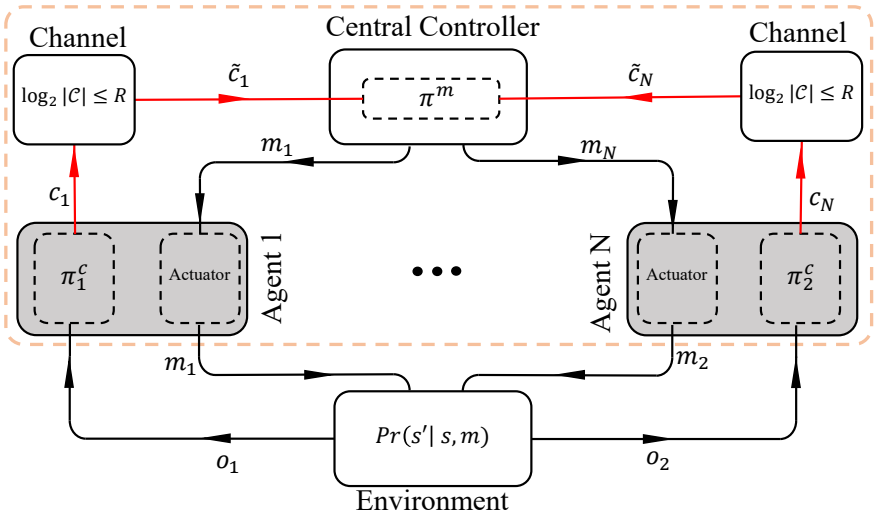


Figure 4.3: Illustration of the interactions of the CC and agents for the control of the environment. The red link shows the communication channels that are bit-budgeted - implying the local (and not global) observability of the CC. The orange dashed box is detailing the same box in Fig. 4.1 and Fig. 4.2 .

tuple $M = \{\mathcal{S}, \mathcal{M}^N, r(\cdot), \gamma, T(\cdot)\}$, where $r(\cdot) : \mathcal{S} \times \mathcal{M}^N \to \mathbb{R}$ is the per-stage reward function and the scalar $0 \le \gamma \le 1$ is the discount factor. The function $T(\cdot) : \mathcal{S} \times \mathcal{M}^N \times \mathcal{S} \to [0,1]$ is a conditional probability mass function (pmf) which represents state transitions such that $T(\mathsf{s}(t+1), \mathsf{s}(t), \mathsf{m}(t)) = \Pr(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t))$. According to the per-stage reward signals, the system's return within the time horizon T' is denoted by

$$\mathbf{g}(t') = \sum_{t=t'}^{T'} \gamma^{t-1} r(\mathbf{o}_1(t), ..., \mathbf{o}_N(t), \mathbf{m}_1(t), ..., \mathbf{m}_N(t)). \tag{4.1}$$

While the system state is jointly observable by the agents [37], each agent i's observation $\mathbf{o}_i(t)$ is local [6]. Once per time step, agent $i \in \mathcal{N}$ is allowed to transmit its local observations through a communication message $\mathbf{c}_i(t)$ to the CC. The communications between agents and the central controller are done in a synchronous (not sequential) and simultaneous (not delayed) fashion [4]. Each agent i generates its communication message $\mathbf{c}_i(t)$ by following its communication policy $\pi_i^c(\cdot) : \Omega \to \mathcal{C}$. In parallel to all other agents, agent i follows the communication policy $\pi_i^c(\cdot)$ to map its current observation $\mathbf{o}_i(t)$ to the communication message

has the same transition model $T(\cdot)$ and reward model $r(\cdot)$ as the environment experienced by our MAS.

[6] In our problem setting, each agent does not see the environment as an MDP due to their local observability. We only assume the presence of an underlying MDP for the environment, which is widely adopted in the literature for the reinforcement learning algorithm, e.g., [266] [257]. We have this assumption as our performance guarantees rely on the optimality of the solution provided for the control task, which is also assumed in [7], [10]. Let us recall that throughout all of our numerical studies, even the CC, given joint observations of all agents, cannot observe the true/nominal state of the environment.

$\mathbf{c}_i(t)$ which will be received by the central controller in the same time-step t. The code-book $\mathcal{C}$ is a set composed of a finite number of communication code-words s $\mathsf{c}, \mathsf{c}', \mathsf{c}'', ..., \mathsf{c}^{(|\mathcal{C}|-1)}$ - we use the same notation to refer to the different members of the action, observation and state spaces too. Agents' communication messages are sent over an error-free finite-rate bit pipe, with its rate constraint to be $R \in \mathbb{R}$ (bits per channel use) or equivalently (bits per time step). As a result, the size of the quantization codebook should follow the inequality $|\mathcal{C}| \leq 2^R$. The CC exploits the received communication messages $\mathbf{c}(t) \triangleq \langle \mathbf{c}_1(t), ..., \mathbf{c}_N(t) \rangle$ within the last d number of time-steps to generate the action signal $\mathsf{m}(t)$ following the control policy $\pi^m(\cdot) : \mathcal{C}^{Nd} \to \mathcal{M}^N$. Based on the above description, the environment from the point of view of the CC as well as from the agent's point of view is not necessarily an MDP - as none is capable of viewing the nominal state of the environment.

4.2.2 Problem statement: Joint Control and Communication Design (JCCD) problem

Now we define the JCCD problem. Let M be the MDP governing the environment and the scalar $R \in \mathbb{R}$ to be the bit-budget of the uplink of all agents. At any time step t', we aim at selecting the tuple $\pi = \langle \pi^m(\cdot), \pi^c \rangle$ with $\pi^c \triangleq \langle \pi^c_1(\cdot), ..., \pi^c_N(\cdot) \rangle$ to solve the following variational dynamic programming

$$\underset{\pi}{\text{argmax}} \;\; \mathbb{E}_\pi \Big\{ \mathbf{g}(t') \Big\}; \;\; \text{s.t.} \;\; |\mathcal{C}| \leq 2^R, \tag{4.2}$$

where the expectation is taken over the joint pmf of the system's trajectory $\{\mathsf{tr}\}_{t'}^{T'} = \mathsf{o}_1(t'), ..., \mathsf{o}_N(t'), m(t'), ..., \mathsf{o}_1(T'), ..., \mathsf{o}_N(T'), \mathsf{m}(T')$, when the agents follow the policy tuple π. In the next section, similar to [9] we will disentangle the design of action and communication policies via action-based quantization of observations. In contrast to [9], here the communication network of the MAS is assumed to follow a star topology. The idea behind this disentanglement is to extract the features of the control design problem that can affect the communication design and to take them into account while designing the communications. Thus our communication design will be aware of the key features of the control task. We extract the key features of the control task using analytical techniques as well as reinforcement learning [4, 9]. In fact, the new communication problem called TODC, will no longer be similar to the conventional communication problems, as it is inspired by the JCCD

problem.

In [9, 35], authors use the value of agents' observations for the given task as the key feature of the control task considered in the communication design. Accordingly, the idea was to cluster together the observation points that have similar values. In contrast to [9, 35], which considers the value of observations as an explicit key feature of the control task, here we consider the optimal control/action values assigned to each observation as the key feature. Accordingly, ABSA clusters the observation values together, whenever the observation points have similar optimal control/action values assigned to them. Action-based state aggregation has been already introduced in the literature of reinforcement learning as a means for reducing the complexity of the reinforcement learning algorithms while maintaining the average return performance [284, 285].

4.3 Action-based Lossless compression of observations

In this section, we will set yet another example - in addition to [9] - for the use of a generic framework to solve JCCD problem. In [9], a similar problem is solved for distributed control and quantization, wherein, the authors disentangle the design of task-oriented communication policies and action policies given the aid of a hypothetical functional Π^{m^*}. In particular, the functional Π^{m^*} is a map from the vector space $\mathcal{K}^c$ of all possible communication policies π^c to the vector space $\mathcal{K}^m$ of optimal corresponding control policy $\pi^{m^*}(\cdot)$. Upon the availability of the functional Π^{m^*}, wherever the function π^m appears in the JCCD problem, it can be replaced with $\Pi^{m^*}(\pi^c)$ resulting in a novel problem in which only the communication policies π^c are to be designed. While in [9], authors use an approximation of $\Pi^{m^*}(\pi^c)$ to obtain a task-oriented quantizer design problem, in the current work we derive an exact solution for a simplified version of (4.3) - where the number of agents communicating with the central controller is limited to one agent. To adapt ABSA to the generic setting of the problem (4.3), in ABSA-2, we will lift this limitation given the aid of an approximation technique.

The JCCD problem can already be formulated as a form of data-quantization problem. Lemma 13, identifies the quantization metric that we aim to optimize in this chapter. It reformulates the JCCD problem as a novel generalized data quantization problem.

Lemma 13. *The JCCD problem (4.2) can also be expressed as a generalized data quantization*

problem as follows

$$\underset{\pi}{argmin} \quad \mathbb{E}_{p(\mathbf{s}(t))}\Big|V^{\pi^*}(\mathbf{s}(t)) - V^{\pi^m}(\mathbf{c}(t))\Big|, \quad s.t. \quad |\mathcal{C}| \leq 2^R, \tag{4.3}$$

where the communication vector $\mathbf{c}(t)$ *generated by* π^c *is a quantized version of the system's state* $\mathbf{s}(t)$.

Proof. Appendix 4.6. ■

In contrast to the classic data-quantization problems, here the distortion metric, measures the difference between two different functions of the original signal and its quantized version - namely $V^{\pi^*(\cdot)}$ and $V^{\pi^m(\cdot)}$ - thus the distortion measure that we aim to optimize by solving (4.3) is not conventional. In fact, the variational minimization problem is solved over the vector space of joint quantization policies π^c and action policy π^m functions.

4.3.1 ABSA-1 Algorithm

The applicability of the proposed ABSA-1, is limited to two mathematically equivalent scenarios: (i) we have a single agent communicating to the CC - consider the Fig. 4.2-a, with only one agent connected to the CC - or (ii) that the agents communicate with the CC through a relay. In the latter scenario, the relay has full access to the agents' communication observation, i.e., $\mathbf{o}_i, \forall i \in \mathcal{N}$, while the relay to CC channel is bit-budgeted. This limited scenario is useful for us to facilitate our analytical studies on the problem (4.3), allowing us to establish theoretical proof for the losslessness of compression in ABSA-1 as well as its optimal average return performance. These statements will be confirmed by Lemma 15 - the results of which will also be useful to design ABSA-2. The central idea of ABSA-1 is to represent any two states $\mathbf{s}^{(i)}, \mathbf{s}^{(j)}$ using the same communication message $\mathbf{c}$ iff $\pi^*(\mathbf{s}^{(i)}) = \pi^*(\mathbf{s}^{(j)})$, where $\pi^*(\cdot) : \mathcal{S} \rightarrow \mathcal{M}^N$ is the optimal control policy of the agents, given the access of observations from all agents. Thus, ABSA-1 and ABSA-2 solve the JCCD problem at three different phases: (i) solving the centralized control problem under perfect communications via reinforcement learning i.e., Q-learning, to find $\pi^*(\cdot)$[7], (ii) solving the task-oriented data quantization problem to find π^c via a form of data clustering, (iii) finding the π^m corresponding to π^c.

[7] ABSA's bottleneck arises from the increasing complexity of Q-learning as agents increase in number N. Similar limitations are in place for any other algorithm that requires a centralized training phase [3, 276]

In order to explain ABSA-1, we introduce the problem of task-oriented data compression with centralized control. TBIC is derived using similar techniques in [9] but for a different setting i.e., the communication network of MAS has a star topology. The TBIC problem is no longer a joint control and communication problem but is a quantization design problem in which the features of the control problem are taken into account. To arrive to TODC problem from the JCCD problem, we use the functional Π^{m^*} to replace $\pi^m(\cdot)$ with $\Pi^{m^*}(\pi^c)$. Upon the availability of Π^{m^*}, by plugging it into the JCCD problem (4.2), we will have a new problem

$$\underset{\pi^c}{\operatorname{argmin}} \qquad \mathbb{E}_{p(\mathbf{s}(t))}\Big|V^{\pi^*}(\mathbf{s}(t)) - V^{\Pi^{m^*}(\pi^c)}(\mathbf{c}(t))\Big|, \quad \text{s.t.} \quad |\mathcal{C}| \leq 2^R, \tag{4.4}$$

where we maximize the system's return with respect to only the communication policies $\pi^c(\cdot)$ of the local relay. The optimal control policy $\pi^{m^*}(\cdot)$ of the CC is automatically computed by the mapping $\Pi^{m^*}(\pi^c(\cdot))$. The problem is called here as the TODC problem. Upon the availability of Π^{m^*}, the JCCD problem (4.2) can be reduced to (4.4). Definition 14 is provided to formalize a precise approach to solve (4.4) via obtaining the communication policy of the relay $\pi^c(\cdot)$ as well as the corresponding Π^{m^*}, to solve (4.2).

Definition 14. ***Quantization and control policies in ABSA-1:***

The communication policy $\pi^{c,ABSA-1}(\cdot)$ designed by ABSA-1 will be obtained by solving the following k-median clustering problem

$$\underset{P}{min} \qquad \sum\nolimits_{i=1}^{|\mathcal{C}|} \sum\nolimits_{\mathbf{s}(t)\in\mathcal{P}_i} \Big|\pi^*(\mathbf{s}(t)) - \mu_i\Big|, \tag{4.5}$$

where $\mathcal{P} = \{\mathcal{P}_1, ..., \mathcal{P}_B\}$ is a partition of $\mathcal{S}$ and μ_i is the centroid of each cluster i. The communication policy of ABSA-1 - $\pi^{c,ABSA-1}(\cdot)$ - is an arbitrary non-injective mapping such that $\forall k \in \{1, ..., B\} : \pi^{c,ABSA-1}(\mathbf{s}) = \mathbf{c}^{(k)}$ if and only if $\mathbf{s} \in \mathcal{P}_k$. Now let C_g be a function composition operator such that $C_g f = g \circ f$. We define the operator $\Pi^{m^*} \triangleq C_g$, with $g = \pi^*(\pi^{c,ABSA-1^{-1}}(\cdot))$[8] .

The optimality of the proposed ABSA-1 algorithm is subsequently provided in Theorem 15.

[8]Note that as $\pi^{c,ABSA-1}(\cdot)$ is non-injective, its inverse would not produce a unique output given any input. Thus, by $\pi^*(\pi^{c,ABSA-1^{-1}}(\mathbf{c}'))$ we mean $\pi^*(\mathbf{s}')$, where $\mathbf{s}'$ can be any arbitrary output of $\pi^{c,ABSA-1^{-1}}(\mathbf{c}')$.

Lemma 15. *The communication policy $\pi^{c,ABSA-1}$ - as described by Definition 14 - will carry out lossless compression of observation data w.r.t. the average return if $|\mathcal{C}| \geq |\mathcal{M}|^N$.*

Proof. Appendix 4.7. ■

Remark: ABSA-1 will also carry out lossless compression of observation data with respect to the distortion measure introduced in problem (4.3). Given the proofs of lemma 2 and lemma 1, the proof of this remark is straightforward and is therefore, omitted.

The losslessness of quantization in ABSA-1 implies that the π^{ABSA-1} will result in no loss of the system's average return, compared with the case where the optimal policy $\pi^*(\cdot)$ is used to control the MAS under perfect communications. Consequently, the control policy $\pi^{m,ABSA-1}(\cdot)$ is optimal. Let us recall once again that here, we do not use a conventional quantization distortion metric, we select a representation of local observation in such a way that the conveyed message maximizes the average task return.

Note that in [7], the authors do not find the higher order function Π^{m^*} that reduces the joint communications and control problem to a task-oriented communication design - instead they solve an approximated version of the task-oriented communication design problem. In this chapter, however, we introduce a closed form Π^{m^*} by ABSA-1 that can map every communication policy $\pi^{c,ABSA-1}$ introduced by ABSA-1, to the exact optimal control policy. This implies that the solutions provided by ABSA-1 are also the optimal solutions of the joint communication and control design (JCCD) problem.

4.3.2 ABSA-2 Algorithm

We saw earlier in lemma 15 that the communication policy obtained by solving the problem 4.5 is optimal and can result in a lossless average return performance when $|\mathcal{C}| \geq |\mathcal{M}|^N$. To solve the problem 4.5, however, we need to know $\pi^*(\mathbf{s}(t))$. This is a limiting assumption that in ABSA-1 can be translated to two different system models which are less general than the system pictured in Fig. 4.3: (i) presence of an extra relay between the agents and the central controller where the relay has perfect downlink channels to agents and a single bit-budgeted channel to the CC. (ii) The MAS is only composed of one single agent and a CC where the uplink of the agent is bit-budgeted but its downlink is a perfect channel.

Our second proposed algorithm ABSA-2 removes the need to know $\pi^*(\mathbf{s}(t))$ and can run

under the more general settings shown in Fig. 4.3. This is done by approximating the local element $\mathbf{m}_i^*(t)$ of $\pi^*(\mathbf{s}(t)) = \langle \mathbf{m}_1 * (t), ..., \mathbf{m}_N * (t) \rangle$ at agent agent i given the local observation of this agent $\mathbf{o}_i(t)$. That is, given a centralized training phase, we will have access to the empirical joint distribution of $p(\mathsf{o}_i, \mathsf{m}_i^*)$, using which we can obtain a numerical MAP estimator of $\hat{\mathsf{m}^*}_i$. Thus ABSA-2 allows for fully distributed communication policies. In particular, the encoding of the communication messages of each agent is carried out separately by them before they communicate with CC or any other agent. This form of encoding is often referred to as distributed encoding. Furthermore, the encoding carried out by ABSA-2 at each agent is a low-complexity and low-power process that requires no inter-agent communications before hands. In this case, each agent directly communicates its encoded observations to the CC via a bit-budgeted communication channel. In order to improve the learning efficiency at CC, it can take into account all the communications received in the time frame $[t-d, t]$ to make a control decision $m(t)$. Therefore, the ABSA-2 algorithm can strike a trade-off between the complexity of the computations carried out at the CC - directly impacted by the value of d - and effectiveness of agents' communications - inversely impacted by the value of $|\mathcal{C}|$. Moreover, ABSA-2 is straightforwardly extendable to the different values of $|\mathcal{C}|$ per each agent i, instead of having only one fixed bit-budget $R = \log_2|\mathcal{C}|$ for all agents.

As illustrated in Fig. 4.4, ABSA-2, each agent i obtains a communication policy function $\pi_i^c(\cdot)$ by solving a clustering problem over its local observation space instead of the global state space, formulated as follows:

$$\min_{P_i} \quad \sum_{j=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_i(t) \in \mathcal{P}_{i,j}} \left| \tilde{\pi}_i^*(\mathsf{o}_i(t)) - \mu_{i,j} \right|, \tag{4.6}$$

where $\mathcal{P}_i = \{\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,|\mathcal{C}|}\}$ is a partition of Ω, and

$$\tilde{\pi}_i^*(\mathsf{o}_i(t)) = \text{argmax}_{\mathbf{m}_i^*} \; p_{\pi^*}(\mathbf{m}_i^* | \mathsf{o}_i(t)), \tag{4.7}$$

and $\mathbf{m}_i^*$ is the optimal action of agent i, which is i-th element of $\mathbf{m}^* \triangleq \pi^*(\mathbf{o}_1(t), ..., \mathbf{o}_N(t))$. Thus $\tilde{\pi}_i^*(\mathsf{o}_i(t))$ is the maximum aposteriori estimator of $\mathbf{m}_i^* = \pi^*(\mathbf{s}(t))$ given the local observation $\mathsf{o}_i(t)$.

Once the clustering in (4.6) is done, each agent i will train its local communication policy $\pi_i^{c,ABSA-2}(\cdot)$, which is any non-injective mapping such that $\forall k \in \{1, ..., |\mathcal{C}|\} : \pi_i^{c,ABSA-2}(\mathbf{o}_i) =$

Algorithm 2 Action Based State Aggregation (ABSA-2)

Initialize replay memory D to capacity 10'000.
Initialize state-action value function $Q(\cdot)$ with random weights θ.
Initialize target state-action value function $Q^t(\cdot)$ with weights $\theta^t = \theta$.
Obtain $\pi^*(\cdot)$ and $Q^*(\cdot)$ by solving (2) using Q-learning [267]*, where $R >> H(\mathbf{o}_i(t))\ \ \forall i \in \mathcal{N}$.
Compute $\pi_i^*(\mathsf{o}_i(t)) = \text{Mode}[\mathbf{m}_i^*|\mathsf{o}_i(t)]$, for $\forall \mathsf{o}_i(t) \in \Omega$, for $i \in \mathcal{N}$.
Solve problem (5) by applying k-median clustering to obtain $\mathcal{P}_i$ and $\pi_i^c(\cdot)$, for $i \in \mathcal{N}$.
for *each episode* $k = 1 :$ *200'000* **do**
    Randomly initialize observation $\mathsf{o}_i(t=0)$, for $i \in \mathcal{N}$
Randomly initialize the message $\mathsf{c}(t=0)$ **for** $t = 1 : T'$ **do**
Select $\mathsf{c}_i(t)$, at agent i, following $\pi_i^c(\cdot)$, for $i \in \mathcal{N}$
Obtain the message $\langle \mathsf{c}_1(t), ..., \mathsf{c}_N(t) \rangle$ at the CC
Follow ϵ-greedy, at CC, to generate the action $\mathsf{m}_i(t)$, for $i \in \mathcal{N}$
Obtain reward $r(t) = R(\mathsf{s}(t), \mathsf{m}(t))$ at the CC
Store the transition $\Big\{\mathsf{c}(t), \mathsf{m}(t), r(t), \mathsf{c}(t+1)\Big\}$ in D
$t \leftarrow t + 1$
end
Sample $D' = \Big\{\mathsf{c}(t'), \mathsf{m}(t'), r(t'), \mathsf{c}(t'+1)\Big\}_{t'=t'_1}^{t'=t'_{62}}$ from D
    for *each transition* $t' = t'_1 : t'_{62}$ *of the mini-batch* D' **do**
        Compute DQN's average loss $L_{t'}(\theta) = \frac{1}{2}\Big(r(t') + \max_{m^*} Q^t(\mathsf{c}(t'+1), \mathsf{m}^*, \theta^t) - \max_{m^*} Q(\mathsf{c}(t'), \mathsf{m}^*, \theta)\Big)^2$,
Perform a gradient descent step on $L_{t'}(\theta)$ w.r.t θ
end
Update the target network $Q^t(\cdot)$ every 1000 steps
end

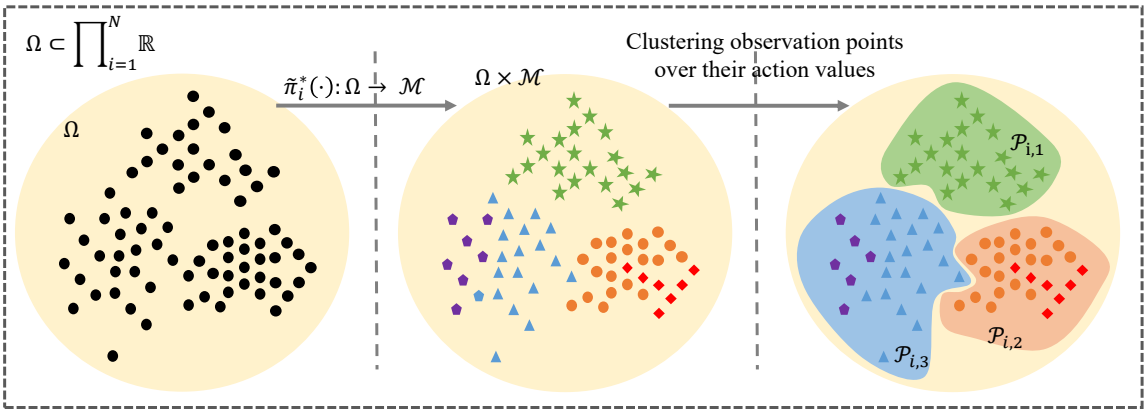


Figure 4.4: Abstract representation of states in ABSA-2 with $|\mathcal{C}|= 3$ and $|\mathcal{M}|= 5$ - $|\mathcal{M}|$ is represented by the number of shapes selected to show the observation points and $|\mathcal{C}|$ is represented by the number of clusters shown in the right subplot. The left subplot shows the observation points prior to aggregation. During a centralized training phase we first compute $\pi^*(\cdot)$ according to which $\pi_i^*(\cdot) : \Omega \rightarrow \mathcal{M}$ can be obtained. We use the surjection $\pi_i^*(\cdot)$ to map a high dimensional/precision observation space to a low dimensional/precision space. The middle subplot shows the observation points together with the action values assigned to them - each unique shape represents a unique action value. **This new representation of the observation points, embeds the features of the control problem into the data quantization problem.** Finally, we carry out the clustering of observation points according to their action values - all observation points assigned to (a set of) action values are clustered together. The right subplot shows the aggregated observation space, where all the observation points in each cluster will be represented using the same communication message. The centralized controller which is run using DQN, observes the environment at each time step, through all these aggregated observations/communications it receives from all the agents.

$\mathbf{c}^{(k)}$ iff $\mathbf{o}_i \in \mathcal{P}_{i,k}$. After obtaining the communication policies $\langle \pi_i^{c,ABSA-2}(\cdot) \rangle_{i=1}^{N}$, to obtain a proper control $\pi^m(\cdot)$ policy at the CC corresponding to the communication policies, we perform a single-agent reinforcement learning. To this end and to manage the complexity of the algorithm for larger values of d, we propose to use DQN architecture [105] at the CC.

4.4 Performance Evaluation

In this section, we evaluate our proposed schemes via numerical results for the popular multi-agent geometric consensus problem[9]. Through indirect design, ABSA-1 and ABSA-2 never rely on explicit domain knowledge about any specific task, such as geometric consensus. Thus, we conjecture that their indirect design allows them to be applied beyond geometric

[9]In our numerical experiments, the discount factor is assumed to be $\gamma = 0.9$. All experiments are done over a grid world of size 8×8, where the goal point of the rendezvous is located at the grid number $\Omega^T = \{22\}$.

consensus problems and to a much wider range of tasks. To make the geometric consensus task suitable for the evaluation of our proposed algorithms, similar to [9], we have introduced a bit constraint to the communication channel between the agents and the CC. After evaluating the proposed algorithms in the context of the rendezvous problem, we attempt to explain the behaviour of all the algorithms via the existing metric - positive listening - for measuring the task-effectiveness of communications. As positive listening falls short in explaining all the aspects of the behaviour of the investigated algorithms, we will also introduce a new metric. Called *task relative information*, the new metric assists to further explain the behaviour of different algorithms with a higher accuracy and reliability.

4.4.1 The geometric consensus problem

Our proposed schemes are evaluated in this section through numerical results for the rendezvous problem [265, 277], which is a specific type of geometric consensus problems under finite observability [208]. Following the instantaneous and synchronous communication model and the star network topology explained in section 4.2.1 and Fig. 4.2 respectively, the rendezvous problem is explained as the following. At each time step t several events happen in the following order. First, an agent i obtains a local observation $\mathsf{o}_i(t)$ - which is equivalent to its own location in the grid-world. The agent i, subsequently, follows its quantization/communication policy to generate a compressed version $\mathsf{c}_i(t)$ of its observation to be communicated to the CC via bit-budgeted communication links. After receiving the quantized observations of all agents, CC follows its control policy to decide and select the joint action vector $\mathsf{m}(t)$ and communicate each agent i's local action $\mathsf{m}_i(t)$ to it accordingly. The local action $\mathsf{m}_i(t) \in \mathcal{M}$ that is communicated back to the agent i via a perfect communication channel is a one directional move in the greed world, i.e, $\mathcal{M} = \{$ left, right, up, down, pause$\}$. Given each agent i's action $\mathsf{m}_i(t)$ the environment evolves and transitions to the next time step $t+1$ where each agent i obtains a new local observation $\mathsf{o}_i(t+1)$. All agents receive a single team reward

$$\mathrm{r}_t = \begin{cases} C_1, & \text{if } \exists\, i, j \in N : \mathsf{o}_i(t) \in \Omega^T \ \ \& \ \ \mathsf{o}_j(t) \notin \Omega^T \\ C_2, & \text{if } \nexists\, i \in N : \mathsf{o}_i(t) \in \Omega - \Omega^T, \\ 0, & \text{otherwise,} \end{cases} \tag{4.8}$$

where $C_1 < C_2$ and Ω^T is the set of terminal observations i.e., the episode terminates if $\exists\, i \in \mathcal{N} : \mathsf{o}_i(t) \in \Omega^T$. Accordingly, when not all agents arrive at the target point, a smaller reward $C_1 = 1$ is obtained, while the larger reward $C_2 = 10$ is attained when all agents visit the goal point at the same time. We compare our proposed ABSA algorithms with the heuristic non-communicative (HNC), heuristic optimal communication (HOC) and SAIC algorithms proposed in [9] which are direct schemes to jointly design the communication and control policies for the specific geometric consensus problem solved here. In contrast to ABSA-1 and ABSA-2 which enjoy an indirect design, the direct design of HOC and HNC does not allow them to be applied in any other problem rather than the specific geometric consensus problem with the finite observability i.e., the rendezvous problem explained here.

4.4.2 Numerical experiment

A constant learning rate $\alpha = 0.07$ is applied when exact Q-learning is used to obtain $\pi^*(\cdot)$ and $\alpha = 0.0007$ when DQN is used to learn $\pi^m(\cdot)$ for ABSA-2. For the exact Q-learning, a UCB[10] exploration rate of $c = 1.25$ considered. The deep neural network that approximates the Q-values is considered to be a fully connected feed-forward network with 10 layers of depth, which is optimized using the Adam optimizer. An experience reply buffer of size 10'000 is used with the mini-batch size of 62. The target Q-network is updated every 1000 steps and for the exploration, decaying ϵ-greedy with the initial $\epsilon = 0.05$ and final $\epsilon = 0.005$ is used [105]. In any figure that the performance of each scheme is reported in terms of the averaged discounted cumulative rewards, the attained rewards throughout training iterations are smoothed using a moving average filter of memory equal to 20,000 iterations. As explained in section 4.3.1, ABSA-1 and ABSA-2 both require a centralized training phase prior to be capable of being executed in a distributed fashion.

For all black curves, one prior centralized training phase to obtain $\pi^*(\cdot)$ is required. As detailed in Section III, the proposed algorithms, ABSA-1 and ABSA-2, leverage $\pi^*(\cdot)$ to design π^c and then π^m afterwards. Dashed curves, HOC and HNC, as proposed by [9] provide heuristic schemes which exploit the domain knowledge of its designer about the rendezvous task making it not applicable to any other task rather than the rendezvous problem. While HOC enjoys a joint control and communication design, HNC runs with no communication.

[10] UCB is a standard scheme used in exact reinforcement learning to strike a trade-off between the exploration and exploitation [267].

Note that HNC & HOC require communication/coordination between agents prior to the starting point of the task - which is not required for any other scheme. These schemes, introduced by [9], are detailed as the following.

- A joint communication and control policy is designed **using domain knowledge** in the rendezvous problem. HNC agents approach the goal point and wait nearby for a sufficient number of time steps to ensure that the other agent has also arrived. Only after that, they will get to the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task, since they have to have had agreed upon this scheme of coordination.
- A joint communication and control policy is designed **using domain knowledge** in the rendezvous problem. HOC agents wait next to the goal point until the other agent informs them that they have also arrived there. Only after that, they will get to the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task, since they have to have had agreed upon this scheme of coordination and communications as well as on the the meaning that each communication message entails.

To obtain the results demonstrated in Fig. 4.5, we have simulated the rendezvous problem for a three-agent system. The black curves illustrate the training phase that is occurring at CC to obtain π^m after π^c is already computed using equations (4.5) and (4.6). We observe the lossless performance of ABSA-1 in achieving the optimal average return without requiring any (2nd round) training. To enable fully decentralized quantization of the observation process, ABSA-2 was proposed which is seen to approach the optimal solution as d grows. All ABSA-2 curves are plotted with $|\mathcal{C}|= 3$, and ABSA-1 curve is plotted with $|\mathcal{C}|= |\mathcal{M}|^N= 125$ in 3 agent scenarios - Fig. 4.5 - and $|\mathcal{C}|= |\mathcal{M}|^N= 25$ in the two agent scenario - Fig. 4.6.

In Fig. 4.5, we see how the performance of ABSA-2 compares with HNC, HOC and SAIC at different rates of quantization. As expected, with the increase in the size of the quantization codebook, the average return performance of ABSA-2 is gradually improved, such that it approaches near-optimal performance at $d = 3$. We also observe the superior performance of ABSA-2 compared with SAIC at very tight bit-budgets where SAIC's performance sees a drastic drop. It is observed that as d grows, ABSA-2 approaches the optimal return

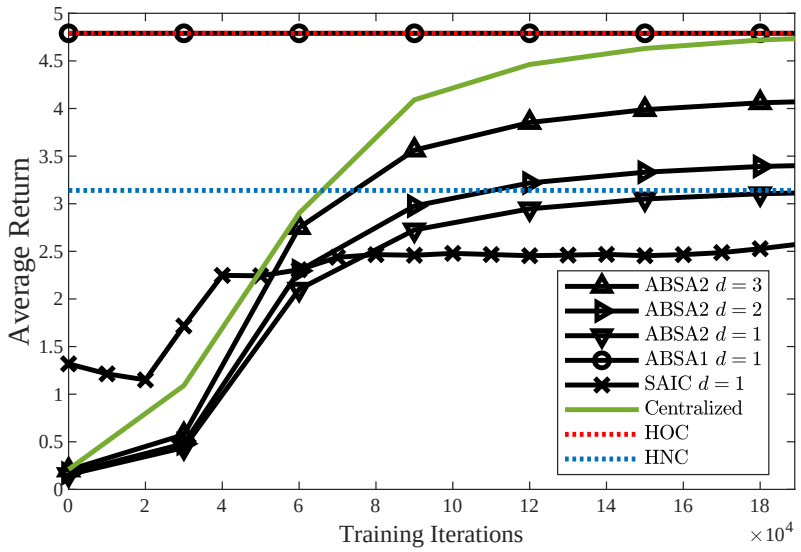


Figure 4.5: Average return comparison made between the proposed schemes and some benchmarks introduced in [9] - the three agent scenario under constant bit-budget values.

performance even under higher rates of quantization, however, higher values of d come at the cost of the increased computational complexity of ABSA-2.

4.4.3 Explainablity of the learned communication policies

One common metric to evaluate the effectiveness of communications in the literature [257] is *positive listening* $I(\mathbf{c}_i(t); \mathbf{m}_j(t))$ $j \in \mathcal{N} - \{i\}$, which is the mutual information between the communication $\mathbf{c}_i(t)$ produced by an agent i and the action $\mathbf{m}_j(t)$ selected by another agent following the receipt of the communication $\mathbf{c}_i(t)$ from agent i. Positive signaling $I(\mathbf{o}_i(t); \mathbf{c}_i(t))$ is another metric proposed by [257], measuring the mutual information between agent i's observation $\mathbf{o}_i(t)$ and its own produced communication message $\mathbf{c}_i(t)$ at the same time step. As to be shown below, however, these metrics are unable to fully capture the underlying performance trends of all schemes. Therefore, we, for the first time, introduce a new metric called *task relevant information* (RI) - allowing us to explain the task-effectiveness of the learned communication policies.

Measuring positive listening is one way to quantify the contribution of the communicated messages of agent i to the action selection of agent j. Positive signalling, on the other hand, measures the consistency as well as the relevance of the communicated messages $\mathbf{c}_i(t)$ and the agent's observations $\mathbf{o}_i(t)$. As SAIC and ABSA use a deterministic mapping of observation

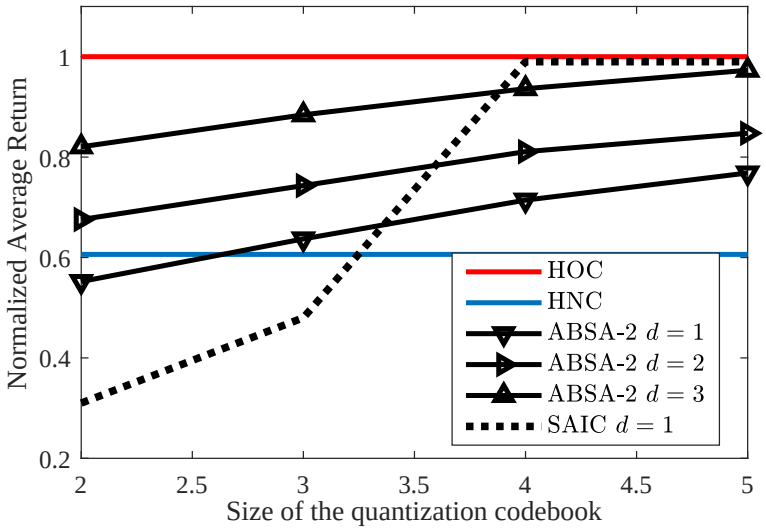


Figure 4.6: The obtained normalized average return as a function of codebook size $|\mathcal{C}|$ is compared across a range of schemes: proposed schemes and some benchmarks introduced in [9] - two-agent scenario.

o_i to produce the communication message c_i, they are always guaranteed to have positive signalling [257] - the degree of which, however, is limited by the uplink channel's bit budget $R = \log_2|\mathcal{C}|$. Thus, among the existing metrics for the measurement of the effectiveness of communications, we limit our numerical studies to the measurement of positive listening. It is known that the higher positive listening is, the stronger (not necessarily better) we expect the coordination between the agents to be. That is, the higher positive listening means higher degree of dependence between agents (their actions and observations) which is not necessarily sufficient for the team agents to fulfill the task.

Figure 4.7 explains how stronger coordination between agents and the CC is often resulting in an increased performance of the MAS in obtaining a higher average return. For instance, the enhancement in the positive-listening performance of SAIC from $|\mathcal{C}|= 3$ to $|\mathcal{C}|= 4$ quantizer in Fig. 4.7 is resulting in an improved average return performance, as shown in Fig. 4.6. This metric also reasonably explains the enhancement of ABSA-2 performance in obtaining higher return by increasing d - the memory of the CC - and the size of the quantization codebook $|\mathcal{C}|$. Moreover, stronger coordination between agents and CC is visible in ABSA-2 when compared with HOC. Thus, we expect better average return performance for ABSA-2 which is in contrast to the results of Fig. 4.5. This event suggests that stronger coordination - measured by positive listening - may not necessarily result in an improved average return performance as the coordination may not be perfectly aligned with task needs.

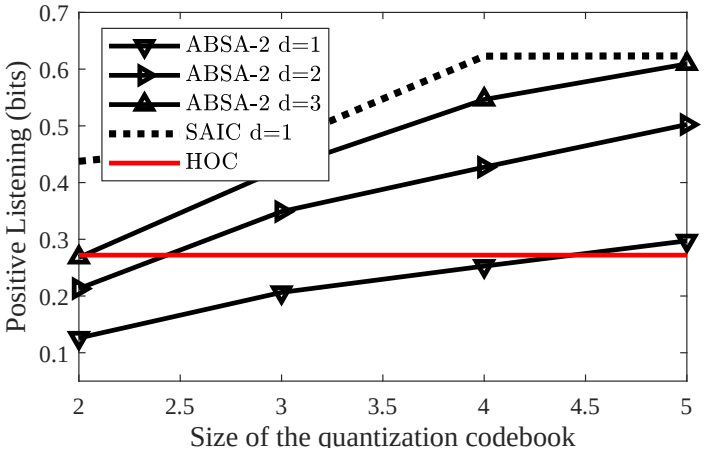


Figure 4.7: Comparing the positive listening $I(\mathbf{c}_i(t); \mathbf{m}_j(t))$ performance across a range of schemes.

The curve concerning the HOC scheme allows us to recall that a positive listening of 0.3 (bit) is sufficient to maintain the coordination required for optimal performance in the aforementioned geometric consensus task. Therefore, in the ABSA-2 and SAIC schemes, there is still an unnecessary influence from the side of the communication messages to the actions selected by the receiving end. In fact, not all the information received from the receiving end has contributed to the higher average return of the system. Accordingly, there is yet, some unnecessary data in the communication messages designed by ABSA that contain no task-specific/useful information.

Thus we believe that positive listening cannot explicitly quantify the effectiveness of the task-oriented communication algorithms; therefore they fall short in explaining the behaviour of these algorithms. Even when positive listening is computed as $I(\mathbf{c}_i(t); \mathbf{m}(t))$ to capture the mutual information between the communication of agent i and the control signals of all agents we arrive at almost similar patterns - Fig. 4.8.

Figure 4.9, investigates the performance of multiple schemes via a novel performance metric: task relevant information (TRI). Here we define the task relevant information metric to be

$$I\Big(\pi^c(\mathbf{o}_i(t)); \pi^*(\mathbf{s}(t))\Big) = I(\mathbf{c}_i(t); \mathbf{m}^*(t)), \tag{4.9}$$

which measures the mutual information (in bits) between the communicated message of agent i and the vector $\mathbf{m}^*(t)$ of joint optimal actions at the CC - which is selected by the optimal

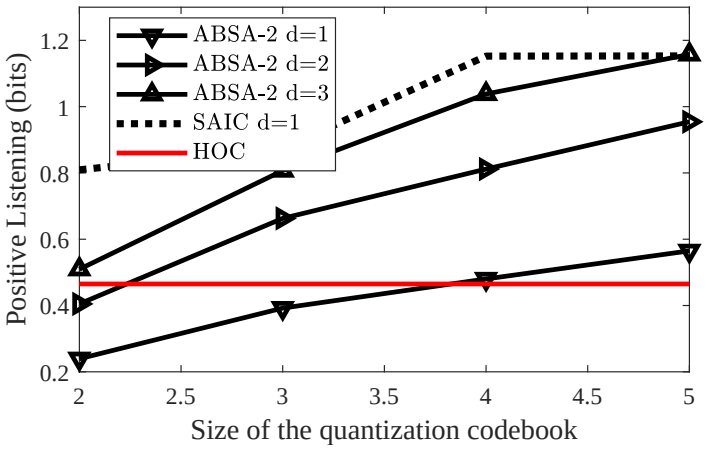


Figure 4.8: Comparing the positive listening $I(\mathbf{c}_i(t); \mathbf{m}(t))$ performance across a range of schemes.

centralized control policy $\pi^*(\cdot)$. As demonstrated by Fig. 4.9, TRI is an indirect metric of the effectiveness of communications that can explain the behaviour of different communication designs. It is also observed that the TRI metric magnifies the performance gap between different schemes as they get closer to the optimal performance. Nevertheless, TRI can be utilized as a standalone measure to quantify the effectiveness of a communication design since it almost perfectly predicts the average return performance of the a communication policy - without the need for the communication to be tested when solving the real task.

Note that, we measure the task-effectiveness of a quantization algorithm based on the average return that can be obtained when using it. Further, to measure the average return that can be obtained under the communication policies $\langle \pi_1^c(\cdot), ..., \pi_N^c(\cdot) \rangle$, we have to design the control policy $\pi^m(\cdot)$ at the CC that selects the control vector $\mathbf{m}(t)$ having access to only the quantized observations of the agents $\mathbf{c}(t)$. Accordingly, we cannot measure the effectiveness of the communication policy of an MAS without having a specific design for their control policy. Even after the design of the control policy of the MAS, it is challenging to understand if the suboptimal performance of the algorithm is caused by an ineffective design of the control policy or the communication policy. In fact, it is hard disentangle the effect of the control and communication policies on the MAS's average return. Our proposed metric TRI can facilitate measuring the performance of any communication policy in isolation and without the effect of the control policy being present in the numerical values of TRI.

Accordingly, the importance of introducing this metric is multi-fold: (i) by using TRI as an indirect metric we can measure the effectiveness of a communication policy for any specific

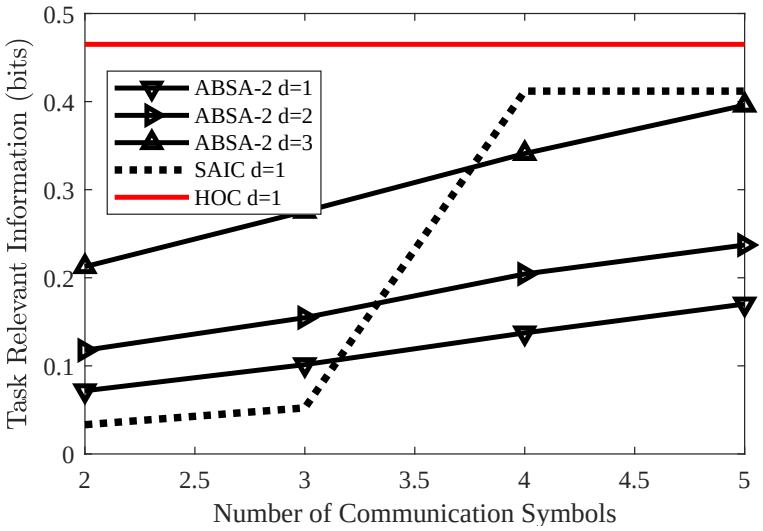


Figure 4.9: Comparing the task relevant information (TRI) performance across a range of schemes. It is observed that TRI can comprehensively explain the behaviour of all task-effective quantization schemes in a certain task without the need to measure their effectiveness via their resulting average return in the task - compare this figure with Fig. 4.6 .

task; (ii) it allows us to measure the effectiveness of the communication scheme prior to the design of any control policy; (iii) it helps to design task effective communication policies in complete separation from the control policy design.

4.5 Conclusion

In this chapter, we have investigated the joint design of control and communications in an MAS under centralized control and distributed communication policies. We first proposed an action-based state aggregation algorithm (ABSA-1) for lossless compression and provided analytical proof of its optimality. Then we proposed ABSA-2, which offers a fully distributed communication policy and can trade computational complexity for communication efficiency. We finally demonstrated the task-effectiveness of the proposed algorithms via numerical experiments performed on a geometric consensus problem via a number of representative metrics. Furthermore, our numerical studies demonstrate the pressing need for further research on finding a metric that can measure/explain the task-effectiveness of communications with more accuracy. And, scalability in task-oriented design is yet another central challenge to be addressed in future research.

4.6 Proof of Lemma 13

Proof. Applying Adam's law on equation (4.2) yields

$$\underset{\pi}{\text{argmax}} \quad \mathbb{E}_{p(\mathbf{c}(t))}\Big\{\mathbb{E}_{p_{\pi^c,\pi^m}(\{\mathbf{tr}\}_{t'}^{T'}|\mathbf{c}(t))}\{\mathbf{g}(t')|\mathbf{c}(t)\}\Big\}, \text{ s.t. } |\mathcal{C}|\leq 2^R \tag{4.10}$$

where $\mathbf{c}(t)$ is generated by the communication policy π^c and the joint pmf of the system's trajectory $\{tr\}_{t'}^{T'}$ is directly influenced by the action policy π^m. The conditional pmf $p_{\pi^c,\pi^m}(\{\mathbf{tr}\}_{t'}^{T'}|\mathbf{c}(t))$ is the joint probability of the trajectory of the system given the received communication $\mathbf{c}(t)$ when policies $\pi^c(\cdot)$ and $\pi^m(\cdot)$ are followed. We proceed by negating the equation (4.10) and adding a second term to the objective function which is constant with respect to the decision variables of the problem to have

$$\begin{aligned}\underset{\pi^c}{\text{argmin}} \quad & \mathbb{E}_{p(\mathbf{s}(t))}\Big\{\mathbb{E}_{p_{\pi^*}(\{\mathbf{tr}\}_{t'}^{T'}|\mathbf{s}(t))}\{\mathbf{g}(t')|\mathbf{s}(t)\}\Big\}- \\ & \mathbb{E}_{p(\mathbf{c}(t))}\Big\{\mathbb{E}_{p_{\pi^c,\pi^m}(\{\mathbf{tr}\}_{t'}^{T'}|\mathbf{c}(t))}\{\mathbf{g}(t')|\mathbf{c}(t)\}\Big\}, \text{ s.t. } |\mathcal{C}|\leq 2^R.\end{aligned} \tag{4.11}$$

We replace the conditional expectation of system return by the value function $V(\cdot)$, [267](Ch. 3.5), and we will have

$$\begin{aligned}\underset{\pi^c}{\text{argmin}} \quad & \mathbb{E}_{p(\mathbf{s}(t))}\Big\{V^{\pi^*}(\mathbf{s}(t))\Big\} - \mathbb{E}_{p(\mathbf{c}(t))}\Big\{V^{\pi^m}(\mathbf{c}(t))\Big\}, \\ \text{s.t.} \quad & |\mathcal{C}|\leq 2^R.\end{aligned} \tag{4.12}$$

Note that the empirical joint distribution of $\mathbf{c}(t)$ can be obtained by following the communication policy π^c on the empirical distribution of $\mathbf{s}(t)$.

$$\begin{aligned}\underset{\pi^c}{\text{argmin}} \quad & \mathbb{E}_{p(\mathbf{s}(t))}\Big\{V^{\pi^*}(\mathbf{s}(t))\Big\} - \mathbb{E}_{p(\mathbf{s}(t))}\Big\{V^{\pi^m}(\mathbf{c}(t))\Big\}, \\ \text{s.t.} \quad & |\mathcal{C}|\leq 2^R.\end{aligned} \tag{4.13}$$

As $V^{\pi^*}(\mathbf{s}(t)) - V^{\pi^m}(\mathbf{c}(t)) \geq 0$ is true for any $\mathbf{s}(t) \in \mathcal{S}$, merging the two expectations results in

$$\underset{\pi^c}{\text{argmin}} \quad \mathbb{E}_{p(\mathbf{s}(t))}\Big|V^{\pi^*}(\mathbf{s}(t)) - V^{\pi^m}(\mathbf{c}(t))\Big|, \quad \text{s.t.} \quad |\mathcal{C}|\leq 2^R, \tag{4.14}$$

which concludes the proof of the lemma. ■

4.7 Proof of Lemma 15

Proof. We depart from the result of lemma 13 - problem (4.3). By taking the expectation over the empirical distribution of $\mathbf{s}(t)$ and applying Bellman optimality equation, we obtain

$$\begin{aligned} &\underset{\pi}{\text{argmin}} && \frac{1}{n}\sum_{t=1}^{n}\left|Q^{\pi^*}(\mathbf{s}(t),\pi^*(\mathbf{s}(t)))-Q^{\pi^m}\Big(\mathbf{c}(t),\pi^m(\pi^c(\mathbf{s}(t)))\Big)\right|, \\ &\text{s.t.} && |\mathcal{C}|\leq 2^R, \end{aligned} \tag{4.15}$$

where the vector $\pi^c(\mathbf{s}(t))$ is of N dimensions and its i-th element is $\mathbf{c}_i(t)$. We proceed by plugging $\pi^{c,ABSA-1}(\cdot)$ and Π^{m^*}, according to the definition 14, into the equation (4.15) to obtain

$$\frac{1}{n}\sum_{t=1}^{n}\left|Q^{\pi^*}(\mathbf{s}(t),\pi^*(\mathbf{s}(t)))-Q^{\pi^*}\Big(\mathbf{c}(t),\pi^*(\mathbf{s}')\Big)\right|, \tag{4.16}$$

where $\mathbf{s}'=\pi^{c,ABSA-1^{-1}}\Big(\pi^{c,ABSA-1}(\mathbf{s}(t))\Big)$, and any possible value for it lies in the same subset $\mathcal{P}_{k'}$ as $\mathbf{s}(t)$ does, while according to the definition of $\mathcal{P}_{k'}$, we know $\pi^*(\mathbf{s}(t))=\pi^*(\mathbf{s}')$, if $|\mathcal{C}|\geq|\mathcal{M}|^N$. Thus, by replacing $\pi^*(\mathbf{s}')$ in with $\pi^*(\mathbf{s}(t))$ in equation (4.17) we get

$$\frac{1}{n}\sum_{t=1}^{n}\left|Q^{\pi^*}(\mathbf{s}(t),\pi^*(\mathbf{s}(t)))-Q^{\pi^*}\Big(\mathbf{s}(t),\pi^*(\mathbf{s}(t))\Big)\right|=0. \tag{4.17}$$

This concludes the proof of theorem 15.

■

Chapter 5

Task-Oriented Communication Design at Scale

5.1 Introduction

Be it the communication of data between distinct agents, or the transmission of information/signals inside a neural network (NN), communication and information exchange have always been an inseparable part of every data-driven learning system. For decades, the role of communication inside an AI system has been less investigated; often resulting in AI systems where communications are assumed to be carried out in a perfect fashion e.g., the perfect communication of signals inside a NN. Nevertheless, communications is an integral part of AI, directly influencing its efficiency and accuracy. This is especially true when we view communications with its modern definitions steaming from the concept of task-oriented communications. In particular, with the rise of task-oriented communications [244, 286], there is a wide consensus about the diverse value of every bit sequence for a specific task [8–10]. When the receiving end of communications intends to carry out a learning task, some bits in a sequence of received communications might prove more useful. Under these circumstances, effective design of communications can (i) significantly reduce the complexity of computations at the receiving end, (ii) improve the accuracy of the learning task in a receiver with limited computational resources and (iii) reduce the power consumed for communications, and (iv) reduce the rate of communications to overcome channel rate-constraints or network resource limitations. While most of these benefits have a direct effect on the complexity as

well as the accuracy of the AI system operating at the receiving end, they are also considered to be different aspects of the mission that task-oriented communications considers for itself - making it an integral part of AI.

Rethinking communications by understanding the value of bits can result in fundamental changes in the building blocks of (distributed) machine learning. While the literature of communications research has numerous examples of how machine learning can be leveraged to solve various communication problems [287], the contribution of communications in optimizing the (distributed) learning systems, which happens to be the main focus of this manuscript, has only recently started to receive attention from the research community [2, 4, 9, 11, 25–29, 286, 288, 289].

In [2, 288], the authors investigate the effect of quantization channels for the transmission of signals from one neuron to another, inside different types of NNs. Federated learning over rate-reduced communication channels is investigated by [289], resulting in a reduced energy consummation for the whole distributed learning system. Direct task-oriented data quantization for an estimation tasks is introduced in [29]. Direct task-oriented communications for a user scheduling task is introduced by [25], achieving superior goodput performance. Indirect [1] design of communications for control tasks is carried out by [4, 9, 11, 290], with [11, 290] being focused on star typologies for the communication network of the agents and [4, 9] on full mesh networks - Fig. 5.1.

By introducing the Dec-POMDPs [106], recently, there has been a shift towards the joint design of communications and control [11]. However, we believe there is, yet, a huge potential in the disentanglement of the two problems, resulting in the introduction of task-effective communication design problems [9, 34][2]. In particular, the separation of the two problems has a multitude of advantages: (i) it drastically reduces the complexity of the original joint problem, (ii) it allows evaluating of the performance of our control and communication solutions in isolation [257, 290], (iii) it allows formulating a larger set of problems - in which agents can also communicate in an instantaneous fashion [37] - as in the joint problem a

[1] By an indirect algorithm here we mean an approach that is not dependent on our knowledge from a particular task. Indirect approaches are applicable to any/(wide range of) tasks. In contrast to indirect schemes, we have direct schemes that are specifically designed for a niche application [29]. As defined by [286]: "the direct schemes aim at guaranteeing or improving the performance of the cyber-physical system at a particular task by designing a task-tailored communication strategy".

[2] According to [9], task-effective communication problem is different from a traditional communication problem in that it capture some important features of the control task.

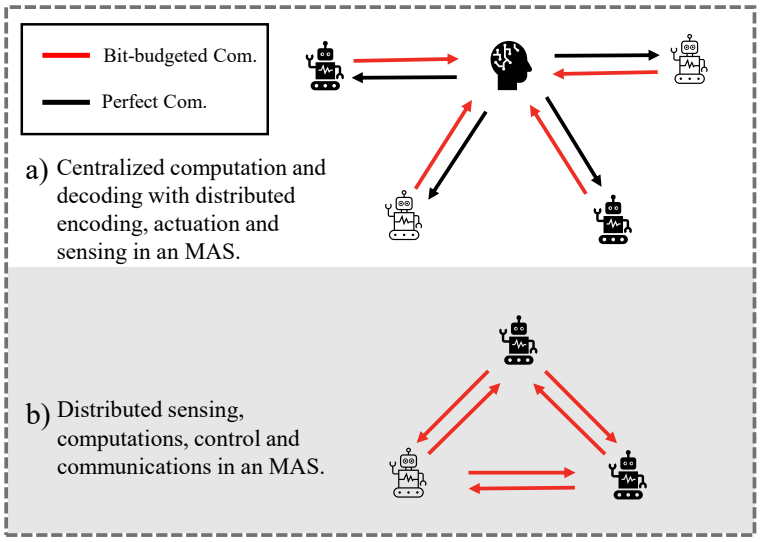


Figure 5.1: The communication network topology assumed in [11] vs. the adopted communication network topology in the current chapter and in [9].

delay in communications is inevitable [106], (iv) solving the joint problem is oblivious to the inefficiencies of the communication solution, as we ultimately measure the effectiveness of the whole system according to the average task's cost/reward obtained by the joint communication and control solution. As per [257], we can obtain a desirable performance in the task while the communications are not effective yet. Further, as [290] suggests, and is shown in Fig. 5.2, the achievable average return of the system can be improved by increasing the memory of the receiving end's controller. However, it leads to much higher complexity at the controller for the selection of suitable actions and also to the potential laziness of the learning algorithms in the transmitting end to obtain an effective communication policy. In fact, in the Dec-POMDP framework, obtaining effective distributed joint communication and control policies relies on processing the history of observations [106, 291], with the complexity of the distributed policies growing as the size of observation histories increase. Therefore, Dec-POMDP approaches can result in near-optimal control policies at the cost of complex computations [291]. The high cost of computations stems from the ineffectiveness of inter-agent communications that makes the decision-making more dependent on the larger history of observations.

The paper [9], is one of the first recent efforts to separate the data quantization and control policies. In contrast with the classic quantization problems [292] where the goal is to minimize the distortion between the original signal and its quantized version, in the task-

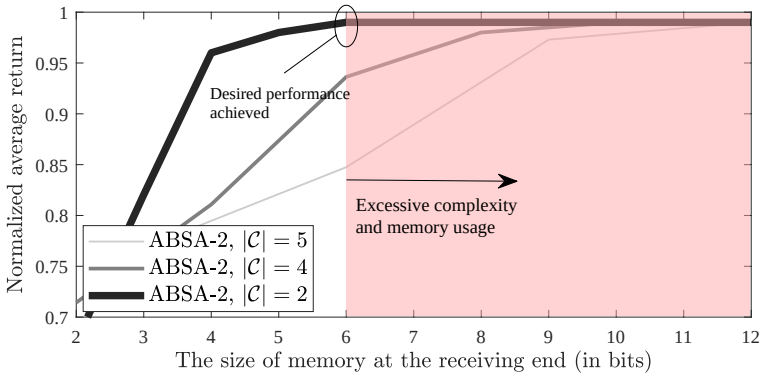


Figure 5.2: Joint design of communications and control can potentially lead to inefficient communication policies whose weakness is compensated in the controller at the cost of radical increase in the complexity its running algorithms. The three curves shown in the figure, demonstrate the performance of Action-Based State Aggregation (ABSA) introduced in [290], at three different sizes of the quantization codebook $|\mathcal{C}|$. When the controller does not have access to the state information, regardless of the method used to design the communications to it, by increasing the memory of the controller, we can increase the average return performance of the system. Although the desired performance can be achieved by increasing the size of memory at the receiving end, this comes at the cost of a significant increase in the complexity of decision making at the receiver.

oriented/semantic quantization problem, the goal is to minimize the distortion between the task-relevant/semantic data available in the original signal vs. the task-relevant/semantic data available in quantized signal [9, 34]. The analysis provided in [34] and similar works [36, 293] are not specific to a certain function that can capture the semantic/task-relevant data inside a given signal, whereas in [9] the authors introduce a particular and indirect measure that can evaluate usefulness/relevance of an observation data for control tasks. The introduced measure, being referred to as value function in dynamic programming and reinforcement learning, is shown to be able to measure the importance of local observation data for a generic multi-agent control task over Markov decision processes (MDP)s [9]. The complexity of computing the value function, however, is multiplied by the size of action-observation space with the addition of every agent to the system - making the value function extremely expensive to compute for multi-agent systems (MAS)s with large number of agents [9, 278][3]. In this direction, the authors in [211] pronounce that: "since the complexity in the state and action space grows exponentially with the number of agents, even modern deep learning approaches may reach their limits.".

[3]Since the computational complexity of Q-learning in the centralized training phase is order of $|\Omega^n \times \mathcal{M}^n|$ time complexity [278], the addition of one single agent will multiply the complexity of the centralized training by $|\Omega \times \mathcal{M}|$.

Due to the importance of the scalability of multi-agent reinforcement learning (MARL) algorithms, the initial efforts to address this issue can be traced back to 1999 [209, 210], where authors introduce reward/value sharing for local optimizations. Since then, there has been a sustained effort to address the problem by other means such as introducing factored MDPs [211] or independent Q-learning [212]. Although the independent Q-learning and similar schemes [212] can scale with the growing number of agents they suffer from sub-optimality caused by the non-stationarity of the environment from each agent's perspective. This issue is addressed by modern MARL algorithms that comprise a centralized training phase, through which the training environment is guaranteed to stay stationary [276]. The complexity of the centralized training phase, however, grows exponentially as the number of agents increases in the MAS. Monotonic Value Function Factorization - QMIX [213] - enforces a monotonicity constraint on the relationship between the local Q-functions and the centralized Q-function to reduce the complexity of factorizing the value decomposition networks. The overall complexity of MARL, however, increases with the addition of each agent to the system. Even with the use of attention mechanisms for the centralized training authors have not been able to go beyond linearly increasing the complexity of the centralized training with respect to the number of agents [214]. To the best of our knowledge, the current thesis is the first to reduce the complexity of the centralized training phase from exponential time complexity - $\mathcal{O}(|\Omega|^N \times |\mathcal{M}|^N)$ - to constant time complexity - $\mathcal{O}(1)$ - with respect to the number of agents. The only caveat is that our proposed scheme to reduce the complexity of the centralized training phase cannot be applied to every multi-agent learning problem but only to design the inter-agent communications in the multi-agent setting. In particular, the contributions of the chapter are as follows.

- We provide analytical studies to show that a two-agent centralized training phase is sufficient to draw the insights we need from the centralized training phase - if the initial conditions of SAIC are met. Regardless of the method used in the centralized training to compute the value of observation space e.g., deep reinforcement learning, exact reinforcement learning or dynamic programming, our analytical results stay valid.
- The proposed analytical studies suggest that the value function obtained from the two-agent centralized training phase is sufficient to cast the task-oriented data quantization problem - even if we do not know the relationship between value function of the two-

agent system versus N-agent system, N being the real number of agents the system is composed of.

- According to these results, we propose an scalable version of SAIC - an existing task-oriented data quantization scheme. While SAIC is very hard to scale, our proposed algorithm - ESAIC - can easily be applied to MASs composed of a large number of agents.

- By carrying out numerical studies on geometrical consensus problem [208], will show that the proposed method in the current chapter is capable of reducing the complexity of the centralized training for hundreds of days - if not years - even in very simple problems, while it maintains the average return performance of the algorithm close to the optimality.

5.1.1 Organization

Section II describes the system model for a cooperative multi-agent task with rate-constrained inter-agent communications. Section II provides a quick overview to SAIC, an exiting algorithm that can solve provide a solution to the joint control and data compression policy design problem. Our goal is to make SAIC computationally less complex in this manuscript. Section III proposes the extended SAIC (ESAIC) a scheme for the joint design of data compression and control policies which is much less complex to run and very similar to SAIC in average return performance. We also provide analytical results on the conditions that ESAIC can maintain the performance of its predecessor. The numerical results and discussions are provided in section IV. Finally, section V concludes the chapter.

We also use the concept of image functions in our analytical studies which is defined as the following. Let $g(\cdot) : \mathcal{D} \to \mathcal{C}$ be a function and $\mathcal{D}' \subset \mathcal{D}$ be a subset of its domain. The image function of $g(\cdot)$ denoted by $\ddot{g}(\cdot) : \mathbb{P}(\mathcal{D}) \to \mathbb{P}(\mathcal{C})$ is defined as $\ddot{g}(\mathcal{D}') \triangleq \{c \in \mathcal{C} \mid g(d) = c \ , d \in \mathcal{D}'\}$. For the sake of the simplicity of the analysis, the arguments of the function may be omitted when no confusion is raised, e.g., we have used $r^{[n]}(\cdot)$ instead of $r^{[n]}(\mathsf{o}_1, ..., \mathsf{o}_n, \mathsf{m}_1, ..., \mathsf{m}_n)$.

Table 5.1: Table of notations

Symbol	Meaning
$\mathbf{x}(t)$	A generic random variable generated at time t
$\mathsf{x}(t)$	Realization of $\mathbf{x}(t)$
$\mathcal{X}$	Alphabet of $\mathbf{x}$(t)
$\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
$\mathbb{P}(\mathcal{X})$	Power set of $\mathcal{X}$
$p_{\mathbf{x}}(\mathsf{x}(t))$	Shorthand for $\Pr(\mathbf{x}(t) = \mathsf{x}(t))$
$H(\mathbf{x}(t))$	Information entropy of $\mathbf{x}(t)$ (bits)
$\mathcal{X}_{-\mathbf{x}}$	$\mathcal{X} - \{\mathbf{x}\}$
$\mathbb{E}_{p(\mathsf{x})}\{\mathbf{x}\}$	Expectation of the random variable X over the probability distribution $p(\mathsf{x})$
$\delta(\cdot)$	Dirac delta function
$\mathsf{tr}(t)$	Realization of the system's trajectory at time t

5.2 Problem Statement

We consider a multiagent system (MAS) in which multiple agents $i \in \mathcal{N} = \{1, 2, ..., N\}$ collaboratively and distributedly execute a task. The system runs on discrete time steps t. The observation of each agent i at time step t is shown by $\mathbf{o}_i(t) \in \Omega$ and the state $\mathbf{s}(t) \in \mathcal{S}$ of the system is defined by the vector of joint observations $\mathbf{s}(t) \triangleq [\mathbf{o}_i(t)]_{i\in\mathcal{N}} \in \Omega^N$. Now let $\mathbf{s}_i(t) \in \{\Omega \cup 0\}^N$ be the vector of agent i's local state, with all its elements being equal to zero except for its i'th element which is equal to $\mathbf{o}_i(t)$. We assume that $\forall i, j \in \mathcal{N}$ the local states $\mathbf{s}_i(t)$ and $\mathbf{s}_j(t)$ are linearly independent. This is also referred to as joint observability of the state [37]. The control action of each agent i at the time t is shown by $\mathbf{m}_i(t) \in \mathcal{M}$, and the action vector $\mathbf{m}(t) \in \mathcal{M}^N$ of the MAS is defined by the joint actions $\mathbf{m}(t) \triangleq \langle \mathbf{m}_1(t), ..., \mathbf{m}_N(t)\rangle$. The observation space Ω, state-space $\mathcal{S}$, and action space $\mathcal{M}$ are all discrete sets. The environment is governed by an underlying Markov Decision Process that is described by the tuple $M = \langle \mathcal{S}, \mathcal{M}^N, r(\cdot), \gamma, T(\cdot)\rangle$, where $r(\cdot) : \mathcal{S} \times \mathcal{M}^N \to \mathbb{R}$ is the per-stage reward function and the scalar $0 \leq \gamma \leq 1$ is the discount factor. Also, the function $r^{[n]}(\cdot) : \Omega^n \to \mathcal{M}^n$ is the reward function of an MAS comprised of n agents. The function $T(\cdot) : \mathcal{S} \times \mathcal{M}^N \times \mathcal{S} \to [0, 1]$ is a conditional probability mass function (PMF) which represents state transitions such that $T(\mathsf{s}(t+1), \mathsf{m}(t), \mathsf{s}(t)) = \Pr(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t))$. The performance of the MAS is measured according to the system's average return defined as the summation of obtained per-stage rewards within the time horizon T':

$$\mathbf{g}(t^{'}) = \sum\nolimits_{t=t^{'}}^{T'} \gamma^{t-1} r(\mathbf{s}(t), \mathbf{m}(t)). \tag{5.1}$$

Once per time step, following the Fig. 5.3, agent $i \in \mathcal{N}$ is allowed to transmit a communication vector $\mathsf{c}_i(t)$ to each agent $j \in \mathcal{N}_{-i} = \mathcal{N}_{-i}$. Conditioned on its observation $\mathsf{o}_i(t)$, agent i transmits a vector of communication messages $\mathsf{c}_i(t) = [\mathsf{c}_{i,j}(t)]_{j\in\mathcal{N}_{-i}} \in \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{i,j}$, in which the element $\mathsf{c}_{i,j}(t)$ denotes the message sent by agent i to agent j, where $\mathsf{c}_{i,j}(t)$ is generated following the communication policy $\pi^c_{i,j}(\cdot) : \Omega \to \mathcal{C}_{i,j}$. The set $\mathcal{C}_{i,j}$ is an alphabet $\{\mathsf{c}_{i,j}, \mathsf{c}'_{i,j}, \mathsf{c}''_{i,j}, ..., \mathsf{c}^{(B_{i,j}-1)}_{i,j}\}$ composed of a finite $B_{i,j}$ number of communication code-words - we use the same notation to refer to the different elements of the action, observation and state spaces too. Agent i's communications are generated by following the tuple $\pi^c_i = \langle \pi^c_{i,j}(\cdot)\rangle_{j\in\mathcal{N}_{-i}}$ which is comprised of $N-1$ different communication policies. Agent i's communications are sent over $N-1$ separate error-free finite-rate bit pipe, with its rate constraint to be $R_{i,j} \in \mathbb{R}$ (bits per channel use) or equivalently (bits per time step). As a result, the cardinality of the communication symbol space $\mathcal{C}_{i,j}$ for each i to j inter-agent communication link should follow the inequality

$$B_{i,j} \leq 2^{R_{i,j}}. \tag{5.2}$$

In the special case that the bit-budget is constant across all the communication links of the system the bit-budget of the channels is simply denoted by R.

Each agent i exploits its observation $\mathbf{o}_i(t)$ together with the received communication messages $\tilde{\mathsf{c}}_i(t) = [\mathsf{c}_{j,i}(t)]_{j\in\mathcal{N}_{-i}} \in \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{j,i}$ within time-step t to select the control signal $\mathsf{m}_i(t)$ following a deterministic control policy $\pi^m_i(\cdot) : \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{j,i} \times \Omega \to \mathcal{M}$.

In contrast to [9], we will not characterize the performance gap caused by the limited connectivity in the communication network of agents. Characterizing the difference between the performance of the MAS that runs over heterogeneous bit-budgets and the MAS that runs over perfect communication channels is differed to the future works. The present chapter, however, will provide some numerical as well as analytical studies on the performance of the proposed scheme - ESAIC - under asymmetrical communication bit-budgets $R_{i,j}$.

Definition 16. ***(Distributed Joint Control and Communication Design (D-JCCD) problem).*** *Let M be the MDP governing the environment and the scalar $R_{i,j} \in \mathbb{R}$ to be the bit-budget of each inter-agent communication channels. At any time step t', we aim at*

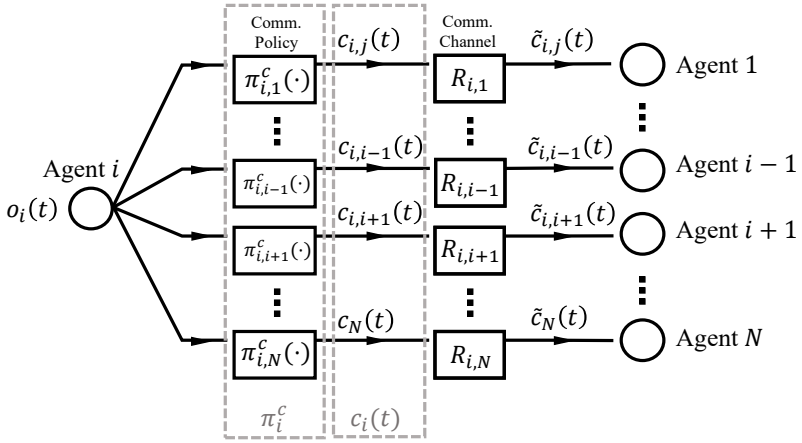


Figure 5.3: Illustration of message transmission (encoding) at agents i. Agent i's observation $\mathsf{o}_i(t)$ at time step t is transmitted to all other agents. Each inter-agent communication channel from agent i to agent j is assumed to be reliable so long as bit-budgets requirements - explained in (5.2) - are respected.

designing the tuple $\pi_i = \langle \pi^m_i(\cdot), \pi^c_i \rangle$ *to solve the following variational dynamic programming*

$$\underset{\pi_i}{\text{argmax}} \;\; \mathbb{E}_{\pi_i}\Big\{\mathbf{g}(t')\Big\}; \quad \text{s.t.} \;\; B_{i,j} \leq 2^R, \;\; \forall i,j \in \mathcal{N} \tag{5.3}$$

where the expectation is taken over the joint pmf of system's trajectory $\{\mathsf{tr}\}_{t'}^{T'} = \mathsf{o}_1(t'), ..., \mathsf{o}_N(t'), \mathsf{m}(t'), ..., \mathsf{o}_1(T'), ..., \mathsf{o}_N(T'), \mathsf{m}(T')$*, when each agent* i *follows the policy* π_i *for all agents* $i \in \mathcal{N}$*.*

5.3 Preliminaries - State Aggregation for Information Compression (SAIC)

Instead of directly solving the D-JCCD problem (5.3), SAIC breaks the problem into three separate parts to enjoy the advantages discussed in section 5.1. This is done such that in a sequence of steps the features of the control task are distilled and captured in a novel communication design problem called task-oriented communication design. First, following a centralized training and distributed execution approach [3, 276], SAIC solves the problem from a centralized point of view where all the communications between agents and a central controller are considered to be perfect:

$$\pi^*(\cdot) = \underset{\pi(\cdot)}{\text{argmax}} \ \mathbb{E}_{\pi}\Big\{\mathbf{g}(t)\Big\}, \tag{5.4}$$

and the policy π can be expressed as a CMF $\pi\Big(\mathsf{m}(t)\Big|\mathsf{s}(t)\Big) = p(\mathsf{m}(t)|\mathsf{s}(t))$. In the centralized problem (5.4), the objective is to design one centralized control strategy $\pi(\cdot) : \Omega^N \rightarrow \mathcal{M}^N$. explain the relation between the value function and the optimal policy.

Subsequently, the knowledge that we obtain by solving the centralized problem is captured within the task-oriented communication design by designing a mapping. The non-injective surjective mapping $V^*(\cdot) : \Omega \rightarrow \mathcal{V} \subset \mathbb{R}$, that is obtained after solving the centralized problem, would allow us to solve the communication problem over the output space of the mapping $V^*(\cdot)$ - value space $\mathcal{V}$ - rather than over the original observation space. This is imperative because of a multitude of reasons: (i) The mapping $V^*(\cdot)$ projects the high-dimensional observation points to the single-dimensional space of $\mathcal{V} \subset \mathbb{R}$, leading to a reduced the complexity for the clustering problem, (ii) the mapping $V^*(\cdot)$ captures the features of the control task and allows us to take these features into account inside our communication design problem - that helps us to separately design communications and control policies, (iii) the clusters in the output space of the $V^*(\cdot)$ are shown to be linearly separable, (iv) last but not least, it is very intuitive to see how the mapping $V^*(\cdot)$ is an indirect/universal measure to quantify the value of each observation for any given task. Accordingly, the observation points are not clustered together based on how similar they are, but based on how similarly valuable they are for the task. After obtaining the communication policies of all agents, within the last training phase, each agent i follows the communication policy learned earlier and learns its control policy $\pi_i^m(\cdot)$. All these steps are briefly explained in the following part of the current section.

It is shown in [35,294] that, after obtaining $\pi^*(\cdot)$ as the optimal solution to (5.4), one can obtain the value of each observation $\mathsf{o}^{(k)}$ for all $k \in \{1, ..., |\Omega|\}$ following the

$$\begin{aligned} &V^{*[N]}(\mathsf{o}_i(t) = \mathsf{o}^{(k)}) = \\ &\sum_{\mathsf{o}_{-i}(t) \in \Omega^{N-1}} \mathbb{E}_{\pi^*}\Big\{\mathbf{g}(t)|\mathbf{s}(t), \pi^*(\mathbf{s}(t))\Big\} p(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t)), \end{aligned} \tag{5.5}$$

where $\mathbf{s}(t) = [\mathsf{o}_1(t), \ldots, \mathsf{o}_N(t)]$ and the summation $\sum_{\mathsf{o}_{-i}(t) \in \Omega^{N-1}}$ is used to denote $N-1$ summations over all possible values for $\mathsf{o}_{-i}(t) = [\mathsf{o}_i(t)]_{i \in \mathcal{N}_{-i}}$. As also shown in the transition

of Fig. 5.4. a to Fig. 5.4. b, by knowing the mapping $V^{*[N]}(\cdot)$ we can map all the observation values to the one-dimensional value space $\mathcal{V}$. Accordingly, the clustering of observation points will no longer be done based on their observation values e.g., $\mathsf{o}_i(t)$, but based on the value function of the observation values, e.g., $V^{*[N]}(\mathsf{o}_i(t))$ - where the superscript $[N]$ illustrates the number of agents in the centralized training phase [4]. This would result in solving the following task-oriented data compression problem in the form of a clustering problem

$$\min_{P_{i,j}} \quad \sum\nolimits_{k=1}^{2^{R_{i,j}}} \sum\nolimits_{\mathsf{o}\in\mathcal{P}_{i,k}} \left| V^{*[N]}(\mathsf{o}_i(t)=\mathsf{o}) - \mu_k^{'} \right|, \tag{5.6}$$

to cluster observations via the partition $\mathcal{P}_{i,j} = \{\mathcal{P}_{i,j,1}, \ldots, \mathcal{P}_{i,j,B_{i,j}}\}$ and learn the communication strategy $\pi^c_{i,j}(\cdot)$, where for each $\mathcal{P}_{i,j}$, all the observations $\mathsf{o}_i(t) \in \mathcal{P}_{i,j,k}, \forall k \in \{1, ..., B_{i,j}\}$ are corresponded to a single unique code-word $\pi^c_{i,j}(\mathsf{o}_i(t))$, equivalently the output of the image function $\ddot{\pi}^c_{i,j}(\mathcal{P}_{i,j,k}), \forall k \in \{1, ..., B_{i,j}\}$ is a single member set. Solving problem (5.6) is illustrated in Fig. 5.4 by a transition from subplot "b" to "c". It was shown in [35] that the optimal solution to (5.6) is the optimal solution to an approximated version of the problem (5.3). Note that for every agent i, the problem (5.6), should be solved N^c_i number of times where N^c_i denotes the of agents $j \in \mathcal{N}_{-i}$ for whom the bit-budget of communications $B_{i,j}$ are distinct. Equivalently, the communication policy of agent i for the two distinct receiving ends $j, j' \in \mathcal{N}_{-i}$ will stay the same if $B_{i,j} = B_{i,j'}$.

Once the clustering problem (5.6) is solved, we have obtained indirect task-effective communication policies $\pi^c_i, \forall i \in \mathcal{N}$ [5]. To completely solve the D-JCCD problem (5.3) via SAIC, we still have to find the optimal control policy $\pi^m_i(\cdot)$ of for agent each agent i. Via the control policy $\pi^m_i(\cdot)$, at any time step t, agent i selects a control signal $\mathbf{m}_i(t)$, conditioned only on the quantized data received from the other agents $\tilde{\mathbf{c}}_i(t) \in \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{j,i}$, together with its own observation $\mathbf{o}_i(t) \in \Omega$. SAIC obtains the control policy $\pi^m_i(\cdot)$ for each agent i, via a distributed training phase, in which agents communicate through bit-budgeted communication channels - following (5.2). To this aim, the communications of each agent $i \in \mathcal{N}$ to each agent $j \in \mathcal{N}_{-i}$ are carried out via the communication policy $\pi^c_{i,j}(\cdot)$ that is obtained by solving (5.6). To obtain asymptotically optimal control policies, SAIC utilizes distributed Q-learning [249] for

[4] Note that, whenever a function/policy - e.g., $\pi^c_{i,j}(\cdot)$ - is obtained via a centralized training which has had N' number of agents in it, the superscript $[N']$ is used for that function/policy - e.g., $\pi^{c,[N']}_{i,j}(\cdot)$.

[5] These quantization policies are indirect since they can be obtained without any prior knowledge about the task. And, they are task-effective, since they can be designed to preserve the accuracy of observation data when the observation data is deemed valuable for the specific task.

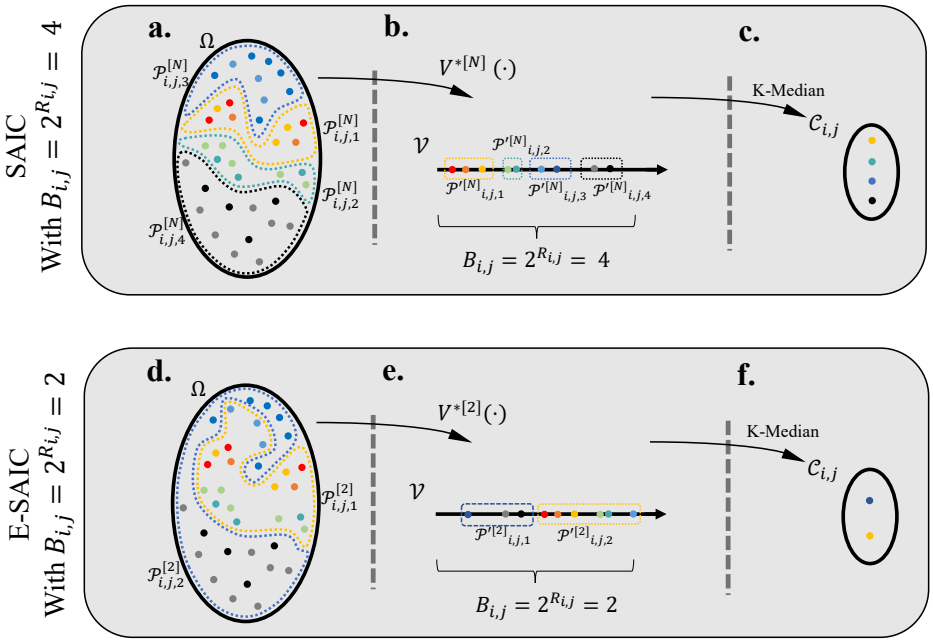


Figure 5.4: Illustration of the steps taken to design the communication policy $\pi^c_{i,j}(\cdot)$ using SAIC and ESAIC.

the distributed training phase.

The computational complexity of the centralized training phase $\mathcal{O}(|\Omega|^N \times |\mathcal{M}|^N)$ for a certain number of agents N grows linearly with the size of observation and action spaces and for a certain size of observation-action space grows exponentially with the number of agents N. This makes computational cost of SAIC for large MASs prohibitively high, limiting its application to MASs composed of only a few agents. Given the exponential time complexity of the centralized training phase and its much higher time complexity compared with the distributed training phase, the centralized training phase is the major computational bottleneck of SAIC.

5.4 Extended State Aggregation for Information Compression in Multiagent Coordination Tasks

In this section, we propose a straightforward extension of SAIC called Extended SAIC (ESAIC) which is capable of drastically reducing its time complexity in the centralized training phase. While the time complexity of the centralized training phase in SAIC grows exponentially with respect to the number of agents, in ESAIC, increasing the number of agents in

the MAS, has no impact on the computational complexity of the centralized training phase - making ESAIC more efficient than SAIC [9] and any other MARL with a central training phase [213, 276, 290]. ESAIC is not just a replacement for SAIC, but introduces the more general idea of reducing the number of agents in the MAS for the centralized training phase. Extended SAIC, proceeds by following the same steps as SAIC to solves the D-JCCD problem: (i) centralized training phase, (ii) task-oriented data compression problem, (iii) distributed training of agents' control policies. The only difference is that the centralized training phase is done with only two agents in the training phase - regardless of the number of agents N for which we want to solve the original D-JCCD problem (5.3).

5.4.1 Centralized Training Phase

Accordingly, in ESAIC, in the first step by carrying out centralized training phase we solve the problem (5.4) for a two-agent system to obtain $V^{*[2]}(\cdot)$. Afterwards, by solving the following task-oriented quantization problem

$$\min_{P_{i,j}^{[2]}} \quad \sum_{k=1}^{2^{R_{i,j}}} \sum_{\mathsf{o} \in \mathcal{P}_{i,k}} \left| V^{*[2]}(\mathsf{o}_i(t) = \mathsf{o}) - \mu_k^{'} \right|, \tag{5.7}$$

we obtain a new partition $\mathcal{P}_{i,j}^{[2]}$ of the observation space that leads to a different, yet effective communication/quantization policy $\pi_{i,j}^{c[2]}(\cdot)$. K-median clustering can be used to solve the above-mentioned problem (5.7). In this direction, to obtain the quantization policy of agent i for its communication to agent j we compute a partition $\mathcal{P}'^{[2]}_{i,j}$ of the set $\mathcal{V}_i^{[2]}$ - where $\mathcal{V}_i^{[2]}$ is the image of Ω under the function $V^{*[2]}(\cdot)$ i.e., $\mathcal{V}_i^{[2]} = \ddot{V}^{*[2]}(\Omega)$. We first solve the following problem

$$\min_{P'^{[2]}_{i,j}} \quad \sum_{k=1}^{2^{B}} \sum_{V^*(\mathsf{o}_i(t)) \in P'^{[2]}_{i,j,k}} \left| V^{*[2]}(\mathsf{o}_i(t)) - \mu_k^{''} \right|.$$

Afterwards, as shown in Figure 5.4, the observation points should be clustered according to the clustering of their corresponding values. That is, any two distinct observation points $\mathsf{o}_i', \mathsf{o}_i'' \in \Omega$ are clustered together in $\mathcal{P}_{i,j,k}$ if and only if their values $V^{*[2]}(\mathsf{o}_i'), V^{*[2]}(\mathsf{o}_i'') \in \mathcal{P}'_{i,j}$ are in the same cluster $\mathcal{P}'_{i,j,k}$.

5.4.2 Distributed Training Phase

After obtaining the communication policies, we solve the following distributed control design problem

$$\underset{\pi_i^m}{\text{argmax}} \ \ \mathbb{E}_{\pi_i}\Big\{\mathbf{g}(t')\Big\}, \quad \forall i \in \mathcal{N} \tag{5.8}$$

through a distributed training phase to obtain the control policy of each agent i, where the expectation is taken over the MAS's trajectory that is influenced by both the control policy $\pi_i^c(\cdot)$ and the communication/quantization policy $\pi_i^m(\cdot)$ of all agents $i \in \mathcal{N}$. The detailed recipe of ESAIC can be found in Algorithm 1 and its performance will be studied both analytically and numerically in sections 5.5 and 5.6, respectively.

As will be shown in section 5.5, the number of agents in the training phase can be reduced, regardless of the specific method used to compute the function $V^{*[2]}(\cdot)$. Accordingly, we conjecture that other schemes such as deep Q-learning [105], deep double Q-learning [295], deep deterministic policy gradient [296] and other similar (deep) reinforcement learning algorithms can be used for a two-agent centralized training phase to approximate the value function $V^{*[2]}(\cdot)$ - as long as the condition of theorem 17 is met.

Algorithm 3 Extended State Aggregation for Information Compression (ESAIC)

Input: γ, α, c
Initialize all-zero Q-table $Q_i^m(\cdot) \leftarrow Q_i^{m,(k-1)}(\cdot)$, for $i = 1 : N$
and all-zero Q-table $Q(\mathsf{s}(t), \mathsf{m}(t))$.
Obtain $\pi^{*[2]}(\cdot)$ & $Q^{*[2]}(\cdot)$ by solving (5.4) using Q-learning [267].
Compute $V^{*[2]}(\mathsf{o}_i(t))$ following eq. (5.5), for $\forall \mathsf{o}_i(t) \in \Omega$.
Obtain $\pi_i^{c[2]}$ by solving the problem (5.7) N_i^c times, for $i = 1 : N$. **for** *each episode* $k = 1 : K$ **do**
Randomly initialize the observation $\mathsf{o}_i(t = 1)$, for $i = 1 : N$ **for** $t_k = 1 : M$ **do**
Select $\mathsf{c}_i(t)$ following $\pi_i^{c[2]}(\cdot)$, for $i = 1 : N$
Obtain message $\tilde{\mathsf{c}}_i(t)$, for $i = 1 : N$
Update $Q_i^m(\mathsf{o}_i(t-1), \tilde{\mathsf{c}}_i(t-1), \mathsf{m}_i(t-1))$, for $i = 1 : N$
Select $\mathsf{m}_i(t) \in \mathcal{M}$ following ϵ-greedy, for $i = 1 : N$
Obtain reward $r(\mathsf{s}(t), \mathsf{m}(t))$, for $i = 1 : N$
Make a local observation $\mathsf{o}_i(t)$, for $i = 1 : N$
$t_k = t_k + 1$
end
Compute $\sum_{t=1}^{M} \gamma^{t-1} r_t$ for the lth episode
update ϵ via: $\epsilon = -0.99k/K + 1$
end
Output: $Q_i^m(\cdot)$ and $\pi_i^m(\mathsf{m}_i(t)|\mathsf{o}_i(t), \tilde{\mathsf{c}}_i(t))$, for $i = 1 : N$

5.5 Analytical study of ESAIC

After introducing the idea of ESAIC, in 5.4, in this section, we provide some analytical studies on its average return performance of it as well as studies on its computational complexity.

5.5.1 Average return performance

The main result of this subsection is to prove that by solving the problem (5.7), one can obtain inter-agent communication/quantization policies which are as effective as the solutions to the problem (5.6). Equivalently, one can reduce the number of agents in the centralized training phase and yet draw enough insights from it to design task-oriented communication policies. The proof provided in this section, therefore, is a testament to how rich is the value function of a two-agent centralized training phase to indirectly incorporate the features of the control task into the task-oriented communication design problem (5.7). These features have been previously extracted e.g., from the control problem through the Eigenvalues of the plant[6] to be controlled [74] - for linear time-invariant plants.

Theorem 17. *Let the bijection $f(\cdot) : \mathcal{V}^{[2]} \rightarrow \mathcal{V}^{[N]}$ be the mapping from the value of observations for a two-agent scenario to the N-agent. For all $i, j \in \mathcal{N}$, the partition $\mathcal{P}^{[2]}_{i,j}$ proposed by ESAIC (that is obtained by solving the problem (5.7)) are the same as the partition $\mathcal{P}^{[N]}_{i,j}$ proposed by SAIC (that is obtained by solving the problem (5.6)) if*

$$c_1 : \quad \forall k \in \{1, \dots, B_{i,j}\} \;\; \exists k' \in \{1, \dots, B_{i,j}\} : \quad \ddot{f}(\mathcal{P}'^{[2]}_{i,j,k}) = \mathcal{P}'^{[N]}_{i,j,k'} \tag{5.9}$$

Proof. Appendix 5.8. ∎

Remark 1: Following the theorem 17, all the guarantees that are presented for the performance of SAIC are in place if $R_{i,j} = R \;\; \forall i, j \in \mathcal{N}$.

Theorem 17, provides a conditional guarantee for the equivalence of the results obtained by SAIC and ESAIC. The condition, however, is not always easy to verify. In the next section, we will provide numerical results also for the cases where the condition of the theorem is violated. The near-optimal performance of ESAIC, even under the violation of c_1 in (5.9),

[6] In the terminology of reinforcement learning, the plant is referred to as the environment.

confirms that the proposed condition is too strong and can be further relaxed in future works. Moreover, once we see certain structures and features in the function $f(\cdot)$ for smaller MASs, they may hold for larger MASs too. This would provide us with an analytical basis to use proof by induction to verify the condition c_1. The following remarks, introduce some of these features.

Remark 2: If the function $f(\cdot)$ is limit-preserving then it meets the condition c_1 [297].

Remark 3: If the function $f(\cdot)$ is strictly monotonic, then it is limit-preserving too [297].

In particular, let the superscript in $f^{[N]}(\cdot)$ determine the superscript of its range $\mathcal{V}^{[N]}$. We will show in lemma 18 that if $f^{[3]}(\cdot)$ is strictly monotonic, so is $f^{[N]}(\cdot)$ - for a specific class of reward functions and observation structures. Accordingly, to verify the condition c_1 for every $N \geq 3$, it will be sufficient to just verify it for $N = 3$.

Lemma 18. *Let the function $f^{[3]}(\cdot) : \mathcal{V}^{[2]} \to \mathcal{V}^{[3]}$ be strictly monotonic, and the conditions c_2 and c_3 met, defined as follows:*

c_2: The discrete derivative of the $r^{[k]}(\cdot)$ with respect to k be a linear function, i.e., there exist scalars $\tau \in \mathbb{R}^+$ and $\zeta \in \mathbb{R}$ such that

$$r^{[k+1]}(t) = \tau r^{[k]} + \zeta, \tag{5.10}$$

c_3: The observations $\mathsf{o}_i(t)$ of the i'th agent are independent of the observations $\mathsf{o}_j(t)$ of the j'th agent $\forall i, j \in \mathcal{N}$ and $\forall \mathsf{o} \in \Omega$.

Then, the Proposition $P(N)$ holds true as follows:

$P(N)$: The function $f^{[N]}(\cdot) : \mathcal{V}^{[2]} \to \mathcal{V}^{[N]}$ is strictly monotonic for all $N \geq 3$.

Proof. Appendix 5.9. ■

remark 4: justify c2 and c3

5.5.2 Computational complexity

As is discussed in [278], the computational complexity of exact Q-learning is proportional to the size of state-action space. Exact Q-learning is used in the centralized and distributed

training phases of SAIC and ESAIC. In the centralized training phase of SAIC, the computational complexity $\mathcal{O}(|\Omega \times \mathcal{M}|^N)$ grows exponentially with the size of MAS N. Accordingly, the addition of each agent to the system multiplies the complexity of the Q-learning by $|\Omega \times \mathcal{M}|$. The complexity $\mathcal{O}(|\Omega \times \mathcal{M}|^2)$ of the centralized training phase in ESAIC with respect to the size of the MAS N, however, is constant time. That is, ESAIC will always execute at the same time (or space) regardless of the size of the MAS N.

The complexity $\mathcal{O}(|\Omega \times \mathcal{C}^{n-1} \times \mathcal{M}|)$ of the Q-learning problem that each agent solves in SAIC, at the decentralized training phase, also grows exponentially with the addition of each agent to the system. Compared with the centralized training phase, in the distributed training phase, SAIC is much less sensitive to the addition of an agent to the system. Although the complexity of the Q-learning at each agent i multiplies by a constant $|\mathcal{C}|$ with the addition of each agent to the system, the size of the communication space $|\mathcal{C}|$ is much smaller than $|\Omega \times \mathcal{M}|$[7]. In the decentralized training phase, ESAIC follows the same complexity patterns.

Remark: If the condition c_1 of theorem 17 is met, ESAIC offers the same performance as SAIC at a much reduced computational cost. Accordingly, for a problem comprised of N agents, the time complexity of SAIC is $|\Omega \times \mathcal{M}|^{N-2}$ times higher than ESAIC.

5.6 Numerical Studies

To evaluate our proposed method, ESAIC, in this section, we leverage numerical experiments on a specific cooperative task i.e., a geometric consensus problem with finite observability, called the rendezvous problem. Geometric consensus problems are emerging in many new applications, such as UAV/vehicle platooning, which makes them a useful application domain for the framework proposed in this paper [208]. Based on the results demonstrated in this section, the proposed framework, ESAIC, has been shown to be a suitable candidate for the distributed control of the large vehicle/UAV platoons under limited communications.

The rendezvous problem, which is a subcategory of the geometric consensus, has already been studied in the literature of multi-agent systems [265, 277], whereas in our case the inter-agent communication channel is set to have a limited bit-budget. The rendezvous is a

[7]To understand why "the size of the communication space $|\mathcal{C}|$ is much smaller than $|\Omega \times \mathcal{M}|$", remember that we solve the problem (5.7) to significantly reduce the size of the communication message space $\mathcal{C}$ of agent i compared with the size of its observation space Ω.

particularly interesting testbed for multi-agent communications as it allows us to consider a cooperative MAS consisting of multiple agents whose coordination is dependent on communication. In particular, as detailed in subsection 5.6.1, if the communication between agents is not efficient, at any time step t each agent i will only have access to its local observation $\mathsf{o}_i(t)$, which is its own location in the case of rendezvous problem. This mere information is insufficient for an agent to attain the larger reward C_2, but is sufficient to attain the smaller reward C_1. Accordingly, compared with cases in which no communication between agents is present, in the set-up of the rendezvous problem, efficient communication policies can increase the attained objective function of the MAS [9]. We consider a variety of grid worlds with different size values N and different locations for the goal-point ω^T. We compare the proposed ESAIC and SAIC with the centralized Q-learning scheme that is guaranteed to achieve the optimal average return performance in the rendezvous problem. Our approach can be straightforwardly applied to other geometric systems e.g., by changing the reward function. In particular, a reward function that encourages the agents to come together as close as possible but not collide with each other can emulate a vehicle platooning scenario. While useful, it is outside the scope of our work to investigate the response of the multi-agent system to different rewarding schemes. Note that, according to [9], regardless of the definition of the reward function, the geometric consensus problem (or in general the joint quantization and control problem) can be solved by SAIC if the necessary conditions are met, and a centralized training phase is feasible.

5.6.1 Rendezvous Problem

As illustrated in Fig. 5.5, in a rendezvous problem, multiple agents operate on an $N \times N$ grid world and aim at arriving at the same time at the goal point on the grid. The system operates in discrete time, with agents taking actions and communicating in each time step $t = 1, 2, \dots$. Each agent $i \in \mathcal{N}$ at any time step t can only observe its own location $\mathsf{o}_i(t) \in \Omega$ on the grid, where the observation space is $\Omega = \{0, 1, ..., n^2 - 1\}$. Each episode terminates as soon as an agent or more visit the goal point which is denoted as $\omega^T \in \Omega$. That is, at any time step t that the observation of each agent $i \in \mathcal{N}$ is a member of Ω^T, the episode will be terminated - so the time horizon M is non-deterministic. The subset $\mathcal{S}^T \subset \mathcal{S}$ also defines all state realizations where one or more agents are in the goal location i.e.,

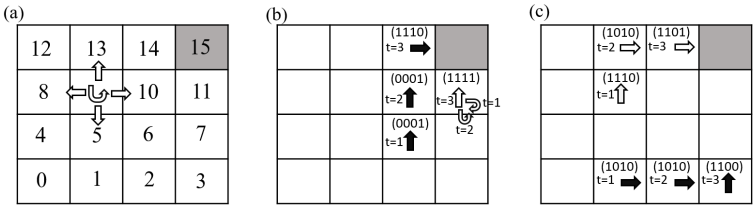


Figure 5.5: The rendezvous problem when $n = 2$, $N = 4$ and $\omega^T = 15$: (a) illustration of the observation space, Ω, i.e., the location on the grid, and the environment action space $\mathcal{M}$, denoted by arrows, and of the goal state ω^T, marked with gray background; (b) demonstration of a sampled episode, where arrows show the environment actions taken by the agents (empty arrows: actions of agent 1, solid arrows: actions of agent 2) and the $B = 4$ bits represent the message sent by each agent. A larger reward $C_2 > C_1$ is given to both agents when they enter the goal point at the same time, as in the example; (c) in contrast, C_1 is the reward accrued by agents when only one agent enters the goal position [4].

$$\mathcal{S}^T = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \exists i \in \mathcal{N} : \mathsf{o}_i(t) \in \omega^T\}.$$

We also define the subset $\mathcal{S}^T_{n'} \subset \mathcal{S}^T$ that includes all the terminal states where only n' number of agents have arrived at the goal location i.e.,

$$\mathcal{S}^T_{n'} = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \forall i \in \mathcal{N}' : \mathsf{o}_i(t) \in \omega^T\},$$

where $\mathcal{N}' \subseteq \mathcal{N}$ is a subset of all agents with size $|\mathcal{N}'| = n'$. Following the same definition for $\mathcal{S}^T_{n'}$, the subset $\mathcal{S}^T_n$ is equivalent to the set of all terminal states where all agents are at the goal location. At time $t = 1$, the initial position of all agents is randomly and uniformly selected amongst the non-goal states, i.e., for each agent $i \in \mathcal{N}$ the initial position of the agent is $\mathsf{o}_i(1) \in \Omega - \{\omega^T\}$.

At any time step $t = 1, 2, ...$ each agent i observes its position, or environment state, and acquires information about the position of the other agents by receiving a communication message vector $\mathsf{c}_{-i}(t)$ sent by the other agents $j \in \mathcal{N}_{-i}$ at the time step t. Based on this information, agent i selects its environment action $\mathsf{m}_i(t)$ from the set $\mathcal{M} = \{\text{Right}, \text{Left}, \text{Up}, \text{Down}, \text{Stop}\}$, where an action $\mathsf{m}_i(t) \in \mathcal{M}$ represent the horizontal/vertical move of agent i on the grid at time step t. For instance, if an agent i is on a grid-world as depicted on Fig. 5.5 (a), and observes $\mathsf{o}_i(t) = 4$ and selects "Up" as its action, the agent's observation at the next time step will be $\mathsf{o}_i(t+1) = 8$. If the position to which the agent should be moved is outside the grid, the environment is assumed to keep the agent in its current position. We assume that all these deterministic state transitions are captured by $T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t))$, which can determine the observations of

agents in the next time step $t+1$ following

$$\langle \mathsf{o}_1(t+1), ..., \mathsf{o}_n(t+1)\rangle = T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)).$$

Accordingly, given observations $\langle \mathsf{o}_i(t+1), ..., \mathsf{o}_n(t+1)\rangle$ and actions $\langle \mathsf{m}_1(t+1), ..., \mathsf{m}_n(t+1)\rangle$, all agents receive a single team reward

$$r(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) = \begin{cases} C_1, & \text{if } P_1 \\ C_2, & \text{if } P_2, \\ 0, & \text{otherwise,} \end{cases} \tag{5.11}$$

where $C_1 < C_2$ and the propositions P_1 and P_2 are defined as $P_1 : T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) \in \mathcal{S}^T - \mathcal{S}_n^T$ and $P_2 : T(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)) \in \mathcal{S}_n^T$. When only a subset $\mathcal{N}'$, $|\mathcal{N}'| = n' < n$ of agent arrives at the target point ω^T, the episode will be terminated with the smaller reward C_1 being obtained, while the larger reward C_2 is attained only when all agents visit the goal point at the same time. Note that this reward signal encourages coordination between agents which in turn can benefit from inter-agent communications.

Furthermore, at each time step t agents choose a communication message to send to the other agent by selecting a communication action $\mathsf{c}_i(t) \in \mathcal{C} = \{0,1\}^R$ of R bits, where R (bits per channel use / per time step) is the fixed bit-budget of all inter-agent communication channels. The goal of the MAS is to maximize the average return by solving the D-JCCD problem (5.3).

5.6.2 Results

ESAIC, SAIC, and centralized schemes are compared by their average return in Fig. 5.6 with the ESAIC curve represented as a dotted red line, the SAIC, introduced by [9], curve as a solid blue line, and the centralized curve as a solid black line. The figure is intended to show the applicability of the ESAIC scheme in more complex geometric consensus environments. The size of the grid world for this figure is 8×8, and the multi-agent system is composed of three agents. The figure demonstrates that the performance of ESAIC closely follows that of SAIC, with almost similar average return performance as well as the speed of convergence. The centralized scheme, which is represented by the solid black curve, achieves

optimal performance but requires virtually twice the time required for the convergence of ESAIC and SAIC. Fig. 5.6 suggests that ESAIC is a promising approach for achieving high average return performance in complex MASs, with similar performance to SAIC and faster convergence time than the centralized scheme.

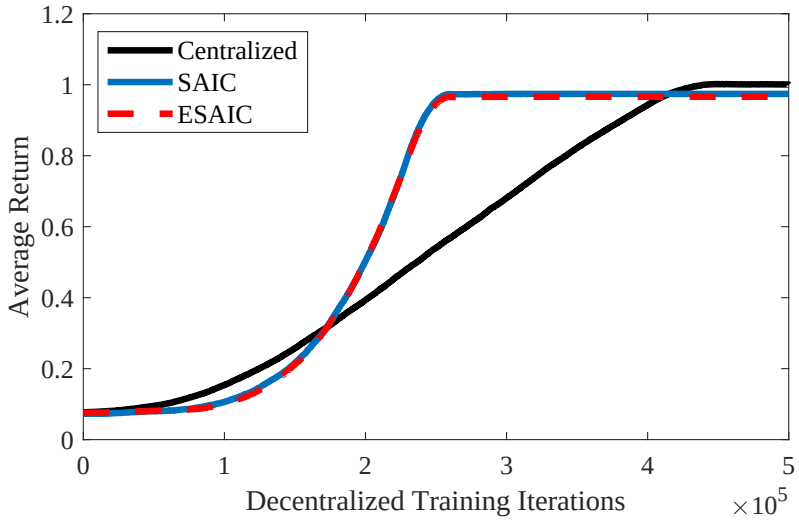


Figure 5.6: Comparison of the obtained average return via SAIC and ESAIC in MAS in the decentralized training phase while the condition c_1 in (5.9) is violated.

Figure 5.6 was comparing the average return performance of ESAIC against SAIC for a three-agent system. The following figure, Figure 5.7, presents a similar comparison for multi-agent systems with a variable number of agents. The figure shows that ESAIC achieves an average return performance that is similar to SAIC, while also offering a remarkable reduction in computational complexity. Due to its extravagant computational complexity, SAIC could not be evaluated for multi-agent systems composed of more than 4 agents. Given the exponential increase in the complexity of SAIC with respect to the number of agents, to be able to study ints performance for a 4-agent system, this figure has been plotted for the grid worlds of smaller size i.e., 3×3 across all schemes.

As discussed earlier in sec. 5.5, SAIC suffers from prohibitively high computational complexity in its centralized training phase. ESAIC is introduced in this chapter to tackle the issue of complexity in the centralized training phase by designing the communication policies only according to a two-agent centralized training. Figure 5.8 compares the time required for the implementation of the centralized training phase in both schemes SAIC and ESAIC - both theoretically and analytically. Similar to Fig. 5.7, this figure as well as the next one have been plotted for the grid worlds of smaller size i.e., 3×3 across all schemes. The

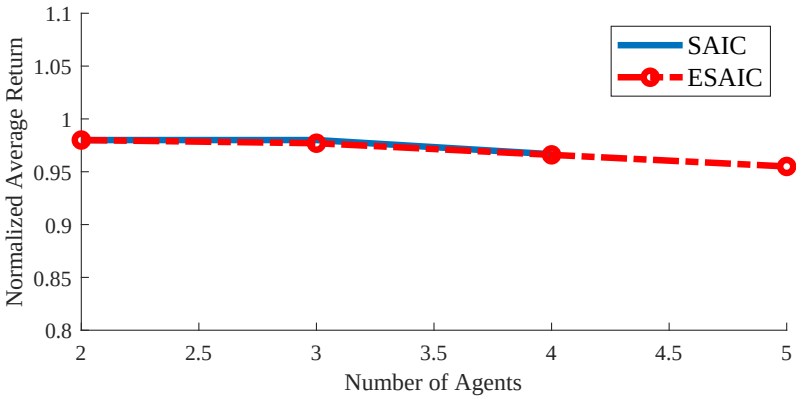


Figure 5.7: Comparison of the obtained average return via SAIC and ESAIC in MAS with varying numbers of agents.

analytical results reflect the explanations provided at 5.5.2.

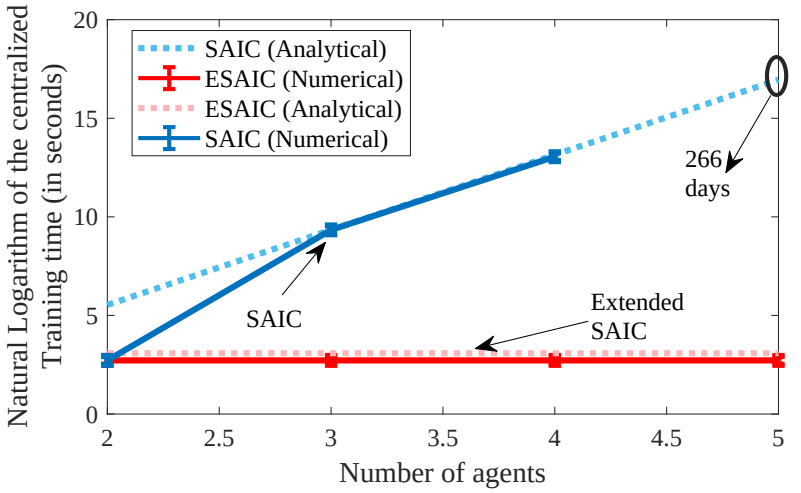


Figure 5.8: Comparison of the average time required to carry out the centralized training phase in both algorithms SAIC and ESAIC.

To realize the end-to-end time required for the training of both algorithms, Fig. 5.9 is brought. This figure illustrates the combined time required to carry out the centralized as well as the decentralized training phase. Inter-agent communications are considered to be $R_{i,j} = 2$ (bits per channel use) across all agents $\forall i, j \in \mathcal{N}$. With an increase in the number of agents, the size of the received communication message space $\mathcal{C}^{n-1}$ increases exponentially leading to an increase in the end-to-end complexity of both algorithms SAIC and ESAIC. Nevertheless, the goal of solving the problems (5.6) and (5.7) is to significantly reduce the size of each agent's communication transmission space $\mathcal{C}$ compared with the observation space Ω. Accordingly, the exponential increase in the size of received communication space has a much

less pronounced impact on the overall complexity of both algorithms. Yet, we expect the size of the received message space $\mathcal{C}^{n-1}$ to be another bottleneck of SAIC that ESAIC can not solve. This bottleneck gets more serious when the number of agents goes double digits. The analytical results reflect the explanations provided at 5.5.2.

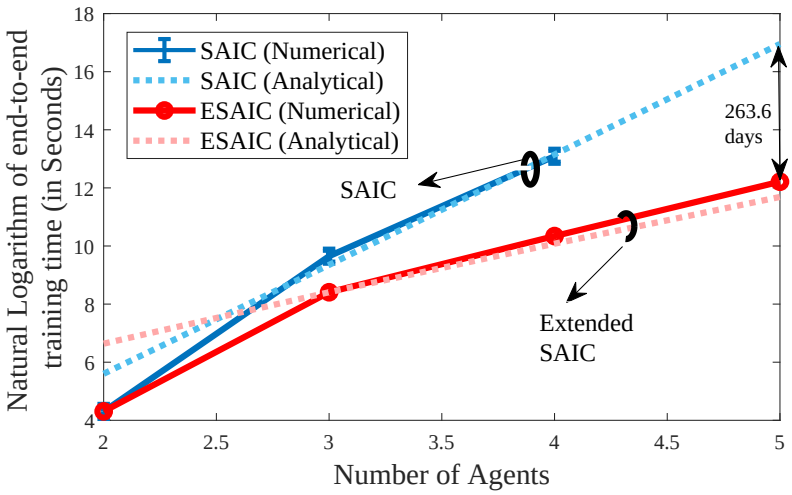


Figure 5.9: Comparison of the average time required to carry out end-to-end training in both algorithms SAIC and ESAIC.

To show that both SAIC and ESAIC can perform well even under heterogeneous bit-budgets, Fig. 5.10 is obtained. This figure studies the average return performance of ESAIC - which is equivalent to that of SAIC for a two-agent system - in an 8×8 grid world with heterogeneous bit-budgets for agents. We observe near-optimal performance for both schemes for all heterogeneous rates $R_{i,j} > 2$.

5.7 Conclusion

In conclusion, this chapter has presented a novel scalable task-oriented quantization algorithm for multi-agent communications over bit-budgeted channels. The proposed algorithm, ESAIC, offers a unique approach to designing the communications of a multi-agent system, regardless of the number of agents involved. The two-agent centralized training phase used in the algorithm has been shown to be effective in designing abstract communications and obtaining near-optimal average returns. We have also demonstrated that any approximate/deep reinforcement learning scheme can be used in the centralized training phase without affecting the reported results - making the results of this chapter more applicable to real-world

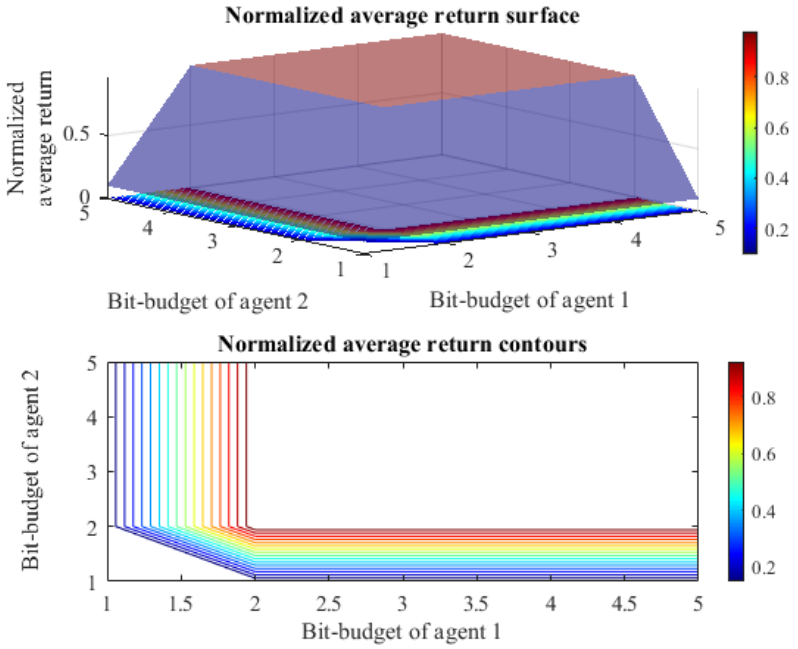


Figure 5.10: Normalized average return of a two-agent system when ESAIC is applied under heterogeneous bit-budgets.

scenarios.

The results of our analytical analysis are strong evidence for the effectiveness of the proposed algorithm. The main theorem proposed by the chapter offers a solid foundation for future research as we expect the condition of the theorem to be further relaxed in future works, leading to even more efficient and effective communication designs for large multi-agent systems. We believe that the proposed algorithm, ESAIC, has significant implications for the design of communication systems in multi-agent systems, with potential applications in areas such as autonomous vehicles, robotics, and wireless sensor networks.

5.8 Proof of Theorem 17

To prove this theorem we first introduce a lemma together with its proof in subsection 5.8.1. We then use the result of this lemma in subsection 5.8.2 to see how one can design the partition $\mathcal{P}_{i,j}^{[N]}$, after a two-agent centralized training but given the help of an oracle that knows a specific function $f(\cdot)$. Subsequently, we complete the proof of Theorem 17, in subsection 5.8.3 leveraging the above-mentioned lemmas with no further need to the knowledge of the function $f(\cdot)$.

5.8.1 An Instrumental Lemma

Lemma 19. *Let $\mathcal{A} \subset \mathbb{R}$ and $\mathcal{B} \subset \mathbb{R}$ be discrete sets, $h(\cdot) : \mathbb{R} \to \mathbb{R}$ be a bijection, $\ddot{h}(\mathcal{A})$ be a discrete set and $c \in \mathbb{R}$ be a constant value. We can state*

$$\ddot{h}(\mathcal{A}) = \mathcal{B} \implies \sum_{a' \in h(\mathcal{A})} |a' - c| = \sum_{b \in \mathcal{B}} |b - c|. \tag{5.12}$$

Proof.

$$\forall\, b \in \mathcal{B}\ \exists a \in \mathcal{A} : h(a) = b, \tag{5.13}$$

where by adding the constant value $-c$ to the side of equality and applying the absolute value will result in

$$\forall\, b \in \mathcal{B}\ \exists a \in \mathcal{A} : |h(a) - c| = |b - c|. \tag{5.14}$$

Which is equivalent to

$$\forall\, b \in \mathcal{B}\ \exists a' \in \ddot{h}(\mathcal{A}) : a' - c = b - c, \tag{5.15}$$

according to the definition of $\ddot{h}(\mathcal{A}) \triangleq \{a' : h^{-1}(a') \in \mathcal{A}\}$. Performing a summation across all the elements of $\ddot{h}(\mathcal{A})$ and $\mathcal{B}$ will result in

$$\sum_{a' \in \ddot{h}(\mathcal{A})} |a' - c| = \sum_{b \in \mathcal{B}} |b - c|. \tag{5.16}$$

■

5.8.2 If an oracle tells us $f(\cdot)$

Lemma 20. *Let the bijection $f(\cdot) : \mathcal{V}^{[2]} \to \mathcal{V}^{[N]}$ be the mapping from the value of observations for a two-agent scenario to the N-agent. For all $i, j \in \mathcal{N}$, the partition $\mathcal{P}^{[2]}_{i,j}$ proposed by ESAIC are the same as the partition $\mathcal{P}^{[N]}_{i,j}$ proposed by SAIC if the function $f(\cdot)$ is known*

and if

$$c_1: \quad \forall k \in \{1,\dots,B_{i,j}\}\ \exists k' \in \{1,\dots,B_{i,j}\}: \quad \ddot{f}(\mathcal{P}'^{[2]}_{i,j,k}) = \mathcal{P}'^{[N]}_{i,j,k'} \tag{5.17}$$

Proof. We start by a clustering problem over the space of values $\mathcal{V}^{[2]}$ that is obtained by ESAIC and we show the problem as well as its solution are equivalent to the problem that is solved by SAIC.

Given the help of an oracle, we know the function $f(\cdot)$ and thus, after obtaining the values $V^{*[2]}(\mathsf{o}) = v^{[2]} \in \mathcal{V}^{[2]}$ by solving the centralized two agent problem, we proceed by obtaining the solution for

$$\underset{P'}{\operatorname{argmin}} \quad \sum_{k'=1}^{2^R} \sum_{v^{[2]} \in \mathcal{P}'^{[2]}_{i,j,k'}} |f(v^{[2]}) - \mu_{k'}|. \tag{5.18}$$

We also know from (5.17) and lemma 19 that for any $k' \in \{1,...,2^R\}$ there is a $k \in \{1,...,2^R\}$ such that

$$\sum_{v^{[2]} \in \mathcal{P}'^{[2]}_{i,j,k'}} |f(v^{[2]}) - \mu_{k'}| = \sum_{v^{[N]} \in \mathcal{P}'^{[N]}_{i,j,k}} |v^{[N]} - \mu_k|. \tag{5.19}$$

By replacing the right-hand term in equality (5.19) with the inner summation of eq. (5.18), we will arrive at

$$\underset{P}{\operatorname{argmin}} \quad \sum_{k=1}^{2^R} \sum_{v^{[N]} \in \mathcal{P}'^{[N]}_{i,j,k}} |v^{[N]} - \mu_k|, \tag{5.20}$$

and since problem (5.24) is the exact problem that is solved by SAIC, the proof is concluded.

■

5.8.3 Proof of theorem 17

Proof. As we assume that we have no knowledge about the function $f(\cdot)$, after obtaining the optimal value function $V^{*[2]}(\cdot)$, we will solve the clustering problem as if we are designing a

communication policy for a two-agent system by SAIC. Accordingly we will have to solve

$$\underset{P'}{\operatorname{argmin}} \quad \sum_{k'=1}^{2^R} \sum_{v^{[2]} \in \mathcal{P}'^{[2]}_{i,j,k'}} |v^{[2]} - \mu_{k'}|. \tag{5.21}$$

Given eq. (5.19) and lemma 19, we know that for any $k' \in \{1, ..., 2^R\}$ there is a $k \in \{1, ..., 2^R\}$ such that

$$\sum_{v^{[2]} \in \mathcal{P}'^{[2]}_{i,j,k'}} |f^{-1}(f(v^{[2]})) - \mu_{k'}| = \sum_{v^{[N]} \in \mathcal{P}'^{[N]}_{i,j,k}} |f^{-1}(v^{[N]}) - \mu_k|. \tag{5.22}$$

Be reminded that the inner summation of eq. (5.21) is equal to the left-hand term in equality (5.22). By replacing the right-hand term in equality (5.22) with the inner summation of eq. (5.21), we will arrive at

$$\underset{P}{\operatorname{argmin}} \quad \sum_{k=1}^{2^R} \sum_{v^{[N]} \in \mathcal{P}^{[N]}_{i,j,k}} |f^{-1}(v^{[N]}) - \mu_k|. \tag{5.23}$$

The inner summation of eq. (5.23) can also be taken over the observation space as the following

$$\underset{P}{\operatorname{argmin}} \quad \sum_{k=1}^{2^R} \sum_{\mathsf{o} \in \mathcal{P}^{[N]}_{i,j,k}} |f^{-1}(V^{*[N]}(\mathsf{o})) - \mu_k|, \tag{5.24}$$

where by applying the function $f(\cdot)$, according to the lemma 19, we will get

$$\underset{P}{\operatorname{argmin}} \quad \sum_{k'=1}^{2^R} \sum_{\mathsf{o} \in \mathcal{P}^{[2]}_{i,j,k'}} |V^{*[2]}(\mathsf{o}) - \mu_{k'}|. \tag{5.25}$$

As shown in Fig. 5.11, from this point onward using a set of known relationships we will try to find the relationship between $\mathcal{P}^{[2]}_{i,j,k'}$ and $\mathcal{P}^{[N]}_{i,j,k}$. It will be demonstrated that $\mathcal{P}^{[2]}_{i,j,k'} = \mathcal{P}^{[N]}_{i,j,k}$ where k is the same index that allows the equality $\ddot{f}(\mathcal{P}'^{[2]}_{i,j,k'}) = \mathcal{P}'^{[N]}_{i,j,k}$ to hold. Then by running the inner summation of eq. 5.23 over all the elements of the set $\mathcal{P}'^{[N]}_{i,j,k}$ - instead of $\mathcal{P}'^{[2]}_{i,j,k'}$ - the proof will be concluded.

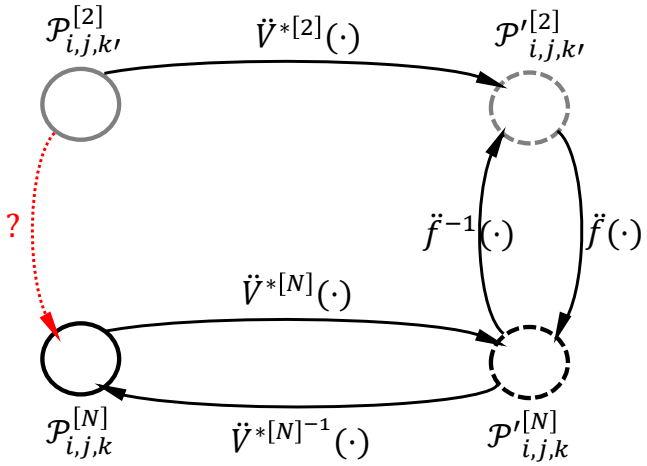


Figure 5.11: Known and unknown relationships between different partitions of observation and value space. We will show how one can get from the partition $\mathcal{P}^{[2]}_{i,j}$ on the observation space to the $\mathcal{P}^{[N]}_{i,j}$ in a few steps.

We know that the the set $\mathcal{P}^{[2]}_{i,j,k'}$ corresponds to $\mathcal{P}'^{[2]}_{i,j,k'}$, i.e.,

$$\ddot{V}^{*\,[2]}(\mathcal{P}^{[2]}_{i,j,k'}) = \mathcal{P}'^{[2]}_{i,j,k'}. \tag{5.26}$$

Despite being a surjection, we define the *inverse image* of $V^{*\,[2]}(\cdot)$ to be

$$[V^{*\,[2]}]^{-1}(v^{[2]}) \triangleq \{\mathsf{o} \,|\, V^{*\,[2]}(\mathsf{o}) = v^{[2]}\}, \tag{5.27}$$

such that

$$[\ddot{V}^{*\,[2]}]^{-1}(\mathcal{P}'^{[2]}_{i,j,k'}) = \mathcal{P}^{[2]}_{i,j,k'}. \tag{5.28}$$

Similar to (5.27) we define the inverse of $V^{*\,[N]}(\cdot)$ to be

$$[V^{*\,[N]}]^{-1}(v^{[N]}) = \{\mathsf{o} \,|\, V^{*\,[N]}(\mathsf{o}) = v^{[N]}\}, \tag{5.29}$$

such that

$$[\ddot{V}^{*\,[N]}]^{-1}(\mathcal{P}'^{[N]}_{i,j,k}) = \mathcal{P}^{[N]}_{i,j,k}. \tag{5.30}$$

To show the equivalnce of $\mathcal{P}^{[2]}_{i,j,k'} = \mathcal{P}^{[N]}_{i,j,k}$, we will show that they both correspond to the same value cluster. Further to the (5.17), we know that there exit a k such that $\ddot{f}(\mathcal{P}'^{[2]}_{i,j,k'}) = \mathcal{P}'^{[N]}_{i,j,k}$.

This together with the eq. (5.30) and (5.26) implies that

$$[\ddot{V}^{*\,[N]}]^{-1}\Big(\ddot{f}(\ddot{V}^{*\,[2]}(\mathcal{P}^{[2]}_{i,j,k'}))\Big) = \mathcal{P}^{[N]_{i,j,k}}. \tag{5.31}$$

In the following we will show that $[\ddot{V}^{*\,[N]}]^{-1}\Big(\ddot{f}(\ddot{V}^{*\,[2]}(\cdot))\Big)$ is an identity image function; i.e.,

$$[\ddot{V}^{*\,[N]}]^{-1}\Big(\ddot{f}(\ddot{V}^{*\,[2]}(\Omega' \subset \Omega))\Big) = \Omega'. \tag{5.32}$$

Since we know from the assumptions of lemma 20 that $V^{*\,[N]}(\mathsf{o}) = f(V^{*\,[2]}(\mathsf{o}))\ \ \forall o \in \Omega$, it could be trivial to show the correctness of ${V^{*\,[N]}}^{-1}\Big(f(V^{*\,[2]}(\mathsf{o}))\Big) = \mathsf{o}\ \ \forall o \in \Omega$ or (5.32); but it is not the case where the functions involved are not bijections and when instead of functions we have image functions. Yet, to prove (5.32), it is sufficient to prove that for all $g(\cdot) : \mathcal{E} \to \mathcal{F}$ and $\mathcal{E}' \subset \mathcal{E}$ we can state that

$$\ddot{g}^{-1}\Big(\ddot{g}(\mathcal{E}')\Big) = \mathcal{E}'. \tag{5.33}$$

In the following, we prove the above-mentioned - while a similar statement is can be found in [298](page 17) without proof given that $[V^{*\,[N]}]^{-1}(\cdot)$ is a surjective mapping. According to the definition of image functions, for every image function $g(\cdot) : \mathbb{P}(\mathcal{E}) \to \mathbb{P}(\mathcal{F})$ and $\mathcal{E}' \subset \mathcal{E}$ we have

$$\epsilon'' \in \ddot{g}^{-1}(\mathcal{E}'') \iff \ddot{g}(\epsilon'') \in \mathcal{E}''. \tag{5.34}$$

To show the equality of the set $\ddot{g}^{-1}\Big(\ddot{g}(\mathcal{E}')\Big)$ and $\mathcal{E}'$ we now have to prove that

$$\epsilon' \in \mathcal{E}' \qquad \iff \epsilon' \in \ddot{g}^{-1}\Big(\ddot{g}(\mathcal{E}')\Big) \tag{5.35}$$

Consider eq. (5.34) and replace the $\ddot{g}(\mathcal{E}')$ in eq. (5.35) with $\mathcal{E}''$ of eq. (5.34)

$$\iff \ddot{g}(\epsilon') \in \ddot{g}(\mathcal{E}') \tag{5.36}$$

$$\iff \epsilon' \in \mathcal{E}'. \tag{5.37}$$

■

5.9 Proof of Lemma 18

Without loss of generality, instead of proving Lemma 3 for strictly monotonic functions, we only prove it for the strictly increasing functions. The strictly decreasing functions also can be treated in a similar manner. Then, Lemma 3 for strictly monotonic functions $f^{[k]}(\cdot)$ is automatically deduced. So, we consider the case $f^{[3]}(\cdot)$ is strictly increasing and we prove $f^{[N]}(\cdot)$, $N > 3$ is also strictly increasing. This is done via induction.

Proof. **Base case:** We assume that $P(3)$ has been verified.

Induction step: We will show that for every $k > 3$, if $P(k)$ holds, then $P(k+1)$ also holds. Equivalently,

$$P(k): \quad \text{If } V^{[2]}(\mathsf{o}') > V^{[2]}(\mathsf{o}) \implies \tag{5.38}$$

$$f^{[k]}(V^{[2]}(\mathsf{o}')) > f^{[k]}(V^{[2]}(\mathsf{o})), \ \forall \mathsf{o} \in \Omega, \tag{5.39}$$

which can also be expressed as

$$P(k): \quad \text{If } V^{[2]}(\mathsf{o}') > V^{[2]}(\mathsf{o}) \implies V^{[k]}(\mathsf{o}') > V^{[k]}(\mathsf{o}), \ \forall \mathsf{o} \in \Omega. \tag{5.40}$$

To prove the statement $P(k+1)$, we represent $V^{[k+1]}(\cdot)$ in terms of $V^{[k]}(\cdot)$ via eq. (5.41) - (5.45), and we will use this relationship to deduce the induction step given (5.40) holding.

The value of observation $\mathsf{o} \in \Omega$ in an k agent system is expressed by

$$V^{[k]}(\mathsf{o}_i = \mathsf{o}) = \sum_{\mathsf{o}_{-i} \in \Omega^{k-1}} p(\mathsf{o}_{-i} = [\mathsf{o}_j]_{j \in \mathcal{N}_{-i}}) \mathbb{E}\{\mathbf{g}^{[k]} | \mathsf{o}_i, \mathsf{o}_{-i}\}, \tag{5.41}$$

where the set of agents $\mathcal{N} = \{1, 2, ..., k\}$ is comprised of k elements and $\mathbf{g}^{[n]}(t') = \sum_{t=t'}^{T'} \gamma^{t-1} r^{[n]}(\mathbf{s}(t), \mathbf{m}(t))$. Similarly, the value of observation $\mathsf{o} \in \Omega$ in an $k+1$ agent system is expressed by

$$V^{[k+1]}(\mathsf{o}_i = \mathsf{o}) = \sum_{\mathsf{o}_{-i} \in \Omega^{k}} p(\mathsf{o}_{-i} = [\mathsf{o}_j]_{j \in \mathcal{N}'_{-i}}) \mathbb{E}\{\mathbf{g}^{[k+1]} | \mathsf{o}_i, \mathsf{o}_{-i}\}, \tag{5.42}$$

where the set of agents $\mathcal{N}' = \{1, 2, ..., k+1\}$ is comprised of $k+1$ elements. Taking condition c_2 into account, it can be readily shown that

$$V^{[k+1]}(\mathsf{o}_i = \mathsf{o}) = \sum_{\mathsf{o}_{-i} \in \Omega^k} p(\mathsf{o}_{-i} = [\mathsf{o}_j]_{j \in \mathcal{N}'_{-i}}) \mathbb{E}\{\tau \mathbf{g}^{[k]} + \zeta | \mathsf{o}_i, \mathsf{o}_{-i}\}. \tag{5.43}$$

By expanding the summation in (5.43) and considering c_3, one can obtain that

$$\begin{aligned} &V^{[k+1]}(\mathsf{o}_i = \mathsf{o}) = \\ &\sum_{\mathsf{o}_{k+1} \in \Omega} p(\mathsf{o}_{k+1}) \sum_{\mathsf{o}_{-i} \in \Omega^{k-1}} p(\mathsf{o}_{-i} = [\mathsf{o}_j]_{j \in \mathcal{N}_{-i}}) \mathbb{E}\{\tau \mathbf{g}^{[k]} + \zeta | \mathsf{o}_i, \mathsf{o}_{-i}, \mathsf{o}_{k+1}\}. \end{aligned} \tag{5.44}$$

By applying the law of iterated expectations on eq. (5.44) we can express $V^{[k+1]}(\mathsf{o}_i = \mathsf{o})$ as

$$V^{[k+1]}(\mathsf{o}_i = \mathsf{o}) = \sum_{\mathsf{o}_{-i} \in \Omega^{k-1}} p(\mathsf{o}_{-i} = [\mathsf{o}_j]_{j \in \mathcal{N}_{-i}}) \mathbb{E}\{\tau \mathbf{g}^{[k]} + \zeta | \mathsf{o}_i, \mathsf{o}_{-i}\},$$

which allows us to directly imply

$$V^{[k+1]}(\mathsf{o}_i = \mathsf{o}) = \tau V^{[k]}(\mathsf{o}_i = \mathsf{o}) + \zeta. \tag{5.45}$$

Upon the correctness of the statement $P(k)$, and for the constant values of $\tau, \zeta \in \mathbb{R}$ one can also state that

$$\begin{aligned} &\text{If } V^{[2]}(\mathsf{o}') > V^{[2]}(\mathsf{o}) \implies \\ &\tau V^{[k]}(\mathsf{o}') + \zeta > \tau V^{[k]}(\mathsf{o}) + \zeta, \ \forall \mathsf{o} \in \Omega. \end{aligned} \tag{5.46}$$

Eq. (5.45) together with eq. (5.46) are sufficient to establish the proof of the induction step, which in turn concludes the proof. ■

Chapter 6

Conclusion

We saw earlier how the booming number of data-hungry services is the driving force for research and innovation in multiple domains: (i) task-effective communication of data on a large scale while relying on a limited overall network capacity and (ii) reducing the complexity of the learning/computing systems that demand large sizes of data. These initial motivations are giving birth to the field of task-oriented communications [286] and data-centric AI [299]. As these underlying reasons are driving research towards task-oriented communication systems, a new disruptive technology will soon be available that can have unforeseen implications. On top of the two above-mentioned motivating forces, we believe that the emergence of task-oriented communication systems can also have a direct impact on the (iii) architectural redesign of CPUs - to optimize the flow of useful data, (iv) architectural design of neural networks - to find theoretical means for pruning neural networks with the aim of reducing their computational complexity, and (v) to change the legal meaning of data privacy and make it dependant to the existing technologies, (vi) rethinking the internal signallings and information exchange at autonomous decision-making systems.

6.1 Application Scenarios

Below we will provide some further application scenarios in addition to the ones introduced in chapter 2.

6.1.1 Platooning of vehicles and UAVs

New research advancements in vehicle-to-vehicle communications (V2V) and vehicle-to-infrastructure communications (V2I) are intended to create cooperative transportation systems that are more effective in different ways such as safety, fuel economy and traffic flow and users' comfort [300]. One particular application of these technologies is platooning which is defined as a group of moving vehicles that are actively coordinated in formation and travel together. Each car cooperatively participates in the platoon as an autonomous decision-making system. The sensory inputs of each vehicle in the system are mostly limited to the measurement of the movements of the neighbouring cars, i.e. there is no "look ahead". Given the local observability of each agent in the system, to control the whole platoon in a desired fashion and to avoid lateral and longitudinal instability in the platoon, it is necessary to use V2V communications [301].

Motivation of Using Task-Oriented Communication Design in This Context

Using task-effective V2V communications in the context of car platooning is particularly important due to the central importance of safety in transportation systems. The safety of autonomous decision-making systems is directly impacted by their agility in responding to changes in the dynamic environment [302, 303]. Note that in our framework, as well as in this application area, each agent understands the environment (and its changes) through its sensory system and the communications it receives (3.11). With the increased size of communications, however, the delay in the communications and processing time of the received data rises significantly - leading to less agility of the system when responding to the dynamics of the environment. In this direction, by reducing the size of inter-agent communications, V2V communications, SAIC and ESAIC can make a car platoon safer and more agile. The only caveat is that by the addition of every vehicle to the system, there will be another increase in the total size of received communications at each vehicle, leading to an exponential increase in the size of the learning task both in the distributed and centralized training phases 5.5.2. ESAIC can be used here to help design task-oriented V2V communications at scale. While SAIC and ESAIC are useful to design the communications and control policies of the vehicles in a decentralized fashion, ABSA can be used in this setting for similar reasons but for V2I scenarios. An example arises when intra-platoon communications are mediated via a hub

node e.g., a UAV [302].

More Technical Details on the Platooning of vehicles and UAVs

Similar to the framework of this thesis, communication is used here as a means to tackle data deficiency for the decision-making of every agent. Moreover, the joint observability assumption (3.1) holds here making it reasonable to use SAIC and ESAIC for V2V communications in this context. Moreover, the numerical results that are reported for SAIC, ESAIC and ABSA are obtained after applying these schemes to geometrical consensus problems. It is shown in [208] that car platooning is yet another special case of the geometrical consensus problems. In fact, as explained in section 3.5, the platooning problem can be simulated via the existing setting of the numerical experiments with the caveat that the reward function must also take agents' speed into account.

6.1.2 Privacy preserving recommender systems

A privacy-preserving recommender system is a type of recommendation system that protects users' privacy by limiting the amount of personal data that is collected and processed. In traditional recommender systems, the recommendation algorithm relies on user data such as items previously clicked, purchased or rated to generate recommendations for that user. However, this data can reveal sensitive information about the user's preferences and behaviour to malicious parties, leading to concerns about data privacy and security [304].

Privacy-preserving recommender systems use techniques such as differential privacy [305], secure multi-party computation, federated learning [306], and homomorphic encryption to generate recommendations without compromising privacy [307]. These systems allow users to receive personalized recommendations without divulging their data to a central server or exposing their sensitive preferences and behaviour to third-party providers. According to the European data protection rules, for each processing activity in scope, the digital service has to implement measures that ensure that the collection of personal data is adequate, relevant and strictly, limited to what is necessary in relation to the purposes for which they are processed. Accordingly, the anonymization of data is not sufficient to adhere to the strict rules of data protection.

Privacy-preserving recommender systems have become essential in online social communities where user data is often sensitive and personal. These systems ensure that users receive personalized recommendations while protecting their privacy and without compromising their sensitive information. These concerns are even more pronounced in the European Union countries with the General Data Protection Regulation (GDPR) coming into effect from 2018. The GDPR is considered to be the toughest privacy and security law in the world, and it regulates how the personal data of individuals in the EU may be processed and transferred. While data privacy is a legal concept according to GDPR, we believe that this meaning is being constantly redefined by developments happening in the area of task-oriented data compression.

Motivation of Using Task-Oriented Communication Design in This Context

The legal definition of GDPR limits the collection of data by digital services, including recommendation systems, to "Adequate, relevant and limited to what is necessary in relation to the purposes" referred to as "data-minimization practices". According to GDPR, for each processing activity in scope, the digital service has to implement measures that ensure that the collection of personal data is adequate, relevant and strictly, limited to what is necessary in relation to the purposes for which they are processed. In particular, the entity has assessed that it cannot achieve the purpose of its processing activity with less (privacy invasive) data (e.g. working with less granular data) Ref: (GDPR Articles 5 and 25) (Recitals 29, 78, 116, 123).

In this context, the design of an average return maximizing recommender system that runs under constraint user data acquisitions can be considered as a joint communication and computation problem [19]. The communication part of the problem is associated with the acquisition of data from users. The rate of communicated data from users to the recommender system has to be limited to give further privacy to the users and the communication network topology is a star topology similar to the scenario investigated in section 4.1.1. The recommender system is considered to be a decision-making system, that has a deficiency of data for a recommendation to a certain user. This deficiency is addressed by acquiring further data from that certain user and others.

While a recommender system is not, in general, a CPS, the computations carried out by

recommender systems are likely to fall under the category of control tasks - being addressed by this thesis. According to the Fig. 1.1, as well as the details provided in chapter 4, in control tasks, there is an influence from the side of actions taken by the MAS on the future observations of the system, whereas such influence is absent in the estimation tasks. Take online advertising systems, which are a subclass of recommender systems, as an example [308]. When operating an online advertisement algorithm using reinforcement learning, it is widely known that deep reinforcement learning algorithms can suffer from highly sub-optimal trade-offs on exploration vs. exploitation [309, 310]. That is to say, most reinforcement learning-based advertising algorithms focus on optimizing the revenue of the system while it might cause a negative impact on the user experience and deteriorate the long-term revenue of the advertisement system [310]. In fact, a well-known observation is that these algorithms may tend to show advertisements too frequently to the user which in turn decreases user satisfaction and impacts the behaviour of the user with the system. Because of all the changes, that different recommendation strategies may cause in user behaviour, the actions taken by recommender systems have the influence, demonstrated on 1.1.b, on the observations of the system and, thus, fall within the scope of this thesis.

Following the data minimization guidelines, the amount of personal data that can be granted to each digital service should be limited to what is deemed necessary. In fact, the necessary amount of data required for a certain computational task changes according to the used technologies/algorithms for task-effective data compression. The fact that data only a small part of users' data can have a major influence on the recommender's output [305] is the reason that task-effective data compression can be a suitable candidate for this application. Following the framework of this thesis e.g., the method proposed in chapter 4, one can optimize the performance of the recommender system while minimizing the amount of data requested from each user.

6.1.3 Collaborative perception

Collaborative perception is an emerging research area in the field of autonomous systems that aims to improve the perception capabilities of individual devices by enabling them to share data and collaboratively process it [280, 311]. The main idea behind collaborative perception is to leverage the collective intelligence of a network of devices to improve the accuracy and

reliability of perception. For example, for object detection tasks, more true positive objects and fewer false positives are expected to be exported through collaboration [312].

Collaborative perception also poses several technical challenges, particularly in the communication of data. Data produced by each radar/Lidar sensor is several hundreds of Megabytes per second. It is, therefore, challenging to develop efficient communication protocols that enable devices to exchange data in real time while minimizing latency and ensuring reliable transmission. Another important challenge in collaborative perception is to develop algorithms that can effectively process and fuse large size of data from multiple devices. These algorithms must be able to handle data with sufficient accuracy and reliability and must be scalable to support large networks of devices.

Collaborative perception systems can improve the perception of individual agents about the environment by complementing the agent's local observations with the local observations made by other agents in the system. The observations of other agents in the system, however, are communicated to an agent over wireless communication channels. The perception of an agent from the environment in a CP system is explained by the (3.11). By controlling the robots/agents one can improve the collaborative perception given the limited resources of the CP system, e.g., by acquiring non-overlapping/statistically independent local observations to minimize the difference between the perceived state of the environment, shown in (3.11), and the true state.

"Wireless-in-the-Loop" is also a concept adjacent to task-effective communications. It refers to a methodology for testing and evaluating wireless communication systems where the communication channel is included within a closed-loop system, allowing for the accurate evaluation of the impact of wireless communication on system performance [313] - similar to the impact of inter-agent communications in Fig. 4.1. In the case of collaborative perception, WiL/task-effective communications enable vehicles or robots to effectively exchange their sensory data and information about their surroundings, such as the data generated by lidar, radar, or camera. By doing so, the importance of the data to be communicated between vehicles/robots will be taken into account leading to further efficiency in the usage of network resources as well as further accuracy in the delivery of the data when it is of specific importance to the task.

Motivation of using Task-Oriented Communication Design in This Context

Using our framework of task-oriented communications to design the data exchange in a CP system is of particular importance due to a number of reasons. The limited data rates of wireless channels may vary across devices and often pose limitations on sensor data sharing, due to limited wireless bandwidth and dynamic channel conditions. To address these challenges, the problem formulation proposed in section 5.2, proposes inter-agent communications over wireless communication channels with heterogeneous bit-budgets (5.3). Moreover, in particular settings such as CP systems that are realized by a fleet of autonomous vehicles, the safety of the system plays a central role. As explained in section 6.1.1, task-oriented communications can improve the safety of autonomous transportation systems by making them more agile. While SAIC and ESAIC are useful to design the communications and control policies of the vehicles in a decentralized fashion, ABSA can be used in this setting for similar reasons but for V2I scenarios.

6.2 Standardization opportunities for Collaborative Perception

The wireless communication resources in autonomous systems are not sufficient to handle the large volume of data generated by Cyber-Physical (CP) systems. In the case of Vehicular Ad Hoc Networks (VANETs), Cellular Vehicle-to-Everything (C-V2X) technology has been developed in two phases: LTE-V2X for basic safety services and NR-V2X for advanced use cases, including CP. The maximum sidelink bandwidth allocated for NR-V2X by 3GPP in Rel 16 at the sub-6 GHz frequency band is 40 MHz [314], while real-time streaming of raw sensor data could require several megabytes per second for a single link. While NR-V2X can utilize mmWave to achieve higher data rates, blockage effects due to higher frequencies need to be mitigated.

In particular, a rich presentation of each vehicle's sensory data over sidelinks is needed, where local observability of the vehicle does not allow to perform the computational/control task in need. We know there are scenarios in the literature where an agent can perform the computation/control task locally and just transmit the result of the computation/control task via a few bits [315]. However, the local observability of a vehicle might cease it from proper

fulfilment of the computation/control task e.g., in an object detection task [316]. Under these circumstances, vehicles are preferred to share richer features of their observations with each other to help overcome local observability.

6.3 Summary of the technical research findings and contributions

The contribution of each chapter was separately mentioned, however, here, we will try to narrate the contributions of this thesis from a different/higher-level perspective. Some of the research findings that are shared in this section are the result of the synergy of the combination of the technical sections that could not have been understood/realized in separation.

6.3.1 Separation of Communication and Control: Task-Oriented Communications

In chapter 3, we developed a generic framework to solve task-oriented communication problems - for a multi-agent system (MAS) with full mesh connectivity. The framework consists, of three conceptual steps: (i) **solve a centralized control design problem** to capture the important features of the control problem (ii) **solve a task-oriented communication design** to integrate the features of the control problem into a communication design problem and obtain a task-oriented communication policy, and (iii) **solve a control policy design problem** to work jointly with task-oriented communication policies with the aim of maximising the MAS's average return. The framework was proposed in chapter 3, to tackle the joint design of communications and control policies of a multi-agent system that is connected through a full mesh network i.e., a star topology.

In chapter 4, however, this framework was adopted to a new problem setting for the design of task-effective communications where agents follow a star network topology for their connectivity in chapter 4. In this direction, chapter 4 transcends the applicability of the proposed framework beyond the specific problem that was solved in 3 and provides further insights into how the framework can be used in wider terms and under a wider range of settings. Chapter 4, generalizes the application of the proposed framework yet in another way. While the first step of the framework in SAIC is carried out by quantizing the value

of observations, ABSA, uses a different approach to capture the important features of the control problem i.e., optimal centralized control policy - see e.g., eq. (4.5) or eq. (4.6) and compare them with eq. (3.13).

As was shown previously, in all chapters 3, 4, 5, this framework allows solving the joint control and communication design problem while communication and control policies are separately designed through step (ii) and (iii) of the framework respectively. This is not a separation in the design of communication and control policies in its classical sense, see e.g., [317] to learn more about the classical separation in the design of communications and control. In fact, in step (i) we acquire some prime features of the control problem and take them into account while casting and solving the task-oriented communication design in step (ii).

6.3.2 Value of observations

In chapter 3, we could quantify the value of observations. That is to say, we could find a quantitative metric to measure how valuable each observation made by agent j is for agent $i \neq j$ in its decision-making. We have not obtained this metric heuristically, but through an analytical process that starts by solving the joint communication and control design and arrives at the task-oriented communication design problem 3.12.

In particular, in subsection 3.7.4, we show that the design of the communication policy by solving the joint problem (3.10) can be approximated as a generalized data quantization problem (3.12). The interesting feature of the problem (3.12) is that it poses a form similar to [34], in which the value function is capturing the semantics of observation data.

In SAIC and ESAIC algorithms, introduced in 3 and 5 respectively, the knowledge that we obtain by solving the centralized problem is captured within the task-oriented communication design by designing a mapping for observations. This mapping is obtained by performing the first step of our framework explained in 6.3.1. The non-injective surjective mapping $V^*(\cdot) : \Omega \to \mathcal{V} \subset \mathbb{R}$, that is obtained after solving the centralized problem, would allow us to solve the communication problem over the output space of the mapping $V^*(\cdot)$ - value space $\mathcal{V}$ - rather than over the original observation space. This is imperative because of a multitude of reasons: (i) The mapping $V^*(\cdot)$ projects the high-dimensional observation points to the single-dimensional space of $\mathcal{V} \subset \mathbb{R}$, leading to a reduced the complexity for the clustering

problem, (ii) the mapping $V^*(\cdot)$ captures the features of the control task and allows us to take these features into account inside our communication design problem - that helps us to separately design communications and control policies, (iii) the clusters in the output space of the $V^*(\cdot)$ are shown to be linearly separable, (iv) last but not least, it is very intuitive to see how the mapping $V^*(\cdot)$ is an indirect/universal measure to quantify the value of each observation for any given task. Accordingly, the observation points are not clustered together based on how similar they are, but based on how similarly valuable they are for the task.

6.3.3 Computational cost of the value function

Despite the central role that the value functions play in the context of this thesis, they are non-closed form/numerical functions that are hard to compute. Computations required to obtain the value function sharply increase with the addition of each agent to the MAS. In particular, the time complexity of the centralized training phase in SAIC grows exponentially with respect to the number of agents. In ESAIC, however, increasing the number of agents in the MAS, has no impact on the computational complexity of the centralized training phase - making ESAIC more efficient than SAIC [9] and any other MARL with a central training phase [213, 276, 290].

6.3.4 Time Complexity VS. Average Return Performance

As is discussed in [278], the computational complexity of exact Q-learning - as a standard approach to solving the centralized control problem - is proportional to the size of the state-action space. Exact Q-learning is used in the centralized and distributed training phases of SAIC and ESAIC. In the centralized training phase of SAIC, the computational complexity $\mathcal{O}(|\Omega \times \mathcal{M}|^N)$ grows exponentially with the size of MAS N. Accordingly, the addition of each agent to the system multiplies the complexity of the Q-learning by $|\Omega \times \mathcal{M}|$. The complexity $\mathcal{O}(|\Omega \times \mathcal{M}|^2)$ of the centralized training phase in ESAIC with respect to the size of the MAS N, however, is constant time. That is, ESAIC will always execute at the same time (or space) regardless of the size of the MAS N. Despite its superiority in time complexity, ESAIC hardly compromises the average return performance of the MAS.

ABSA-2, on the other hand, offers a trade-off between average return performance and its time complexity by tuning the memory size of the receiver. As is displayed in Fig. 4.6, when

using ABSA-2, an increase in the size of the memory of the receiver improves the average return while incurring computational cost at the decentralized training phase. Accordingly, ABSA-2 offers some fluidity in how computationally demanding it is.

6.3.5 New KPIs to Measure Task-effectiveness of Communications

In the context of this thesis - and many other references e.g., [3, 11] - the task-effectiveness of the inter-agent communications is measured based on the average return that can be obtained when using it. Further, to measure the average return that can be obtained under the communication policies π^c, we have to design the control policy π^m that selects the control vector $\mathbf{m}(t)$ having access to only the quantized observations of the agents $\mathbf{c}(t)$. Accordingly, we cannot measure the effectiveness of the communication policy of an MAS without having a specific design for its control policy. Even after the design of the control policy of the MAS, it is challenging to understand if the suboptimal performance of the algorithm is caused by an ineffective design of the control policy or the communication policy. In fact, it is hard to disentangle the effect of the control and communication policies on the MAS's average return. Our proposed metric TRI can facilitate measuring the performance of any communication policy in isolation and without the effect of the control policy being present in the numerical values of TRI - chapter 4.

Accordingly, the importance of introducing this metric is multi-fold: (i) by using TRI as an indirect metric we can measure the effectiveness of a communication policy for any specific task; (ii) it allows us to measure the effectiveness of the communication scheme prior to the design of any control policy; (iii) it helps to design task effective communication policies in complete separation from the control policy design.

6.4 Discussion of limitations and future research avenues

The common assumptions and limitations which are in place across the thesis are three. (i) **Existence of an underlying MDP**, implying that the system state and transitions can be described by a single agent MDP, that has the same transition model and reward function as experienced by agents [106]. (ii) **Access to this underlying MDP for the sake of its simulation** at the computer to train our agent in - and not to use this information tailor our

policy designs accordingly. This assumption is almost always in place when reinforcement learning is used to train a (number of) decision-making agent(s) in a virtual environment - see e.g. [105] and the experiments within. (iii) **Joint observability of the environment** [37], which implies that the state of the underlying MDP can be obtained when combining all observations of agents - note that we never find the chance to have access to the state of the underlying MDP at each agent, since agents cannot share their observations with each other through perfect communication channels. The imperfections of the communication channels are the reason why agents cannot have access to the state of underlying MDP, despite assuming the joint observability of the environment.

In addition to the above-mentioned common assumptions, there are further assumptions in place in chapters 3 and 5. In chapter 3 and 5, we also assume the lumpability of the underlying MDP - see definition 6 and the condition to follow. While this condition is somewhat limiting, it can be further relaxed by including a memory in the receiving ends - as we did in chapter 4. Including a memory at the receiving end within the framework of SAIC/ESAIC, allows us to provide guarantees on the performance of the system while the perceived state of the environment, demonstrated by (3.11), does not follow Markov property. This can be an avenue for future research - making the results of the chapter 3, applicable to a wider set of applications.

Another assumption that is in place both in chapter 3 and 5 is the environment to be deterministic. This assumption was mainly there to guarantee the optimal result of a distributed Q-learning to solve the decentralized control problem - in step (iii) of the framework 6.3.1. Non-stationarity of the environment from the perspective of an agent in an MAS, is the main challenge here ceasing us to achieve optimal results in solving the decentralized control problem. One way to tackle the issue is again to consider placing a memory at each agent that saves the past observations and received communications of the agent. Authors of [106] have introduced distributed control design algorithms that are able to achieve nearly optimal results when agents possess memory in a similar setting. While being insightful, the results reported by [106] are not directly applicable to our setting as their framework is not directly applicable to multi-agent systems with instantaneous communications - the assumed setting for the communication between agents in this thesis.

Bibliography

[1] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, "Federated learning based audio semantic communication over wireless networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[2] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2020.

[3] J. Foerster, Y. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, 2016.

[4] A. Mostaani, O. Simeone, S. Chatzinotas, and B. Ottersten, "Learning-based physical layer communications for multiagent collaboration," in *2019 IEEE Intl. Symp. on Personal, Indoor and Mobile Radio Communications*, Sep. 2019.

[5] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4525–4540, 2019.

[6] C. E. Shannon and W. Weaver, "The mathematical theory of communication [1949]. urbana, il," 1959.

[7] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[8] R. A. Howard, "Information value theory," *IEEE Transactions on systems science and cybernetics*, vol. 2, no. 1, pp. 22–26, 1966.

[9] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Task-oriented data compression for multi-agent communications over bit-budgeted channels," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1867–1886, 2022.

[10] T. Soleymani, J. S. Baras, and S. Hirche, "Value of information in feedback control: Quantification," *IEEE Transactions on Automatic Control*, vol. 67, no. 7, pp. 3730–3737, 2021.

[11] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2590–2603, 2021.

[12] G. He, S. Cui, Y. Dai, and T. Jiang, "Learning task-oriented channel allocation for multi-agent communication," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 12 016–12 029, 2022.

[13] D. Kim, S. Moon, D. Hostallero, W. J. Kang, T. Lee, K. Son, and Y. Yi, "Learning to schedule communication in multi-agent reinforcement learning," in *Intl. Conf. on Learning Representations*, 2019.

[14] L. S. Vailshery, "Number of internet of things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030," Aug 2022. [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[15] D. Gündüz, Z. Qin, I. Estella Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Guest editorial special issue on beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 1–4, 2023.

[16] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[17] G. P. Fettweis and H. Boche, "6g: The personal tactile internet—and open questions for information theory," *IEEE BITS the Information Theory Magazine*, vol. 1, no. 1, pp. 71–82, 2021.

[18] S. K. Sharma, I. Woungang, A. Anpalagan, and S. Chatzinotas, "Toward tactile internet in beyond 5G era: Recent advances, current issues, and future directions," *IEEE Access*, vol. 8, pp. 56 948–56 991, 2020.

[19] M. A. Uusitalo, M. Ericson, B. Richerzhagen, E. U. Soykan, P. Rugeland, G. Fettweis, D. Sabella, G. Wikström, M. Boldi, M.-H. Hamon *et al.*, "Hexa-X the european 6G flagship project," in *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2021, pp. 580–585.

[20] H. Zou, C. Zhang, S. Lasaulce, L. Saludjian, and H. V. Poor, "Goal-oriented quantization: Analysis, design, and application to resource allocation," *IEEE Journal on Selected Areas in Communications*, 2022.

[21] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, 2018.

[22] L. Hu, G. Wu, Y. Xing, and F. Wang, "Things2vec: Semantic modeling in the internet of things with graph representation learning," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1939–1948, 2020.

[23] J. Cai, W. Zhong, and J. Luo, "Seminer: Side-information-based semantics miner for proprietary industrial control protocols," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22 796–22 810, 2022.

[24] A. Mostaani, T. X. Vu, S. K. Sharma, V.-D. Nguyen, Q. Liao, and S. Chatzinotas, "Task-oriented communication design in cyber-physical systems: A survey on theory and applications," *IEEE Access*, 2022.

[25] M. P. Mota, A. Valcarce, J.-M. Gorce, and J. Hoydis, "The emergence of wireless MAC protocols with multi-agent reinforcement learning," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[26] N. Shlezinger and Y. C. Eldar, "Deep task-based quantization," *Entropy*, vol. 23, no. 1, p. 104, 2021.

[27] M. A. Gutierrez-Estevez, Y. Wu, and C. Zhou, "Learning to communicate with intent: An introduction," *arXiv preprint arXiv:2211.09613*, 2022.

[28] C. Zhang, H. Zou, S. Lasaulce, W. Saad, M. Kountouris, and M. Bennis, "Goal-oriented communications for the IoT and application to data compression," *IEEE Internet of Things Magazine*, vol. 5, no. 4, pp. 58–63, 2022.

[29] N. Shlezinger and Y. C. Eldar, "Task-based quantization with application to MIMO receivers," *arXiv preprint arXiv:2002.04290*, 2020.

[30] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96–102, 2021.

[31] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, 2021, pp. 1–5.

[32] R. Carnap, Y. Bar-Hillel *et al.*, "An outline of a theory of semantic information," 1952.

[33] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, "Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 170–185, 2022.

[34] P. A. Stavrou and M. Kountouris, "A rate distortion approach to goal-oriented communication," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 590–595.

[35] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "State aggregation for multiagent communication over rate-limited channels," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–7.

[36] J. Liu, S. Shao, W. Zhang, and H. V. Poor, "An indirect rate-distortion characterization for semantic sources: General model and the case of gaussian observation," *arXiv preprint arXiv:2201.12477*, 2022.

[37] D. V. Pynadath and M. Tambe, "The communicative multiagent team decision problem: Analyzing teamwork theories and models," *Journal of Artificial Intelligence Research*, vol. 16, pp. 389–423, Jun. 2002.

[38] U. Cisco, "Cisco annual internet report (2018–2023) white paper," *Cisco: San Jose, CA, USA*, 2020.

[39] R. Dobrushin, "General formulation of shannon's basic theorems of information theory," *AMS Translations*, vol. 33, pp. 323–438, 1959.

[40] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 44–54, 1995.

[41] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity," *J. the Indian Institute of Science*, vol. 100, no. 2, pp. 435–443, 2020.

[42] W. Weaver, "Recent contributions to the mathematical theory of communication," in *The Mathematical Theory of Communication*, C. E. Shannon and W. Weaver, Eds. The University of Illinois Press, 1949, ch. 10.

[43] E. C. Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[44] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, 2021.

[45] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," 2021. [Online]. Available: https://arxiv.org/pdf/2007.11579

[46] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*. IEEE, 2021, pp. 1–5.

[47] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wire. Commun.*, vol. 29, no. 1, pp. 210–219, 2022.

[48] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gunduz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2590–2603, 2021.

[49] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, B. Soret *et al.*, "Semantic communications in networked systems: A data significance perspective," *IEEE Network*, vol. 36, no. 4, pp. 233–240, 2022.

[50] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *2011 IEEE Network Science Workshop*, 2011, pp. 110–117.

[51] M. Kalfa, M. Gok, A. Atalik, B. Tegin, T. M. Duman, and O. Arikan, "Towards goal-oriented semantic signal processing: Applications and future challenges," *Digital Signal Processing*, p. 103134, 2021.

[52] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software-defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1–18, 2015.

[53] O. Alliance, "O-RAN use cases and deployment scenarios," *White Paper, Feb*, 2020.

[54] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3gpp release 15," *IEEE Access*, vol. 7, pp. 127 639–127 651, 2019.

[55] A. Aijaz, "Private 5G: The future of industrial wireless," *IEEE Industrial Electronics Magazine*, vol. 14, no. 4, pp. 136–145, 2020.

[56] Q. Zhou, C.-X. Wang, S. McLaughlin, and X. Zhou, "Network virtualization and resource description in software-defined wireless networks," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 110–117, 2015.

[57] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

[58] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "Flexran: A flexible and programmable platform for software-defined radio access networks," in *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 427–441.

[59] A. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Transactions on Information Theory*, vol. 17, no. 5, pp. 508–513, 1971.

[60] M. Sudan, H. Tyagi, and S. Watanabe, "Communication for generating correlation: A unifying survey," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 5–37, 2019.

[61] P. W. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4181–4206, 2010.

[62] M. Mylonakis, P. A. Stavrou, and M. Skoglund, "Empirical coordination subject to a fidelity criterion," in *2019 IEEE Information Theory Workshop (ITW)*. IEEE, 2019, pp. 1–5.

[63] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—part i: Scalar systems," *IEEE transactions on Information Theory*, vol. 52, no. 8, pp. 3369–3395, 2006.

[64] N. Martins, M. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems using a stochastic digital link," in *2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601)*, vol. 2, 2004, pp. 1889–1895 Vol.2.

[65] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 595–609.

[66] R. A. Amjad and B. C. Geiger, "Learning representations for neural network-based classification using the information bottleneck principle," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[67] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *2007 IEEE International Symposium on Information Theory*. IEEE, 2007, pp. 566–570.

[68] D. Gondek and T. Hofmann, "Conditional information bottleneck clustering," in *3rd ieee international conference on data mining, workshop on clustering large data sets*. Citeseer, 2003, pp. 36–42.

[69] B. Larrousse, S. Lasaulce, and M. R. Bloch, "Coordination in distributed networks via coded actions with application to power control," *IEEE Trans. on Information Theory*, vol. 64, no. 5, pp. 3633–3654, 2018.

[70] S. Lasaulce and S. Tarbouriech, "Information constraints in multiple agent problems with i.i.d. states," in *Control subject to computational and communication constraints*. Springer, 2018, pp. 311–323.

[71] J. W. Overstreet and A. Tzes, "An internet-based real-time control engineering laboratory," *IEEE Control Systems Magazine*, vol. 19, no. 5, pp. 19–34, 1999.

[72] B. Aktan, C. A. Bohus, L. A. Crowl, and M. H. Shor, "Distance learning applied to control engineering laboratories," *IEEE Transactions on education*, vol. 39, no. 3, pp. 320–326, 1996.

[73] R. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1279–1289, 2000.

[74] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on automatic control*, vol. 49, no. 7, pp. 1056–1068, 2004.

[75] G. N. Nair and R. J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, 2004.

[76] N. C. Martins, M. A. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems in the presence of a direct link," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 438–447, 2006.

[77] P. Minero, M. Franceschetti, S. Dey, and G. Nair, "Data rate theorem for stabilization over fading channels," in *Proc. 45th Ann. Allerton Conf. on Communic., Control and Comput*, 2007.

[78] M. Andreasson, D. V. Dimarogonas, H. Sandberg, and K. H. Johansson, "Distributed control of networked dynamical systems: Static feedback, integral action and consensus," *IEEE Transactions on Automatic Control*, vol. 59, no. 7, pp. 1750–1764, 2014.

[79] D. Antunes and W. P. M. H. Heemels, "Rollout event-triggered control: Beyond periodic control performance," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3296–3311, 2014.

[80] W. Liu, G. Nair, Y. Li, D. Nesic, B. Vucetic, and H. V. Poor, "On the latency, rate and reliability tradeoff in wireless networked control systems for IIoT," *IEEE Internet of Things Journal*, 2020.

[81] K. Huang, W. Liu, Y. Li, A. Savkin, and B. Vucetic, "Wireless feedback control with variable packet length for industrial IoT," *IEEE Wireless Communications Letters*, 2020.

[82] A. S. Matveev and A. V. Savkin, *Estimation and control over communication networks*. Springer Science & Business Media, 2009.

[83] G. N. Nair, S. Dey, and R. J. Evans, "Infimum data rates for stabilising Markov jump linear systems," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 2. IEEE, 2003, pp. 1176–1181.

[84] Y. Song, J. Yang, M. Zheng, and C. Peng, "Disturbance attenuation for Markov jump linear system over an additive white Gaussian noise channel," *International Journal of Control*, vol. 89, no. 12, pp. 2482–2491, 2016.

[85] H. Zhenglong, S. Yang, Z. Min, J. Li, and T. Yang, "Stabilization of Markov jump linear systems over Gaussian relay channel," in *2016 UKACC 11th International Conference on Control (CONTROL)*. IEEE, pp. 1–6.

[86] C. Zhang, K. Chen, and G. E. Dullerud, "Stabilization of Markovian jump linear systems with limited information-a convex approach," in *2009 American Control Conference*. IEEE, 2009, pp. 4013–4019.

[87] V. S. Borkar, S. K. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of Markov sources," *SIAM journal on control and optimization*, vol. 40, no. 1, pp. 135–148, 2001.

[88] D. P. Bertsekas and D. A. Castanon, "Adaptive aggregation methods for infinite horizon dynamic programming," *IEEE Transactions on Automatic Control*, vol. 34, no. 6, pp. 589–598, June 1989.

[89] D. Bertsekas, "Biased aggregation, rollout, and enhanced policy improvement for reinforcement learning," *arXiv preprint arXiv:1910.02426*, 2019.

[90] M. L. Puterman and M. C. Shin, "Modified policy iteration algorithms for discounted Markov decision problems," *Management Science*, vol. 24, no. 11, pp. 1127–1137, 1978.

[91] M. Puterman and M. C. Shin, "Action elimination procedures for modified policy iteration algorithms," *Operations Research*, vol. 30, no. 2, pp. 301–318, 1982.

[92] M. Voelkel, A.-L. Sachs, and U. W. Thonemann, "An aggregation-based approximate dynamic programming approach for the periodic review model with random yield," *European Journal of Operational Research*, vol. 281, no. 2, pp. 286–298, 2020.

[93] M. A. Voelkel, A.-L. Sachs, and U. W. Thonemann, "An aggregation-based approximate dynamic programming approach for the periodic review model with random yield," *European Journal of Operational Research*, vol. 281, no. 2, pp. 286–298, 2020.

[94] F. Chatelin and W. L. Miranker, "Acceleration by aggregation of successive approximation methods," *Linear Algebra and its Applications*, vol. 43, pp. 17–47, 1982.

[95] E. Shafieepoorfard, M. Raginsky, and S. P. Meyn, "Rationally inattentive control of Markov processes," *SIAM Journal on Control and Optimization*, vol. 54, no. 2, pp. 987–1016, 2016.

[96] D. Maity, M. H. Mamduhi, S. Hirche, K. H. Johansson, and J. S. Baras, "Optimal LQG control under delay-dependent costly information," *IEEE control systems letters*, vol. 3, no. 1, pp. 102–107, 2018.

[97] S. Yüksel and T. Linder, "Optimization and convergence of observation channels in stochastic control," *SIAM Journal on Control and Optimization*, vol. 50, no. 2, pp. 864–887, 2012.

[98] C. A. Sims, "Implications of rational inattention," *Journal of monetary Economics*, vol. 50, no. 3, pp. 665–690, 2003.

[99] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE transactions on automatic control*, vol. 42, no. 5, pp. 674–690, 1997.

[100] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: an overview," in *Proceedings of 1995 34th IEEE Conference on Decision and Control*, vol. 1. IEEE, 1995, pp. 560–564.

[101] M. Riedmiller, "Neural fitted Q iteration: first experiences with a data efficient neural reinforcement learning method," in *European Conference on Machine Learning*. Springer, 2005, pp. 317–328.

[102] A. Antos, C. Szepesvári, and R. Munos, "Fitted Q-iteration in continuous action-space MDPs," in *Advances in neural information processing systems*, 2008, pp. 9–16.

[103] F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 664–671.

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

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

[106] F. A. Oliehoek, M. T. Spaan, and N. Vlassis, "Optimal and approximate Q-value functions for decentralized POMDPs," *Journal of Artificial Intelligence Research*, vol. 32, pp. 289–353, 2008.

[107] F. A. Oliehoek and N. Vlassis, "Q-value functions for decentralized POMDPs," in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 2007, pp. 1–8.

[108] F. A. Oliehoek, S. Whiteson, M. T. Spaan *et al.*, "Lossless clustering of histories in decentralized POMDPs." in *AAMAS (1)*, 2009, pp. 577–584.

[109] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, 2017, pp. 6382–6393.

[110] T. Chen, Y. Sun, and W. Yin, "LASG: Lazily Aggregated Stochastic Gradients for Communication-Efficient Distributed Learning," *arXiv e-prints*, p. arXiv:2002.11360, Feb. 2020.

[111] Y. Liu, Y. Sun, and W. Yin, "Decentralized learning with lazy and approximate dual gradients," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1362–1377, 2021.

[112] D. Abel, D. Hershkowitz, and M. Littman, "Near optimal behavior via approximate state abstraction," ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, 2016, pp. 2915–2923.

[113] L. Li, T. J. Walsh, and M. L. Littman, "Towards a unified theory of state abstraction for MDPs," in *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2006, Fort Lauderdale, Florida, USA, January 4-6, 2006*, 2006. [Online]. Available: http://anytime.cs.umass.edu/aimath06/proceedings/P21.pdf

[114] O. Nachum, S. Gu, H. Lee, and S. Levine, "Near-optimal representation learning for hierarchical reinforcement learning," *arXiv preprint arXiv:1810.01257*, 2018.

[115] P. Ioannou and J. Sun, "Theory and design of robust direct and indirect adaptive-control schemes," *International Journal of Control*, vol. 47, no. 3, pp. 775–813, 1988.

[116] K. S. Narendra and L. S. Valavani, "Direct and indirect adaptive control," *IFAC Proceedings Volumes*, vol. 11, no. 1, pp. 1981–1987, 1978.

[117] S. Sedighi, K. V. Mishra, M. B. Shankar, and B. Ottersten, "Localization with one-bit passive radars in narrowband internet-of-things using multivariate polynomial optimization," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2525–2540, 2021.

[118] J. Liu, W. Zhang, and H. V. Poor, "A rate-distortion framework for characterizing semantic information," *arXiv preprint arXiv:2105.04278*, 2021.

[119] A. Kipnis, S. Rini, and A. J. Goldsmith, "The rate-distortion risk in estimation from compressed data," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2910–2924, 2021.

[120] J. Dommel, Z. Utkovski, O. Simeone, and S. Stańczak, "Joint source-channel coding for semantics-aware grant-free radio access in IoT fog networks," *IEEE Signal Processing Letters*, vol. 28, pp. 728–732, 2021.

[121] E. Raei, M. Alaee-Kerahroodi, and M. B. Shankar, "Spatial- and range- ISLR trade-off in MIMO radar via waveform correlation optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 3283–3298, 2021.

[122] Z. Cheng, S. Shi, Z. He, and B. Liao, "Transmit sequence design for dual-function radar-communication system with one-bit DACs," *IEEE Trans. Wire. Commun.*, vol. 20, no. 9, pp. 5846–5860, 2021.

[123] P. Minero, M. Franceschetti, S. Dey, and G. N. Nair, "Data rate theorem for stabilization over time-varying feedback channels," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 243–255, 2009.

[124] W. Liu, P. Popovski, Y. Li, and B. Vucetic, "Wireless networked control systems with coding-free data transmission for industrial IoT," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1788–1801, 2020.

[125] W. Liu, P. Popovski, Y. Li, and B. Vucetic, "Real-time wireless networked control systems with coding-free data transmission," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[126] M. Pezzutto, F. Tramarin, S. Dey, and L. Schenato, "Adaptive transmission rate for LQG control over Wi-Fi: A cross-layer approach," *Automatica*, vol. 119, p. 109092, 2020.

[127] Y. Wu, H.-N. Dai, and H. Tang, "Graph neural networks for anomaly detection in industrial internet of things," *IEEE Internet of Things J.*, pp. 1–1, 2021.

[128] V.-D. Nguyen and O.-S. Shin, "Cooperative prediction-and-sensing-based spectrum sharing in cognitive radio networks," *IEEE Trans. Cognitive Commun. and Netw.*, vol. 4, no. 1, pp. 108–120, 2018.

[129] Y. Zhen, W. Chen, L. Zheng, X. Li, and D. Mu, "Multi-agent cooperative caching policy in industrial internet of things," *IEEE Internet of Things J.*, pp. 1–1, 2022.

[130] D. Ullmann, S. Rezaeifar, O. Taran, T. Holotyak, B. Panos, and S. Voloshynovskiy, "Information bottleneck classification in extremely distributed systems," *Entropy*, vol. 22, no. 11, p. 1237, 2020.

[131] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," *arXiv preprint arXiv:1712.01887*, 2017.

[132] F. Seide, H. Fu, J. Droppo, G. Li, and D. 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.

[133] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Advances in neural information processing systems*, 2017, pp. 1509–1519.

[134] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *arXiv preprint arXiv:1606.06160*, 2016.

[135] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *arXiv preprint arXiv:2008.11141*, 2020.

[136] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[137] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[138] Z. Li, C. Wang, and C.-J. Jiang, "User association for load balancing in vehicular networks: An online reinforcement learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2217–2228, 2017.

[139] A. Mohajer, M. Bavaghar, and H. Farrokhi, "Mobility-aware load balancing for reliable self-organization networks: Multi-agent deep reinforcement learning," *Reliability Engineering & System Safety*, vol. 202, p. 107056, 2020.

[140] Y. Xu, W. Xu, Z. Wang, J. Lin, and S. Cui, "Load balancing for ultradense networks: A deep reinforcement learning-based approach," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9399–9412, 2019.

[141] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, "Reinforcement learning for self organization and power control of two-tier heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3933–3947, 2019.

[142] S. S. Mwanje, L. C. Schmelz, and A. Mitschele-Thiel, "Cognitive cellular networks: A Q-learning framework for self-organizing networks," *IEEE Transactions on Network and Service Management*, vol. 13, no. 1, pp. 85–98, 2016.

[143] P. Muñoz, R. Barco, and I. de la Bandera, "Load balancing and handover joint optimization in LTE networks using fuzzy logic and reinforcement learning," *Computer Networks*, vol. 76, pp. 112–125, 2015.

[144] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5488–5502, 2020.

[145] J. Dong, Y. Shi, and Z. Ding, "Blind over-the-air computation and data fusion via provable wirtinger flow," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1136–1151, 2020.

[146] M. Frey, I. Bjelakovic, and S. Stanczak, "Over-the-air computation for distributed machine learning," *arXiv preprint arXiv:2007.02648*, 2020.

[147] P. G. Otanez, J. R. Moyne, and D. M. Tilbury, "Using deadbands to reduce communication in networked control systems," in *Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301)*, vol. 4. IEEE, 2002, pp. 3015–3020.

[148] E. Steinbach, S. Hirche, J. Kammerl, I. Vittorias, and R. Chaudhari, "Haptic data compression and communication," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 87–96, 2011.

[149] S. Hirche and M. Buss, "Transparent data reduction in networked telepresence and teleaction systems. part ii: Time-delayed communication," *Presence: Teleoperators and Virtual Environments*, vol. 16, no. 5, pp. 532–542, 2007.

[150] M. Mukherjee, M. Guo, J. Lloret, and Q. Zhang, "Leveraging intelligent computation offloading with fog/edge computing for tactile internet: Advantages and limitations," *IEEE Network*, vol. 34, no. 5, pp. 322–329, 2020.

[151] U. Aßmann, C. Baier, C. Dubslaff, D. Grzelak, S. Hanisch, A. P. P. Hartono, S. Köpsell, T. Lin, and T. Strufe, "Tactile computing: Essential building blocks for the tactile internet," in *Tactile Internet*. Elsevier, 2021, pp. 293–317.

[152] C. Shahabi, A. Ortega, and M. R. Kolahdouzan, "A comparison of different haptic compression techniques," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1. IEEE, 2002, pp. 657–660.

[153] P. Hinterseer, S. Hirche, S. Chaudhuri, E. Steinbach, and M. Buss, "Perception-based data reduction and transmission of haptic data in telepresence and teleaction systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 588–597, 2008.

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

[155] J. Hu, H. Zhang, and L. Song, "Reinforcement learning for decentralized trajectory design in cellular UAV networks with sense-and-send protocol," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6177–6189, 2019.

[156] Q. Wang, W. Zhang, Y. Liu, and Y. Liu, "Multi-UAV dynamic wireless networking with deep reinforcement learning," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2243–2246, 2019.

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

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

[159] N. H. Chu, D. T. Hoang, D. N. Nguyen, N. Van Huynh, and E. Dutkiewicz, "Joint speed control and energy replenishment optimization for UAV-assisted IoT data collection with deep reinforcement transfer learning," *IEEE Internet of Things J.*, pp. 1–1, 2022.

[160] Z. Chang, H. Deng, L. You, G. Min, S. Garg, and G. Kaddoum, "Trajectory design and resource allocation for multi-UAV networks: Deep reinforcement learning approaches," *IEEE Transactions on Network Science and Engineering*, pp. 1–1, 2022.

[161] W. Wang, Y. Liu, R. Srikant, and L. Ying, "3M-RL: Multi-resolution, multi-agent, mean-field reinforcement learning for autonomous UAV routing," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2021.

[162] T. Ren, J. Niu, B. Dai, X. Liu, Z. Hu, M. Xu, and M. Guizani, "Enabling efficient scheduling in large-scale UAV-assisted mobile edge computing via hierarchical reinforcement learning," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[163] H. Chang, Y. Chen, B. Zhang, and D. Doermann, "Multi-UAV mobile edge computing and path planning platform based on reinforcement learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–10, 2021.

[164] R. Zhang, Q. Zong, X. Zhang, L. Dou, and B. Tian, "Game of drones: Multi-UAV pursuit-evasion game with online motion planning by deep reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2022.

[165] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and A. Nallanathan, "Deep reinforcement learning based dynamic trajectory control for UAV-assisted mobile edge computing," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[166] C. Zhan and Y. Zeng, "Energy minimization for cellular-connected UAV: From optimization to deep reinforcement learning," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[167] X. Zhong and Y. Zhou, "A reinforcement learning trained fuzzy neural network controller for maintaining wireless communication connections in multi-robot systems," in *Machine Intelligence and Bio-inspired Computation: Theory and Applications VIII*, vol. 9119. International Society for Optics and Photonics, 2014, p. 91190A.

[168] S. H. Alsamhi, O. Ma, M. Ansari *et al.*, "Convergence of machine learning and robotics communication in collaborative assembly: mobility, connectivity and future perspectives," *Journal of Intelligent & Robotic Systems*, vol. 98, no. 3, pp. 541–566, 2020.

[169] D. O. Pop, A. Rogozan, C. Chatelain, F. Nashashibi, and A. Bensrhair, "Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction," *IEEE Access*, vol. 7, pp. 149 318–149 327, 2019.

[170] X. Na and D. J. Cole, "Modelling of a human driver's interaction with vehicle automated steering using cooperative game theory," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 5, pp. 1095–1107, 2019.

[171] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep Q-network for UAV-assisted online power transfer and data collection," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 215–12 226, 2019.

[172] S. Tomashevich and B. Andrievsky, "Improved adaptive coding procedure for transferring the navigation data between uavs in formation," in *AIP Conference Proceedings*, vol. 2046, no. 1. AIP Publishing LLC, 2018, p. 020102.

[173] M. Hüttenrauch, A. Sosic, and G. Neumann, "Deep reinforcement learning for swarm systems," *ArXiv*, vol. abs/1807.06613, 2019.

[174] D. Baldazo, J. Parras, and S. Zazo, "Decentralized multi-agent deep reinforcement learning in swarms of drones for flood monitoring," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[175] Y. Sung, A. K. Budhiraja, R. K. Williams, and P. Tokekar, "Distributed assignment with limited communication for multi-robot multi-target tracking," *Autonomous Robots*, vol. 44, no. 1, pp. 57–73, 2020.

[176] S. W. Loke, "Cooperative automated vehicles: A review of opportunities and challenges in socially intelligent vehicles beyond networking," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 4, pp. 509–518, 2019.

[177] S. Kar, J. M. F. Moura, and H. V. Poor, "$\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through $\mathsf{m}consensus + \mathsf{m}innovations$," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1848–1862, April 2013.

[178] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of control and estimation over lossy networks," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 163–187, 2007.

[179] A. S. Matveev and A. V. Savkin, "The problem of lqg optimal control via a limited capacity communication channel," *Systems & control letters*, vol. 53, no. 1, pp. 51–64, 2004.

[180] S. Chai and V. K. N. Lau, "Online trajectory and radio resource optimization of cache-enabled uav wireless networks with content and energy recharging," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1286–1299, 2020.

[181] G. Faraci, C. Grasso, and G. Schembra, "Design of a 5G network slice extension with mec uavs managed with reinforcement learning," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2020.

[182] F. Wu, H. Zhang, J. Wu, and L. Song, "Cellular UAV-to-device communications: Trajectory design and mode selection by multi-agent deep reinforcement learning," *IEEE Transactions on Communications*, pp. 1–1, 2020.

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

[184] Z. Fang, J. Wang, Y. Ren, Z. Han, H. V. Poor, and L. Hanzo, "Age of information in energy harvesting aided massive multiple access networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 5, pp. 1441–1456, 2022.

[185] W. Wei, J. Wang, Z. Fang, J. Chen, Y. Ren, and Y. Dong, "3U: Joint design of UAV-USV-UUV networks for cooperative target hunting," *IEEE Transactions on Vehicular Technology*, 2022.

[186] F. Dressler, F. Klingler, M. Segata, and R. L. Cigno, "Cooperative driving and the tactile internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 436–446, Feb 2019.

[187] A. Benloucif, A. Nguyen, C. Sentouh, and J. Popieul, "Cooperative trajectory planning for haptic shared control between driver and automation in highway driving," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9846–9857, 2019.

[188] M. During and K. Lemmer, "Cooperative maneuver planning for cooperative driving," *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 3, pp. 8–22, Fall 2016.

[189] M. Segata and *et al.*, "Toward communication strategies for platooning: Simulative and experimental evaluation," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5411–5423, Dec 2015.

[190] A. Ghosh and S. Huang, "Cooperative traffic control where autonomous cars meet human drivers," in *2019 SoutheastCon*, 2019, pp. 1–6.

[191] K. Sonoda and T. Wada, "Displaying system situation awareness increases driver trust in automated driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 185–193, 2017.

[192] F. Yan, K. Wang, B. Zou, L. Tang, W. Li, and C. Lv, "LiDAR-based multi-task road perception network for autonomous vehicles," *IEEE Access*, vol. 8, pp. 86 753–86 764, 2020.

[193] S. Li and R. Li, "Task allocation based on task deployment in autonomous vehicular cloud," in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 2019, pp. 450–454.

[194] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[195] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[196] S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós, and A. Smola, "Aide: Fast and communication efficient distributed optimization," *arXiv preprint arXiv:1608.06879*, 2016.

[197] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[198] D. Wen, X. Li, Q. Zeng, J. Ren, and K. Huang, "An overview of data-importance aware radio resource management for edge machine learning," *Journal of Communications and Information Networks*, vol. 4, no. 4, pp. 1–14, 2019.

[199] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 265–278, 2020.

[200] A. Molin, H. Esen, and K. H. Johansson, "Scheduling networked state estimators based on value of information," *Automatica*, vol. 110, p. 108578, 2019.

[201] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[202] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[203] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[204] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Coordinating handover parameter optimization and load balancing in LTE self-optimizing networks," in *2011 IEEE 73rd vehicular technology conference (VTC Spring)*. IEEE, 2011, pp. 1–5.

[205] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Toward haptic communications over the 5g tactile internet," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3034–3059, 2018.

[206] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.

[207] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 518–521, 1980.

[208] A. Barel, R. Manor, and A. M. Bruckstein, "Come together: Multi-agent geometric consensus," *arXiv preprint arXiv:1902.01455*, 2017.

[209] J. Schneider, W.-K. Wong, A. Moore, and M. Riedmiller, "Distributed value functions," 1999.

[210] D. H. Wolpert, K. R. Wheeler, and K. Tumer, "General principles of learning-based multi-agent systems," in *Proceedings of the third annual conference on Autonomous Agents*, 1999, pp. 77–83.

[211] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2022.

[212] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PloS one*, vol. 12, no. 4, p. e0172395, 2017.

[213] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7234–7284, 2020.

[214] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *International conference on machine learning*. PMLR, 2019, pp. 2961–2970.

[215] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6382–6393.

[216] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99. Citeseer, 1999, pp. 278–287.

[217] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," 2018.

[218] X. V. Lin, R. Socher, and C. Xiong, "Multi-hop knowledge graph reasoning with reward shaping," 2018.

[219] H. Zhang, W. Chen, Z. Huang, M. Li, Y. Yang, W. Zhang, and J. Wang, "Bi-level actor-critic for multi-agent coordination," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7325–7332.

[220] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[221] E. Vinogradov, H. Sallouha, S. De Bast, M. M. Azari, and S. Pollin, "Tutorial on UAV: A blue sky view on wireless communication," *arXiv preprint arXiv:1901.02306*, 2019.

[222] M. M. Azari, F. Rosas, K.-C. Chen, and S. Pollin, "Ultra reliable UAV communication using altitude and cooperation diversity," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 330–344, 2018.

[223] Y. Yuan, L. Lei, T. X. Vu, S. Chatzinotas, S. Sun, and B. Ottersten, "Energy minimization in UAV-aided networks: Actor-critic learning for constrained scheduling optimization," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2021.

[224] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019. [Online]. Available: https://arxiv.org/abs/1912.04977

[225] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. of the IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[226] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S. L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. of the IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[227] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. International Conference on Learning Representations*, 2019.

[228] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *Journal of Machine Learning Research*, vol. 14, no. 1, p. 3321–3363, 2013.

[229] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Select. Areas in Commun.*, vol. 37, no. 6, pp. 1205–1221, June 2019.

[230] V. D. Nguyen, S. K. Sharma, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Efficient federated learning algorithm for resource allocation in wireless IoT networks," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[231] V.-D. Nguyen, S. Chatzinotas, B. Ottersten, and T. Q. Duong, "FedFog: Network-aware optimization of federated learning over wireless fog-cloud systems," *IEEE Trans. Wire. Commun.*, pp. 1–18, 2022.

[232] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *J. ACM*, vol. 61, no. 6, pp. 1–57, Dec. 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2666468

[233] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," 2018. [Online]. Available: https://arxiv.org/abs/1710.06963

[234] Z. Cao, P. Zhou, R. Li, S. Huang, and D. Wu, "Multiagent deep reinforcement learning for joint multichannel access and task offloading of mobile-edge computing in industry 4.0," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6201–6213, 2020.

[235] 3GPP, "TS 32.500 Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements, Rel-16," July 2020. [Online]. Available: https://www.3gpp.org

[236] H. Xu, S. Feng, Y. Zhang, and L. Li, "A grouping-based cooperative driving strategy for cavs merging problems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6125–6136, 2019.

[237] J. He and et al, "Cooperative connected autonomous vehicles (CAV): Research, applications and challenges," in *2019 IEEE 27th International Conference on Network Protocols (ICNP)*, 2019, pp. 1–6.

[238] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–20, 2020.

[239] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, 2021.

[240] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Deep inverse reinforcement learning for behavior prediction in autonomous driving: Accurate forecasts of vehicle motion," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 87–96, 2021.

[241] J. García, Fern, and o Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 42, pp. 1437–1480, 2015.

[242] K. Zhang, Z. Yang, and T. Başar, *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*. Cham: Springer International Publishing, 2021, pp. 321–384. [Online]. Available: https://doi.org/10.1007/978-3-030-60990-0_12

[243] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multi-agent systems: A review of challenges, solutions, and applications," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[244] D. Gunduz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," 2022. [Online]. Available: https://arxiv.org/abs/2207.09353

[245] M. R. Palattella and N. Accettura, "Enabling internet of everything everywhere: Lpwan with satellite backhaul," in *2018 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, 2018, pp. 1–5.

[246] L. Chaari, M. Fourati, and J. Rezgui, "Heterogeneous lorawan & leo satellites networks concepts, architectures and future directions," in *2019 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, 2019, pp. 1–6.

[247] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. M. Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani *et al.*, "Evolution of non-terrestrial networks from 5g to 6g: A survey," *IEEE Communications Surveys & Tutorials*, 2022.

[248] G. N. Nair and R. J. Evans, "Exponential stabilisability of finite-dimensional linear systems with limited data rates," *Automatica*, vol. 39, no. 4, pp. 585–593, 2003.

[249] M. Lauer and M. A. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Proc. Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2000.

[250] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[251] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for reinforcement learning: From a single agent to cooperative agents," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 123–135, 2020.

[252] C. Zhang and V. Lesser, "Coordinating multi-agent reinforcement learning with limited communication," in *Conference on Autonomous Agents and Multi-agent Systems*, St. Paul, Minnesota, May 2013, pp. 1101–1108.

[253] F. Fischer, M. Rovatsos, and G. Weiss, "Hierarchical reinforcement learning in communication-mediated multiagent coordination," in *Proc. IEEE Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, New York, Jul. 2004, pp. 1334–1335.

[254] T. Kasai, H. Tenmoto, and A. Kamiya, "Learning of communication codes in multi-agent reinforcement learning problem," in *Soft Computing in Industrial Applications, 2008. SMCia'08. IEEE Conf. on.* IEEE, 2008, pp. 1–6.

[255] F. Wu, S. Zilberstein, and X. Chen, "Online planning for multi-agent systems with bounded communication," *Artificial Intelligence*, vol. 175, no. 2, pp. 487–511, Feb. 2011.

[256] A. Amini, A. Asif, and A. Mohammadi, "CEASE: A collaborative event-triggered average-consensus sampled-data framework with performance guarantees for multi-agent systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 23, pp. 6096–6109, 2018.

[257] R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, and Y. Dauphin, "On the pitfalls of measuring emergent communication," in *Intl. Conf. on Autonomous Agents and MultiAgent Systems*, 2019.

[258] D. P. Bertsekas, "Feature-based aggregation and deep reinforcement learning: A survey and some new implementations," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 1–31, 2018.

[259] D. Abel, D. Hershkowitz, and M. Littman, "Near optimal behavior via approximate state abstraction," in *International Conference on Machine Learning.* PMLR, 2016, pp. 2915–2923.

[260] G. Rubino, "On weak lumpability in markov chains," *Journal of Applied Probability*, vol. 26, no. 3, pp. 446–457, 1989.

[261] H. Zou, C. Zhang, S. Lasaulce, and et al, "Decision-oriented communications: Application to energy-efficient resource allocation," in *Intl. Conf. on Wireless Networks and Mobile Communications*. IEEE, 2018.

[262] H. Mao, Z. Zhang, Z. Xiao, Z. Gong, and Y. Ni, "Learning agent communication under limited bandwidth by message pruning," *arXiv preprint arXiv:1912.05304*, 2019.

[263] S. Sukhbaatar, R. Fergus *et al.*, "Learning multiagent communication with backpropagation," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, 2016, pp. 2244–2252.

[264] G. E. Monahan, "State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms," *Management science*, vol. 28, no. 1, pp. 1–16, 1982.

[265] P. Xuan, V. Lesser, and S. Zilberstein, "Communication decisions in multi-agent cooperation: Model and experiments," in *Proceedings of the Fifth International Conference on Autonomous Agents*, ser. AGENTS '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 616–623. [Online]. Available: https://doi.org/10.1145/375735.376469

[266] F. A. Oliehoek, M. T. Spaan, N. Vlassis *et al.*, "DEC-PoMDPs with delayed communication," in *Proc. Multi-agent Sequential Decision-Making in Uncertain Domains*, Honolulu, Hawaii, May 2007.

[267] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*, 2nd ed. MIT Press, Nov. 2017, vol. 135.

[268] Y. Rizk, M. Awad, and E. W. Tunstel, "Decision making in multiagent systems: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 514–529, 2018.

[269] C. Boutilier, "Multiagent systems: Challenges and opportunities for decision-theoretic planning," *AI magazine*, vol. 20, no. 4, pp. 35–35, 1999.

[270] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Advances in neural information processing systems*, 1994, pp. 703–710.

[271] F. Heylighen, "Stigmergy as a universal coordination mechanism i: Definition and components," *Cognitive Systems Research*, vol. 38, pp. 4–13, 2016.

[272] F. A. Oliehoek, C. Amato *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.

[273] S. Yüksel, "Jointly optimal lqg quantization and control policies for multi-dimensional systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1612–1617, 2013.

[274] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[275] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.

[276] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[277] C. Amato, J. S. Dibangoye, and S. Zilberstein, "Incremental policy generation for finite-horizon dec-pomdps," in *Nineteenth International Conference on Automated Planning and Scheduling*, 2009.

[278] M. G. Azar, R. Munos, M. Ghavamzadaeh, and H. J. Kappen, "Speedy Q-learning," 2011.

[279] C.-M. Chou, C.-Y. Li, W.-M. Chien, and K.-c. Lan, "A feasibility study on vehicle-to-infrastructure communication: Wifi vs. wimax," in *2009 tenth international conference on mobile data management: systems, services and middleware*. IEEE, 2009, pp. 397–398.

[280] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.

[281] Z. Ding, W. Hong, L. Zhu, T. Huang, and Z. Lu, "Sequential communication in multi-agent reinforcement learning," 2021.

[282] J. Albowicz, A. Chen, and L. Zhang, "Recursive position estimation in sensor networks," in *Proceedings Ninth International Conference on Network Protocols. ICNP 2001.* IEEE, 2001, pp. 35–41.

[283] S. Dorvash and S. Pakzad, "Stochastic iterative modal identification algorithm and application in wireless sensor networks," *Structural Control and Health Monitoring*, vol. 20, no. 8, pp. 1121–1137, 2013.

[284] L. Li, T. J. Walsh, and M. L. Littman, "Towards a unified theory of state abstraction for MDPs." in *AI&M*, 2006.

[285] A. K. McCallum, *Reinforcement learning with selective perception and hidden state.* University of Rochester, 1996.

[286] A. Mostaani, T. X. Vu, S. K. Sharma, V.-D. Nguyen, Q. Liao, and S. Chatzinotas, "Task-oriented communication design in cyber-physical systems: A survey on theory and applications," *IEEE Access*, vol. 10, pp. 133 842–133 868, 2022.

[287] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3039–3071, 2019.

[288] D. Lin, S. Talathi, and S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *International conference on machine learning.* PMLR, 2016, pp. 2849–2858.

[289] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[290] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Task-effective compression of observations for the centralized control of a multi-agent system over bit-budgeted channels," 2023. [Online]. Available: https://arxiv.org/abs/2301.01628

[291] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2022.

[292] Y. Yamada, S. Tazaki, and R. Gray, "Asymptotic performance of block quantizers with difference distortion measures," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 6–14, 1980.

[293] J. Liu, W. Zhang, and H. V. Poor, "A rate-distortion framework for characterizing semantic information," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE Press, 2021, p. 2894–2899. [Online]. Available: https://doi.org/10.1109/ISIT45174.2021.9518240

[294] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Task-based information compression for multi-agent communication problems with channel rate constraints," *arXiv preprint arXiv:2005.14220*, 2020.

[295] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[296] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[297] B. A. Davey and H. A. Priestley, *Complete lattices and Galois connections*, 2nd ed. Cambridge University Press, 2002, p. 145–174.

[298] J. R. Munkres, *Topology (Classic Version), 2nd edition*. Pearson, 2023.

[299] M. Motamedi, N. Sakharnykh, and T. Kaldewey, "A data-centric approach for training deep neural networks with less data," *arXiv preprint arXiv:2110.03613*, 2021.

[300] C. Bergenhem, S. Shladover, E. Coelingh, C. Englund, and S. Tsugawa, "Overview of platooning systems," in *Proceedings of the 19th ITS World Congress, Oct 22-26, Vienna, Austria (2012)*, 2012.

[301] C. Bergenhem, E. Hedin, and D. Skarin, "Vehicle-to-vehicle communication for a platooning system," *Procedia-Social and Behavioral Sciences*, vol. 48, pp. 1222–1233, 2012.

[302] Y. Sun, K. Zheng, and Y. Tang, "Control efficient power allocation of uplink noma in uav-aided vehicular platooning," *IEEE Access*, vol. 9, pp. 139 473–139 488, 2021.

[303] J. Lunze, "Design of the communication structure of cooperative adaptive cruise controllers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4378–4387, 2019.

[304] S. Badsha, X. Yi, and I. Khalil, "A practical privacy-preserving recommender system," *Data Science and Engineering*, vol. 1, pp. 161–177, 2016.

[305] F. McSherry and I. Mironov, "Differentially private recommender systems: Building privacy into the netflix prize contenders," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 627–636.

[306] W. Ali, R. Kumar, Z. Deng, Y. Wang, and J. Shao, "A federated learning approach for privacy protection in context-aware recommender systems," *The Computer Journal*, vol. 64, no. 7, pp. 1016–1027, 2021.

[307] S. Badsha, X. Yi, I. Khalil, and E. Bertino, "Privacy preserving user-based recommender system," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, 2017, pp. 1074–1083.

[308] H. Cai, K. Ren, W. Zhang, K. Malialis, J. Wang, Y. Yu, and D. Guo, "Real-time bidding by reinforcement learning in display advertising," in *Proceedings of the tenth ACM international conference on web search and data mining*, 2017, pp. 661–670.

[309] X. Zhao, C. Gu, H. Zhang, X. Yang, X. Liu, J. Tang, and H. Liu, "Dear: Deep reinforcement learning for online advertising impression in recommender systems," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 750–758.

[310] C. Du, Z. Gao, S. Yuan, L. Gao, Z. Li, Y. Zeng, X. Zhu, J. Xu, K. Gai, and K.-C. Lee, "Exploration in online advertising systems with deep uncertainty-aware learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2792–2801.

[311] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 2022, pp. 316–332.

[312] S. Su, Y. Li, S. He, S. Han, C. Feng, C. Ding, and F. Miao, "Uncertainty quantification of collaborative detection for self-driving," *arXiv preprint arXiv:2209.08162*, 2022.

[313] M. Pezzutto, M. Farina, R. Carli, and L. Schenato, "Remote mpc for tracking over lossy networks," *IEEE Control Systems Letters*, vol. 6, pp. 1040–1045, 2022.

[314] M. Harounabadi, D. M. Soleymani, S. Bhadauria, M. Leyh, and E. Roth-Mandutz, "V2X in 3GPP standardization: NR sidelink in release-16 and beyond," *IEEE Communications Standards Magazine*, vol. 5, no. 1, pp. 12–21, 2021.

[315] S.-W. Kim, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "The impact of cooperative perception on decision making and planning of autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 3, pp. 39–50, 2015.

[316] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNET: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.

[317] C. D. Charalambous and A. Farhadi, "LQG optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels," *Automatica*, vol. 44, no. 12, pp. 3181–3188, 2008.