ON CENTRALIZED AND DECENTRALIZED DECISION-MAKING PROBLEMS WITH PARTIAL INFORMATION

by

Aditya Deepak Dave

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mechanical Engineering

Summer 2023

ON CENTRALIZED AND DECENTRALIZED DECISION-MAKING PROBLEMS WITH PARTIAL INFORMATION

by

Aditya Deepak Dave

Approved: ____

Ajay Prasad, Ph.D.
Chair of the Department of Mechanical Engineering

Approved: ____

Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: ____

Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Andreas A. Malikopoulos, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Aditya Mahajan, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Ioannis Poulakakis, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Adam Wickenheiser, Ph.D.
Member of dissertation committee

I dedicate this dissertation to my parents
both of whom have taught me everything I know,
whose love and support have been unwavering,
who have shaped me into the person I am.

I also begin this dissertation with a prayer
for any reader who has felt out of their depth
during a labor of love:
I hope your efforts succeed.

ACKNOWLEDGEMENTS

A Ph.D. is a long undertaking, full of many twists and turns. I would like to begin by expressing my utmost gratitude towards my advisor, Dr. Andreas Malikopoulos, for his constant guidance, technical inputs, many insights, and support over the course of these years. He has helped shape the subjects of my research and my approach towards research by constantly stimulating my thinking and challenging me to do my best. I would also like to acknowledge the members of my dissertation committee. Dr. Aditya Mahajan has always made himself available to guide me and has provided me with insights into my work at time when I have struggled to find them on my own. Dr. Ioannis Poulakakis has also been invaluable in offering his unique perspective over the course of many helpful discussions and Dr. Adam Wickenheiser who has guided me in considering more applied aspects of my work through his suggestions.

I would also like to sincerely all my family for their consistent patience and support, for all the joy, cheer and love they have brought into my life throughout these years, whether in-person or over long distances. I would also like to thank my friends back in India and here in the US, whose presence has made my time all the more memorable. I would like to specifically highlight the contributions of many current and previous members of the IDS lab, without whose presence and support this journey would not have been possible. Their company, help, constant feedback and suggestions have been instrumental towards the eventual form my research and this dissertation have taken.

TABLE OF CONTENTS

Appendix

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The advent of cyber-physical systems has revolutionized numerous applications, including connected and automated vehicles, medicine and healthcare, the Internet of Things, social media platforms, and robotic swarms. These systems require new approaches that can utilize the improved computational capabilities of the cyber core to optimally control their physical components, while accounting for various forms of incomplete information and uncertain disturbances during real-world implementation. This dissertation primarily focuses on two areas of research: (1) centralized worst-case control and learning with partial observations, and (2) decentralized control of a team of cooperative agents. Additionally, the dissertation presents a mechanism design approach to effectively coordinate the actions of competing agents, specifically in a misinformation filtering problem involving competing social media platforms and a democratic government.

The first contribution of this dissertation is to introduce a general non-stochastic theory of approximate information states, addressing the computational challenges of worst-case control and facilitating worst-case reinforcement learning in partially observed systems. An important feature of the proposed framework is that approximate information states can be constructed using output data in control problems and can be learned from output data in reinforcement learning problems. Then, these states facilitate the efficient computation of approximate control strategies while only conceding a bounded loss in worst-case performance. Thus, our proposed framework provides a principled approach for approximately optimal worst-case control and worst-case reinforcement learning in systems with partially observed states.

The second contribution of this dissertation is towards the theory of decentralized decision-making for teams with one-directional information sharing, in both stochastic and non-stochastic formulations. A general information structure, called nested accessible information, is introduced for teams of two agents with one-directional communication. This information structure is analyzed in the stochastic setting to derive structural results and a dynamic programming decomposition that computes optimal control strategies. Then, this information structure is extended to multiple agents residing within nested subsystems, which is analyzed in the non-stochastic setting. As before, structural results and a dynamic programming decomposition are presented for control strategies that optimize the worst-case performance of the team. The effectiveness of the results is illustrated using a numerical example in the non-stochastic setting.

The third contribution of this dissertation is in the application of a mechanism design approach to a specific problem concerning how a democratic government can incentivize social media platforms to filter misinformation. This problem is modeled by drawing upon theoretical and empirical conclusions from political science and sociology. Then, a mechanism is proposed that takes a fixed budget of the government, creates incentives for social media platforms and distributes these incentives to encourage an optimal level of filtering. Key properties of the proposed mechanism, such as budget balance, voluntary participation of all agents, and implementation of a globally optimal level of filtering given a budget, are derived.

Collectively, this dissertation contributes towards an improved understanding of decision-making in cyber-physical systems with partial information. The theoretical results presented here have potential applications in various cyber-physical systems subject to uncontrolled disturbances, adversarial attacks, and information asymmetry.

Chapter 1

INTRODUCTION

The only thing that is certain
is that nothing is certain.

Pliny the Elder, 70 CE

1.1 Motivation

Cyber-physical systems have caused a major paradigm shift in the way we interact with physical processes [1]. They pervade various important applications including, but not limited to, connected and automated vehicles [2, 3], medicine and healthcare [4, 5], internet of things [6, 7], social media platforms [8, 9], and robotic swarms [10, 11]. Cyber-physical systems consist of physical components monitored and controlled by powerful digital cores. Thus, they can capture the large amounts of data generated by the system and utilize it in complex models to generate optimal decisions. However, this increased capacity for complex decision-making has necessitated new approaches to optimally control the evolution of their physical components while accounting for various uncertainties. For example, consider a centralized decision-making problem where an agent must sequentially select actions to control the system across multiple instances of time to achieve a predetermined objective [12]. In cyber-physical systems, these problems are often challenging because of a paucity of information about the system state during operation, or because of incomplete knowledge of the system dynamics a priori. An agent may only obtain noisy and infrequent observations and they may be unable to accurately predict the impact of an action on the state's evolution. In general, deriving an optimal decision-making strategy that accounts for such incomplete information is a computationally challenging task [13].

The computational challenge imposed by partial information is compounded in decentralized systems consisting of multiple agents. In decentralized problems, multiple agents act either cooperatively in a team to achieve a shared objective [14], or compete with each other to achieve individual objectives [15]. Different agents in such a system may not have access to the same information at any instance of time. For example, agents seeking to act cooperatively in a team may suffer from restricted communication or agents competing to achieve their individual objectives may deliberately withhold private information. This information asymmetry amongst multiple agents adds the burden of effectively coordinating the actions to the computational challenges imposed by partial information. Thus, generating an optimal decision-making strategy is even more computationally challenging in decentralized problems than it is in centralized problems. Motivated by these challenges to decision-making in cyber-physical systems under partial information, this dissertation's primary focus is on the following two areas of research: (1) centralized worst-case control and learning with partial observations in Chapter 2, and (2) decentralized control of a team of cooperative agents in Chapter 3. Subsequently, Chapter 4 reports results on effective coordination of the actions of competing agents using mechanism design. The results in Chapter 4 are developed for a specific misinformation filtering problem involving competing social media platforms and a democratic government.

1.2 Overview and Literature Review

This section presents an overview and a literature review pertaining to the two primary areas of focus of this dissertation, namely, centralized worst-case decision making and decentralized decision making in teams. The relevant literature for mechanism design and misinformation filtering are presented in the introduction to Chapter 4.

1.2.1 Centralized Worst-Case Control and Learning With Partial Observations

A typical centralized decision-making problem considers an agent that sequentially selects actions to control the evolution of a dynamic system using only partial observations at each instance of time, while simultaneously accounting for interference from uncontrolled disturbances. The most common modeling paradigm for such decision-making problems is *the stochastic approach,* where all disturbances to the system are considered to be random variables with known distributions, and the agent aims to select a decision-making strategy that minimizes the *expected incurred cost* [12]. Stochastic models have been utilized for problems in both control theory [16–20] and reinforcement learning [14,21–24]. A decision-making strategy derived using the stochastic approach performs optimally on average when it is implemented consistently across numerous operations of the system. However, this performance degrades rapidly when there is a mismatch between the distribution of disturbances considered in modeling and the realizations encountered during implementation [25]. To mitigate this drawback, alternative formulations have been considered in the literature, including *(1) robust stochastic formulations,* where an agent minimizes the worst-case expected cost given a set of feasible probability distributions [26, 27]; and *(2) risk-averse formulations,* where an agent minimizes a combination of both the expected cost and the cost variance [16,28]. While these formulations improve the performance of a strategy under a distribution mismatch, many safety-critical applications require further guarantees on the worst-case performance of a strategy against either adversarial attacks or system failure.

A *non-stochastic formulation* provides a more appropriate modeling paradigm for such safety-critical systems, where all disturbances are considered to belong to known sets with unknown distributions. The agent aims to select a robust decision-making strategy that minimizes the *worst-case incurred cost* across a given time horizon [29, 30]. Because this approach focuses on robustness against worst-case realizations of the disturbances, the resulting strategy yields more conservative decisions than the stochastic approach. At the expense of average performance, such a robust strategy provides concrete guarantees on the worst-case performance during each operation of the system. Thus, this approach has been widely applied to systems under attack from an adversary, e.g., cyber-security [31] or generic cyber-physical systems [32], and systems where a single failure can be damaging, e.g., water reservoirs [33], or power systems [34]. This non-stochastic formulation is maximally robust for a given set of uncertainties with any feasible set of distributions. It has also been shown to be the most risk-averse limit in various approaches [28,35,36]. Most of the literature relevant to the problem of worst-case decision making in a non-stochastic formulation can be broadly categorized as either *control theory* or as *reinforcement learning*. Presented next is a review of the literature relevant to each of these categories.

1) Control theory: There have been numerous research efforts in control theory to study dynamic decision-making problems given the system dynamics. For both stochastic and non-stochastic models, an agent can derive an optimal decision-making strategy offline using a dynamic programming (DP) decomposition of the problem [37, 38]. For systems with perfectly observed states, it is known that, at each instance of time, the agent's optimal action is simply a function of the state. Using this property in a DP facilitates the efficient computation of an optimal control strategy [39, 40]. In contrast, for systems with partially observed states, any optimal action is generally a function of the agent's entire memory of past observations and actions, which grows in size with time [41]. Subsequently, the domain of the optimal control strategy grows in size with time, and the corresponding DP decomposition of the problem requires a large number of computations for long time horizons [42]. This concern is alleviated in the literature for both stochastic and non-stochastic formulations by constructing an optimal DP decomposition for the problem using, instead of the memory, an *information state* that takes values in a time-invariant space [43].

The most commonly used information state in stochastic control is the *belief state*, i.e., a distribution on the state space conditioned on the agent's memory [44,45]. A general notion of information states for stochastic control was recently defined in [46].

For non-stochastic control problems, the DP decomposition has been simplified using two well known information states: (1) the *conditional range*, which is the set of feasible states at any time consistent with the agent's memory [47] and can be used in both terminal cost [48–51] and instantaneous cost problems [52–54]; and (2) the *maximum cost-to-come*, which is the maximum accrued cost at any time for each state in the conditional range [35] and can be used in additive cost problems [36, 55, 56]. The advantage of using information states is that in many applications they do not grow in size with time. Thus, they generally yield a more computationally efficient DP decomposition than the entire memory.

Information states have also been derived for mixed problems considering both stochastic and non-stochastic objectives in [57]. Robust stochastic formulations have been considered for perfectly observed problems in [26, 27, 58, 59] and for partially observed problems using variations of information states in [60–63]. Information states have also been derived for various forms of partially observed risk-sensitive control problems [16,35,36,64,65]. A complimentary approach is to view the worst-case control problem as a zero-sum games with two players, one being the agent and one being nature who acts as an adversary [27,30]. While this perspective has not been explicitly considered within dissertation, more details concerning this approach can be found in [30,66–68] and the references therein.

2) Reinforcement learning: The literature on reinforcement learning is concerned with decision-making when the agent may not have prior knowledge of the system's dynamics [69]. For systems with perfectly observed states, these problems have been addressed using a variety of approaches [70]. In the stochastic formulation, both model-based [71, 72] and model-free approaches [73] have been utilized. These include off-policy approaches such as Q-learning [74–77] and soft-actor critic methods [78,78], and on policy approaches [79] such as policy gradient methods [80–82] and actor-critic methods [83]. In the non-stochastic formulation, the worst-case reinforcement learning problem was formulated and analyzed in [84]. Worst-case Q-learning was proposed for reinforcement learning problems in [85–88] and extended to problems with output-feedback and partially known dynamics in [89]. Actor-critic methods [90] and model-based off-policy learning approaches [91] have also been developed for robust control. For perfectly observed systems with a robust stochastic formulation, robust Q-learning algorithms [92, 93] and robust policy gradient methods [93, 94] have been developed. Similarly, model-based approaches have also been developed for robust stochastic reinforcement learning [95]. Adaptive approaches for worst-case control of linear systems were presented in [96] and alternate non-stochastic adaptive algorithms in [97, 98]. However, in general, reinforcement learning is challenging when the agent can only access partial observations, since without knowledge of the system dynamics, the information state must be learned from data [99].

In the stochastic formulation, *approximate information states* was presented in [100] to address the challenges of control and learning with partial observations. Approximate information states can improve the computational tractability of control problems with large state spaces at the cost of a bounded loss in performance [101]. The explicit performance bounds of a finite-memory based approximate information state were derived in [102]. In reinforcement learning, approximate information states can be learned from output data and function as surrogate states to compute approximately optimal strategies, whose performance has been empirically validated in robotics [103] and medical care [104]. A related approach to accommodate partial observations in reinforcement learning is a *world model* [105] and a *predictive state representation* [106]. However, these do not provide any theoretical performance bounds and have only been validated empirically [107, 108].

1.2.2 Decentralized Control of a Team of Cooperative Agents

The framework of a team comprises of several agents who collaborate to achieve the same shared objective [109, 110]. The basic components of a team decision problem are [111]: (1) the number of agents $K \in \mathbb{N}$; (2) the decision of each agent; (3) the information available to each agent; (4) the communication pattern and restrictions for the agents; and (5) a common objective for all agents. Examples of team decision-making problems are systems with interdependent subsystems and limited or delayed communication, such as robot swarms [10], power systems [34], large hierarchical organizations [112], connected and automated vehicles [2] and vehicle platoons [113]. The communication restrictions among the agents in a team leads to decentralization of information where different agents have access to different information [114]. Generally, in a decentralized problem, no single agent has both: (1) access to all information in the system and (2) the ability to assign all actions. Thus, as each agent must select an action at any time using only locally available information, the optimal coordination of actions among agents is a challenging problem for teams. This is in sharp contrast to centralized problems with a single agent, where a key assumption is that the agent perfectly recalls all past control actions and observations. Note that in a manner similar to centralized problems, agents in decentralized problems also suffer from computational intractability resulting from partial observations.

The literature on decentralized control has predominantly focused on a stochastic formulation with the goal of minimizing the expected cost shared by all agents. The research originated with the study of static team problems where the information received by each agent is independent of the actions of all other agents [111, 115, 116]. The dynamic team problem or decentralized control problem considers the more general case where at least one agent's information is affected by the decisions of other agents [117, 118] and the agents select actions sequentially across multiple time steps. Dynamic teams are characterized by their *information structure*, which describes who knows what at any instance of time [119]. The information structure can designate the complexity of a problem [120] and have implications on the tractability of computing optimal decisions [121]. Different information structures can be classified as [19, 20]: *(1) Classical,* where all agents communicate perfectly and recall all information they have received [17, 122, 123]; *(2) Quasi-classical,* if agent 1 can affect the state of agent 2, and the information available to agent 1 is *also* available to agent 2 [124]; and *(3)*

Non-classical, where agents can affect each others' states with incomplete information [45,125]. Non-classical systems typically suffer from doubly exponential growth in computations required to generate optimal control strategies with an increase in the planning horizon [126]. Furthermore, the growth in the memory of each agent within the team makes it impractical to derive control strategies offline and to implement them online. Recall that in centralized control, this problem was alleviated using the notion of an information state which takes values in a time invariant space and thus, can be utilized to construct a DP decomposition of the problem. However, deriving such information states for decentralized problems with non-classical information structures is not trivial. The following approaches have been proposed in the literature to derive optimal control strategies using DP in problems with non-classical information structures:

1) The person-by-person approach aims to transform the problem into a centralized stochastic control problem from the point of view of a single agent. This is done by arbitrarily fixing the control policies for all agents except for one agent i. The control policy of the chosen agent is then optimized for this new problem, assuming the fixed strategies used by the other agents are known to agent i. This allows for the use of techniques from centralized stochastic control including derivation of structural results and DP. This process of fixing all other strategies and optimizing the individual strategy is then repeated for all agents until a stable equilibrium is attained. A control policy leading to this equilibrium is called a person-by-person optimal strategy. In general, this solution is not globally optimal, but it was proved in [127, 128] that a person-by-person strategy is globally optimal for quasi-classical information structures with linear dynamics, Gaussian disturbances and quadratic costs. An alternate approach is to use a Girasinov transformation on the probability measure to transform a dynamic team problem into a static team problem and solve for person-by-person optimality [129,130]. The person-by-person approach can be used to identify structural properties for globally optimal strategies, given the fact that every globally optimal strategy must necessarily be person-by-person optimal. Structural properties often guarantee the existence of an optimal control policy where all control strategies have a time-invariant domain. An early application of this approach was found in problems of real-time communication using encoders and decoders [131–135]. It has also been used in decentralized hypothesis testing and quickest detection problems [136, 137], broadcast information structures [138], in networked control systems [139], and team decision problems with partially nested information structures [140–142].

2) The designer's approach addresses decentralized control from the point of view of a designer with knowledge of the system model and statistics of all sources of randomness in the system. The designer's action at any instance of time is the optimal control law mapping the memory of each agent to their control actions at that instance of time. This is resolved by first transforming the problem into a centralized planning problem from the designer's point of view, and then using DP to derive the optimal strategy as a centralized open-loop problem. This approach yields a globally optimal solution, however, the DP involves solving an optimization problem over a space of control laws, which are functions, at each instance of time. This renders the designer's approach computationally challenging. This approach was first introduced in [143] for a centralized system with one agent and later extended to decentralized systems in [144] and [145]. The person-by-person has been used in conjunction with the designer's approach in real-time communication problems [132,146–148], in systems with a broadcast information structure [138], and in networked control systems [139].

3) The common information approach was formalized in problems with partial history sharing [45], where agents share a subset of their past observations and decisions to a shared memory accessible by all other agents. The solution is derived by reformulating the problem from the viewpoint of a coordinator with access only to the shared information (the common information), whose task is to provide prescriptions to each agent. At each instance of time, the prescription corresponding to any agent maps the private history of that agent's observations and decisions to their optimal action at that time. This yields a centralized DP decomposition from the perspective of the coordinator which involves optimizing over a space of partially evaluated functions, i.e., prescriptions. This approach has been used in problems with delayed information sharing [149], control sharing information structures [125], mean-field sharing information structures [150], belief sharing [118] and other non-classical information structures [151–153]. This has also been used in conjunction with the person-by-person approach to address teams with partially nested information [124], one-directional communication [154, 155] and unreliable communication [156, 157]. For such systems, current methods focus on identifying specific dynamics and information structures which yield computationally tractable solutions [125,155]. More recently, the common information approach has been used to also compress the private information of each agent [158] and to approximately compress both the common and private information [24].

4) The manager's approach and team approach are the most recent development in the theory of decentralized teams [20]. The manager's approach views the team from the perspective of a manager with access to the entire history of observation and control actions and which can assign control actions to the various agents. This yields a centralized DP decomposition that yields a globally optimal control strategy for each agent. Then, the team approach utilizes the information available at each agent to define an offline DP decomposition for each agent. It is shown that the strategies derived using the individual DP decomposition for each agent yields the same performance as that of the manager. This approach has also been extended to problems involving both learning and control [14].

While the literature in decentralized control described above has predominantly focused on minimizing the expected cost shared by all agents, an alternate formulation involves minimizing a given norm of a shared cost in linear time-invariant systems [140, 159, 160] for specific information structures. More recently, there has also been an interest in decentralized control problems with a worst-case cost. The common information approach has been extended to worst-case problems with a terminal cost [50], and has been used to analyse the worst-case pointwise performance of real-time communication [54].

1.3 Contributions of This Dissertation

Over the course of my Ph.D., I have primarily produced results in the two fields of study described in the literature review. In this section, I first explicitly summarize the contributions of this dissertation towards advancing the state of the art in each of these research areas. Subsequently, I also summarize the contributions to mechanism design for competing agents, for which the literature review is given in Chapter 4.

1.3.1 Centralized Worst-Case Control and Learning With Partial Observations

As highlighted in the literature review, worst-case decision for partially observed systems is computationally challenging because the default DP decomposition utilizes the memory, which grows in size with time. Thus, research efforts have primarily focused on deriving information states with time-invariant domains that yield a DP decomposition more tractable than a memory-based DP decomposition to compute an optimal control strategy. However, in problems with large state spaces, utilizing information states may not sufficiently simplify the DP to be practical [161, 162]. Furthermore, information states cannot be accurately determined in systems with unknown dynamics or unknown state-space models. Instead, they must be learned from output data. Both of these challenges were addressed in stochastic problems by introducing the notion of approximate information states [100] which can facilitate control and reinforcement learning with a bounded loss in expected performance. However, no analogous notions and results exist for either worst-case control or reinforcement learning problems.

Chapter 2 of this dissertation addresses this research gap by presenting a general non-stochastic theory of approximate information states. The theoretical development utilizes the mathematical framework of uncertain variables [163] with set-valued uncertainties, which distinguishes the presented theory from the corresponding work in the stochastic formulation. Non-stochastic approximate information states can be constructed from output data when the system dynamics are known and thus, they can facilitate computationally efficient control with bounded loss in worst-case performance. Furthermore, they can be learned from output data in problems with unknown dynamics, and thus, they provide a principled approach to reinforcement learning using partial observations. Specifically, the contributions of this dissertation within towards worst-case control and learning are: (1) the introduction of a general notion of *information states* which yields an optimal DP decomposition for worst-case control with both pointwise and additive performance criteria; (2) the introduction of the notion of *approximate information states* that can either be constructed from output variables or learned from output data; (3) the formulation of an approximate DP which computes a control strategy with a bounded loss of optimality; (4) the exposition of examples of approximate information states along with theoretical approximation bounds; (5) the extension of the notion of information states to infinite horizon problems with a discounted performance criterion; (6) an extension of the notion of approximate information states and derivation of performance bounds for infinite horizon problems in systems with perfectly observed costs, which facilitates partially observed reinforcement learning; and (7) the illustration of the proposed approach in various worst-case control and reinforcement learning problems for both finite and infinite time horizons using numerical examples. The contributions of Chapter 2 were previously published in the following articles:

- [164]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "Approximate information states for worst-case control of uncertain systems," in *Proceedings of the 61th IEEE Conference on Decision and Control (CDC)*, pages 4945–4950, 2022.

- [165]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "On robust control of partially observed uncertain systems with additive costs," in *Proceedings of the 2023 American Control Conference (ACC)*, pp. 4639-4644, 2023.

- [166]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "Approximate Information States for Worst-Case Control and Learning in Uncertain Systems," *arXiv:2301.05089*, 2023 (in review).

- [167]: **Aditya Dave**, Ioannis Faros, Nishanth Venkatesh, and Andreas A. Malikopoulos, "Worst-Case Control and Learning Using Partial Observations Over an Infinite Time Horizon," *arXiv:2303.16321*, 2023 (in review).

1.3.2 Decentralized Control of a Team of Cooperative Agents

The common information approach has been the default approach towards decentralized control of teams [45,50,168] for both stochastic and non-stochastic formulations. When the private information of all agents does not grow in size with time, this approach yields control strategies with time-invariant domains which improves the computational tractability of the resulting DP decomposition. However, teams with one-directional information sharing such as large hierarchical organizations [112], connected and automated vehicles [2] and vehicle platoons [113] may not satisfy the restriction on private information. Thus, the common information approach cannot be directly used to compute optimal control strategies for such teams. Research efforts on decision-making in such teams in the stochastic formulation have focused primarily on identifying specific information structures that can yield control strategies with time-invariant domains, and a corresponding DP decomposition [124,154–157,169]. Furthermore, in the non-stochastic formulation, information structures with one-directional communication have only been studied for real time communication problems [54]. Chapter 3 of this dissertation advances the state of the art on decentralized decision making in teams with one-directional information sharing for both stochastic and non-stochastic formulations.

A stochastic formulation is analyzed in Section 3.1. In this section, I consider a team of two agents and identify a general information structure, called *nested accessible information*. This information structure subsumes as special cases previously studied information structures, including instantaneous or delayed one-directional communication [154], partially nested systems [124,156], transmission of data using an unreliable communication channel [157], real time communication from one agent to another [155]; and decentralized control with unreliable communication among agents [169]. Then, for this general information structure, I derive results on the structure of optimal control strategies and present a DP decomposition which can improve the tractability of computing the optimal control strategy. The specific contributions of this section of Chapter 3 to the state of the art are: (1) the establishment of a structural form for optimal control strategies in systems with nested accessible information which restricts their domain to a space which does not grow in size with time; (2) a DP decomposition utilizing these structural results to obtain optimal control strategies; (3) a simplification of the structural results for teams with decoupled dynamics for each agent; and (4) an approximation scheme which can improve the computational tractability of the resulting DP decomposition. While I restrict my attention to a team of two agents to simplify the exposition, these results can be easily extended to systems with multiple agents which follow similar information structures. This contribution of Chapter 3 was previously published in the following article:

- [170]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "On decentralized control of two agents with nested accessible information," in *Proceedings of the 2022 American Control Conference (ACC)*, pages 3423–3430, 2022.

Next, I analyze a non-stochastic formulation in Section 3.2. Here, I consider an extension of the nested accessible information structure to a team of multiple agents organized within *nested subsystems* that seek to minimize a worst-case terminal cost. Teams with nested subsystems allow for one-directional communication amongst various subsystems and thus, they generalize the information structures previously studied for non-stochastic control of teams [50, 54]. The specific contributions of this section of Chapter 3 to the state of the art are: (1) the derivation of a structural form for optimal control strategies in decentralized worst-case problems with nested subsystems which restricts their domain to a space which does not grow in size with time; (2) a DP decomposition utilizing these structural results to obtain optimal control strategies; and (3) an extension of our results to worst-case additive cost problems. The effectiveness of these results is also illustrated using a numerical example. This contribution of Chapter 3 was previously published in the following article:

- [171]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "On decentralized minimax control with nested subsystems," in *Proceedings of the 2022 American Control Conference (ACC)*, pages 3437–3444. 2022.

In addition to the above two contributions, in the course of my Ph.D. I have also derived and published other results for decentralized stochastic control. I have excluded them from this dissertation because they are closely related to the ones included in Chapter 3. These contributions have been previously published in the following articles:

- [172]: **Aditya Dave** and Andreas A. Malikopoulos, "Decentralized stochastic control in partially nested information structures," in *IFAC-PapersOnLine*, 52(20):97–102, 2019.

- [173]: **Aditya Dave** and Andreas A. Malikopoulos, "Structural results for decentralized stochastic control with a word-of-mouth communication," in *Proceedings of the 2020 American Control Conference (ACC)*, pages 2796–2801. 2020.

- [174]: **Aditya Dave** and Andreas A. Malikopoulos, "A dynamic program for a team of two agents with nested information," in *Proceedings of the 2021 IEEE Conference on Decision and Control (CDC)*, pages 3768–3773. 2021.

- [175]: **Aditya Dave** and Andreas A. Malikopoulos, "The prescription approach for decentralized stochastic control with word-of-mouth communication," *arXiv preprint, arXiv:1907.12125*, 2021.

1.3.3 Mechanism Design for Competing Agents With Private Information

Mechanism design is the study of incentives to implement system-wide optimal solutions in problems involving multiple strategic agents (or alternatively, selfish agents) with conflicting interests, each with private information about preferences [176]. Typically, when strategic agents take actions to optimize their individual utilities, the net result is not optimal for the group as a whole [177]. Thus, a mechanism design approach is different from decentralized control with information asymmetry [19, 45] because agents may not follow actions recommended by a mechanism and need to be appropriately incentivized to do so [178]. The fact that mechanism design optimizes the behavior of competing agents has led to broad applications spanning different fields including economics, communication networks, and resource allocation problems [179–184].

Chapter 4 of this dissertation applies a mechanism design approach to a specific problem of how a democratic government can incentivize social media platforms to filter misinformation. In the formulation, social media platforms are strategic agents that seek to maximize user engagement at all costs and the government is a strategic agent who seeks to maximize the trust of the users in governmental institutions. In the model, the engagement of social media platforms decreases with an increase in misinformation filtering [9] whereas the trust in governmental institutions increases with more misinformation filtering [185–188]. Consequently, increasing filtering of misinformation by the social media platforms increases the utility of the government and the government is willing to make an investment to incentivize the social media platforms to filter misinformation. Chapter 4 proposes a mechanism to distribute this investment amongst the platforms optimally, and in return, implement an optimal level of filtering. The specific contributions of Chapter 4 are (1) a mechanism to incentivize social media platforms to filter misleading information, and (2) derivations of key properties of the proposed mechanism such as budget balance, voluntary participation of all agents, and implementation of a globally optimal level of filtering given a budget. These contributions have been previously published in the following article:

- [189]: **Aditya Dave**, Ioannis Vasileios Chremos, and Andreas A. Malikopoulos, "Social media and misleading information in a democracy: A mechanism design approach," *IEEE Transactions on Automatic Control*, 67(5), pp. 2633-2639, 2022.

This dissertation primarily includes the main contributions of my research. However, the publications listed below have also resulted from my research and collaborations with other colleagues:

- [190]: Nishanth Venkatesh, Viet-Anh Le, **Aditya Dave**, and Andreas A. Malikopoulos, "Connected and automated vehicles in mixed-traffic: Learning human driver behavior for effective on-ramp merging," *arxiv:2304.00397*, 2023 (in review).

- [191]: Ioannis Vasileios Chremos, Heeseung Bang, **Aditya Dave**, Viet-Anh Le, and Andreas A. Malikopoulos, "Modeling travel behavior in mobility systems with an atomic routing game and prospect theory," *arXiv:2303.17790*, 2023 (in review).

1.4 Dissertation Outline

This chapter presented a review of the relevant literature and identified the main contributions of this dissertation. Chapter 2 develops the theoretical framework of approximate information states for worst-case control and learning in partially observed systems. Then, Chapter 3 presents the analysis and results on the structural form of optimal control strategies for decentralized teams of agents with one-directional information sharing, in both stochastic and non-stochastic formulations. Subsequently, Chapter 4 formulates the misinformation filtering game and present a mechanism using which a democratic government can incentivize social media platforms to filter misinformation. Finally, Chapter 5 summarizes the main contributions of this dissertation and provides some directions for future research.

Chapter 2

CENTRALIZED WORST-CASE CONTROL AND LEARNING WITH PARTIAL OBSERVATIONS

2.1 Approximate Information States for Worst-Case Control and Learning in Uncertain Systems

2.1.1 Notation and Preliminaries

1) Uncertain Variables: We utilize the mathematical framework for *uncertain variables* from [163, 192] which was introduced for non-stochastic information theory. An uncertain variable is a non-stochastic analogue of a random variable with set-valued uncertainty. For a sample space Ω and a set $\mathcal{X}$, an uncertain variable is a mapping $X : \Omega \to \mathcal{X}$. For any $\omega \in \Omega$, it has the realization $X(\omega) = x \in \mathcal{X}$. The *marginal range* of X is the set $[[X]] := \{X(\omega) \mid \omega \in \Omega\}$. For two uncertain variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, their *joint range* is $[[X, Y]] := \{\big(X(\omega), Y(\omega)\big) \mid \omega \in \Omega\}$. For a given realization y of Y, the *conditional range* of X is $[[X|y]] := \{X(\omega) \mid Y(\omega) = y,\ \omega \in \Omega\}$ and, generally, $[[X|Y]] := \{[[X|y]] \mid y \in [[Y]]\}$.

2) Hausdorff Distance: Consider that the feasible sets $\mathcal{X}, \mathcal{Y}$ are nonempty subsets of a metric space $(\mathcal{S}, \eta)$, where $\eta(x, y)$ is the distance between any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, we define a distance between the two sets as follows.

Definition 1. The *Hausdorff distance* between $\mathcal{X}$ and $\mathcal{Y}$ is

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) := \max\left\{\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{Y}} \eta(x, y), \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} \eta(x, y)\right\}. \tag{2.1}$$

When the two sets $\mathcal{X}, \mathcal{Y}$ are bounded, the Hausdorff distance in (2.1) constitutes a pseudo-metric, i.e., $\mathcal{H}(\mathcal{X}, \mathcal{Y}) = 0$ if and only if $closure(\mathcal{X}) = closure(\mathcal{Y})$ [193, Appendix]. When both $\mathcal{X}, \mathcal{Y}$ are compact, the Hausdorff distance is a metric, i.e.,

$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = 0$ if and only if $\mathcal{X} = \mathcal{Y}$ [194, Chapter 1.12]. In both cases, the distance $\mathcal{H}$ satisfies the triangle inequality.

3) L-invertible Functions: Consider a function $f : \mathcal{X} \to \mathcal{Y}$. For any $y \in \mathcal{Y}$, the pre-image of the function is $f^{-1}(y) = \big\{x \in \mathcal{X} \mid f(x) = y\big\}$. Then, we use the Hausdorff distance to define the notion of an L-invertible function as follows.

Definition 2. A function $f : \mathcal{X} \to \mathcal{Y}$ is called *L-invertible* if there exists a constant $L_{f^{-1}} \in \mathbb{R}_{\geq 0}$ such that for all $y^1, y^2 \in \mathcal{Y}$:

$$\mathcal{H}\big(f^{-1}(y^1), f^{-1}(y^2)\big) \leq L_{f^{-1}} \cdot \eta(y^1, y^2). \tag{2.2}$$

For uncertain variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ such that $Y = f(X)$, the pre-image of f given a realization $y \in [[Y]]$ equals the conditional range $[[X|y]]$, i.e., $f^{-1}(y) = [[X|y]]$. Thus, if f is *L-invertible*, we equivalently state that for all $y^1, y^2 \in [[Y]]$:

$$\mathcal{H}\big([[X|y^1]], [[X|y^2]]\big) \leq L_{X|Y} \cdot \eta(y^1, y^2), \tag{2.3}$$

where $L_{X|Y} = L_{f^{-1}}$.

2.1.2 Problem Formulation

We consider an agent which seeks to control the trajectory of an uncertain system by selecting actions over $T \in \mathbb{N}$ discrete time steps. At each time $t = 0, \dots, T$, the agent receives an observation from the system, denoted by the uncertain variable $Y_t \in \mathcal{Y}_t$, and generates a control action denoted by the uncertain variable $U_t \in \mathcal{U}_t$. After generating the action at each t, the agent incurs a cost denoted by the uncertain variable $C_t \in \mathcal{C}_t \subset \mathbb{R}_{\geq 0}$. To account for the case that the agent may have no knowledge of a state-space model, we describe the system dynamics using an *input-output model*, as follows. At each $t = 0, \dots, T$, the system receives two inputs: the control action U_t, and an uncontrolled disturbance denoted by the uncertain variable $W_t \in \mathcal{W}_t$. We consider that the uncontrolled disturbances $\{W_t : t = 0, \dots, T\}$ constitute a sequence of independent uncertain variables. After receiving the inputs at each $t = 0, \dots, T$, the system generates two outputs:

$$Y_{t+1} = h_{t+1}(W_{0:t}, U_{0:t}), \tag{2.4}$$

$$C_t = d_t(W_{0:t}, U_{0:t}), \tag{2.5}$$

for some observation function $h_{t+1} : \prod_{\ell=0}^{t} \mathcal{W}_\ell \times \prod_{\ell=0}^{t} \mathcal{U}_\ell \to \mathcal{Y}_{t+1}$ and cost function $d_t : \prod_{\ell=0}^{t} \mathcal{W}_\ell \times \prod_{\ell=0}^{t} \mathcal{U}_\ell \to \mathcal{C}_t$. The initial observation is generated as $Y_0 = h_0(W_0)$.

The agent has perfect recall of the history of observations and control actions. The memory of the agent at each t is denoted by the uncertain variable $M_t := (Y_{0:t}, U_{0:t-1})$, which takes values in the set $\mathcal{M}_t := \prod_{\ell=0}^{t} \mathcal{Y}_\ell \times \prod_{\ell=0}^{t-1} \mathcal{U}_\ell$. The agent uses the memory M_t and a control law $g_t : \mathcal{M}_t \to \mathcal{U}_t$ at each t to generate the action $U_t = g_t(M_t)$. We denote the control strategy by $\boldsymbol{g} := (g_0, \dots, g_T)$ and the set of all feasible control strategies by $\mathcal{G}$. The performance of a strategy $\boldsymbol{g} \in \mathcal{G}$ is measured by the *worst-case or maximum instantaneous cost*

$$\mathcal{J}(\boldsymbol{g}) := \max_{t=0,\dots,T} \sup_{w_{0:t} \in [[W_{0:t}]]} C_t. \tag{2.6}$$

Problem 1. The optimization problem of the agent is to derive the control strategy $\boldsymbol{g} \in \mathcal{G}$ such that $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the marginal ranges $\{[[U_t]], [[W_t]], [[C_t]], [[Y_t]] \mid t = 0, \dots, T\}$ and the functions $\{h_t, d_t \mid t = 0, \dots, T\}$.

If there exists a strategy $\boldsymbol{g}^* \in \mathcal{G}$ that achieves the optimal performance in Problem 1, i.e., $\boldsymbol{g}^* = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, we refer to it as an optimal control strategy for Problem 1. Our aim is to tractably compute an optimal strategy if one exists. In our modeling framework, we impose the following assumptions:

Assumption 1. We consider that the sets $\{\mathcal{U}_t, \mathcal{W}_t, \mathcal{Y}_t \mid t = 0, \dots, T\}$ and $\{\mathcal{C}_t \mid t = 0, \dots, T\}$ are all bounded subsets of a metric space $(\mathcal{S}, \eta)$ and $\mathbb{R}_{\geq 0}$, respectively.

Assumption 1 allows for both continuous and finite valued feasible sets, while ensuring that the marginal range of each uncertain variable in the problem formulation is also bounded.

Assumption 2. The observation functions $\{h_t \mid t = 0, \dots, T\}$ of the system are both Lipschitz and L-invertible, whereas the cost functions $\{d_t \mid t = 0, \dots, T\}$ are Lipschitz continuous.

Assumption 2 is satisfied by a large class of observation functions, including: (1) all functions with compact domains and finite co-domains; and (2) bi-Lipschitz functions, like linear functions, with compact domains and compact co-domains (see Appendix A). We will require both assumptions in Section 2.1.4 when deriving the main results.

Remark 1. In our exposition, we also consider a special case of (2.6), called the *maximum terminal cost criterion*, given by

$$\mathcal{J}^{\mathrm{tm}}(\boldsymbol{g}) := \sup_{w_{0:T} \in [[W_{0:T}]]} C_T. \tag{2.7}$$

In addition to the general results for Problem 1, we often present results specifically for systems which utilize (2.7) as the performance measure. This serves two purposes: (1) the results are often easier to interpret for a terminal cost problem; and (2) these results can be extended to *additive cost problems*. We explicitly present this extension in Subsection 2.1.3.3.

Remark 2. We derive our results for Problem 1 with known dynamics. However, our main results in Section 2.1.4 can also be used in learning problems with unknown dynamics. We illustrate this application with an example in Subsection 2.1.5.2.

2.1.3 Dynamic Programs and Information States

In this section, we first present a memory-based DP decomposition for Problem 1 which computes the optimal value of the performance criterion 2.6. This will serve as a reference to analyze subsequent DPs in the paper. Then, we highlight the DP's computational challenges and present information states in Subsections 2.1.3.1 and 2.1.3.2 to alleviate them. In Subsection 2.1.3.3 we present examples of information states.

To arrive at the memory-based DP, we construct a “new” perfectly observed system whose state at each $t = 0, \dots, T$ is the memory M_t, which evolves as $M_{t+1} =$

(M_t, U_t, Y_{t+1}). Furthermore, for given realizations $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$, the maximum incurred cost at time t can be written as

$$\sup_{w_{0:t} \in [[W_{0:t}]]} C_t = \sup_{c_t \in [[C_t]]^{\boldsymbol{g}}} c_t = \sup_{m_t, u_t \in [[M_t, U_t]]^{\boldsymbol{g}}} \sup_{c_t \in [[C_t | m_t, u_t]]^{\boldsymbol{g}}} c_t, \tag{2.8}$$

for all $t = 0, \ldots, T$, where $[[C_t]]^{\boldsymbol{g}}$, $[[M_t, U_t]]^{\boldsymbol{g}}$ and $[[C_t|m_t, u_t]]^{\boldsymbol{g}}$ are the respective marginal ranges and the conditional range induced by strategy $\boldsymbol{g}$. Recall that $m_t = (y_{0:t},\, u_{0:t-1})$ and thus, we can expand the conditional range as

$$\begin{aligned}[[C_t|m_t, u_t]]^{\boldsymbol{g}} = \big\{c_t \in \mathcal{C}_t \;\big|\; & \exists\, w_{0:t} \in [[W_{0:t}]] \text{ such that } c_t = d_t(w_{0:t}, u_{0:t}), \\ & y_\ell = h_t(w_{0:\ell}, u_{0:\ell-1}),\ \forall \ell = 0, \ldots, t\big\} \\ = [[C_t|m_t, u_t]], & \end{aligned} \tag{2.9}$$

which shows that $[[C_t|m_t, u_t]]^{\boldsymbol{g}}$ is independent of the choice of strategy $\boldsymbol{g}$, hence we can drop $\boldsymbol{g}$. Next, we define $e_t(m_t, u_t) := \sup_{c_t \in [[C_t|m_t,u_t]]} c_t$, independent of $\boldsymbol{g}$, and state that

$$\begin{aligned}\sup_{m_t, u_t \in [[M_t, U_t]]^{\boldsymbol{g}}} \sup_{c_t \in [[C_t|m_t,u_t]]^{\boldsymbol{g}}} c_t &= \sup_{m_t, u_t \in [[M_t, U_t]]^{\boldsymbol{g}}} e_t(m_t, u_t) \\ &= \sup_{w_{0:t} \in [[W_{0:t}]]} e_t(M_t, U_t),\end{aligned} \tag{2.10}$$

where, in the second equality, note that the marginal range of external disturbances $[[W_{0:t}]]$ is independent of the strategy $\boldsymbol{g}$. Since $e_t(M_t, U_t)$ is a function of the new state M_t and control action U_t, it serves as an incurred cost at each $t = 0, \ldots, T$ in our new perfectly observed system [53]. The new instantaneous performance criterion is $\mathcal{E}(\boldsymbol{g}) := \max_{t=0,\ldots,T} \sup_{w_{0:t} \in [[W_{0:t}]]} e_t(M_t,\, U_t)$ and from (2.10), $\mathcal{E}(\boldsymbol{g}) = \mathcal{J}(\boldsymbol{g})$ for any $\boldsymbol{g}$. Subsequently, any strategy which achieves the optimal performance in the new system is optimal for Problem 1. If such an optimal strategy exists, we can compute it using a standard DP for perfectly observed systems, as follows. For all $t = 0, \ldots, T$, for each $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$ we recursively define the *value functions*

$$Q_t(m_t, u_t) := \max\Big\{ \sup_{c_t \in [[C_t|m_t,u_t]]} c_t,\ \sup_{m_{t+1} \in [[M_{t+1}|m_t,u_t]]} V_{t+1}(m_{t+1}) \Big\}, \tag{2.11}$$

$$V_t(m_t) := \inf_{u_t \in [[U_t]]} Q_t(m_t, u_t), \tag{2.12}$$

where $V_{T+1}(m_{T+1}) := 0$, identically. We define the extra value function V_{T+1} to ensure that the right hand side (RHS) of (2.11) is well defined at time T. Then, we can show using standard arguments [49,53] that the optimal value of Problem 1 is $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g}) = \sup_{m_0 \in [[M_0]]} V_0(m_0)$. Furthermore, at any $t = 0, \dots, T$, if there exists an action $u_t^* \in [[U_t]]$ which achieves the infimum in the RHS of (2.12), then $g_t^*(m_t) := \arg\min_{u_t \in [[U_t]]} Q_t(m_t, u_t)$ gives an optimal control law at time t. If the infimum is achieved at each t, the control strategy $\boldsymbol{g}^* = (g_0^*, \dots, g_T^*)$ is optimal for this system and Problem 1.

Remark 3. The DP (2.11) - (2.12) can be specialized to the terminal cost criterion (2.7) by defining for all $t = 0, \dots, T-1$,

$$Q_t^{\text{tm}}(m_t, u_t) := \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}^{\text{tm}}(m_{t+1}), \tag{2.13}$$

$$V_t^{\text{tm}}(m_t) := \inf_{u_t \in [[U_t]]} Q_t^{\text{tm}}(m_t, u_t), \tag{2.14}$$

where $Q_T^{\text{tm}}(m_T, u_T) := \sup_{c_T \in [[C_T | m_T, u_T]]} c_T$ and $V_T^{\text{tm}}(m_T) := \inf_{u_T \in [[U_t]]} Q_T^{\text{tm}}(m_T, u_T)$. We will use this terminal cost DP to simplify the exposition in Section 2.1.4.

Remark 4. A valid argument referring to the minimum of the RHS of (2.12) at each $t = 0, \dots, T$ is both a necessary and sufficient condition to ensure the existence of an optimal control strategy in Problem 1 [49, 195]. Consider that marginal ranges of all uncertain variables are *compact* rather than just bounded. From Assumption 2, the observation and cost functions at each t are Lipschitz. Using these properties in (2.13) - (2.14), we can show that the value functions are continuous and the conditional ranges are compact for all t, which implies that the minimum is achieved in the RHS of (2.12). Thus, compactness of all marginal ranges and Assumptions 1 - 2 constitute sufficient conditions for existence of an optimal solution to Problem 1, which is consistent with the conditions given in [196]. However, we continue using sup and inf in our exposition since we use only Assumptions 1 - 2 to establish our results without assuming compactness.

Remark 5. In the RHS of (2.12) at each t, we are required to solve an optimization for each $m_t \in [[M_t]]$. This is computationally challenging for longer horizons as the size of the set $[[M_t]]$ increases with time t with addition of new data. This concern motivates our search for an alternate DP decomposition which can derive an optimal control strategy while potentially achieving more favourable computational properties. We present such a DP decomposition in Subsection 2.1.3.1 by identifying an uncertain variable, called an *information state*, which can be used to generate an optimal control action at each time step instead of the memory.

2.1.3.1 Information States

In this subsection, we define information states for partially observed uncertain systems, use them in a DP decomposition, and prove it yields the optimal value for Problem 1.

Definition 3. An *information state* for Problem 1 at each $t = 0, \dots, T$ is an uncertain variable $\Pi_t = \sigma_t(M_t)$ taking values in a bounded set $\mathcal{P}_t$ and generated by a function $\sigma_t : \mathcal{M}_t \to \mathcal{P}_t$. Furthermore, for all $t = 0, \dots, T$, and for all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$, it satisfies the following properties:

1) *Sufficient to evaluate cost:*

$$\sup_{c_t \in [[C_t | m_t, u_t]]} c_t = \sup_{c_t \in [[C_t | \sigma_t(m_t), u_t]]} c_t. \tag{2.15}$$

2) *Sufficient to predict itself:*

$$[[\Pi_{t+1} | m_t, u_t]] = [[\Pi_{t+1} | \sigma_t(m_t), u_t]], \tag{2.16}$$

where both conditional ranges in (2.16) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

We can use the information states from Definition 3 directly in a DP, as follows. For all $t = 0, \dots, T$, for all $\pi_t \in [[\Pi_t]]$ and $u_t \in [[U_t]]$, we recursively define the value functions

$$\bar{Q}_t(\pi_t, u_t) := \max\Big\{ \sup_{c_t \in [[C_t | \pi_t, u_t]]} c_t, \sup_{\pi_{t+1} \in [[\Pi_{t+1} | \pi_t, u_t]]} \bar{V}_{t+1}(\pi_{t+1}) \Big\}, \tag{2.17}$$

$$\bar{V}_t(\pi_t) := \inf_{u_t \in [[U_t]]} \bar{Q}_t(\pi_t, u_t), \tag{2.18}$$

where $\bar{V}_{T+1}(\pi_{T+1}) := 0$ identically. If the minimum in the RHS of (2.18) exists at each $t = 0, \dots, T$, then this DP yields a control law at time t as $\bar{g}_t^*(\pi_t) := \arg\min_{u_t \in [[U_t]]} \bar{Q}_t(\pi_t, u_t)$. Next, we prove that the DP (2.17) - (2.18) computes the same value as the optimal DP (2.11) - (2.12).

Theorem 1. *Let $\Pi_t = \sigma_t(M_t)$ be an information state at any t. Then, for all t, and for all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$,*

$$Q_t(m_t, u_t) = \bar{Q}_t\big(\sigma_t(m_t), u_t\big) \text{ and } V_t(m_t) = \bar{V}_t\big(\sigma_t(m_t)\big). \tag{2.19}$$

Proof. Let $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$ be given realizations of M_t and U_t, respectively, for all $t = 0, \dots, T$. We prove the result by mathematical induction starting at the last time step. At time $T+1$, (2.19) holds trivially because $V_{T+1}(m_{T+1}) = \bar{V}_{T+1}(\sigma_{T+1}(m_{T+1})) = 0$. This forms the basis of our induction. Next, for any $t = 0, \dots, T$, we consider the induction hypothesis that $V_{t+1}(m_{t+1}) = \bar{V}_{t+1}(\sigma_{t+1}(m_{t+1}))$. Given the hypothesis, we first prove that $Q_t(m_t, u_t) = \bar{Q}_t(\sigma_t(m_t), u_t)$ by comparing the RHS of (2.11) to the RHS of (2.17) term by term. The first terms are equal by direct application of (2.15) from Definition 3. Next, we use the induction hypothesis for the second term in the RHS of (2.11), to state that

$$\begin{aligned}
\sup_{m_{t+1} \in [[M_{t+1}|m_t, u_t]]} V_{t+1}(m_{t+1}) &= \sup_{m_{t+1} \in [[M_{t+1}|m_t, u_t]]} \bar{V}_{t+1}(\sigma_{t+1}(m_{t+1})) \\
&= \sup_{\sigma_{t+1}(m_{t+1}) \in [[\Pi_{t+1}|\sigma_t(m_t), u_t]]} \bar{V}_{t+1}(\sigma_{t+1}(m_{t+1})),
\end{aligned} \tag{2.20}$$

where, in the second equality, we use the fact that $[[\Pi_{t+1}|m_t, u_t]] = \big\{\sigma_{t+1}(m_{t+1}) \in \mathcal{P}_{t+1} \big| m_{t+1} \in [[M_{t+1}|m_t, u_t]]\big\}$ and (2.16) from Definition 3. This establishes that the second term in the RHS of (2.11) equals the second term in the RHS of (2.17) and subsequently, that given the induction hypothesis for time $t+1$, we have $Q_t(m_t, u_t) = \bar{Q}_t(\sigma_t(m_t), u_t)$. Next, we minimize both sides of the equality with respect to $u_t \in [[U_t]]$, and use the definitions of the value functions in (2.12) and (2.18) to write that

$$V_t(m_t) = \inf_{u_t \in \mathcal{U}_t} Q_t(m_t, u_t) = \inf_{u_t \in \mathcal{U}_t} \bar{Q}_t\big(\sigma_t(m_t), u_t\big) = V_t\big(\sigma_t(m_t)\big), \tag{2.21}$$

which proves the induction hypothesis at time t. Thus, starting at time $T+1$, the result follows for all t using mathematical induction. □

Theorem 1 implies that (2.17) - (2.18) is an optimal DP decomposition for Problem 1, i.e., if an optimal strategy exists for this DP, it yields an optimal solution to Problem 1 as follows. Consider a control strategy $\bar{\boldsymbol{g}}^* = (\bar{g}_0^*, \dots, \bar{g}_T^*)$ computed using (2.17) - (2.18). We can construct a corresponding memory-based strategy $\boldsymbol{g} = (g_0, \dots, g_T)$ by defining $g_t(m_t) := \bar{g}_t^*(\sigma_t(m_t))$ for all $m_t \in [[M_t]]$ and $t = 0, \dots, T$. Then, using Theorem 1, we conclude that $\boldsymbol{g}$ achieves the infimum value at each t and thus, constitutes an optimal solution to Problem 1.

Remark 6. In practice, using an information state to construct the DP decomposition is useful computationally only if, for most time steps in $t = 0, \dots, T$, either the value functions in (2.17) - (2.18) have useful properties like concavity, or the set $\mathcal{P}_t$ is *smaller* than $\mathcal{M}_t$ for some measure of size. Potentially useful measures of sizes for sets include the number of elements, set diameter, and set dimension. We present some examples of information states for different systems in Subsection 2.1.3.3.

2.1.3.2 Alternate Characterization of Information States

When exploring whether an uncertain variable is a valid candidate to be considered an information state, it may be difficult to verify the second property (2.16) in Definition 3. In this subsection, we present two *stronger* conditions to replace (2.16). Specifically, at each $t = 0, \dots, T$, to establish that $\Pi_t = \sigma_t(M_t)$ is a valid information state, it is sufficient to satisfy the following conditions instead of (2.16):

1) State-like evolution: There exists a function $\bar{f}_t : \mathcal{P}_t \times \mathcal{U}_t \times \mathcal{Y}_{t+1} \to \mathcal{P}_{t+1}$, independent of the strategy $\boldsymbol{g}$, such that

$$\sigma_{t+1}(M_{t+1}) = \bar{f}_t(\sigma_t(M_t), U_t, Y_{t+1}). \tag{2.22}$$

2) Sufficient to predict observations: For all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}_t$,

$$[[Y_{t+1} | m_t, u_t]] = [[Y_{t+1} | \sigma_t(m_t), u_t]], \tag{2.23}$$

where both conditional ranges in (2.23) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

Next, we prove that these two conditions, in addition to (2.15) from Definition 3 are sufficient to identify an information state.

Lemma 1. *For all $t = 0, \dots, T$, if an uncertain variable $\Pi_t = \sigma_t(M_t)$ satisfies* (2.22) - (2.23)*, it also satisfies* (2.16)*.*

Proof. For all $t = 0, \dots, T$ and $m_t \in \mathcal{M}_t$, suppose that $\pi_t = \sigma_t(m_t)$ satisfy (2.22) - (2.23). Then, we substitute (2.22) into the left hand side (LHS) of (2.16) to state that

$$\begin{aligned}[[\Pi_{t+1} | m_t, u_t]] &= [[\bar{f}_t(\sigma_t(m_t), u_t, Y_{t+1}) \mid m_t, u_t]] \\ &= \big\{\bar{f}_t(\sigma_t(m_t), u_t, y_{t+1}) \in \mathcal{P}_{t+1} \bigm| y_{t+1} \in [[Y_{t+1}|m_t, u_t]]\big\},\end{aligned} \tag{2.24}$$

where, in the second equality, we write the conditional range as a set. Next, using (2.23) on the range of observations in the conditioning of (2.24), we can state that

$$\begin{aligned}&\big\{\bar{f}_t(\sigma_t(m_t), u_t, y_{t+1}) \in \mathcal{P}_{t+1} \bigm| y_{t+1} \in [[Y_{t+1}|m_t, u_t]]\big\} \\ &\quad = \big\{\bar{f}_t(\sigma_t(m_t), u_t, y_{t+1}) \in \mathcal{P}_{t+1} \bigm| y_{t+1} \in [[Y_{t+1}|\sigma_t(m_t), u_t]]\big\} \\ &\quad = [[\bar{f}_t(\sigma_t(m_t), u_t, Y_{t+1}) \in \mathcal{P}_{t+1} \mid \sigma_t(m_t), u_t]] \\ &\quad = [[\Pi_{t+1}|\sigma_t(m_t), u_t]],\end{aligned} \tag{2.25}$$

which is equal to the RHS of (2.16). □

2.1.3.3 Examples of Information States

In this subsection, we present examples of information states which satisfy the conditions in Definition 3 for systems with a given state-space model to describe their evolution. At each $t = 0, \dots, T$, let $\mathcal{X}_t$ be a known set of feasible states and let the system's state be denoted by an uncertain variable $X_t \in \mathcal{X}_t$. The agent's observation is given by $Y_t = h_t(X_t, N_t)$, where $N_t \in \mathcal{N}_t$ is a noise in observation, and the agent incurs a cost $C_t = d_t(X_t, U_t)$ when they implement an action $U_t \in \mathcal{U}_t$. Starting at $X_0 \in \mathcal{X}_0$, the state evolution is given by $X_{t+1} = f_t(X_t, U_t, W_t)$ for all t. Each uncertain variable in $\{X_0, W_t, N_t \mid t = 0, \dots, T\}$ is independent of all other uncertain variables in that set. Next, we present information states for different cases which may offer computational advantages over using the entire memory:

1) Systems with perfectly observed states: Consider that $Y_t = X_t$ for all $t = 0, \dots, T$. Then, an information state at each t is $\Pi_t = X_t$, i.e., the state itself [49]. It takes values in the set $\mathcal{X}_t$ and satisfies (2.15) - (2.16) for all t. Note that it is always computationally advantageous to construct a DP decomposition using the state at each time step instead of the entire memory of the agent.

2) Systems with partially observed states: Generally in a partially observed system with a known state space, an information state at each $t = 0, \dots, T$ is the conditional range $\Pi_t = [[X_t|M_t]]$, which is a set-valued uncertain variable [53]. Explicitly, for a given realization of the memory $m_t \in \mathcal{M}_t$ at time t, the conditional range takes the realization $P_t := \big\{x_t \in \mathcal{X}_t \bigm| \exists x_0 \in \mathcal{X}_0, w_{0:t-1} \in \prod_{\ell=0}^{t-1} \mathcal{W}_\ell,\ n_{0:t} \in \prod_{\ell=0}^{t} \mathcal{N}_\ell$ such that $y_t = h_t(x_t, n_t),\ x_{\ell+1} = f_\ell(x_\ell, u_\ell, w_\ell), y_\ell = h_\ell(x_\ell, n_\ell)$ for all $\ell = 0, \dots, t-1\big\}$. We denote the realization by P_t instead of π_t to highlight that it is a set. To establish that the conditional range is a valid information state, it is easier to verify the alternate conditions (2.22) and (2.23) instead of property (2.16) in Definition 3. Generally, it is computationally advantageous to construct a DP decomposition using the conditional range instead of the memory for systems with longer time horizons.

3) Systems with additive costs: Consider a system with partially observed states with an additive performance criterion $\mathcal{J}^{\mathrm{ad}}(\boldsymbol{g}) := \sup_{x_0, w_{0:T}, n_{0:T}} \sum_{t=0}^{T} d_t(X_t, U_t)$. We can construct a DP and an information state for an additive cost problem by recasting it as a terminal cost problem [49]. At $t = 0$, we define $A_0 := 0$ and for all $t = 1, \dots, T$, we recursively define an uncertain variable $A_t \in \mathcal{A}_t$ as $A_t := A_{t-1} + d_{t-1}(X_{t-1}, U_{t-1})$. Note that A_t tracks the cost incurred by the system up to time t, i.e., before the action U_t has been implemented. Then, at each t, we consider an augmented state for the system, $S_t = (X_t, A_t)$ and note that it evolves as $S_{t+1} = \big(f_t(X_t, U_t, W_t), A_t + d_t(X_t, U_t)\big)$. Furthermore, this augmentation yields a terminal cost problem with the cost $A_T + c_T(X_T, U_T)$. Thus, we can derive an optimal control strategy using the terminal cost DP and, as in case 2, an information state at each t is the conditional range $\Pi_t = [[X_t, A_t | M_t]]$. Generally, this information state is useful for systems with longer time horizons.

Remark 7. The conditions in Definition 3 can help us identify information states for systems with known dynamics and simplify the DP decomposition. However, many applications with large state spaces may require a further improvement in computational tractability, even at the cost of optimality. Moreover, in certain applications, we need to learn a representation of the information state using limited observations with incomplete knowledge of the dynamics. Information states are insufficient to account for these cases. Next, in Section 2.1.4, we introduce approximate information states that can address the above concerns.

2.1.4 Approximate Information States

In this section, we define approximate information states by relaxing the conditions given in Definition 3, and utilize them to develop an approximate DP decomposition which computes a sub-optimal control strategy for Problem 1. In Subsection 2.1.4.1, we derive the preliminary results required to establish useful properties of approximate information states. Then, in Subsection 2.1.4.2, we prove these properties, namely, the Lipischitz continuity of approximate value functions, and the following error bounds: (1) an upper bound on the error when the optimal value functions are estimated using approximate value functions, and (2) an upper bound on the loss in performance when control actions are generated using a sub-optimal control strategy instead of an optimal strategy.

Definition 4. An *approximate information state* for Problem 1 at each $t = 0, \dots, T$ is an uncertain variable $\hat{\Pi}_t = \hat{\sigma}_t(M_t)$ taking values in a bounded set $\hat{\mathcal{P}}_t$ and generated by an L-invertible function $\hat{\sigma}_t : \mathcal{M}_t \to \hat{\mathcal{P}}_t$. Furthermore, for all $t = 0, \dots, T$, there exist parameters $\epsilon_t, \delta_t, \lambda_t \in \mathbb{R}_{\geq 0}$ such that for all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$, it satisfies the properties:

1) Sufficient to approximate cost:

$$\Big| \sup_{c_t \in [[C_t | m_t, u_t]]} c_t - \sup_{c_t \in [[C_t | \hat{\sigma}_t(m_t), u_t]]} c_t \Big| \leq \epsilon_t. \tag{2.26}$$

2) Sufficient to approximate evolution: There exist the sets $\mathcal{K}_{t+1} := [[\hat{\Pi}_{t+1} \mid m_t, u_t]]$ and $\hat{\mathcal{K}}_{t+1} := [[\hat{\Pi}_{t+1} \mid \hat{\sigma}_t(m_t), u_t]]$ independent of the strategy $\boldsymbol{g}$, and it holds that

$$\mathcal{H}(\mathcal{K}_{t+1}, \hat{\mathcal{K}}_{t+1}) \leq \delta_t, \tag{2.27}$$

where recall that $\mathcal{H}$ is the Hausdorff distance in (2.1).

3) Lipschitz-like evolution: For all $\hat{\pi}_t^1, \hat{\pi}_t^2 \in [[\hat{\Pi}_t]]$,

$$\mathcal{H}\big([[\hat{\Pi}_{t+1} | \hat{\pi}_t^1, u_t]], [[\hat{\Pi}_{t+1} | \hat{\pi}_t^2, u_t]]\big) \leq \lambda_t \cdot \eta(\hat{\pi}_t^1, \hat{\pi}_t^2), \tag{2.28}$$

where η is an appropriate metric on $\hat{\mathcal{P}}_t$.

Using the approximate information state in Definition 4, we can construct a DP as follows. For all t, for all $\hat{\pi}_t \in [[\hat{\Pi}_t]]$ and $u_t \in [[U_t]]$, we recursively define the value functions

$$\hat{Q}_t(\hat{\pi}_t, u_t) := \max \Big\{ \sup_{c_t \in [[C_t | \hat{\pi}_t, u_t]]} c_t, \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\pi}_t, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}) \Big\}, \tag{2.29}$$

$$\hat{V}_t(\hat{\pi}_t) := \inf_{u_t \in [[U_t]]} \hat{Q}_t(\hat{\pi}_t, u_t), \tag{2.30}$$

where $\hat{V}_{T+1}(\hat{\pi}_{T+1}) := 0$ identically. If there exists a minimizing argument in the RHS of (2.30) at each $t = 0, \dots, T$, then $\hat{g}_t^*(\hat{\pi}_t) := \arg\min_{u_t \in \mathcal{U}_t} \hat{Q}_t(\hat{\pi}_t, u_t)$ constitutes an approximate control law at time t. Furthermore, we call $\hat{\boldsymbol{g}}^* = (\hat{g}_0^*, \dots, \hat{g}_T^*)$ an approximately optimal strategy for Problem 1. In Subsection 2.1.4.2, we derive performance guarantees on the approximate DP and control strategy.

Remark 8. As we showed in Section 2.1.3, we can specialize this DP for terminal cost problems, with the value functions for all $t = 0, \dots, T-1$ given by

$$\hat{Q}_t^{\text{tm}}(\hat{\pi}_t, u_t) := \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\pi}_t, u_t]]} \hat{V}_{t+1}^{\text{tm}}(\hat{\pi}_{t+1}), \tag{2.31}$$

$$\hat{V}_t^{\mathrm{tm}}(\hat{\pi}_t) := \inf_{u_t \in \mathcal{U}_t} \hat{Q}_t^{\mathrm{tm}}(\hat{\pi}_t, u_t), \tag{2.32}$$

and $\hat{Q}_T^{\mathrm{tm}}(\hat{\pi}_T, u_T) := \sup_{c_T \in [[C_T|\hat{\pi}_T, u_T]]} c_T$ and $\hat{V}_T^{\mathrm{tm}}(\hat{\pi}_T) := \inf_{u_T \in \mathcal{U}_T} \hat{Q}_T^{\mathrm{tm}}(\hat{\pi}_T, u_T)$ at time T.

Remark 9. The conditions in Definition 4 can be investigated using only output variables. Thus, an approximate information state can be learned from output data without knowledge of dynamics, as illustrated in Subsection 2.1.5.2.

2.1.4.1 Preliminary Results

In this subsection, we derive results necessary to prove the properties of the approximate DP in Subsection 2.1.4.2.

Lemma 2. *Consider three bounded subsets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ of a metric space $(\mathcal{S}, \eta)$. Let $X \in \mathcal{X}$, $Y \in \mathcal{Y}$ and $Z \in \mathcal{Z}$ be uncertain variables satisfying $Y = g(X)$, where $g : \mathcal{X} \to \mathcal{Y}$ is L-invertible, and $Z = h(X)$, where $h : \mathcal{X} \to \mathcal{Z}$ is Lipschitz. Then, there exists an $L_{Z|Y} \in \mathbb{R}_{\geq 0}$ such that:*

$$\mathcal{H}([[Z|y^1]],[[Z|y^2]]) \leq L_{Z|Y} \cdot \eta(y^1, y^2), \ \forall y^1, y^2 \in [[Y]]. \tag{2.33}$$

Proof. We prove the result by constructing a feasible constant $L_{Z|Y} \in \mathbb{R}_{\geq 0}$ which ensures that (2.33) is satisfied for all $y^1, y^2 \in [[Y]]$. We begin by using the definition of the Hausdorff distance in (2.1) to expand the LHS of (2.33) as

$$\mathcal{H}\big([[Z|y^1]], [[Z|y^2]]\big) = \max\Big\{ \sup_{x^1 \in g^{-1}(y^1)} \inf_{x^2 \in g^{-1}(y^2)} \eta\big(h(x^1), h(x^2)\big), \sup_{x^2 \in g^{-1}(y^2)} \inf_{x^1 \in g^{-1}(y^1)} \eta\big(h(x^1), h(x^2)\big)\Big\}, \tag{2.34}$$

where, note that $[[Z|y]] = \big\{z \in \mathcal{Z} \mid z = h(x), \forall x \in g^{-1}(y)\big\}$ for any realization $y \in [[Y]]$. Next, recall that h is Lipschitz continuous with a constant $L_h \in \mathbb{R}_{\geq 0}$. Substituting this property into the RHS of (2.34), we write that

$$\mathcal{H}\big([[Z|y^1]], [[Z|y^2]]\big)$$

$$\leq L_h \cdot \max\Big\{ \sup_{x^1 \in g^{-1}(y^1)} \inf_{x^2 \in g^{-1}(y^2)} \eta(x^1, x^2), \sup_{x^2 \in g^{-1}(y^2)} \inf_{x^1 \in g^{-1}(y^1)} \eta(x^1, x^2) \Big\}$$

$$= L_h \cdot \mathcal{H}\big(g^{-1}(y^1), g^{-1}(y^2)\big) = L_h \cdot L_{g^{-1}} \cdot \eta(y^1, y^2), \quad (2.35)$$

where, in the second equality, we use the L-invertibile property of g. Then, the result follows by selecting $L_{Z|Y} := L_h \cdot L_{g^{-1}}$. □

Lemma 3. *Consider a bounded set $\mathcal{X}$ and two functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$. Then,*

$$|\sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x)| \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)|, \quad (2.36)$$

$$|\inf_{x \in \mathcal{X}} f(x) - \inf_{x \in \mathcal{X}} g(x)| \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)|. \quad (2.37)$$

Proof. First, we prove (2.36) by considering two mutually exclusive cases which cover all possibilities. Case 1: We consider $\sup_{x \in \mathcal{X}} f(x) \geq \sup_{x \in \mathcal{X}} g(x)$, which implies $|\sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x)| = \sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x)$. For any infinitesimally small $\beta > 0$, we define $x(\beta) \in \mathcal{X}$ as an element which satisfies $f(x(\beta)) + \beta \geq \sup_{x \in \mathcal{X}} f(x)$. Then, $\sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x) \leq f(x(\beta)) + \beta - \sup_{x \in \mathcal{X}} g(x) \leq f(x(\beta)) + \beta - g(x(\beta)) \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)| + \beta$ for all $\beta > 0$. Therefore, $\sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x) \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. Case 2: $\sup_{x \in \mathcal{X}} f(x) < \sup_{x \in \mathcal{X}} g(x)$. The proof can be completed using similar arguments as in Case 1. Then, (2.37) follows from similar arguments as (2.36). □

Lemma 4. *For any four scalars $a, b, c, d \in \mathbb{R}$,*

$$|\max\{a, b\} - \max\{c, d\}| \leq \max\{|a - c|, |b - d|\}. \quad (2.38)$$

Proof. We prove this result by considering four cases which are mutually exclusive but cover all possibilities. Case 1: For $a \geq b$ and $c \geq d$: The result holds trivially. Case 2: For $a < b$ and $c \geq d$: The LHS can be expanded as $|\max\{a, b\} - \max\{c, d\}| = |b - c|$. Next, if $b \geq c$, we use $c \geq d$ to conclude that $|b - c| < |b - d|$, else if $c > b$, we use $b > a$ to conclude that $|c - b| < |c - a|$. Thus, $|\max\{a, b\} - \max\{c, d\}| \leq \max\{|a - c|, |b - d|\}$. Case 3: For $a < b$ and $c < d$: The result holds trivially. Case 4: For $a \geq b$ and $c < d$: The proof follows from the same sequence of arguments as Case 2. □

Lemma 5. *Consider two bounded subsets $\mathcal{A}, \mathcal{B}$ of a metric space $(\mathcal{X}, \eta)$. Let $f : \mathcal{X} \to \mathbb{R}$ be a bounded continuous function with a Lipschitz constant $L_f \in \mathbb{R}_{\geq 0}$ on $\mathcal{X}$. Then,*

$$\big|\sup_{a\in\mathcal{A}} f(a) - \sup_{b\in\mathcal{B}} f(b)\big| \leq L_f \cdot \mathcal{H}(\mathcal{A}, \mathcal{B}). \tag{2.39}$$

Proof. We prove this result by considering two cases which are mutually exclusive but cover all the possibilities. Case 1: $\sup_{a\in\mathcal{A}} f(a) \geq \sup_{b\in\mathcal{B}} f(b)$, which implies $|\sup_{a\in\mathcal{A}} f(a) - \sup_{b\in\mathcal{B}} f(b)| = \sup_{a\in\mathcal{A}} f(a) - \sup_{b\in\mathcal{B}} f(b)$. We define the non-empty set $\mathcal{A}^1(\beta) := \{a \in \mathcal{A} \mid f(a) + \beta \geq \sup_{b\in\mathcal{B}} f(b)\}$ for any infinitesimal $\beta > 0$. Then, $\sup_{a\in\mathcal{A}} f(a) - \sup_{b\in\mathcal{B}} f(b) \leq \sup_{a\in\mathcal{A}^1(\beta)} f(a) + \beta - \sup_{b\in\mathcal{B}} f(b) \leq \sup_{a\in\mathcal{A}^1(\beta)} \inf_{b\in\mathcal{B}}(f(a) - f(b)) + \beta \leq \sup_{a\in\mathcal{A}} \inf_{b\in\mathcal{B}} |f(a) - f(b)| + \beta \leq L_f \cdot \sup_{a\in\mathcal{A}} \inf_{b\in\mathcal{B}} \eta(a, b) + \beta$ for all $\beta > 0$. This implies that $|\sup_{a\in\mathcal{A}} f(a) - \sup_{b\in\mathcal{B}} f(b)| \leq L_f \cdot \sup_{a\in\mathcal{A}} \inf_{b\in\mathcal{B}} \eta(a, b) \leq L_f \cdot \mathcal{H}(\mathcal{A}, \mathcal{B})$, where, in the second inequality, we invoke the definition of the Hausdorff distance in (2.1) to complete the proof. Case 2: $\sup_{a\in\mathcal{A}} f(a) < \sup_{b\in\mathcal{B}} f(b)$ and we can prove the result using the same sequence of arguments as case 1. □

As a direct consequence of Lemma 5, we can also establish the following property. Consider two bounded subsets $\mathcal{Y}, \mathcal{Z}$ of $\mathbb{R}^n$, $n \in \mathbb{N}$. For two uncertain variables $Y \in \mathcal{Y}$ and $Z \in \mathcal{Z}$, let the conditional range $[[Z|y]]$ satisfy $\mathcal{H}\big([[Z|y^1]], [[Z|y^2]]\big) \leq L_{Z|Y} \cdot \eta(y^1, y^2)$ for all realizations $y^1, y^2 \in \mathcal{Y}$ of Y. Then, for a continuous function $f : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ with a Lipschitz constant L_f, we can use (2.39) from Lemma 5 to state that for all $y^1, y^2 \in [[Y]]$:

$$\Big|\sup_{z^1\in[[Z|y_1]]} f(z^1) - \sup_{z^2\in[[Z|y_2]]} f(z^2)\Big| \leq L_{Z|Y} \cdot L_f \cdot \eta(y^1, y^2). \tag{2.40}$$

2.1.4.2 Properties of Approximate Information States

In this subsection, we present several properties of the approximate DP (2.29) - (2.30). To begin, we prove in Theorem 2 that each approximate value function is Lipschitz continuous. This property subsequently allows us to establish error bounds.

Theorem 2. *In the approximate DP (2.29) - (2.30), the value functions $\hat{Q}_t(\hat{\pi}_t, u_t)$ and $\hat{V}_t(\hat{\pi}_t)$ are Lipschitz continuous with respect to $\hat{\pi}_t \in [[\hat{\Pi}_t]]$ for all $u_t \in [[U_t]]$ and $t = 0, \dots, T$.*

Proof. We prove the Lipschitz continuity of the value functions by constructing a valid candidate for the Lipschitz constant $L_{\hat{V}_t}$ at each $t = 0, \dots, T$, using mathematical induction. At time $T+1$, recall that $\hat{V}_{T+1}(\hat{\pi}_{T+1}) = 0$ identically and thus, $\hat{V}_{T+1}(\hat{\pi}_{T+1})$ is trivially Lipschitz continuous with a constant $L_{\hat{V}_{T+1}} = 0$. This forms the basis of our induction. Then, at each $t = 0, \dots, T$, we consider the induction hypothesis that $\hat{Q}_{t+1}(\hat{\pi}_{t+1}, u_{t+1})$ and $\hat{V}_{t+1}(\hat{\pi}_{t+1})$ are Lipschitz continuous with respect to $\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1}]]$ for all $u_{t+1} \in [[U_{t+1}]]$, and denote the constant by $L_{\hat{V}_{t+1}} \in \mathbb{R}_{\geq 0}$.

At time t, we first prove the result for the value function $\hat{Q}_t(\hat{\pi}_t, u_t)$. Let $\hat{\pi}_t^1, \hat{\pi}_t^2 \in [[\hat{\Pi}_t]]$ be two possible realizations of $\hat{\Pi}_t$. Then, using the definition (2.29) of $\hat{Q}_t(\hat{\pi}_t, u_t)$ and (2.38) from Lemma 4, we state that

$$|\hat{Q}_t(\hat{\pi}_t^1, u_t) - \hat{Q}_t(\hat{\pi}_t^2, u_t)| \leq \max \Bigg\{ \Bigg| \sup_{c_t^1 \in [[C_t | \hat{\pi}_t^1, u_t]]} c_t^1 - \sup_{c_t^2 \in [[C_t | \hat{\pi}_t^2, u_t]]} c_t^2 \Bigg|, \\ \Bigg| \sup_{\hat{\pi}_{t+1}^1 \in [[\hat{\Pi}_{t+1} | \hat{\pi}_t^1, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}^1) - \sup_{\hat{\pi}_{t+1}^2 \in [[\hat{\Pi}_{t+1} | \hat{\pi}_t^2, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}^2) \Bigg| \Bigg\}. \tag{2.41}$$

We consider the RHS of (2.41) term by term. In the first term, we note that for all $\hat{\pi}_t \in [[\hat{\Pi}_t]]$,

$$\sup_{c_t \in [[C_t | \hat{\pi}_t, u_t]]} c_t = \sup_{m_t \in [[M_t | \hat{\pi}_t]]} \Bigg(\sup_{c_t \in [[C_t | m_t, u_t]]} c_t \Bigg). \tag{2.42}$$

In the RHS of (2.42), recall from Assumption 2 that the uncertain variable C_t is a Lipschitz function of $(W_{0:t}, U_{0:t})$, and (M_t, U_t) is an L-invertible function of $(W_{0:t}, U_{0:t})$. Thus, using (2.33) from Lemma 2, there exists a constant $L_{C|M,U}$ such that

$$\mathcal{H}([[C_t | m_t^1, u_t]], [[C_t | m_t^2, u_t]]) \leq L_{M|C,U} \cdot \eta(m_t^1, m_t^2), \tag{2.43}$$

for all $m^1, m^2 \in [[M_t]]$. Furthermore, we use (2.40) to state that

$$\Bigg| \sup_{c_t^1 \in [[C_t | m_t^1, u_t]]} c_t^1 - \sup_{c_t^2 \in [[C_t | m_t^2, u_t]]} c_t^2 \Bigg| \leq L_{M|C,U} \cdot L_{c_t} \cdot \eta(m_t^1, m_t^2). \tag{2.44}$$

Then, consider a function $e_t : \mathcal{M}_t \times \mathcal{U}_t \to \mathbb{R}_{\geq 0}$ defined as $e_t(m_t, u_t) := \sup_{c_t \in [[C_t | m_t, u_t]]} c_t$. As a direct consequence of (2.44), e_t is Lipschitz continuous with respect to m_t with a constant $L_{e_t} := L_{M|C,U} \cdot L_{c_t}$. Using (2.42) and the definition of e_t in the first term in the RHS of (2.41),

$$\Big| \sup_{c_t^1 \in [[C_t|\hat{\pi}_t^1, u_t]]} c_t^1 - \sup_{c_t^2 \in [[C_t|\hat{\pi}_t^2, u_t]]} c_t^2 \Big| \\ = \Big| \sup_{m_t^1 \in [[M_t|\hat{\pi}_t^1]]} e_t(m_t^1, u_t) - \sup_{m_t^2 \in [[M_t|\hat{\pi}_t^2]]} e_t(m_t^2, u_t) \Big|. \quad (2.45)$$

In (2.45), recall that the uncertain variable $\hat{\Pi}_t$ is an L-invertible function of M_t and thus, the conditional range $[[M_t|\hat{\pi}_t]]$ satisfies (2.3). Then, we use (2.40) once more to state that

$$\Big| \sup_{c_t^1 \in [[C_t|\hat{\pi}_t^1, u_t]]} c_t^1 - \sup_{c_t^2 \in [[C_t|\hat{\pi}_t^2, u_t]]} c_t^2 \Big| \leq L_{M_t|\hat{\Pi}_t} \cdot L_{e_t} \cdot \eta(\hat{\pi}_t^1, \hat{\pi}_t^2). \quad (2.46)$$

In the second term in the RHS of (2.41), we use the induction hypothesis and (2.39) from Lemma 5 to write that

$$\Big| \sup_{\hat{\pi}_{t+1}^1 \in [[\hat{\Pi}_{t+1}|\hat{\pi}_t^1, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}^1) - \sup_{\hat{\pi}_{t+1}^2 \in [[\hat{\Pi}_{t+1}|\hat{\pi}_t^2, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}^2) \Big| \\ \leq L_{\hat{V}_{t+1}} \cdot \mathcal{H}\big([[\hat{\Pi}_{t+1}|\hat{\pi}_t^1, u_t]], [[\hat{\Pi}_{t+1}|\hat{\pi}_t^2, u_t]]\big) \\ \leq L_{\hat{V}_{t+1}} \cdot \lambda_t \cdot \eta(\hat{\pi}_t^1, \hat{\pi}_t^2), \quad (2.47)$$

where, in the second inequality, we use the third property (2.28) of approximate information states in Definition 4. Then, the proof for $\hat{Q}_t(\hat{\pi}_t, u_t)$ is complete by substituting (2.46) and (2.47) into the RHS of (2.41) and defining $L_{\hat{Q}_t} := \max\big\{L_{M_t|\hat{\Pi}_t} \cdot L_{e_t}, L_{\hat{V}_{t+1}} \cdot \lambda_t\big\}$. To prove the result for $\hat{V}_t(\hat{\pi}_t)$, we use (2.37) from Lemma 3 to state that

$$\big|\hat{V}_t(\hat{\pi}_t^1) - \hat{V}_t(\hat{\pi}_t^2)\big| = \big| \inf_{u_t \in [[U_t]]} \hat{Q}_t(\hat{\pi}_t^1, u_t) - \inf_{u_t \in [[U_t]]} \hat{Q}_t(\hat{\pi}_t^2, u_t) \big| \\ \leq \sup_{u_t \in [[U_t]]} \big|\hat{Q}_t(\hat{\pi}_t^1, u_t) - \hat{Q}_t(\hat{\pi}_t^2, u_t)\big| \leq L_{\hat{Q}_t} \cdot \eta(\hat{\pi}_t^1, \hat{\pi}_t^2), \quad (2.48)$$

which proves the induction hypothesis at time t. Thus, the result holds using mathematical induction. □

Next, we establish an upper bound on the approximation error when the value functions of the optimal DP (2.11) - (2.12) are estimated using the approximate DP (2.29) - (2.30) at each t.

Theorem 3. *Let $L_{\hat{V}_{t+1}}$ be the Lipschitz constant of $\hat{V}_{t+1}$ for all $t = 0, \dots, T$. Then, for all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$,*

$$|Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| \leq \alpha_t, \tag{2.49}$$

$$|V_t(m_t) - \hat{V}_t(\hat{\sigma}_t(m_t))| \leq \alpha_t, \tag{2.50}$$

where $\alpha_t = \max(\epsilon_t, \alpha_{t+1} + L_{\hat{V}_{t+1}} \cdot \delta_t)$ for all $t = 0, \dots, T$ and $\alpha_{T+1} = 0$.

Proof. For all $t = 0, \dots, T$, let $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$ be realizations of M_t and U_t, respectively. We prove both results by mathematical induction, starting with time step $T + 1$. At $T + 1$, by definition, $V_{T+1}(m_{T+1}, u_{T+1}) = V_{T+1}(\hat{\sigma}_{T+1}(m_{T+1})) = 0$. This forms the basis of our mathematical induction. Then, at each $t = 0, \dots, T$, we consider the induction hypothesis $|V_{t+1}(m_{t+1}) - \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1}))| \leq \alpha_{t+1}$. At time t, we first prove (2.49). Using (2.38) from Lemma 4 in the LHS of (2.49) to state that

$$\begin{aligned}|Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| \leq \max \Bigg\{ &\Bigg| \sup_{c_t \in [[C_t | m_t, u_t]]} c_t - \sup_{c_t \in [[C_t | \hat{\sigma}_t(m_t), u_t]]} c_t \Bigg|, \\ &\Bigg| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}(m_{t+1}) - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}) \Bigg| \Bigg\}. \end{aligned} \tag{2.51}$$

We consider the RHS of (2.51) term-by-term. By direct application of (2.26) in Definition 4, the first term in the RHS satisfies

$$\Bigg| \sup_{c_t \in [[C_t | m_t, u_t]]} c_t - \sup_{c_t \in [[C_t | \hat{\sigma}_t(m_t), u_t]]} c_t \Bigg| \leq \epsilon_t. \tag{2.52}$$

For the second term in the RHS of (2.51), we use the triangle inequality to write that

$$\begin{aligned} &\Bigg| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}(m_{t+1}) - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}) \Bigg| \\ &\quad \leq \Bigg| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}(m_{t+1}) - \sup_{\hat{\sigma}_{t+1}(m_{t+1}) \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) \Bigg| \\ &\qquad + \Bigg| \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}) - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}) \Bigg|. \end{aligned} \tag{2.53}$$

For the first term in the RHS of (2.53), we first note that

$$\sup_{\hat{\sigma}_{t+1}(m_{t+1}) \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) = \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) \tag{2.54}$$

because $[[\hat{\Pi}_{t+1} \mid m_t, u_t]] = \{\hat{\sigma}_{t+1}(m_{t+1}) \in \hat{\mathcal{P}}_t \mid m_{t+1} \in [[M_{t+1} \mid m_t, u_t]]\}$. Then, we can state that

$$\Big| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}(m_{t+1}) - \sup_{\hat{\sigma}_{t+1}(m_{t+1}) \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1}))\Big| \\ \leq \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \big|V_{t+1}(m_{t+1}) - \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1}))\big| \leq \alpha_{t+1}, \quad (2.55)$$

where, in the first inequality, we use (2.36) from Lemma 3; and, in the second inequality, we use the induction hypothesis for time $t+1$. Using (2.39) from Lemma 5 and (2.27) from Definition 4, the second term in the RHS of (2.53) satisfies

$$\Big| \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1}) - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1})\Big| \leq L_{\hat{V}_{t+1}} \cdot \delta_t. \quad (2.56)$$

Substituting the respective inequalities for each term in the RHS of (2.53) yields $\big| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}(m_{t+1}) - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{V}_{t+1}(\hat{\pi}_{t+1})\big| \leq \alpha_{t+1} + L_{\hat{V}_{t+1}} \cdot \delta_t$. We complete the proof for (2.49) by substituting the inequalities in the RHS of (2.52) and (2.53) into the RHS of (2.51). Next, we prove (2.50) at time t. Using the definition of the value functions in the LHS of (2.50), we write that

$$\begin{aligned} |V_t(m_t) - \hat{V}_t(\hat{\sigma}_t(m_t))| &= \Big| \inf_{u_t \in [[U_t]]} Q_t(m_t, u_t) - \inf_{u_t \in [[U_t]]} \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)\Big| \\ &\leq \sup_{u_t \in [[U_t]]} |Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| \\ &\leq \max\{\epsilon_t, \alpha_{t+1} + L_{\hat{V}_{t+1}} \cdot \delta_t\}, \end{aligned} \quad (2.57)$$

where in the first inequality, we use (2.37) from Lemma 3; and in the second inequality, we use (2.49). Thus, the results hold for all $t = 0, \dots, T$ using mathematical induction.

□

After bounding the approximation error for value functions, we also seek to bound the maximum performance loss in the implementation of an approximately optimal strategy. Consider an approximate strategy $\hat{\boldsymbol{g}}^* := (\hat{g}_0^*, \dots, \hat{g}_T^*)$ computed using (2.29) - (2.30), where $\hat{g}_t^*(\hat{\pi}_t) = \arg\min_{u_t \in [[U_t]]} \hat{Q}_t(\hat{\pi}_t, u_t)$ for all $t = 0, \dots, T$. We can construct an approximate memory-based strategy $\boldsymbol{g}^{\text{ap}} = (g_0^{\text{ap}}, \dots, g_T^{\text{ap}})$ by selecting the control law $g_t^{\mathrm{ap}}(m_t) := \hat{g}_t^*(\hat{\sigma}_t(m_t))$ for all $t = 0, \dots, T$. Note that $\boldsymbol{g}^{\mathrm{ap}}$ is equivalent to $\hat{\boldsymbol{g}}^*$ because they generate the same actions at each t and subsequently, yield the same performance. Thus, we evaluate the performance of $\boldsymbol{g}^{\mathrm{ap}}$ to determine the quality of approximation. To this end, for all $t = 0, \dots, T$, for all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$, we define

$$\Theta_t(m_t, u_t) := \max\Big\{ \sup_{c_t \in [[C_t | m_t, u_t]]} c_t, \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \Lambda_{t+1}(m_{t+1}) \Big\}, \tag{2.58}$$

$$\Lambda_t(m_t) := \Theta_t(m_t, g_t^{\mathrm{ap}}(m_t)), \tag{2.59}$$

where $\Lambda_{T+1}(m_{T+1}) := 0$, identically. Then, the performance of the memory-based approximate strategy $\boldsymbol{g}^{\mathrm{ap}}$ is $\Lambda_0(m_0)$. In contrast, recall that the performance of an optimal strategy $\boldsymbol{g}^*$ is the optimal value $V_0(m_0)$ computed using (2.11) - (2.12). Next, we bound the difference in performance between $\boldsymbol{g}^{\mathrm{ap}}$ and $\boldsymbol{g}^*$.

Theorem 4. *Let $L_{\hat{V}_{t+1}}$ be the Lipschitz constant of $\hat{V}_{t+1}$ for all $t = 0, \dots, T$. Then, for all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$,*

$$|Q_t(m_t, u_t) - \Theta_t(m_t, u_t)| \leq 2\alpha_t, \tag{2.60}$$

$$|V_t(m_t) - \Lambda_t(m_t)| \leq 2\alpha_t. \tag{2.61}$$

where $\alpha_t = \max(\epsilon_t, \alpha_{t+1} + L_{\hat{V}_{t+1}} \cdot \delta_t)$ for all $t = 0, \dots, T$ and $\alpha_{T+1} = 0$.

Proof. We begin by recursively defining the value functions that compute the performance of the strategy $\hat{\boldsymbol{g}}$. For all $t = 0, \dots, T$ and for each $\hat{\pi}_t \in [[\hat{\Pi}_t]]$ and $u_t \in [[U_t]]$, let

$$\hat{\Theta}_t(\hat{\pi}_t, u_t) := \max\Big\{ \sup_{c_t \in [[C_t | \hat{\pi}_t, u_t]]} c_t, \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\pi}_t, u_t]]} \hat{\Lambda}_{t+1}(\hat{\pi}_{t+1}) \Big\}, \tag{2.62}$$

$$\hat{\Lambda}_t(\hat{\pi}_t) := \hat{\Theta}_t(\hat{\pi}_t, \hat{g}_t(\hat{\pi}_t)), \tag{2.63}$$

where $\hat{\Lambda}_{T+1}(\hat{\pi}_{T+1}) := 0$, identically. Note that

$$\hat{\Theta}_t(\hat{\pi}_t, u_t) = \hat{Q}_t(\hat{\pi}_t, u_t) \quad \text{and} \quad \hat{\Lambda}_t(\hat{\pi}_t) = \hat{V}_t(\hat{\pi}_t), \tag{2.64}$$

for all $t = 0, \dots, T$, since $\hat{g}_t(\hat{\pi}_t) = \arg\min_{u_t \in \mathcal{U}_t} \hat{Q}_t(\hat{\pi}_t, u_t)$.

We first prove (2.60) for all $t = 0, \dots, T$. At time t, using the triangle inequality and (2.64) in the LHS of (2.60):

$$\begin{aligned} |Q_t(m_t, u_t) - \Theta_t(m_t, u_t)| \leq & |Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| + |\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)| \\ \leq & \alpha_t + |\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)|, \end{aligned} \tag{2.65}$$

where, in the second inequality, we use (2.49) from Theorem 3. Then, to prove (2.60), it suffices to show that

$$|\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)| \leq \alpha_t. \tag{2.66}$$

Then, we use mathematical induction starting at time $T+1$ to prove (2.66) in addition to $|\hat{\Lambda}_t(\hat{\sigma}_t(m_t)) - \Lambda_t(m_t)| \leq \alpha_t$ for all $t = 0, \dots, T$. At time $T+1$, using the definitions it holds that $\hat{\Lambda}_{T+1}(\hat{\sigma}_{T+1}(m_{T+1})) = \Lambda_{T+1}(m_{T+1}) = 0$. This forms the basis of our induction. Next, for all $t = 0, \dots, T$, we consider the induction hypothesis that

$$|\hat{\Lambda}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) - \Lambda_{t+1}(m_{t+1})| \leq \alpha_{t+1}. \tag{2.67}$$

Then, using the definitions of the value functions in (2.58) and (2.62) in addition to (2.38) from Lemma 4 in the LHS of (2.66):

$$\begin{aligned} & |\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)| \\ & \quad \leq \max \Bigg\{ \Bigg| \sup_{c_t \in [[C_t | m_t, u_t]]} c_t - \sup_{c_t \in [[C_t | \hat{\sigma}_t(m_t), u_t]]} c_t \Bigg|, \Bigg| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \Lambda_{t+1}(m_{t+1}) \\ & \qquad \qquad - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{\Lambda}_{t+1}(\hat{\pi}_{t+1}) \Bigg| \Bigg\}. \end{aligned} \tag{2.68}$$

We consider the RHS of (2.68) term by term. By direct application of (2.26) in Definition 4, the first term in the RHS satisfies

$$\Bigg| \sup_{c_t \in [[C_t | m_t, u_t]]} c_t - \sup_{c_t \in [[C_t | \hat{\sigma}_t(m_t), u_t]]} c_t \Bigg| \leq \epsilon_t. \tag{2.69}$$

For the second term in the RHS of (2.68), using the triangle inequality and the fact that $\sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{\Lambda}_{t+1}(\hat{\pi}_{t+1}) = \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \hat{\Lambda}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1}))$, we write that

$$
\begin{aligned}
&|\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)| \\
&\quad \leq \Big| \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | \hat{\sigma}_t(m_t), u_t]]} \hat{\Lambda}_{t+1}(\hat{\pi}_{t+1}) - \sup_{\hat{\pi}_{t+1} \in [[\hat{\Pi}_{t+1} | m_t, u_t]]} \hat{\Lambda}_{t+1}(\hat{\pi}_{t+1}) \Big| \\
&\qquad + \Big| \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \Lambda_{t+1}(m_{t+1}) - \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \hat{\Lambda}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) \Big| \\
&\hspace{8cm} \leq L_{\hat{V}_{t+1}} \cdot \delta_t + \alpha_{t+1}, \quad (2.70)
\end{aligned}
$$

where, in the second inequality, the first term is upper bounded using (2.39) from Lemma 5 and then using (2.27) from Definition 4, whereas the second term is bounded using (2.36) from Lemma 3 and the induction hypothesis. Thus, given the induction hypothesis, (2.66) can be proved by substitution of (2.69) and (2.70) into the RHS of (2.68). Next, using the definitions of the value functions from (2.59) and (2.63), we write that

$$
\begin{aligned}
|\hat{\Lambda}_t(\hat{\sigma}_t(m_t)) - \Lambda_t(m_t)| &= |\hat{\Theta}_t(\hat{\sigma}_t(m_t), \hat{g}_t(\hat{\sigma}_t(m_t)) - \Theta_t(m_t, g_t(m_t))| \\
&= |\hat{\Theta}_t(\hat{\sigma}_t(m_t), \hat{u}_t) - \Theta_t(m_t, \hat{u}_t)| \leq \alpha_t, \quad (2.71)
\end{aligned}
$$

where, in the second equality, we use the definition of the control law to write that $g_t(m_t) = \hat{g}_t(\hat{\sigma}_t(m_t)) =: \hat{u}_t$; and in the inequality, we use (2.66). This proves the induction hypothesis for time t given the hypothesis for time $t+1$. Thus, using mathematical induction (2.66) holds for all $t = 0, \dots, T$. Subsequently, we complete the proof for (2.60) for all $t = 0, \dots, T$ by substituting (2.66) into the RHS of (2.65). Furthermore, note that (2.61) follows directly from (2.60) using the same sequence of arguments used to prove (2.71). □

Remark 10. We can specialize the results of both Theorem 3 and Theorem 4 to terminal cost problems, where the optimal DP is given by (2.13) - (2.14) and the approximate DP is given by (2.31) - (2.32). The approximation bounds in both theorems hold for terminal cost problems with a recursively defined constant $\alpha_t := \alpha_{t+1} + L_{\hat{V}^{\text{tm}}_{t+1}} \cdot \delta_t$ for all $t = 0, \dots, T-1$ and $\alpha_T := \epsilon_T$.

2.1.4.3 Alternate Characterization

In this subsection, we provide stronger but simpler conditions which can identify an approximate information state as alternatives to (2.27) and (2.28). These conditions prescribe that an approximate information state $\hat{\Pi}_t = \hat{\sigma}_t(M_t)$ must satisfy for all $t = 0, \dots, T$:

1) State-like evolution: There exists a Lipschitz continuous function $\hat{f}_t : \hat{\mathcal{P}}_t \times \mathcal{U}_t \times \mathcal{Y}_{t+1} \to \mathcal{P}_{t+1}$, independent of the strategy $\boldsymbol{g}$, such that

$$\hat{\sigma}_{t+1}(M_{t+1}) = \hat{f}_t(\hat{\sigma}_t(M_{t+1}), U_t, Y_{t+1}). \tag{2.72}$$

2) Sufficient to approximate observations: For all $m_t \in [[M_t]]$ and $u_t \in [[U_t]]$, there exist the sets $\mathcal{K}^{\text{ob}}_{t+1} := [[Y_{t+1} \mid m_t, u_t]]$ and $\hat{\mathcal{K}}^{\text{ob}}_{t+1} := [[Y_{t+1} \mid \hat{\sigma}_t(m_t), u_t]]$ independent of the strategy $\boldsymbol{g}$, and it holds that

$$\mathcal{H}(\mathcal{K}^{\text{ob}}_{t+1}, \hat{\mathcal{K}}^{\text{ob}}_{t+1}) \leq \delta^{\text{ob}}_t, \tag{2.73}$$

where $\delta^{\text{ob}}_t \in \mathbb{R}_{\geq 0}$ is a known constant.

3) Lipschitz-like observation prediction: There exists a constant $\lambda^{\text{ob}}_t \in \mathbb{R}_{\geq 0}$ such that for all $\hat{\pi}^1_t, \hat{\pi}^2_t \in [[\hat{\Pi}_t]]$,

$$\mathcal{H}\big([[Y_{t+1}|\hat{\pi}^1_t, u_t]], [[Y_{t+1}|\hat{\pi}^2_t, u_t]]\big) \leq \lambda^{\text{ob}}_t \cdot \eta(\hat{\pi}^1_t, \hat{\pi}^2_t), \tag{2.74}$$

where η is an appropriate metric on $\hat{\mathcal{P}}_t$.

Next, we prove that in addition to (2.26) in Definition 4, the conditions (2.72) - (2.74) are sufficient to characterize an approximate information state instead of (2.27) and (2.28).

Lemma 6. *For all $t = 0, \dots, T$, if an uncertain variable $\hat{\Pi}_t = \hat{\sigma}_t(M_t)$ satisfies* (2.72) - (2.73)*, it also satisfies* (2.27).

Proof. Let $m_t \in [[M_t]]$ be a given realization of M_t and let $\hat{\pi}_t = \hat{\sigma}_t(m_t)$ satisfy (2.72) - (2.73), for all $t = 0, \dots, T$. Then, using (2.72), we can write the LHS in (2.27) as

$$\begin{aligned}\mathcal{H}(\mathcal{K}_{t+1},\hat{\mathcal{K}}_{t+1}) &= \mathcal{H}\Big([[\hat{f}_t(\hat{\sigma}_t(m_t),u_t,Y_{t+1})|m_t,u_t]],[[\hat{f}_t(\hat{\sigma}_t(m_t),u_t,Y_{t+1})|\hat{\sigma}_t(m_t),u_t]]\Big)\\ &= \max\Big\{\sup_{y_{t+1}\in\mathcal{K}^{\text{ob}}_{t+1}}\ \inf_{\hat{y}_{t+1}\in\hat{\mathcal{K}}^{\text{ob}}_{t+1}} \eta\big(\hat{f}_t(\hat{\sigma}_t(m_t),u_t,y_{t+1}),\hat{f}_t(\hat{\sigma}_t(m_t),u_t,\hat{y}_{t+1})\big),\\ &\qquad \sup_{\hat{y}_{t+1}\in\hat{\mathcal{K}}^{\text{ob}}_{t+1}}\ \inf_{y_{t+1}\in\mathcal{K}^{\text{ob}}_{t+1}} \eta\big(\hat{f}_t(\hat{\sigma}_t(m_t),u_t,y_{t+1}),\hat{f}_t(\hat{\sigma}_t(m_t),u_t,\hat{y}_{t+1})\big)\Big\},\end{aligned} \tag{2.75}$$

where, in the second equality, we use the definition of the Hausdorff distance from (2.1). Note that $\hat{f}_t$ is globally Lipschitz from the alternate characterization of the approximate information state. This implies that $\eta\big(\hat{f}_t(\hat{\sigma}_t(m_t),u_t,y_{t+1}),\ \hat{f}_t(\hat{\sigma}_t(m_t),u_t,\hat{y}_{t+1})\big) \leq L_{\hat{f}_t}\cdot \eta(y_{t+1},\ \hat{y}_{t+1})$, and thus

$$\begin{aligned}&\mathcal{H}(\mathcal{K}_{t+1},\hat{\mathcal{K}}_{t+1})\\ &\quad\leq L_{\hat{f}_t}\max\Big\{\sup_{y_{t+1}\in\mathcal{K}^{\text{ob}}_{t+1}}\ \inf_{\hat{y}_{t+1}\in\hat{\mathcal{K}}^{\text{ob}}_{t+1}}\eta(y_{t+1},\hat{y}_{t+1}),\ \sup_{\hat{y}_{t+1}\in\hat{\mathcal{K}}^{\text{ob}}_{t+1}}\ \inf_{y_{t+1}\in\mathcal{K}^{\text{ob}}_{t+1}}\eta(y_{t+1},\hat{y}_{t+1})\Big\}\\ &\quad= L_{\hat{f}_t}\cdot\mathcal{H}(\mathcal{K}^{\text{ob}}_{t+1},\hat{\mathcal{K}}^{\text{ob}}_{t+1}) \leq L_{\hat{f}_t}\cdot\delta^{\text{ob}}_t.\end{aligned} \tag{2.76}$$

□

Lemma 7. *For all $t=0,\dots,T$, if an uncertain variable $\hat{\Pi}_t=\hat{\sigma}_t(M_t)$ satisfies (2.72) - (2.74), it also satisfies (2.28).*

Proof. Let $\hat{\pi}^1_t,\hat{\pi}^2_t\in[[\hat{\Pi}_t]]$ be two possible realizations of an approximate information state $\hat{\Pi}_t$, which satisfies (2.72) - (2.74), for all $t=0,\dots,T$. Then, using (2.72), we can write the LHS in (2.28) as

$$\begin{aligned}&\mathcal{H}\big([[\Pi_{t+1}|\hat{\pi}^1_t,u_t]],[[\Pi_{t+1}|\hat{\pi}^2_t,u_t]]\big)\\ &\qquad= \mathcal{H}\big([[\hat{f}_t(\hat{\pi}^1_t,u_t,Y_{t+1})|\hat{\pi}^1_t,u_t]],[[\hat{f}_t(\hat{\pi}^2_t,u_t,Y_{t+1})|\hat{\pi}^2_t,u_t]]\\ &\qquad\leq L_{\hat{f}_t}\cdot\big(\eta(\hat{\pi}^1_t,\hat{\pi}^2_t)+\mathcal{H}\big([[Y_{t+1}|\hat{\pi}^1_t,u_t]],[[Y_{t+1}|\hat{\pi}^2_t,u_t]]\big)\big)\\ &\qquad\leq L_{\hat{f}_t}\cdot(1+\lambda^{\text{ob}}_t)\cdot\eta(\hat{\pi}^1_t,\hat{\pi}^2_t),\end{aligned} \tag{2.77}$$

where, in the first inequality, we use the Lipschitz continuity of the function $\hat{f}_t$ along with the triangle inequality; and in the second inequality, we use (2.74). This completes the proof by defining $\lambda_t := L_{\hat{f}_t}\cdot(1+\lambda^{\text{ob}}_t)$. □

2.1.4.4 Examples

In this subsection, we present two state-quantized [197] approximate information states which satisfy Definition 4. Consider a system as described in Subsection 2.1.3.3 with compact feasible sets $\{\mathcal{X}_t, \mathcal{N}_t, \mathcal{W}_t \mid t = 0, \dots, T\}$ in a metric space $(\mathcal{S}, \eta)$. Recall that $\mathcal{X}_t$ is the state space at any t. Then, a finite subset $\hat{\mathcal{X}}_t \subset \mathcal{X}_t$ is a *set of quantized states* with parameter $\gamma_t \in \mathbb{R}_{\geq 0}$ if $\max_{x_t \in \mathcal{X}_t} \min_{\hat{x}_t \in \hat{\mathcal{X}}_t} \eta(x_t, \hat{x}_t) \leq \gamma_t$. The corresponding *quantization function* $\mu_t : \mathcal{X}_t \to \hat{\mathcal{X}}_t$ is defined as $\mu_t(x_t) := \arg\min_{\hat{x}_t \in \hat{\mathcal{X}}_t} \eta(x_t, \hat{x}_t)$. Note that by construction, $\eta(x_t, \mu_t(x_t)) \leq \gamma_t$ for all $x_t \in \mathcal{X}_t$, for all t.

1) Perfectly Observed Systems: Consider a system where $Y_t = X_t$ for all $t = 0, \dots, T$. Recall from Subsection 2.1.3.3 that the $\Pi_t = X_t \in \mathcal{X}_t$ for all t. Then, a feasible approximate information state for such a system is the quantized state $\hat{\Pi}_t := \mu_t(X_t)$, which satisfies Definition 4 with $\epsilon_t = 2L_{d_t} \cdot \gamma_t$ and $\delta_t = 2\gamma_{t+1} + 2L_{f_t} \cdot \gamma_t$, where $\gamma_{T+1} = 0$, and L_{d_t} and L_{f_t} are the Lipschitz constants for d_t and f_t, respectively (proof in Appendix B). Note that because $\hat{\Pi}_t$ takes values in a finite set, it trivially satisfies (2.28) in Definition 4.

2) Partially Observed Systems: For a partially observed system, recall from Section 2.1.3.3 that an information state is given by the conditional range $\Pi_t = [[X_t | m_t]]$. We construct an approximate conditional range by quantizing each element in Π_t. Thus, the approximation is generated by the mapping $\nu_t : \mathcal{B}(\mathcal{X}_t) \to 2^{\hat{\mathcal{X}}_t}$, where $\mathcal{B}(\mathcal{X}_t)$ is the set of all compact subsets of $\mathcal{X}_t$ and $2^{\hat{\mathcal{X}}_t}$ is the power set of $\hat{\mathcal{X}}_t$. This transformation yields the approximate range $\nu_t(\Pi_t) := \{\mu_t(x_t) \in \hat{\mathcal{X}}_t \mid x_t \in \Pi_t\}$. Then, the approximate range $\hat{\Pi}_t = \nu_t(\Pi_t)$ is an information state for partially observed systems for all $t = 0, \dots, T$ with $\epsilon_t = 2L_{d_t} \cdot \gamma_t$ and $\delta_t = 2\gamma_{t+1} + 2L_{\bar{f}_t} \cdot L_{h_{t+1}} \cdot L_{f_t} \cdot \gamma_t$, where $\gamma_{T+1} = 0$, and $L_{\bar{f}_t}$, $L_{h_{t+1}}$ and L_{f_t} are Lipschitz constants of $\bar{f}_t$, h_{t+1}, and f_t, respectively (proof in Appendix C).

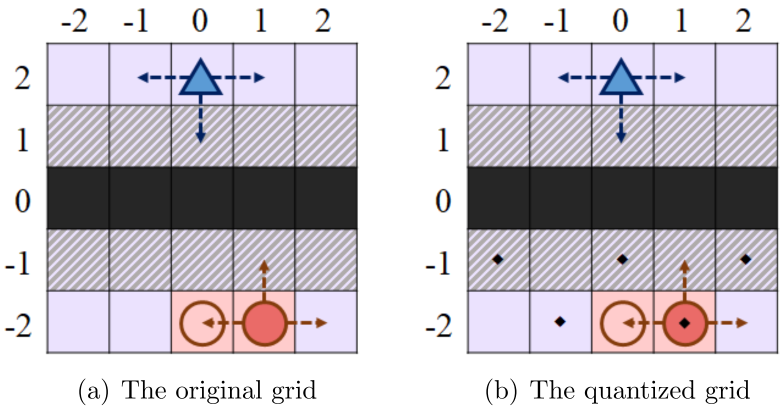
2.1.5 Numerical Examples

We present two numerical examples to illustrate our approach: *(1) The Wall Defense Problem:* a worst-case control problem with partial observations, and *(2) The*

Pursuit Evasion Problem: a worst-case reinforcement learning problem with partly unknown dynamics and partial observations.

2.1.5.1 The Wall Defense Problem

In the wall defense problem, we consider an agent who defends a wall in a 5×5 grid world from an attacker over a time horizon T. The wall is located across the central row of the grid. We illustrate the wall defense problem for one initial condition in Fig. 2.1(a). Here, the black colored cells constitute the wall and the grey hatched cells are adjacent to the wall. The solid blue triangle, solid red circle and red ring are the agent, attacker and observation, respectively, at $t = 0$. The pink cells are feasible positions of the attacker given the observation. The attacker moves within the bottom two rows of the grid and damages a wall cell when positioned in an adjacent cell. At each $t = 0, \ldots, T$, we denote the position of the attacker by $X_t^{\text{at}} \in \mathcal{X}^{\text{at}} = \{(-2,-1), \ldots, (2,-1), (-2,-2), \ldots, (2,-2)\}$. In contrast, the agent moves within the top two rows of the grid and repairs a wall cell when positioned in an adjacent cell. At each t, we denote the position of the agent by $X_t^{\text{ag}} \in \mathcal{X}^{\text{ag}} = \{(-2,1), \ldots, (2,1), (-2,2), \ldots, (2,2)\}$. The state of the wall at each t is the accumulated damage denoted by $D_t = (D_t^{-2}, \ldots, D_t^{2})$, where $D_t^i \in \mathcal{D}_t^i = \{0,1,2,3\}$ for all $i = -2, \ldots, 2$ and $\mathcal{D}_t = \times_{i=-2}^{2} \mathcal{D}_t^i$. The attacker starts at the position $X_0^{\text{at}} \in \mathcal{X}^{\text{at}}$, which evolves for all t as $X_{t+1}^{\text{at}} = \mathbb{I}(X_t^{\text{at}} + W_t \in \mathcal{X}^{\text{at}}) \cdot (X_t^{\text{at}} + W_t) + (1 - \mathbb{I}(X_t^{\text{at}} + W_t \in \mathcal{X}^{\text{at}})) \cdot X_t^{\text{at}}$, where $\mathbb{I}$ is the indicator function and $W_t \in \mathcal{W}_t$ is an uncontrolled disturbance with $\mathcal{W}_t = \{(-1,0), (1,0), (0,0), (0,1), (0,-1)\}$. At each t, the agent observes their own position and the wall's state. The agent also partially observes the attacker's position as $Y_t = \mathbb{I}(X_t^{\text{at}} + N_t \in \mathcal{X}^{\text{at}}) \cdot (X_t^{\text{at}} + N_t) + (1 - \mathbb{I}(X_t^{\text{at}} + N_t \in \mathcal{X}^{\text{at}})) \cdot X_t^{\text{at}}$, where $N_t \in \mathcal{N}_t = \{(0,0), (0,1)\}$ is the measurement noise. Given the history of observations, the agent selects an action $U_t \in \mathcal{U}_t = \mathcal{W}_t$ at each t. Starting with $X_0^{\text{ag}} \in \mathcal{X}^{\text{ag}}$, the agent moves as $X_{t+1}^{\text{ag}} = \mathbb{I}(X_t^{\text{ag}} + U_t \in \mathcal{X}^{\text{ag}}) \cdot (X_t^{\text{ag}} + U_t) + (1 - \mathbb{I}(X_t^{\text{ag}} + U_t \in \mathcal{X}^{\text{ag}})) \cdot X_t^{\text{ag}}$. Starting with $D_0 = (0,0,0,0,0)$, the state of the wall evolves as $D_{t+1}^i = \min\big\{3, \max\big\{0, D_t^i + \mathbb{I}(X_t^{\text{at}} = (i,-1)) - \mathbb{I}(X_t^{\text{ag}} = (i,1))\big\}\big\}$ for all t and $i = -2, \ldots, 2$. At each t, after selecting the action, the agent incurs a cost for the damage to the wall, i.e., $c_t(D_t) = \sum_{i=-2}^{2} D_t^i$. The agent's aim is to minimize the maximum instantaneous damage to the wall, i.e., $\mathcal{J}(\boldsymbol{g}) = \max_{t=0,\dots,T} \max_{x_0, w_{0:T}, n_{0:T}} c_t(D_t)$.

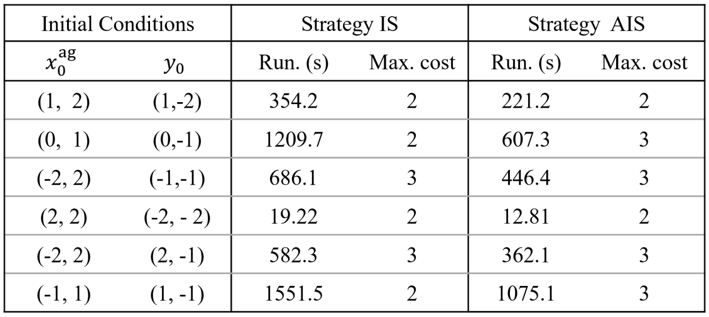


(a) The original grid

(b) The quantized grid

Figure 2.1: The wall defense problem with the initial conditions $x_0^{\text{ag}} = (0, 2)$ and $y_0 = (0, -2)$.

Recall from Subsection 2.1.3.3 that an information state at time t is $\Pi_t = \big(X_t^{\text{ag}}, D_t, [[X_t^{\text{at}}|M_t]]\big)$. We construct an approximation of the conditional range $[[X_t^{\text{at}}|M_t]]$ at time t using the quantization approach from Subsection 2.1.4.4 and define the approximate range $\hat{A}_t = \big\{\mu_t(x_t) \in \hat{\mathcal{X}}^{\text{at}} | x_t \in [[X_t^{\text{at}}|M_t]]\big\}$. The set of quantized cells $\hat{\mathcal{X}}^{\text{at}}$, with $\gamma_t = 1$ for all t, is marked in Fig. 2.1(b) with dots. We consider the approximate information state $\hat{\Pi}_t = \big(X_t^{\text{ag}}, D_t, \hat{A}_t, Y_0\big)$ for all t. The initial observation Y_0 in $\hat{\Pi}_t$ improves the prediction of $\hat{A}_{t+1}$. For five initial conditions, we compute the best control strategy for $T = 6$ using both the information state (IS) and the approximate information state (AIS). In Fig. 2.2, we present the computational times (Run.) for both the DPs in seconds. Note that the approximate DP has a faster run-time in all cases. We also implement both strategies with random disturbances in the system with $T = 6$. In Fig. 2.2, we also present the *actual* worst-case costs across 5×10^3 implementations of both strategies and note that the AIS has a bounded deviation from the IS.

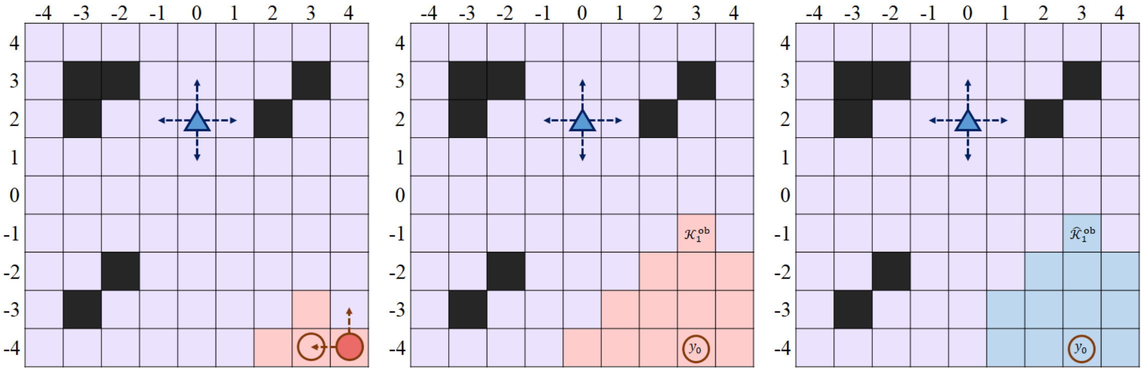


Initial Conditions		Strategy IS		Strategy AIS	
x_0^{ag}	y_0	Run. (s)	Max. cost	Run. (s)	Max. cost
(1, 2)	(1,-2)	354.2	2	221.2	2
(0, 1)	(0,-1)	1209.7	2	607.3	3
(-2, 2)	(-1,-1)	686.1	3	446.4	3
(2, 2)	(-2, - 2)	19.22	2	12.81	2
(-2, 2)	(2, -1)	582.3	3	362.1	3
(-1, 1)	(1, -1)	1551.5	2	1075.1	3

Figure 2.2: Costs and run-times for 5×10^3 simulations and $T = 6$.

2.1.5.2 Pursuit Evasion Problem

In the pursuit evasion problem, we consider an agent who chases a moving target in a 9×9 grid world with static obstacles. The agent aims to get close to the target over a time horizon T. For each $t = 0, \ldots, T$, we denote the position of the agent by $X_t^{\mathrm{ag}} \in \mathcal{X}$ and that of the target by $X_t^{\mathrm{ta}} \in \mathcal{X}$, where $\mathcal{X} = \big\{(-4,-4), \ldots, (4,4)\big\} \setminus \mathcal{O}$ is the set of feasible grid cells and $\mathcal{O} \subset \mathcal{X}$ is the set of obstacles. The target starts at the position $X_0^{\mathrm{ta}} \in \mathcal{X}$, which is updated as $X_{t+1}^{\mathrm{ta}} = \mathbb{I}(X_t^{\mathrm{ta}} + W_t \in \mathcal{X}) \cdot (X_t^{\mathrm{ta}} + W_t) + (1 - \mathbb{I}(X_t^{\mathrm{ta}} + W_t \in \mathcal{X})) \cdot X_t^{\mathrm{ta}}$, where $W_t \in \mathcal{W}_t = \{(-1,0), (1,0), (0,0), (0,1), (0,-1)\}$ is the disturbance. At each t, the agent perfectly observes their own position and nosily observes the target's position as $Y_t = \mathbb{I}(X_t^{\mathrm{ta}} + N_t \in \mathcal{X}) \cdot (X_t^{\mathrm{ta}} + N_t) + (1 - \mathbb{I}(X_t^{\mathrm{ta}} + N_t \in \mathcal{X}))\ \cdot X_t^{\mathrm{ta}}$, where $N_t \in \mathcal{N}_t = \mathcal{W}_t$ is the measurement noise. Next, starting with $X_0^{\mathrm{ag}} \in \mathcal{X}$, the agent selects an action $U_t \in \mathcal{U}_t = \mathcal{W}_t$ to move as $X_{t+1}^{\mathrm{ag}} = \mathbb{I}(X_t^{\mathrm{ag}} + U_t \in \mathcal{X}) \cdot (X_t^{\mathrm{ag}} + U_t) + (1 - \mathbb{I}(X_t^{\mathrm{ag}} + U_t \in \mathcal{X})) \cdot X_t^{\mathrm{ag}}$. At time T, the agent selects no action and observes the target's position X_T^{ta} and incurs a cost $c_T(X_T^{\mathrm{ta}}, X_T^{\mathrm{ag}}) = \eta(X_T^{\mathrm{ta}}, X_T^{\mathrm{ag}}) \in \mathbb{R}_{\geq 0}$, where η is the shortest distance between two cells, while avoiding obstacles. The distance between two adjacent cells is 1 unit. The agent seeks to minimize the worst-case terminal cost *without* prior knowledge of either the observation function or the target's evolution dynamics. Note that this is a reinforcement learning generalization of Problem 1. We illustrate the grid and one initial set up in Fig. 2.3(a). Here, the black cells are obstacles. The solid blue triangle, solid red circle and red ring are the agent, target, and observation, respectively, at $t = 0$. The pink cells are feasible positions of the target given the observation.


-4 -3 -2 -1 0 1 2 3 4
4 3 2 1 0 -1 -2 -3 -4
$\mathcal{K}_1^{\text{ob}}$
$\hat{\mathcal{K}}_1^{\text{ob}}$
y_0

(a) The original problem (b) Actual observation prediction (c) Learned observation prediction

Figure 2.3: The pursuit evasion problem with the initial conditions $x_0^{\text{ag}} = (0, 2)$ and $y_0 = (3, -4)$.

We consider that the agent has access to 3×10^7 observation trajectories from the target which are used to learn an approximate information state representation offline, as characterized in Subsection 2.1.4.3. First, we use the data on observation trajectories to construct estimates of the conditional range $\mathcal{K}_{t+1}^{\text{ob}} = [[Y_{t+1}|Y_{0:t}]]$ for all $t = 0, \dots, T-2$ and $\mathcal{K}_T^{\text{ob}} = [[X_T^{\text{ta}}|Y_{0:T-1}]]$. Then, taking inspiration from [46], we set-up a deep neural network with an encoder-decoder structure for each $t = 0, \dots, T$, as illustrated in Fig. 2.4. At each t, the encoder ψ_t comprises of 3 layer neural network with sizes $(2, 14)$, $(14, 12)$, $(12 + 24, 24)$ and ReLU activation for the first two layers, where the inputs are a 2-d vector of coordinates for observation Y_t and a 24-d vector for the previous approximate information state $\hat{\Pi}_{t-1}$. The encoder compresses these inputs to a 24-d vector representing the approximate information state $\hat{\Pi}_t$. At each t, the decoder ϕ_t is a 4 layer neural network of size $(24, 48)$, $(48, 56)$, $(56, 64)$, $(64, 74)$ with ReLU activation for the first three layers and sigmoid activation for the last layer. Its input is $\hat{\Pi}_t$ and its output is a 74-d vector with each component taking values in $[0, 1]$. Each component of the 74-d output gives a set-inclusion value for a specific feasible cell in the 9×9 grid, excluding obstacles. The output is thus interpreted as the conditional range $\hat{\mathcal{K}}^{\text{ob}}_{t+1} = [[Y_{t+1} \mid \hat{\Pi}_t]]$ for all $t = 0, \ldots, T-2$ and $\hat{\mathcal{K}}^{\text{ob}}_T = [[X^{\text{ta}}_T \mid \hat{\Pi}_{T-1}]]$. We consider a set-inclusion threshold of 0.5 for inclusion in $\hat{\mathcal{K}}^{\text{ob}}_{t+1}$ at each t.

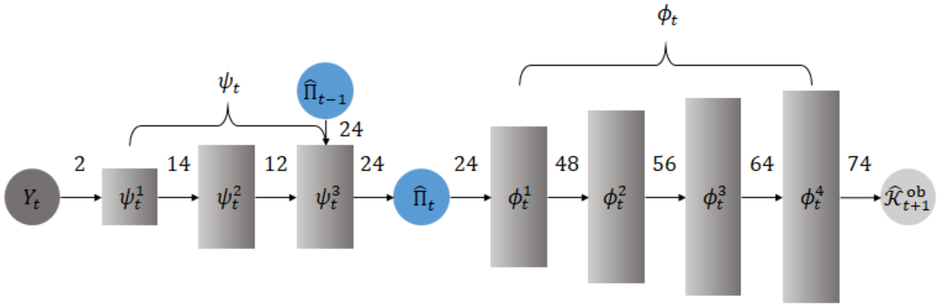
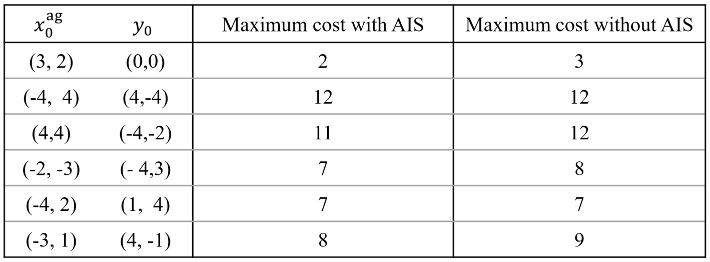


Figure 2.4: The neural network architecture for approximate information states at any $t = 0, \ldots, T-1$.

The learning objective of our neural network at each t is to minimize $\mathcal{H}(\mathcal{K}^{\text{ob}}_{t+1}, \hat{\mathcal{K}}^{\text{ob}}_{t+1})$, which is consistent with the characterization of approximate information states in Subsection 2.1.4.3. Note that at the terminal time step, this objective also minimizes the difference in maximum costs. Since the Hausdorff distance is not differentiable, we adapt the first surrogate function proposed in [198] as a learning objective to train the network weights. We train the network for 40 epochs using 90% of the available data with a learning rate of 0.0003 and test it against the other 10%. To illustrate the training results, consider an out-of-sample initial observation $y_0 = (3, -4)$. Then, the set $\mathcal{K}^{\text{ob}}_1$ constructed using data is shown by pink cells in 2.3(b) and the set $\hat{\mathcal{K}}^{\text{ob}}_1$ generated by of the trained network is shown by blue cells in 2.3(c). Note that the trained network's output matches the conditional range constructed from data accurately except for one cell $(0, -4)$. We train a neural network for each t up to $T = 4$ to learn a complete approximate information state representation for the problem. Then, at each t, the agent uses the state $(X^{\text{ag}}_t, \hat{\Pi}_t)$ in the approximate DP (2.29) - (2.30) to compute an approximately optimal control strategy.

We compare the performance of this approximate strategy with a baseline strategy that uses the observation Y_t at each t instead of $\hat{\Pi}_t$. Thus, for this baseline we train a network to match the prediction $[[Y_{t+1} \mid Y_t]]$ to $[[Y_{t+1} \mid Y_{0:t}]]$ for all $t = 0, \ldots, T-2$ and $[[X^{\text{ta}}_T \mid Y_{T-1}]]$ to $[[X^{\text{ta}}_T \mid Y_{0:T-1}]]$ at time $T-1$. The neural network structure is the same as before except for a lack of $\hat{\Pi}_{t-1}$ in the encoder input at each t and we use the same training parameters as before. Subsequently, the agent computes an approximately optimal strategy using the approximate DP with the state (X_t^{ag}, Y_t) at each t.

For six initial conditions, we present in Fig. 2.5 the worst case costs obtained when implementing both the approximately optimal strategy (Maximum cost with AIS) and the baseline strategy (Maximum cost without AIS) for $T = 4$. Across 10^4 simulations with randomly generated uncertainties, we note that using the learned approximate information state consistently improves worst-case performance when compared to the baseline. Thus, learning an approximate information state representation is a viable approach for worst-case reinforcement learning. In general, we expect our approach to outperform the baseline more for longer time horizons.



x_0^{ag}	y_0	Maximum cost with AIS	Maximum cost without AIS
(3, 2)	(0,0)	2	3
(-4, 4)	(4,-4)	12	12
(4,4)	(-4,-2)	11	12
(-2, -3)	(- 4,3)	7	8
(-4, 2)	(1, 4)	7	7
(-3, 1)	(4, -1)	8	9

Figure 2.5: Worst-case costs for 10^4 simulations and $T = 4$.

2.1.6 Appendix A – L-invertible Functions

In this appendix, we present two classes of functions which are L-invertible: 1) all bi-Lipschitz functions which have a compact domain and a compact co-domain, and 2) all functions with a compact domain and a finite co-domain.

Lemma 8. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two compact subsets of a metric space $(\mathcal{S}, \eta)$. Then, any bi-Lipischitz function $f : \mathcal{X} \to \mathcal{Y}$ is L-invertible.*

Proof. We begin by considering the pre-image set for any $y \in \mathcal{Y}$ under the function f. Note that the function f is continuous because it is bi-Lipschitz and the singleton $\{y\}$ is a compact subset of a metric space. Consequently, the pre-image $f^{-1}(y)$ is a bounded subset of $\mathcal{X}$. Next, let $\mathcal{B}(\mathcal{X})$ denote the set of all bounded subsets of $\mathcal{X}$. Given the first result, we can consider a set-valued mapping $f^{-1} : \mathcal{Y} \to \mathcal{B}(\mathcal{X})$ which returns the pre-image for each $y \in \mathcal{Y}$. Then, for any $y^1, y^2 \in \mathcal{Y}$, using the definition of the Hausdorff distance in (2.1):

$$\mathcal{H}\big(f^{-1}(y^1), f^{-1}(y^2)\big) = \max\Big\{ \sup_{x^1 \in f^{-1}(y^1)} \inf_{x^2 \in f^{-1}(y^2)} \eta(x^1, x^2), \sup_{x^2 \in f^{-1}(y^2)} \inf_{x^1 \in f^{-1}(y^1)} \eta(x^1, x^2)\Big\}. \tag{2.78}$$

In the RHS of (2.78), the bi-Lipschitz property of f implies that there exist constants $\underline{L}_f, \overline{L}_f \in \mathbb{R}_{>0}$ such that $\underline{L}_f \eta(x^1, x^2) \leq |f(x^1) - f(x^2)| \leq \overline{L}_f \eta(x^1, x^2)$, for all $x^1, x^2 \in \mathcal{X}$. Thus, for all $x^1 \in g^{-1}(y^1)$ and $x^2 \in g^{-1}(y^2)$, we write that

$$\eta(x^1, x^2) \leq \underline{L}_f^{-1} \cdot \eta(y^1, y^2). \tag{2.79}$$

The proof is complete by substituting (2.79) into (2.78) and defining the constant $L_{f^{-1}} := \underline{L}_f^{-1}$. □

Lemma 9. *Let $\mathcal{X}$ be a compact subset and $\mathcal{Y}$ be a finite subset of $(\mathcal{S}, \eta)$. Then, any function $f : \mathcal{X} \to \mathcal{Y}$ is L-invertible.*

Proof. Let $||\mathcal{Y}|| > 0$ denote the minimum distance between two distinct elements in the finite, non-empty set $\mathcal{Y}$. Then, for any $y^1, y^2 \in \mathcal{Y}$ such that $y^1 \neq y^2$:

$$\frac{\mathcal{H}\big(f^{-1}(y^1), f^{-1}(y^2)\big)}{\eta(y^1, y^2)} \leq \sup_{y^1, y^2 \in \mathcal{Y}} \frac{\mathcal{H}\big(f^{-1}(y^1), f^{-1}(y^2)\big)}{||\mathcal{Y}||} =: L_{f^{-1}}, \tag{2.80}$$

where $L_{f^{-1}} \in \mathbb{R}_{\geq 0}$ is guaranteed to be finite because the set $\mathcal{X}$ is bounded and thus, so is the numerator. Thus, the function f is L-invertible as defined in (2.78). □

2.1.7 Appendix B – Approximation Bounds for Perfectly Observed Systems

In this appendix, we derive the values of ϵ_t and δ_t for all $t = 0, \dots, T$ when an approximate information state is constructed using state quantization for a perfectly observed system, as described in Subsection 2.1.4.4. We first state a property of the Hausdorff distance which we will use in our derivation.

Lemma 10. *Let $\mathcal{X}$ be a metric space with compact subsets $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D} \subset \mathcal{X}$. Then, it holds that*

$$\mathcal{H}\big(\mathcal{A} \cup \mathcal{B}, \mathcal{C} \cup \mathcal{D}\big) \leq \max\Big\{\mathcal{H}\big(\mathcal{A}, \mathcal{C}\big), \mathcal{H}\big(\mathcal{B}, \mathcal{D}\big)\Big\}. \tag{2.81}$$

Proof. The proof for this result is given in [194, Theorem 1.12.15]. □

Next, we state and prove the main result of this appendix.

Theorem 5. *Consider a perfectly observed system, i.e., $Y_t = X_t$, for all $t = 0, \dots, T$. Let $\mu_t : \mathcal{X}_t \to \hat{\mathcal{X}}_t$ such that $\max_{x_t \in \mathcal{X}_t} \eta(x_t, \mu_t(x_t)) \leq \gamma_t$ at each t. Then, $\hat{\Pi}_t = \mu_t(X_t)$ is an approximate information state which satisfies (2.26) with $\epsilon_t = 2L_{d_t} \cdot \gamma_t$ and (2.27) with $\delta_t = 2\gamma_{t+1} + 2L_{f_t} \cdot \gamma_t$ for all t, where $\gamma_{T+1} = 0$, and where L_{d_t}, and L_{f_t} are Lipschitz constants for d_t and f_t, respectively.*

Proof. For all $t = 0, \dots, T$, let $m_t = (x_{0:t}, u_{0:t-1})$ be the realization of M_t and let the approximate information state be $\hat{x}_t = \mu_t(x_t)$. We first derive the value of ϵ_t in the RHS of (2.26). At time t, can expand the conditional ranges to write that $[[X_t|m_t]] = [[X_t|x_t]] = \{x_t\}$ and $[[X_t|\hat{x}_t]] = \{x_t \in \mathcal{X} \mid \eta(x_t, \hat{x}_t) \leq \gamma_t\}$. On substituting these into the LHS of (2.26), we state that

$$\begin{aligned}
\Big| \sup_{c_t \in [[C_t|m_t,u_t]]} c_t - \sup_{c_t \in [[C_t|\mu_t(x_t),u_t]]} c_t \Big| \\
&= \Big| d_t(x_t, u_t) - \sup_{\bar{x}_t \in [[X_t|\mu_t(x_t)]]} d_t(\bar{x}_t, u_t) \Big| \\
&\leq \sup_{\bar{x}_t \in [[X_t|\mu_t(x_t)]]} |d_t(x_t, u_t) - d_t(\bar{x}_t, u_t)| \\
&\leq L_{d_t} \cdot \sup_{\bar{x}_t \in [[X_t|\mu_t(x_t)]]} \eta(x_t, \bar{x}_t) \qquad (2.82) \\
&\leq L_{d_t} \cdot \big(\eta(x_t, \mu_t(x_t)) + \sup_{\bar{x}_t \in [[X_t|\mu_t(x_t)]]} \eta(\mu_t(x_t), \bar{x}_t)\big), \\
&\leq 2L_{d_t} \cdot \gamma_t =: \epsilon_t,
\end{aligned}$$

where, in the third inequality, we use the triangle inequality. Next, to derive the value of δ_t, we expand the LHS of (2.27) as

$$\begin{aligned}
&\mathcal{H}\big([[\hat{X}_{t+1}|x_t,u_t]],[[\hat{X}_{t+1}|\mu_t(x_t),u_t]]\big)\\
&=\mathcal{H}\Big(\big\{\mu_{t+1}(f_t(x_t,u_t,w_t))|w_t\in\mathcal{W}_t\big\},\big\{\mu_{t+1}(f_t(\bar{x}_t,u_t,w_t))|\bar{x}_t\in[[X_t|\mu_t(x_t)]],w_t\in\mathcal{W}_t\big\}\Big)\\
&\qquad\leq\sup_{w_t\in\mathcal{W}_t}\mathcal{H}\Big(\{\mu_{t+1}(f_t(x_t,u_t,w_t))\},\{\mu_{t+1}(f_t(\bar{x}_t,u_t,w_t))|\bar{x}_t\in[[X_t|\mu_t(x_t)]]\}\Big),
\end{aligned}\tag{2.83}$$

where, in the inequality, we use (2.81) from Lemma 10 and the fact that

$$\big\{\mu_{t+1}(f_t(x_t,u_t,w_t))\in\hat{\mathcal{P}}\mid w_t\in\mathcal{W}_t\big\}=\cup_{w_t\in\mathcal{W}_t}\big\{\mu_{t+1}(f_t(\bar{x}_t,u_t,w_t))\big\}.\tag{2.84}$$

Once again using (2.81) in the RHS of (2.83), we conclude that

$$\begin{aligned}
&\mathcal{H}\big([[\hat{X}_{t+1}|x_t,u_t]],[[\hat{X}_{t+1}|\mu_t(x_t),u_t]]\big)\\
&\quad\leq\sup_{w_t\in\mathcal{W}_t,\bar{x}_t\in[[X_t|\mu_t(x_t)]]}\eta\big(\mu_{t+1}\big(f_t(x_t,u_t,w_t)\big),\mu_{t+1}\big(f_t(\bar{x}_t,u_t,w_t)\big)\big)\\
&\qquad\leq\sup_{w_t\in\mathcal{W}_t,\bar{x}_t\in[[X_t|\mu_t(x_t)]]}\Big(\eta\big(\mu_{t+1}(f_t(x_t,u_t,w_t)),f_t(x_t,u_t,w_t)\big)+\eta\big(f_t(x_t,u_t,w_t),\\
&\qquad\qquad f_t(\bar{x}_t,u_t,w_t)\big)+\eta\big(f_t(\bar{x}_t,u_t,w_t),\mu_{t+1}(f_t(\bar{x}_t,u_t,w_t))\big)\Big)\\
&\qquad\qquad\leq\gamma_{t+1}+2L_{f_t}\cdot\gamma_t+\gamma_{t+1}=:\delta_t,
\end{aligned}\tag{2.85}$$

where, in the second inequality, we use the triangle inequality. □

2.1.8 Appendix C – Approximation Bounds for Partially Observed Systems

In this appendix, we derive the values of ϵ_t and δ_t for all $t=0,\dots,T$, when an approximate information state is constructed using state quantization for a partially observed system, as described in Subsection 2.1.4.4.

Theorem 6. *Consider a partially observed system with $Y_t=h_t(X_t,N_t)$ for all $t=0,\dots,T$. Let $\mu_t:\mathcal{X}_t\to\hat{\mathcal{X}}_t$ such that $\sup_{x_t\in\mathcal{X}_t}\eta(x_t,\mu_t(x_t))\leq\gamma_t$ at each t. Then, $\hat{\Pi}_t=\nu_t(\Pi_t)$ is an approximate information state with $\epsilon_t=2L_{d_t}\cdot\gamma_t$ and $\delta_t=2\gamma_{t+1}+2L_{\bar{f}_t}\cdot L_{h_{t+1}}\cdot L_{f_t}\cdot\gamma_t$ for all t, where $\gamma_{T+1}=0$, and where L_{d_t}, $L_{\bar{f}_t}$, $L_{h_{t+1}}$, and L_{f_t} are Lipschitz constants for the respective functions in the subscripts.*

Proof. For all $t = 0, \dots, T$, let $m_t \in [[M_t]]$, $P_t = [[X_t|m_t]] \in \mathcal{P}_t$, and $\hat{P}_t = \nu_t(P_t) \in \hat{\mathcal{P}}_t$ be the realizations of the memory M_t, the conditional range Π_t and the approximate information state $\hat{\Pi}_t$, respectively. Note that the conditional range P_t satisfies (2.15) and (2.16) from Definition 3. Next, to derive the value of ϵ_t, we write the LHS of (2.26) using (2.15) as

$$\begin{aligned}\Big|\sup_{c_t \in [[C_t|m_t,u_t]]} c_t - \sup_{c_t \in [[C_t|\nu_t(P_t),u_t]]} c_t\Big| =& \Big|\sup_{x_t \in P_t} d_t(x_t,u_t) - \sup_{\bar{x}_t \in [[X_t|\nu_t(P_t)]])} d_t(\bar{x}_t,u_t)\Big| \\ \leq& L_{d_t} \cdot \mathcal{H}(P_t, [[X_t|\nu_t(P_t)]]) \\ \leq& L_{d_t} \cdot \big(\mathcal{H}(P_t, \nu_t(P_t)) + \mathcal{H}(\nu_t(P_t), [[X_t|\nu_t(P_t)]])\big),\end{aligned} \tag{2.86}$$

where, in the equality, we use (2.15); in the first inequality, we use (2.39) from Lemma 5; and in the second inequality we use the triangle inequality for the Hausdorff distance. We can expand the first term in the RHS of (2.86) as

$$\begin{aligned}&\mathcal{H}(P_t, \nu_t(P_t)) = \mathcal{H}(P_t, \{\mu_t(x_t) \in \hat{\mathcal{X}}_t \mid x_t \in P_t\}) \\ &\qquad = \mathcal{H}\big(\cup_{x_t \in P_t} \{x_t\}, \cup_{x_t \in P_t}\{\mu_t(x_t) \in \hat{\mathcal{X}}_t\}\big) \leq \sup_{x_t \in P_t} \eta(x_t, \mu_t(x_t)) \leq \gamma_t,\end{aligned} \tag{2.87}$$

where we use (2.81) from Lemma 10 in the first inequality. We can also expand the second term in the RHS of (2.86) as

$$\begin{aligned}\mathcal{H}(\nu_t(P_t), [[X_t|\nu_t(P_t)]])) =& \mathcal{H}\Big(\nu_t(P_t), \{x_t \in \mathcal{X}_t| \inf_{\bar{x}_t \in \nu_t(P_t)} \eta(x_t, \bar{x}_t) \leq \gamma_t\}\Big) \\ =& \sup_{x_t \in [[X_t|\nu_t(P_t)]]} \inf_{\bar{x}_t \in \nu_t(P_t)} \eta(x_t, \bar{x}_t) \leq \gamma_t,\end{aligned} \tag{2.88}$$

where the second equality holds by expanding the Hausdorff distance and noting that $\nu_t(P_t) \subseteq [[X_t|\nu_t(P_t)]]$. The proof is complete by substituting the results for both terms in the RHS of (2.86).

Next, to derive the value of δ_t, we note that $P_t = \sigma_t(m_t)$. Then, using the triangle inequality in the LHS of (2.27), it holds that

$$\mathcal{H}\Big([[\nu_{t+1}(\Pi_{t+1})|m_t, u_t]], [[\nu_{t+1}(\Pi_{t+1})|\nu_t(\sigma_t(m_t)), u_t]]\Big)$$

$$
\begin{aligned}
&\leq \mathcal{H}\Big([[\nu_{t+1}(\Pi_{t+1})|m_t,u_t]],[[\Pi_{t+1}|m_t,u_t]]\Big)+\mathcal{H}\Big([[\Pi_{t+1}|m_t,u_t]],[[\Pi_{t+1}|\nu_t(\sigma_t(m_t)),u_t]]\Big)\\
&\quad+\mathcal{H}\Big([[\Pi_{t+1}|\nu_t(\sigma_t(m_t)),u_t]],[[\nu_{t+1}(\Pi_{t+1})|\nu_t(\sigma_t(m_t)),u_t]]\Big)\\
&\leq 2\gamma_{t+1}+\mathcal{H}\Big([[\Pi_{t+1}|m_t,u_t]],[[\Pi_{t+1}|\nu_t(\sigma_t(m_t)),u_t]]\Big),
\end{aligned}
\tag{2.89}
$$

where, in the second inequality we use the fact that $\mathcal{H}\big(P_{t+1},\nu_{t+1}(P_{t+1})\big)\leq\gamma_{t+1}$, which was proved above. We can write the second term in the RHS of (2.89) using (2.16) as

$$
\mathcal{H}\Big([[\Pi_{t+1}|m_t,u_t]],[[\Pi_{t+1}|\nu_t(\sigma_t(m_t)),u_t]]\Big)=\mathcal{H}\Big([[\Pi_{t+1}|P_t,u_t]],[[\Pi_{t+1}|\nu_t(P_t),u_t]]\Big).
\tag{2.90}
$$

Furthermore, note that $[[\Pi_{t+1}|\nu_t(P_t),u_t]]=\big\{\tilde{P}_{t+1}\in[[\Pi_{t+1}|\tilde{P}_t,u_t]]\;|\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]\big\}=\cup_{\tilde{P}_t\in[[\Pi_t|\nu(P_t)]]}[[\Pi_{t+1}|\tilde{P}_t,u_t]]$. Next, we use (2.81) from Lemma 10 to write that

$$
\begin{aligned}
&\mathcal{H}\big([[\Pi_{t+1}|P_t,u_t]]),[[\Pi_{t+1}|\nu_t(P_t),u_t]]\big)\\
&\qquad\leq\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]}\mathcal{H}\big([[\Pi_{t+1}|P_t,u_t]]),[[\Pi_{t+1}|\tilde{P}_t,u_t]]\big)\\
&\qquad\leq L_{\bar{f}_t}\cdot\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]}\mathcal{H}\big([[Y_{t+1}|P_t,u_t]]),[[Y_{t+1}|\tilde{P}_t,u_t]]\big)\\
&\qquad\leq L_{\bar{f}_t}\cdot L_{h_{t+1}}\cdot\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]}\mathcal{H}\big([[X_{t+1}|P_t,u_t]]),[[X_{t+1}|\tilde{P}_t,u_t]]\big),
\end{aligned}
\tag{2.91}
$$

where, in the second inequality we use the same arguments as in Lemma 6 and the third inequality can be proven by substituting $Y_{t+1}=h_{t+1}(X_{t+1},V_{t+1})$ into the equation. We can further expand the third term in the RHS of (2.91) and use (2.81) to write that

$$
\begin{aligned}
&\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]}\mathcal{H}\big([[X_{t+1}|P_t,u_t]]),[[X_{t+1}|\tilde{P}_t,u_t]]\big)\\
&\quad\leq\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]],w_t\in\mathcal{W}_t}\mathcal{H}\big(\{f_t(x_t,u_t,w_t)|x_t\in P_t\},\{f_t(x_t,u_t,w_t)|x_t\in\tilde{P}_t\}\big)\\
&\quad\leq L_{f_t}\cdot\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]}\mathcal{H}\big(P_t,\tilde{P}_t\big)\\
&\quad\leq L_{f_t}\cdot\sup_{\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]}\big(\mathcal{H}(P_t,\nu_t(P_t))+\mathcal{H}(\nu_t(P_t),\tilde{P}_t)\big)\\
&\quad\leq 2L_{f_t}\cdot\gamma_t,
\end{aligned}
\tag{2.92}
$$

where, in the third inequality, we use the triangle inequality and in the fourth inequality we use the fact that for all $\nu_t(\tilde{P}_t)=\nu_t(P_t)$, for all $\tilde{P}_t\in[[\Pi_t|\nu_t(P_t)]]$. □

2.2 On Robust Control of Partially Observed Uncertain Systems with Additive Costs

2.2.1 Notation and Preliminaries

We use the non-stochastic framework of *uncertain variables* from [163]. For a sample space Ω and a set $\mathscr{X}$, an uncertain variable is a mapping $\mathsf{X} : \Omega \to \mathscr{X}$ written concisely as $\mathsf{X} \in \mathscr{X}$. For any $\omega \in \Omega$, its realization is $\mathsf{X}(\omega) = \mathsf{x} \in \mathscr{X}$. The *marginal range* of an uncertain variable X is the set $[[\mathsf{X}]] := \{\mathsf{X}(\omega) \mid \omega \in \Omega\}$. The *joint range* of two uncertain variables $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ is $[[\mathsf{X}, \mathsf{Y}]] := \{(\mathsf{X}(\omega), \mathsf{Y}(\omega)) \mid \omega \in \Omega\}$. The *conditional range* of X given a realization y of Y is $[[\mathsf{X}|\mathsf{y}]] := \{\mathsf{X}(\omega) \mid \mathsf{Y}(\omega) = \mathsf{y}, \omega \in \Omega\}$, and $[[\mathsf{X}|\mathsf{Y}]] := \{[[\mathsf{X}|\mathsf{y}]] \mid \mathsf{y} \in [[\mathsf{Y}]]\}$. Next, consider two compact, nonempty subsets $\mathscr{X}, \mathscr{Y}$ of a metric space $(\mathscr{S}, d)$, where $d(\cdot, \cdot)$ is the metric. Then, the Hausdorff distance [194, Chapter 1.12] between the sets is

$$\mathcal{H}(\mathscr{X}, \mathscr{Y}) := \max\{\max_{\mathsf{x}\in\mathscr{X}} \min_{\mathsf{y}\in\mathscr{Y}} d(\mathsf{x}, \mathsf{y}), \max_{\mathsf{y}\in\mathscr{Y}} \min_{\mathsf{x}\in\mathscr{X}} d(\mathsf{x}, \mathsf{y})\}. \tag{2.93}$$

2.2.2 Problem Formulation

We consider an agent who controls the evolution of a system over $T \in \mathbb{N}$ discrete time steps. At any time $t = 0, \dots, T$, the system is denoted by an uncertain variable $X_t \in \mathcal{X}$ and the agent's action is denoted by an uncertain variable $U_t \in \mathcal{U}$. At each t, the system also receives an uncontrolled disturbance $W_t \in \mathcal{W}$. Starting with an initial state $X_0 \in \mathcal{X}$, the state evolves as $X_{t+1} = f_t\left(X_t, U_t, W_t\right)$ for all $t = 0, \dots, T-1$. Before selecting the control action at each t, the agent partially observes the system state as $Y_t = h_t(X_t, N_t) \in \mathcal{Y}$, where $N_t \in \mathcal{N}$ is a noise.

Remark 11. We denote generic uncertain variables by sans-serif upper case alphabets $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$, whereas, we denote the state and observation at any t by italicized upper-case alphabets $X_t \in \mathcal{X}$ and $Y_t \in \mathcal{Y}$, respectively.

At each $t = 0, \dots, T$, the agent stores the history of observations and control actions in their memory, denoted by $M_t := (Y_{0:t}, U_{0:t-1}) \in \mathcal{M}_t$, where $Y_{0:t} := (Y_0, \dots, Y_t)$. Then, the agent selects an action $U_t = g_t(M_t)$ using a control law $g_t : \mathcal{M}_t \to \mathcal{U}$ and incurs a cost $c_t(X_t, U_t) \in \mathbb{R}_{\geq 0}$. We denote the control strategy by $\boldsymbol{g} := (g_0, \dots, g_T) \in \mathcal{G}$ and measure its performance using the *worst-case criterion:*

$$\mathcal{J}(\boldsymbol{g}) := \max_{\substack{x_0 \in \mathcal{X}, n_{0:T} \in \mathcal{N}^T, \\ w_{0:T-1} \in \mathcal{W}^{T-1}}} \sum_{t=0}^{T} c_t(X_t, U_t). \tag{2.94}$$

In (2.94), we maximize the total cost over all feasible realizations of the *uncontrolled inputs*, i.e., initial state X_0, noises $\{N_t \mid t = 0, \dots, T\}$, and disturbances $\{W_t \mid t = 0, \dots, T-1\}$ because they determine all other variables in the system. Next, we state the agent's optimization problem.

Problem 2. We seek to efficiently compute an optimal strategy $\boldsymbol{g}^* = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the sets $\{\mathcal{X}, \mathcal{U}, \mathcal{Y}, \mathcal{W}, \mathcal{N}\}$ and the functions $\{f_t, h_t, c_t \mid t = 0, \dots, T\}$.

We impose the following assumptions on our model:

Assumption 3. Each uncontrolled input is independent (see [163, Definintion 2.1]) of all other uncontrolled inputs.

Assumption 3 ensures that the system evolution is Markovian in a non-stochastic sense (see [163, Definintion 2.2]). This assumption will help develop our results.

Assumption 4. Each feasible set $\{\mathcal{X}, \mathcal{U}, \mathcal{Y}, \mathcal{W}, \mathcal{N}\}$ is a finite subset of a metric space $(\mathcal{S}, d)$.

Assumption 4 ensures that all extrema are well defined and that an optimal solution to Problem 2 exists. We will use the metric $d(\cdot, \cdot)$ in Section IV to quantify the distance between two elements in any set.

Assumption 5. All uncertain variables and the cost $c_t(X_t, U_t)$ have a finite maximum value at each t.

Assumption 5, in addition to the finiteness of all feasible sets, ensures that the functions $\{f_t, h_t, c_t \mid t = 0, \dots, T\}$ are globally Lipschitz. To this end, we will denote the Lipschitz constant of a function f_t by $L_{f_t} \in \mathbb{R}_{\geq 0}$.

2.2.3 Dynamic Programs and Information States

In this section, we first present a standard terminal cost DP which can obtain the optimal strategy in Problem 2. Then, in Subsection 2.2.3.2, we construct a DP which is specialized to the additive cost criterion in (2.94), and in Subsection 2.2.3.3, we define information states to simplify it. To begin, we transform Problem 2 into a terminal cost problem by augmenting the state X_t at each t with the *accrued cost*

$$A_t := \sum_{\ell=0}^{t-1} c_\ell(X_\ell, U_\ell), \tag{2.95}$$

which takes values in a finite set $\mathcal{A}_t \subset \mathbb{R}_{\geq 0}$. Starting with $A_0 := 0$, the accrued cost evolves as $A_{t+1} = A_t + c_t(X_t, U_t)$ for all $t = 0, \dots, T-1$. Thus, the augmented state (X_t, A_t) evolves as a controlled Markov chain. Furthermore, note that the performance criterion (2.94) can be written as a function of the terminal augmented state (X_T, A_T), i.e., $\mathcal{J}(\boldsymbol{g}) = \max_{x_0, n_{0:T}, w_{0:T-1}} \big(c_T(X_T, U_T) + A_T\big)$. This construction yields a terminal cost optimization problem in $\boldsymbol{g} \in \mathcal{G}$, where the optimal strategy can be computed using a memory based terminal cost DP [164], as follows. For all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$, for all $t = 0, \dots, T-1$, we define the value functions

$$Q_t^{\text{tm}}(m_t, u_t) := \max_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} V_{t+1}^{\text{tm}}(m_{t+1}), \tag{2.96}$$

$$V_t^{\text{tm}}(m_t) := \min_{u_t \in \mathcal{U}} Q_t^{\text{tm}}(m_t, u_t), \tag{2.97}$$

where, at time T, $Q_T^{\text{tm}}(m_T, u_T) := \max_{a_T, x_T \in [[A_T, X_T | m_T, u_T]]} (c_T(x_T, u_T) + a_T)$ and $V_T^{\text{tm}}(m_T) := \min_{u_T \in \mathcal{U}} Q_T(m_T, u_T)$. The control law at each t is $g_t^{\text{tm}}(m_t) := \arg\min_{u_t \in \mathcal{U}} Q_t(m_t, u_t)$. Using standard arguments, we can conclude that the resulting control strategy $\boldsymbol{g}^{\text{tm}} = (g_0^{\text{tm}}, \dots, g_T^{\text{tm}})$ is an optimal solution to the terminal cost problem as well as Problem 2 [49]. However, note that the right hand side (RHS) of (2.97) involves solving a minimization problem for each possible realization $m_t \in \mathcal{M}_t$, at each t. The number of possible realizations $|\mathcal{M}_t|$ increases with time as the agent receives more observations, and consequently, the DP requires a large number of computations for a longer horizon T. To address this, we formulate a DP specialized for additive cost problems in Subsection 2.2.3.2 and simplify it using *information states* in Subsection 2.2.3.3. We will show (Remark 12) that the specialized DP allows us to define more computationally efficient information states than (2.96) - (2.97). To this end, we present a theory of cost distributions in the next subsection which is required to construct the specialized DP.

2.2.3.1 Cost distributions

In this subsection, we develop the mathematical framework of *cost distributions* for finite uncertain variables. Cost measures and cost distributions were originally defined for $(\max, +)$ algebra [199], and applied to robust control problems [43, 53] independently from the framework of uncertain variables. A cost measure is the non-stochastic analogue of a probability measure. Specifically, for a finite sample space Ω with a sigma algebra $\mathcal{B}(\Omega)$, a cost measure is a function $q : \mathcal{B}(\Omega) \to \{-\infty\} \cup (-\infty, 0]$ satisfying the properties: (1) $q(\Omega) = 0$, (2) $q(\emptyset) = -\infty$, and (3) $q(B) = \max_{\omega \in B} q(\omega)$ for all $B \in \mathcal{B}(\Omega)$, where, by convention, the maximum over an empty set is $-\infty$. Furthermore, for two sets $B^1, B^2 \in \mathcal{B}(\Omega)$ with $q(B^2) > -\infty$, the conditional cost measure of B^1 given B^2 is $q(B^1|B^2) := q(B^1, B^2) - q(B^2)$, where $q(B^1, B^2) = \max_{\omega \in B^1 \cap B^2} q(\omega)$. Next, we extend this definition to define cost distributions on finite uncertain variables.

Definition 5. Let $\mathsf{X} : \Omega \to \mathscr{X}$ and $\mathsf{Y} : \Omega \to \mathscr{Y}$ be two finite uncertain variables. The *cost distribution* for any realization $\mathsf{x} \in \mathscr{X}$ is $q(\mathsf{x}) := \max_{\omega \in \{\Omega | \mathsf{X}(\omega) = \mathsf{x}\}} q(\omega)$, and that for any $\mathsf{x} \in \mathscr{X}$ given a realization $\mathsf{y} \in \mathscr{Y}$ with $q(\mathsf{y}) > -\infty$ is $q(\mathsf{x}|\mathsf{y}) = q(\mathsf{x}, \mathsf{y}) - q(\mathsf{y})$, where $q(x, y) = \max_{\omega \in \{\Omega | X(\omega) = x, Y(\omega) = y\}} q(\omega)$.

Any cost distribution given by Definition 5 satisfies the following useful properties.

Lemma 11. *Let $(\Omega, \mathcal{B}(\Omega))$ have a cost distribution $q : \mathcal{B}(\Omega) \to \{-\infty\} \cup (-\infty, 0]$. Let $\mathsf{X} : \Omega \to \mathscr{X}$ and $\mathsf{Y} : \Omega \to \mathscr{Y}$ be two finite uncertain variables and let $f : \mathscr{X} \to \mathscr{Y}$ such that $\mathsf{Y} = f(\mathsf{X})$ and $f^{-1}(\mathsf{y}) \neq \emptyset$ for all $\mathsf{y} \in \mathscr{Y}$. Then,*

$$q(\mathsf{y}) = \max_{\mathsf{x} \in \{\mathscr{X} | f(\mathsf{x}) = \mathsf{y}\}} q(\mathsf{x}), \quad \forall \mathsf{y} \in \mathscr{Y}, \tag{2.98}$$

and furthermore, for any function $g : \mathscr{Y} \to \mathbb{R}_{\geq 0}$,

$$\max_{\mathsf{x} \in \mathscr{X}} \big(g(f(\mathsf{x})) + q(\mathsf{x}) \big) = \max_{\mathsf{y} \in \mathscr{Y}} \big(g(\mathsf{y}) + q(\mathsf{y}) \big). \tag{2.99}$$

Proof. Using Definition 5, $q(\mathsf{y}) = \max_{\omega \in \{\Omega | \mathsf{Y}(\omega) = \mathsf{y}\}} q(\omega)$, where $\{\Omega \mid \mathsf{Y}(\omega) = \mathsf{y}\} = \cup_{\mathsf{x} \in \{\mathscr{X} | f(\mathsf{x}) = \mathsf{y}\}} \{\Omega \mid \mathsf{X}(\omega) = \mathsf{x}\}$. This implies that $q(\mathsf{y}) = \max_{\mathsf{x} \in \{\mathscr{X} | f(\mathsf{x}) = \mathsf{y}\}} \max_{\omega \in \{\Omega | \mathsf{X}(\omega) = \mathsf{x}\}} q(\omega) = \max_{\mathsf{x} \in \{\mathscr{X} | f(\mathsf{x}) = \mathsf{y}\}} q(\mathsf{x})$, where, in the second equality, we used Definition 5. This proves (2.98). Next, we use (2.98) in the RHS of (2.99) as $\max_{\mathsf{y} \in \mathscr{Y}} (g(\mathsf{y}) + q(\mathsf{y})) = \max_{\mathsf{y} \in \mathscr{Y}} (g(\mathsf{y}) + \max_{\mathsf{x} \in \{\mathscr{X} | f(\mathsf{x}) = \mathsf{y}\}} \; q(\mathsf{x})) = \max_{\mathsf{y} \in \mathscr{Y}} \max_{\mathsf{x} \in \{\mathscr{X} | f(\mathsf{x}) = \mathsf{y}\}} (g(f(\mathsf{x})) + q(\mathsf{x})) = \max_{\mathsf{x} \in \mathscr{X}} \; (g(f(\mathsf{x})) + q(\mathsf{x}))$, which completes the proof for (2.99). □

2.2.3.2 Specialized Dynamic Program

In this subsection, we construct a specialized DP decomposition for Problem 2 using two specific cost distributions, the first of which is an indicator function.

Definition 6. Let $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ be two finite uncertain variables. The *indicator function* for any $\mathsf{x} \in \mathscr{X}$ is given by

$$\mathbb{I}(\mathsf{x}) := \begin{cases} 0, & \text{if } \mathsf{x} \in [[\mathsf{X}]], \\ -\infty, & \text{if } \mathsf{x} \notin [[\mathsf{X}]], \end{cases} \tag{2.100}$$

and the conditional indicator function for any $\mathsf{x} \in \mathscr{X}$ given a realization $\mathsf{y} \in \mathscr{Y}$ with $\mathbb{I}(\mathsf{y}) > -\infty$ is

$$\mathbb{I}(\mathsf{x}|\mathsf{y}) := \begin{cases} 0, & \text{if } \mathsf{x} \in [[\mathsf{X}|\mathsf{y}]], \\ -\infty, & \text{if } \mathsf{x} \notin [[\mathsf{X}|\mathsf{y}]]. \end{cases} \tag{2.101}$$

The indicator function $\mathbb{I}$ can be shown to satisfy the conditions in Definition 5 and thus, it constitutes a valid cost distribution. In addition to Lemma 11, for two uncertain variables $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ and any function $f : \mathscr{X} \to \mathbb{R}$,

$$\max_{\mathsf{x} \in [[\mathsf{X}|\mathsf{y}]]} f(\mathsf{x}) = \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + \mathbb{I}(\mathsf{x}|\mathsf{y}) \big), \quad \forall \mathsf{y} \in \mathscr{Y}. \tag{2.102}$$

We also require the *accrued distribution* for an uncertain variable at each t, defined using the accrued cost $A_t \in \mathcal{A}_t$.

Definition 7. Let $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ be two finite uncertain variables and let $A_t \in \mathcal{A}_t$ be the accrued cost at any $t = 0, \dots, T$. An *accrued distribution* at any t for any $\mathsf{x} \in \mathscr{X}$ is a function $r_t : \mathscr{X} \to \{-\infty\} \cup [-a_t^{\max}, 0]$, given by

$$r_t(\mathsf{x}) := \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(\mathsf{x}, a_t)\big) - \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(a_t)\big), \tag{2.103}$$

and for $\mathsf{x} \in \mathscr{X}$ given a realization $\mathsf{y} \in \mathscr{Y}$, $\mathbb{I}(\mathsf{y}) > -\infty$, it is a function $r_t : \mathscr{X} \times \mathscr{Y} \to \{-\infty\} \cup [-a_t^{\max}, 0]$, given by

$$r_t(\mathsf{x}|\mathsf{y}) := \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(\mathsf{x}, a_t \mid \mathsf{y})\big) - \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(a_t \mid \mathsf{y})\big), \tag{2.104}$$

where $a_t^{\max} := \max \mathcal{A}_t$.

At each $t = 0, \dots, T$, note that the accrued distribution $r_t(\mathsf{x}|\mathsf{y}) = -\infty$ if $\mathsf{x} \notin [[\mathsf{X}|\mathsf{y}]]$ whereas $r_t(\mathsf{x}|\mathsf{y}) \in [-a_t^{\max}, 0]$ if $\mathsf{x} \in [[\mathsf{X}|\mathsf{y}]]$. It satisfies the properties to be a valid cost distribution. Furthermore, we can compute the conditional range $[[X_t, M_{t+1}|m_t, u_t]]$ at any t given the realizations $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$. Subsequently, we can use Definitions 6 - 7 to derive the accrued distribution $r_t(x_t, m_{t+1}|m_t, u_t)$, for all $x_t \in \mathcal{X}$ and $m_{t+1} \in \mathcal{M}_{t+1}$. Then, we use it in the specialized DP decomposition for Problem 2 as follows. For all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$, for all $t = 0, \dots, T-1$, we define

$$Q_t(m_t, u_t) := \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big(c_t(x_t, u_t) + V_{t+1}(m_{t+1}) + r_t(x_t, m_{t+1} \mid m_t, u_t)\big), \tag{2.105}$$

$$V_t(m_t) := \min_{u_t \in \mathcal{U}} Q_t(m_t, u_t), \tag{2.106}$$

where, at time T, $Q_T(m_T, u_T) := \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + r_T(x_T \mid m_T)\big)$ and $V_T(m_T) := \min_{u_T \in \mathcal{U}} Q_T(m_T, u_T)$. We define the corresponding control law at time t as $g_t^*(m_t) := \arg\min_{u_t \in \mathcal{U}} Q_t(m_t, u_t)$ and the control strategy as $\boldsymbol{g}^* = (g_0^*, \dots, g_T^*)$. Next, we show that solving the DP (2.105) - (2.106) computes the optimal performance and control strategy.

Theorem 7. *For all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$, for all $t = 0, \dots, T$,*

$$Q_t^{tm}(m_t, u_t) = Q_t(m_t, u_t) + \max_{a_t \in [[A_t|m_t]]} a_t, \tag{2.107}$$

$$V_t^{tm}(m_t) = V_t(m_t) + \max_{a_t \in [[A_t|m_t]]} a_t, \tag{2.108}$$

and furthermore, $\boldsymbol{g}^*$ *is an optimal solution to Problem 2.*

Proof. For all $t = 0, \dots, T$, let $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$ be given realizations of M_t and U_t, respectively. We prove the result by mathematical induction starting at the last time step T. We use (2.102) to expand the left hand side (LHS) of (2.107) as

$$\begin{aligned} Q_T^{\text{tm}}(m_T, u_T) &= \max_{a_T, x_T \in [[A_T, X_T|m_T, u_T]]} \big(c_T(x_T, u_T) + a_T\big) \\ &= \max_{a_T \in \mathcal{A}_T, x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + a_T + \mathbb{I}(a_T, x_T \mid m_T, u_T)\big). \end{aligned} \tag{2.109}$$

In the RHS, we add and subtract $\max_{a_T \in \mathcal{A}_T}(a_T + \mathbb{I}(a_T \mid m_T))$ to write that

$$\begin{aligned} Q_T^{\text{tm}}(m_T, u_T) &= \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + \max_{a_T \in \mathcal{A}_T} \big(a_T + \mathbb{I}(a_T, x_T \mid m_T)\big) \\ &\quad - \max_{a_T \in \mathcal{A}_T} \big(a_T + \mathbb{I}(a_T \mid m_T)\big) + \max_{a_T \in \mathcal{A}_T} \big(a_T + \mathbb{I}(a_T \mid m_T)\big) \\ &= \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + r_T(x_T \mid m_T)\big) + \max_{a_T \in [[A_T|m_T]]} a_T, \end{aligned} \tag{2.110}$$

where, in the second equality, we use (2.104) from Definition 7. Thus, using the definition of $Q_T(m_T, u_T)$, we complete the proof for (2.107) at time T. We can prove (2.108) at time T directly by minimizing both sides of (2.107) with respect to $u_T \in \mathcal{U}$. Furthermore, note that $g_T(m_T) = \arg\inf_{u_T \in \mathcal{U}} Q_T(m_T, u_T) = \arg\inf_{u_T \in \mathcal{U}} Q_T^{\text{tm}}(m_T, u_T)$, i.e., $u_T = g_T^*(m_T)$ minimizes $Q_T^{\text{tm}}(m_T, u_T)$. This forms the basis of our induction.

Next, for all $t = 0, \dots, T-1$, we consider the induction hypothesis $V_{t+1}^{\text{tm}}(m_{t+1}) = V_{t+1}(m_{t+1}) + \max_{a_{t+1} \in [[A_{t+1}|m_{t+1}]]} a_{t+1}$. Then, using the hypothesis

$$\begin{aligned} Q_t^{\text{tm}}(m_t, u_t) &= \max_{m_{t+1} \in [[M_{t+1}|m_t, u_t]]} V_{t+1}^{\text{tm}}(m_{t+1}) \\ &= \max_{m_{t+1} \in [[M_{t+1}|m_t, u_t]]} \big(V_t(m_{t+1}) + \max_{a_{t+1} \in [[A_{t+1}|m_{t+1}]]} a_{t+1}\big) \\ &= \max_{m_{t+1}, a_{t+1} \in [[M_{t+1}, A_{t+1}|m_t, u_t]]} \big(V_t(m_{t+1}) + a_{t+1}\big) \\ &= \max_{m_{t+1}, x_t, a_t \in [[M_{t+1}, X_t, A_t|m_t, u_t]]} \big(V_t(m_{t+1}) + c_t(x_t, u_t) + a_t\big) \\ &= \max_{m_{t+1} \in \mathcal{M}_{t+1}, x_t \in \mathcal{X}_{t+1}, a_t \in \mathcal{A}_{t+1}} \big(V_t(m_{t+1}) + c_t(x_t, u_t) + a_t + \mathbb{I}(x_t, m_{t+1}, a_t \mid m_t, u_t)\big), \end{aligned} \tag{2.111}$$

where, in the third equality, we use the fact that $[[A_{t+1}|m_{t+1}]] = [[A_{t+1}|m_{t+1}, m_t, u_t]]$ because $m_{t+1} = (m_t, u_t, y_{t+1})$; in the fourth equality, we use the definition of a_{t+1}; and in the fifth equality, we use the property of the of the indicator function. Then, as for time T, we add and subtract $\max_{a_t \in \mathcal{A}_t}(a_t + \mathbb{I}(a_t|m_t))$ in the RHS and use (2.104) from Defintion 7 to conclude that

$$\begin{aligned} &Q_t^{\text{tm}}(m_t, u_t) \\ &= \max_{m_{t+1} \in \mathcal{M}_{t+1}, x_t \in \mathcal{X}_{t+1}} \big(V_t(m_{t+1}) + c_t(x_t, u_t) + r_t(x_t, m_{t+1} \mid m_t, u_t)\big) - \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(a_t \mid m_t)\big) \\ &= Q_t(m_t, u_t) + \max_{a_t \in [[A_t|m_t]]} a_t, \end{aligned} \tag{2.112}$$

which proves (2.107) at time t. We can prove (2.108) at time t directly by minimizing both sides of (2.107) respect to $u_t \in \mathcal{U}$, and furthermore,

$$g_t^*(m_t) = \arg \inf_{u_t \in \mathcal{U}} Q_t(m_t, u_t) = \arg \inf_{u_t \in \mathcal{U}} Q_t^{\text{tm}}(m_t, u_t). \tag{2.113}$$

This proves the induction hypothesis at time t and thus, the result holds for all $t = 0, \dots, T$ using mathematical induction. □

Thoerem 7 establishes that the specialized DP (2.105) - (2.106) computes an optimal solution to Problem 2. Note that at each t, the optimization in the RHS of (2.106) must still be solved for each possible $m_t \in \mathcal{M}_t$, in a manner similar to (2.96) - (2.97) Thus, we still require a large number of computations for longer time horizons. In the next subsection, we define *information states* to address this concern.

2.2.3.3 Information States

In this subsection, we introduce information states to construct an optimal DP decomposition for Problem 2.

Definition 8. An *information state* at any $t = 0, \dots, T$ is an uncertain variable $\Pi_t = \sigma_t(M_t)$ taking values in a finite set $\mathcal{P}_t$, where $\sigma_t : \mathcal{M}_t \to \mathcal{P}_t$. Furthermore, for all t, for all $m_t \in \mathcal{M}_t$, $u_t \in \mathcal{U}$, $x_t \in \mathcal{X}$ and $\pi_{t+1} \in \mathcal{P}_{t+1}$, it satisfies:

$$r_t(x_t, \pi_{t+1} \mid m_t, u_t) = r_t(x_t, \pi_{t+1} | \sigma_t(m_t), u_t), \quad t = 0, \dots, T-1, \tag{2.114}$$

$$r_T(x_T|m_T) = r_T(x_T \mid \sigma_t(m_T)), \tag{2.115}$$

where the conditional accrued distributions in (2.114) and (2.115) can be evaluated independent of the strategy $\boldsymbol{g}$.

In the corresponding DP, for all $\pi_t \in \mathcal{P}_t$ and $u_t \in \mathcal{U}$, for all $t = 0, \ldots, T-1$, we define the value functions

$$\bar{Q}_t(\pi_t, u_t) := \max_{x_t \in \mathcal{X}, \pi_{t+1} \in \mathcal{P}_{t+1}} \big(\bar{V}_{t+1}(\pi_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \pi_{t+1} \mid \pi_t, u_t)\big), \tag{2.116}$$

$$\bar{V}_t(\pi_t) := \min_{u_t \in \mathcal{U}} \bar{Q}_t(\pi_t, u_t), \tag{2.117}$$

where, at time T, $\bar{Q}_T(\pi_T, u_T) := \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + r_T(x_T \mid \pi_T)\big)$ and $\bar{V}_T(\pi_T) := \min_{u_T \in \mathcal{U}} \bar{Q}_T(\pi_T, u_T)$. The control law at each t is $\bar{g}^*_t(\pi_t) := \arg\min_{u_t \in \mathcal{U}} \bar{Q}_t(\pi_t, u_t)$. Next, we prove that the information state based DP (2.116) - (2.117) yields the same value as the specialized DP (2.105) - (2.106).

Theorem 8. *Let $\Pi_t = \sigma_t(M_t)$ be an information state at each $t = 0, \ldots, T$. Then, for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$, $Q_t(m_t, u_t) = \bar{Q}_t(\sigma_t(m_t), u_t)$ and $V_t(m_t) = \bar{V}_t(\sigma_t(m_t))$.*

Proof. Let $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$ be given realizations of M_t and U_t, respectively, for all $t = 0, \ldots, T$. We prove the result using mathematical induction starting with T, where

$$\begin{aligned} Q_T(m_T, u_T) &= \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + r_T(x_T \mid m_T)\big) \\ &= \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + r_T(x_T \mid \sigma_t(m_T))\big) = \bar{Q}_T(\sigma_T(m_T), u_T) \end{aligned} \tag{2.118}$$

holds as a direct consequence of (2.115) in Definition 8. Subsequently, by taking the minimum on both sides with respect to $u_t \in \mathcal{U}$, it holds that $V_T(m_T) = \bar{V}_T(\sigma_T(m_T))$. With this as the basis, for each $t = 0, \ldots, T-1$, we consider the induction hypothesis $V_{t+1}(m_{t+1}) = \bar{V}_{t+1}(\sigma_{t+1}(m_{t+1}))$.

Next, we prove that $Q_t(m_t, u_t) = \bar{Q}_t(\sigma_t(m_t), u_t)$ at time t by showing that the RHS of (2.105) is equal to the RHS of (2.116). Using the induction hypothesis in the RHS of (2.105),

$$\begin{aligned}Q_t(m_t,u_t) &= \max_{x_t\in\mathcal{X},m_{t+1}\in\mathcal{M}_{t+1}} \big(V_{t+1}(m_{t+1}) + c_t(x_t,u_t) + r_t(x_t,m_{t+1} \mid m_t,u_t)\big)\\ &= \max_{x_t\in\mathcal{X},m_{t+1}\in\mathcal{M}_{t+1}} \big(\bar{V}_{t+1}(\sigma_{t+1}(m_{t+1})) + c_t(x_t,u_t) + r_t(x_t,m_{t+1} \mid m_t,u_t)\big)\\ &= \max_{x_t\in\mathcal{X},\pi_{t+1}\in\mathcal{P}_{t+1}} \big(\bar{V}_{t+1}(\pi_{t+1}) + c_t(x_t,u_t) + r_t(x_t,\pi_{t+1} \mid \pi_t,u_t)\big),\end{aligned} \tag{2.119}$$

where, in the second equality, we use result 2 from Lemma 11 and (2.114). Thus, at time t, it holds that $Q_t(m_t,u_t) = \bar{Q}_t(\sigma_t(m_t),u_t)$. Subsequently, we can prove $V_t(m_t) = \bar{V}_t(\sigma_t(m_t))$ by minimizing both sides with $u_t \in \mathcal{U}$. This proves the induction hypothesis at time t, and the result follows by mathematical induction. □

From Theorem 8, the strategy $\bar{\boldsymbol{g}}^* = (\bar{g}_0^*,\dots,\bar{g}_T^*)$ using information states is an optimal solution to Problem 2. In practice, using information states to compute $\bar{\boldsymbol{g}}^*$ is more tractable than using the memory to compute $\boldsymbol{g}^*$ only when the set $\mathcal{P}_t$ has fewer elements than $\mathcal{M}_t$ for most instances of t. This is usually true for systems with long time horizons.

2.2.3.4 Examples of Information States

In this subsection, we present examples of information states which satisfy the conditions in Definition 8.

1) Partially observed systems: Generally, at each $t = 0,\dots,T$ a valid information state which satisfies Definition 8 is the function valued uncertain variable $\Pi_t : \mathcal{X} \to \{-\infty\} \cup [-a_t^{\max}, 0]$. At time t, for a given $m_t \in \mathcal{M}_t$, the realization of Π_t is $p_t(x_t) := r_t(x_t|m_t) = \max_{a_t\in\mathcal{A}_t}\big(a_t + \mathbb{I}(x_t,a_t|m_t)\big) - \max_{a_t\in\mathcal{A}_t}\big(a_t + \mathbb{I}(a_t|m_t)\big)$ for all $x_t \in \mathcal{X}$. Note that this can be interpreted as a normalization [43] of the standard information state from [35,53].

2) Perfectly observed systems: Consider a system where $Y_t = X_t$ for all t. An information state for such a system is $\Pi_t = X_t$ at each t, i.e, the state itself. This information state is simpler than the one in Case 1.

3) Systems with action dependent costs: Consider a partially observed system where at each t the cost has the form $c_t(U_t) \in \mathbb{R}_{\geq 0}$, and the terminal cost is $c_T(X_T,U_T)$.

Then, an information state is the conditional range $\Pi_t = [[X_t|M_t]]$ at each t (see Appendix A). This is simpler than the one presented in Case 1.

Remark 12. From [164], we know that the terminal DP (2.96) - (2.97) can be used to derive another information state $\Xi_t = [[X_t, A_t|M_t]]$ for each t for Case 1. The conditional range Ξ_t can take $2^{|\mathcal{A}_t|\times|\mathcal{X}|}$ feasible values whereas Π_t from Case 1 can take $|\mathcal{A}_t|^{|\mathcal{X}|}$ values. As $|\mathcal{A}_t|$ grows in size with time t, the number of feasible values of Π_t increases at a slower rate than the number of feasible values of Ξ_t. Thus, Π_t yields a more computationally tractable DP than Ξ_t. This illustrates that constructing information states using the specialized DP (2.105) - (2.106) is better than using the terminal DP (2.96) - (2.97).

Remark 13. Using Definition 8 we can identify simpler information states for systems with special properties, as shown in Cases 2 - 3. However, in many applications, merely using an information state may not sufficiently improve the tractbility optimal strategies. Thus, we extend Definition 8 to include approximate information states in Section 2.2.4.

2.2.4 Approximate Information States

In this section, we define approximate information states and utilize them to develop an approximate DP. We begin by defining a distance between two cost distributions.

Definition 9. Let $\mathscr{X}$ be a finite subset of a metric space $(\mathscr{S}, d)$, with an uncertain variable $X \in \mathscr{X}$ and two distributions $r : \mathscr{X} \to \{-\infty\} \cup [-a^1, 0]$ and $q : \mathscr{X} \to \{-\infty\} \cup [-a^2, 0]$, $a^1, a^2 \in \mathbb{R}_{\geq 0}$. Then:

1) The *finite domains* of r and q are the sets $\mathscr{X}^r := \{\mathsf{x} \in \mathscr{X} \,|r(\mathsf{x}) \neq -\infty\}$ and $\mathscr{X}^q := \{\mathsf{x} \in \mathscr{X} | q(\mathsf{x}) \neq -\infty\}$, respectively.

2) For any $\mathsf{x} \in \mathcal{X}^r \cup \mathcal{X}^q$, the *nearest finite inputs* for r and q are given by $\psi^r(\mathsf{x}) := \arg\min_{\hat{\mathsf{x}} \in \mathscr{X}^r} d(\hat{\mathsf{x}}, \mathsf{x})$, and $\psi^q(\mathsf{x}) := \arg\min_{\hat{\mathsf{x}} \in \mathscr{X}^q} d(\hat{\mathsf{x}}, \mathsf{x})$, respectively.

3) The *distance* between the distributions r and q is

$$\mathcal{R}\big(r,q\big) := \max\Big(\mathcal{H}(\mathscr{X}^r,\mathscr{X}^q),\max_{\mathsf{x}\in\mathscr{X}^r\cup\mathscr{X}^q}|r(\psi^r(\mathsf{x}))-q(\psi^q(\mathsf{x}))|\Big), \tag{2.120}$$

where $\mathcal{H}$ is the Hausdorff metric.

Remark 14. Because any cost distribution cannot identically return $-\infty$ for all $\mathsf{x}\in\mathscr{X}$, the sets $\mathscr{X}^r$ and $\mathscr{X}^q$ are non-empty for all distributions r,q on X. Consequently, the distance $\mathcal{R}(r,q)$ always returns a finite value.

Note that $\mathcal{R}$ is the maximum of a metric on a set-space and a metric on a function-space. Thus, it can quantify the distance between two different *accrued distributions* on an uncertain variable $\mathsf{X}\in\mathscr{X}$. Specifically, let $\mathsf{Y}\in\mathscr{Y}$ and $\mathsf{Z}\in\mathscr{Z}$ take realizations $y\in\mathcal{Y}$ and $z\in\mathcal{Z}$, respectively, such that $[[\mathsf{X},A_t|\mathsf{y}]]\neq\emptyset$ and $[[\mathsf{X},A_t|\mathsf{z}]]\neq\emptyset$ for some time t. Then, we denote the functional forms of the conditional distributions on X given y and given z as $r_t(\mathsf{X}|\mathsf{y})$ and $r_t(\mathsf{X}|\mathsf{z})$, respectively, and quantify the distance between them as

$$\mathcal{R}\big(r_t(\mathsf{X}|\mathsf{y}),r_t(\mathsf{X}|\mathsf{z})\big) := \max\Big(\mathcal{H}\big([[\mathsf{X}|\mathsf{y}]],[[\mathsf{X}|\mathsf{z}]]\big),\\ \max_{\mathsf{x}\in[[\mathsf{X}|\mathsf{y}]]\cup[[\mathsf{X}|\mathsf{z}]]}\big|r_t\big(\psi^{\mathsf{y}}(\mathsf{x})|\mathsf{y}\big)-r_t\big(\psi^{\mathsf{z}}(\mathsf{x})|\mathsf{z}\big)\big|\Big), \tag{2.121}$$

where, the finite domains are $\{\mathsf{x}\in\mathscr{X}\mid r_t(\mathsf{x}|\mathsf{y})\neq-\infty\}=[[\mathsf{X}|\mathsf{y}]]$ and $\{\mathsf{x}\in\mathscr{X}\mid r_t(\mathsf{x}|\mathsf{z})\neq-\infty\}=[[\mathsf{X}|\mathsf{z}]]$; and for any $\mathsf{x}\in[[\mathsf{X}|\mathsf{y}]]\cup[[\mathsf{X}|\mathsf{z}]]$, the nearest finite inputs are $\psi^{\mathsf{y}}(\mathsf{x}):=\arg\min_{\hat{\mathsf{x}}\in[[\mathsf{X}|\mathsf{y}]]}d(\hat{\mathsf{x}},\mathsf{x})$ and $\psi^{\mathsf{z}}(\mathsf{x}):=\arg\min_{\hat{\mathsf{x}}\in[[\mathsf{X}|\mathsf{z}]]}d(\hat{\mathsf{x}},\mathsf{x})$. Next, using $\mathcal{R}$ to quantify the approximation gap, we define approximate information states for Problem 2.

Definition 10. An *approximate information state* at any $t=0,\dots,T$ is an uncertain variable $\hat{\Pi}_t=\hat{\sigma}_t(M_t)$ taking values in a finite subset $\hat{\mathcal{P}}_t$ of some metric space, where $\hat{\sigma}_t:\mathcal{M}_t\to\hat{\mathcal{P}}_t$. Furthermore, for all t, there exists a parameter $\epsilon_t\in\mathbb{R}_{\geq0}$ such that for all $m_t\in\mathcal{M}_t$ and $u_t\in\mathcal{U}$, it satisfies:

$$\mathcal{R}\big(r_t(X_t,\hat{\Pi}_{t+1}\mid m_t,u_t),r_t(X_t,\hat{\Pi}_{t+1}\mid\hat{\sigma}_t(m_t),u_t)\big)\leq\epsilon_t,\quad t=0,\dots,T-1, \tag{2.122}$$

$$\mathcal{R}\big(r_T(X_T\mid m_T),r_T(X_T\mid\hat{\sigma}_T(m_T))\big)\leq\epsilon_T, \tag{2.123}$$

where each conditional accrued distribution in (2.122) and (2.123) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

In the approximate DP, for all $t = 0, \dots, T-1$, for all $\hat{\pi}_t \in \hat{\mathcal{P}}_t$ and $u_t \in \mathcal{U}$, we recursively define the value functions

$$\hat{Q}_t(\hat{\pi}_t, u_t) := \max_{x_t \in \mathcal{X}, \hat{\pi}_{t+1} \in \hat{\mathcal{P}}_{t+1}} \big(\hat{V}_{t+1}(\hat{\pi}_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \hat{\pi}_{t+1} \mid \hat{\pi}_t, u_t)\big), \tag{2.124}$$

$$\hat{V}_t(\hat{\pi}_t) := \min_{u_t \in \mathcal{U}} \hat{Q}_t(\hat{\pi}_t, u_t), \tag{2.125}$$

where, at time T, $\hat{Q}_T(\hat{\pi}_T, u_T) := \max_{x_T \in \mathcal{X}} \big(c_T(x_T, u_T) + r_T(x_T | \hat{\pi}_T, u_T)\big)$ and $\hat{V}_T(\hat{\pi}_T) := \min_{u_T \in \mathcal{U}} \hat{Q}_T(\hat{\pi}_T, u_T)$. The control law at each t is $\hat{g}^*_t(\hat{\pi}_t) := \arg\min_{u_t \in \mathcal{U}} \hat{Q}_t(\hat{\pi}_t, u_t)$ and the approximate control strategy is $\hat{\boldsymbol{g}}^* := (\hat{g}^*_0, \dots, \hat{g}^*_T)$. Next, we bound the performance loss from implementing the approximate control strategy $\hat{\boldsymbol{g}}^*$ in Problem 2. We begin with a preliminary result which will be required subsequently.

Lemma 12. *Let $\mathscr{X}$ be a finite subset of a metric space $(\mathscr{S}, d)$ and consider two cost distributions $r : \mathscr{X} \to \{-\infty\} \cup [-a^1, 0]$ and $q : \mathscr{X} \to \{-\infty\} \cup [-a^2, 0]$, where $a^1, a^2 \in \mathbb{R}_{\geq 0}$. Then, for a Lipschitz function $f : \mathscr{X} \to \mathbb{R}$:*

$$\begin{aligned} &\big| \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big) - \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big) \big| \\ &\qquad \leq L_f \cdot \mathcal{H}(\mathcal{X}^r, \mathcal{X}^q) + \max_{x \in \mathcal{X}^r \cup \mathcal{X}^q} |r(\psi^r(x)) - q(\psi^q(x))| \\ &\qquad \leq (L_f + 1) \cdot \mathcal{R}\big(r, q\big). \end{aligned} \tag{2.126}$$

Proof. We prove this result by considering two cases which are mutually exclusive but cover all the possibilities.

Case 1: $\max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big) \geq \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big)$, which implies that $\big| \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big) - \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big) \big| = \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big) - \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big)$. We define a variable $\mathsf{x}^* \in \mathscr{X}^r$ such that $\mathsf{x}^* := \arg\max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big)$ and a function $\psi^i : \mathscr{X} \to \mathscr{X}^i$ such that $\psi^i(\mathsf{x}) := \arg\min_{\tilde{\mathsf{x}} \in \mathscr{X}^i} d(\mathsf{x}, \tilde{\mathsf{x}})$ for each $i = r, q$. Then,

$$\max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big) - \max_{\mathsf{x} \in \mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big)$$

$$
\begin{aligned}
&= f(\mathsf{x}^*) + r(\mathsf{x}^*) - \max_{\mathsf{x}\in\mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big) \\
&\leq f(\mathsf{x}^*) + r(\mathsf{x}^*) - f\big(\psi^q(\mathsf{x}^*)\big) - q\big(\psi^q(\mathsf{x}^*)\big) \\
&\leq L_f \cdot d(\mathsf{x}^*, \psi^q(\mathsf{x}^*)) + \big|r(\mathsf{x}^*) - q\big(\psi^q(\mathsf{x}^*)\big)\big| \\
&\leq L_f \cdot \mathcal{H}(\mathscr{X}^r, \mathscr{X}^q) + \max_{\mathsf{x}\in\mathscr{X}^r\cup\mathscr{X}^q} \big|r\big(\psi^r(\mathsf{x})\big) - q\big(\psi^q(\mathsf{x})\big)\big| \\
&\leq L_f \cdot \mathcal{R}\big(r,q\big) + \mathcal{R}\big(r,q\big),
\end{aligned} \tag{2.127}
$$

where, in the first inequality, we use the fact that $q(\psi^q(\mathsf{x}^*)) \neq -\infty$; in the second inequality, we use the Lipschitz continuity of f; in the third inequality, we use the definition of the Hausdorff metric and the fact that $\max_{x\in\mathcal{X}^r} |r(\mathsf{x}) - q(\psi^q(\mathsf{x}))| = \max_{\mathsf{x}\in\mathscr{X}^r\cup\mathscr{X}^q} |r(\psi^r(\mathsf{x})) - q(\psi^q(\mathsf{x}))|$; and in the fourth inequality, we use (2.120) from Definition 9.

Case 2: $\max_{\mathsf{x}\in\mathscr{X}} \big(f(\mathsf{x}) + r(\mathsf{x})\big) < \max_{\mathsf{x}\in\mathscr{X}} \big(f(\mathsf{x}) + q(\mathsf{x})\big)$, where the result holds using the same arguments as Case 1. □

Next, we bound the maximum error when approximating the value functions in the optimal DP (2.96) - (2.97) with the value functions in the approximate DP (2.124) - (2.125).

Theorem 9. *Let $L_{\hat{V}_{t+1}}$ be the Lipschitz constant of $\hat{V}_{t+1}$ for all $t = 0, \dots, T-1$. Then, for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$,*

$$
|Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| \leq \alpha_t, \tag{2.128}
$$

$$
|V_t(m_t) - \hat{V}_t(\hat{\sigma}_t(m_t))| \leq \alpha_t, \tag{2.129}
$$

where $\alpha_t = \alpha_{t+1} + (2L_t + 1) \cdot \epsilon_t$, where $L_t := \max\{L_{\hat{V}_{t+1}}, L_{c_t}\}$, for all $t = 0, \dots, T-1$ and $\alpha_T = (L_{c_T} + 1) \cdot \epsilon_T$.

Proof. For all $t = 0, \dots, T$, let $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$ be realizations of M_t and U_t, respectively. We prove both results by mathematical induction, starting with time step T. At T, we directly use (2.126) from Lemma 12 and (2.123) from Definition 10 to conclude that $|Q_T(m_T, u_T) - \hat{Q}_T(\hat{\sigma}_t(m_T), u_T)| \leq (L_{c_T} + 1) \cdot \epsilon_T$. Furthermore, minimizing both terms in the LHS of (2.128) yields $|V_T(m_T) - \hat{V}_T(\hat{\sigma}_t(m_T))| \leq \max_{u_T \in \mathcal{U}} |Q_T(m_T, u_T) - \hat{Q}_T(\hat{\sigma}_t(m_T), u_T)| \leq (L_{c_T} + 1) \cdot \epsilon_T$.

This forms the basis of our mathematical induction. Then, at each $t = 0, \dots, T-1$, we consider the induction hypothesis $|V_{t+1}(m_{t+1}) - \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1}))| \leq \alpha_{t+1}$ and first prove (2.128). Using the triangle inequality,

$$\begin{aligned}
&\big|Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)\big| \\
&\qquad \leq \Big| \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big(V_{t+1}(m_{t+1}) + c_t(x_t, u_t) + r_t(x_t, m_{t+1} \mid m_t, u_t)\big) \\
&\qquad - \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big(\hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) + c_t(x_t, u_t) + r_t(x_t, m_{t+1} \mid m_t, u_t)\big)\Big| \\
&\qquad + \Big| \max_{x_t \in \mathcal{X}, \hat{\pi}_{t+1} \in \hat{\mathcal{P}}_{t+1}} \big(\hat{V}_{t+1}(\hat{\pi}_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \hat{\pi}_{t+1} \mid m_t, u_t)\big) \\
&\qquad - \max_{x_t \in \mathcal{X}, \hat{\pi}_{t+1} \in \hat{\mathcal{P}}_{t+1}} \big(\hat{V}_{t+1}(\hat{\pi}_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \hat{\pi}_{t+1} \mid \hat{\sigma}_t(m_t), u_t)\big)\Big|. \qquad (2.130)
\end{aligned}$$

Here, for the first term in the RHS,

$$\begin{aligned}
&\Big| \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big(V_{t+1}(m_{t+1}) + c_t(x_t, u_t) + r_t(x_t, m_{t+1} | m_t, u_t)\big) \\
&\qquad - \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big(\hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1})) + c_t(x_t, u_t) + r_t(x_t, m_{t+1} | m_t, u_t)\big)\Big| \\
&\qquad\qquad \leq \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big|V_{t+1}(m_{t+1}) - \hat{V}_{t+1}(\hat{\sigma}_{t+1}(m_{t+1}))\big| \leq \alpha_{t+1}, \qquad (2.131)
\end{aligned}$$

where in the second inequality, we use the induction hypothesis. Furthermore, in the second term in the RHS, we directly use (2.126) from 12 and (2.122) from Definition 10 to conclude that

$$\begin{aligned}
&\Big| \max_{x_t \in \mathcal{X}, \hat{\pi}_{t+1} \in \hat{\mathcal{P}}_{t+1}} \big(\hat{V}_{t+1}(\hat{\pi}_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \hat{\pi}_{t+1} | m_t, u_t)\big) \\
&\qquad - \max_{x_t \in \mathcal{X}, \hat{\pi}_{t+1} \in \hat{\mathcal{P}}_{t+1}} \big(\hat{V}_{t+1}(\hat{\pi}_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \hat{\pi}_{t+1} | \hat{\sigma}_t(m_t), u_t)\big)\Big| \\
&\qquad\qquad \leq \big(2L_t + 1\big) \cdot \epsilon_t, \qquad (2.132)
\end{aligned}$$

where, $L_t = \max\{L_{\hat{V}_{t+1}}, L_{c_t}\}$ and $2L_t$ is the Lipschitz constant for the function $\hat{V}_{t+1}(\hat{\pi}_{t+1})$ $+c_t(x_t, u_t)$ with respect to the variables $(\hat{\pi}_{t+1}, x_t)$ for all $u_t \in \mathcal{U}$, where $\eta\big((\hat{\pi}^1_{t+1}, x^1_t),$

$(\hat{\pi}^1_{t+1}, x^1_t)\big) := \max\{\eta(\hat{\pi}^1_{t+1}, \hat{\pi}^2_{t+1}), \eta(x^1_t, x^2_t)\}$. Combining results for each term in the RHS completes the proof for (2.128) at time t. Next, we prove (2.129). Using the definition of the value functions in the LHS of (2.129),

$$|V_t(m_t) - \hat{V}_t(\hat{\sigma}_t(m_t))| = |\min_{u_t \in \mathcal{U}} Q_t(m_t, u_t) - \min_{u_t \in \mathcal{U}} \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)|$$
$$\leq \max_{u_t \in \mathcal{U}} |Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| \leq \alpha_t, \quad (2.133)$$

where in the second inequality, we use (2.128). This proves the induction hypothesis at time t. Thus, the results hold for all $t = 0, \ldots, T$ using mathematical induction. □

Next, we bound the maximum difference in the performance of an approximate control strategy $\hat{\boldsymbol{g}}^* := (\hat{g}^*_0, \ldots, \hat{g}^*_T)$ and optimal strategy $\boldsymbol{g}^*$. Recall that $\hat{g}^*_t(\hat{\pi}_t) = \arg\min_{u_t \in \mathcal{U}} \hat{Q}_t(\hat{\pi}_t, u_t)$ for all $t = 0, \ldots, T$. Then, the equivalent strategy $\boldsymbol{g} = (g_0, \ldots, g_T)$, which utilizes the memory but yield the same actions and performance as $\hat{\boldsymbol{g}}^*$, is constructed as $g_t(m_t) := \hat{g}^*_t(\hat{\sigma}_t(m_t))$ for all t. To compute the performance of $\boldsymbol{g}$ (and consequently, of $\hat{\boldsymbol{g}}^*$), we define for all $t = 0, \ldots, T-1$, for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$,

$$\Theta_t(m_t, u_t) := \max_{x_t \in \mathcal{X}, m_{t+1} \in \mathcal{M}_{t+1}} \big(\Lambda_{t+1}(m_{t+1}) + c_t(x_t, u_t) + r_t(x_t, m_{t+1} \mid m_t, u_t)\big), \quad (2.134)$$
$$\Lambda_t(m_t) := \Theta_t(m_t, g_t(m_t)), \quad (2.135)$$

where, at time T, $\Theta_T(m_T, u_T) := \max_{x_T \in \mathcal{X}} (c_T(x_T, u_T) + r_T(x_T \mid m_T, u_T))$ and $V_T(m_T) = \Theta_T(m_T, g_T(m_T))$. Recursively evaluating the value functions (2.134) - (2.135) computes the performance of $\boldsymbol{g}$ as $\Lambda_0(m_0)$, where $m_0 = y_0$. Note that the performance of $\boldsymbol{g}^*$ is simply the optimal value. Next, we bound the difference in the performances of $\boldsymbol{g}$ and $\boldsymbol{g}^*$.

Theorem 10. *Let $L_{\hat{V}_{t+1}}$ be the Lipschitz constant of $\hat{V}_{t+1}$ for all $t = 0, \ldots, T-1$. Then, for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$,*

$$|Q_t(m_t, u_t) - \Theta_t(m_t, u_t)| \leq 2\alpha_t, \quad (2.136)$$
$$|V_t(m_t) - \Lambda_t(m_t)| \leq 2\alpha_t. \quad (2.137)$$

where $\alpha_t = \alpha_{t+1} + (2L_t + 1) \cdot \epsilon_t$ *with* $L_t := \max\{L_{\hat{V}_{t+1}},\, L_{c_t}\}$ *for all* $t = 0, \ldots, T-1$ *and* $\alpha_T = (L_{c_T} + 1) \cdot \epsilon_T$.

Proof. We begin by recursively defining the value functions which compute the performance of the strategy $\hat{\boldsymbol{g}}$. For all $t = 0, \ldots, T-1$ and for each $\hat{\pi}_t \in \hat{\mathcal{P}}_t$ and $u_t \in \mathcal{U}$, let

$$\hat{\Theta}_t(\hat{\pi}_t, u_t) := \max_{x_t \in \mathcal{X}, \hat{\pi}_{t+1} \in \hat{\mathcal{P}}_{t+1}} \big(\hat{\Lambda}_{t+1}(\hat{\pi}_{t+1}) + c_t(x_t, u_t) + r_t(x_t, \hat{\pi}_{t+1} | \hat{\pi}_t, u_t)\big) \tag{2.138}$$

$$\hat{\Lambda}_t(\hat{\pi}_t) := \hat{\Theta}_t(\hat{\pi}_t, \hat{g}_t(\hat{\pi}_t)), \tag{2.139}$$

where, at time T, $\hat{\Theta}_T(\hat{\pi}_T, u_T) := \max_{x_T \in \mathcal{X}}(c_T(x_T, u_T) + r_T(x_T | m_T, u_T))$ and $\hat{\Lambda}_T(\hat{\pi}_T) := \hat{\Theta}_T(\hat{\pi}_t, \hat{g}_T(\hat{\pi}_T))$. Note that $\hat{\Theta}_t(\hat{\pi}_t, u_t) = \hat{Q}_t(\hat{\pi}_t, u_t)$ and $\hat{\Lambda}_t(\hat{\pi}_t) = \hat{V}_t(\hat{\pi}_t)$, for all $t = 0, \ldots, T$, since $\hat{g}_t(\hat{\pi}_t) = \arg\min_{u_t \in \mathcal{U}} \hat{Q}_t(\hat{\pi}_t, u_t)$. Next, we use the triangle inequality in the LHS of (2.136) at any t to write

$$\begin{aligned} |Q_t(m_t, u_t) - \Theta_t(m_t, u_t)| &\leq |Q_t(m_t, u_t) - \hat{Q}_t(\hat{\sigma}_t(m_t), u_t)| + |\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)| \\ &\leq \alpha_t + |\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)|, \end{aligned} \tag{2.140}$$

where, in the second inequality, we use (2.128) from Theorem 9. Then, to prove (2.136), it suffices to show that $|\hat{\Theta}_t(\hat{\sigma}_t(m_t), u_t) - \Theta_t(m_t, u_t)| \leq \alpha_t$. We can show this in addition to $|\hat{\Lambda}_t(\hat{\sigma}_t(m_t)) - \Lambda_t(m_t)| \leq \alpha_t$ for all $t = 0, \ldots, T$ using mathematical induction and following the same arguments as in Theorem 9. □

2.2.4.1 Examples of Approximate Information States

In this subsection, we present two examples of approximate information states constructed by state quantization. For the set of states $\mathcal{X}$, a subset $\hat{\mathcal{X}} \subset \mathcal{X}$ is a *set of quantized states* with parameter $\gamma \in \mathbb{R}_{\geq 0}$ if $\max_{x_t \in \mathcal{X}} \min_{\hat{x}_t \in \hat{\mathcal{X}}} d(x_t,\, \hat{x}_t) \leq \gamma$ and $\mu(x_t) = \arg\min_{\hat{x}_t \in \hat{\mathcal{X}}} d(x_t, \hat{x}_t)$ is the *quantization function*. We apply this approach to two cases.

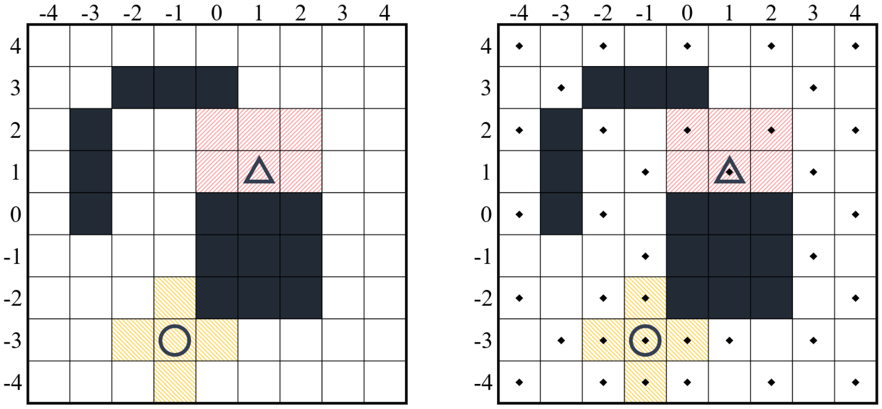
1) Perfectly observed systems: Consider a system where $Y_t = X_t$ for all $t = 0, \ldots, T$. Recall from Subsection 2.1.3.3 that the information state is simply $\Pi_t = X_t$

and it takes values in $\mathcal{X}$ for all t. Then, an approximate information state for such a system can be defined as the quantized state $\hat{\Pi}_t := \mu(X_t)$. The corresponding constants in Definition 10 are $\epsilon_t = 4\gamma + 2L_{f_t} \cdot \gamma$ for all $t = 0, \dots, T-1$ and $\epsilon_T = 2\gamma$, where L_{f_t} is the Lipschitz constant for f_t. The derivation for ϵ_t can be found in Appendix B.

2) Systems with action dependent costs: For such systems, recall from Section 2.1.3.3 that an information state is given by the conditional range $\Pi_t = [[X_t | M_t]]$. We approximate the conditional range by quantizing each element in Π_t using a mapping $\nu : 2^{\mathcal{X}} \to 2^{\hat{\mathcal{X}}}$ such that $\nu(\Pi_t) := \{\mu(x_t) \in \hat{\mathcal{X}} | x_t \in \Pi_t\}$. Then, we can consider that $\hat{\Pi}_t = \nu(\Pi_t)$ is an approximate information state. We illustrate its performance in the next section using a numerical example.

2.2.5 Numerical Example

For our numerical example, we consider an agent pursuing a target across a 9×9 grid with obstacles. At each $t = 0, \dots, T$, the agent's position is X_t^{ag} and the target's position is X_t^{ta}, each of which takes values in the set of grid cells $\mathcal{X} = \big\{(-4,-4), (-4,-3), \dots, (3,4), (4,4)\big\} \setminus \mathcal{O}$, where $\mathcal{O} \subset \mathcal{X}$ is a known set of obstacle cells. Let $\mathcal{W} = \mathcal{N} = \{(-1,0), (1,0), (0,0), (0,1), (0,-1)\}$ and $\mathcal{D} := \{(-1,1), (1,1), (1,-1), (-1,-1)\}$. Starting at $X_0^{\text{ta}} \in \mathcal{X}$, the target's position evolves as $X_{t+1}^{\text{ta}} = \delta(X_t^{\text{ta}} + W_t \in \mathcal{X}) \cdot (X_t^{\text{ta}} + W_t) + (1 - \delta(X_t^{\text{ta}} + W_t \in \mathcal{X})) \cdot X_t^{\text{ta}}$, where $W_t \in \mathcal{W}$ and $\delta(\cdot)$ returns 1 if the condition in the argument holds and 0 otherwise. At each t, the agent observes their own position perfectly and the target's position as $Y_t = \delta(X_t^{\text{ta}} + N_t \in \mathcal{X}) \cdot (X_t^{\text{ta}} + N_t) + (1 - \delta(X_t^{\text{ta}} + N_t \in \mathcal{X})) \cdot X_t^{\text{ta}}$, where $N_t \in \mathcal{N}$. Then, the agent selects an action $U_t \in \mathcal{U} = \mathcal{W} \cup \mathcal{D}$ and moves as $X_{t+1}^{\text{ag}} = \delta(X_t^{\text{ag}} + U_t \in \mathcal{X}) \cdot (X_t^{\text{ag}} + U_t) + (1 - \delta(X_t^{\text{ag}} + U_t \in \mathcal{X})) \cdot X_t^{\text{ag}}$. The agent incurs an interim cost $c_t(U_t) := 0.5 \cdot \delta(U_t \in \mathcal{D})$ only if it moves diagonally, and a terminal cost $d(X_T^{\text{ta}}, X_T^{\text{ag}})$ corresponding to the final distance from the target. We illustrate this in Fig. 2.6(a), where: (1) the black cells are obstacles, (2) the black triangle is the initial position of the agent and the hatched region around it indicates the available actions, and (3) the black circle is the initial observation of

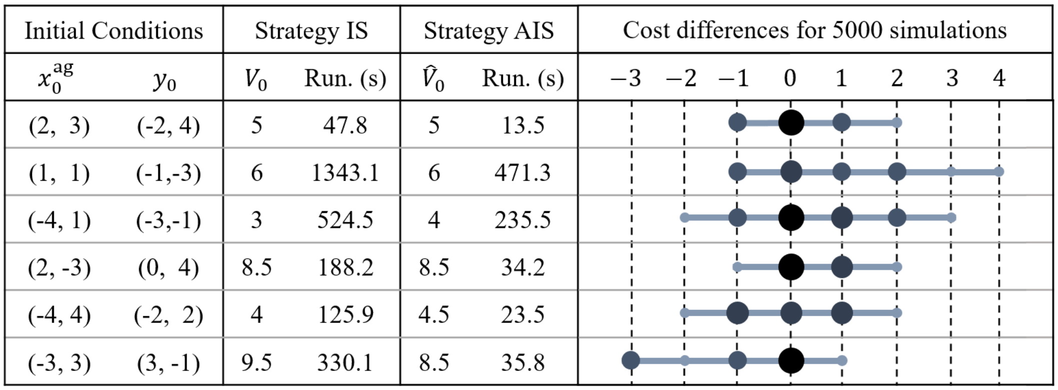


(a) The original grid (b) The quantized grid

Figure 2.6: The gridworld pursuit problem with the initial conditions $x_0^{\text{ag}} = (1,1)$ and $y_0 = (-1,-3)$.

the agent and the hatched region around it indicates the possible initial positions of the target.

This setup constitutes a system with action dependent costs as described in Subsection 2.2.3.4. For such a system, an information state at time t is $\Pi_t = (X_t^{\text{ag}}, \Lambda_t)$, where $\Lambda_t = [[X_t^{\text{ta}} | M_t]]$. We construct an approximation of Λ_t at each t using the quantization approach from Subsection 2.2.4.1. First, we define a static set of quantized states $\hat{\mathcal{X}}$, with $\gamma = 1$ and quantization function $\mu(x_t) : \mathcal{X} \to \hat{\mathcal{X}}$, using the initial observation of the agent, as illustrated using dots in Fig. 2.1(b). Note that we use a finer quantization around the point of initial observation and sparser quantization elsewhere. Then, the approximate range at time t is $\hat{\Lambda}_t = \{\mu(x_t) \in \hat{\mathcal{X}} \mid x_t \in \Lambda_t\}$ and the approximate information state is $\hat{\Pi}_t = \big(X_t^{\text{ag}}, \hat{\Lambda}_t, Y_0\big)$. We include Y_0 in $\hat{\Pi}_t$ because it facilitates the update of $\hat{\Lambda}_t$ to $\hat{\Lambda}_{t+1}$. For six initial conditions, we computed the best control strategy using both the optimal DP and approximate DP for $T = 6$. In Fig. 2.7, we have tabulated the worst-case values (V_0 and $\hat{V}_0$) and run-times in seconds (Run.) for both DPs. We also evaluated the difference between *actual* costs incurred by the approximate strategy and the optimal strategy, respectively, by implementing both of them in 5000 simulations with randomly generated disturbances. We have marked these differences in Fig. 2.7 and indicated the frequency of each cost difference by the size of the disc marking it. While the approximate strategy is faster to compute than the optimal strategy for all cases, we note that it admits bounded deviations in actual costs.



Initial Conditions		Strategy IS		Strategy AIS		Cost differences for 5000 simulations
x_0^{ag}	y_0	V_0	Run. (s)	$\hat{V}_0$	Run. (s)	−3 −2 −1 0 1 2 3 4
(2, 3)	(-2, 4)	5	47.8	5	13.5	
(1, 1)	(-1,-3)	6	1343.1	6	471.3	
(-4, 1)	(-3,-1)	3	524.5	4	235.5	
(2, -3)	(0, 4)	8.5	188.2	8.5	34.2	
(-4, 4)	(-2, 2)	4	125.9	4.5	23.5	
(-3, 3)	(3, -1)	9.5	330.1	8.5	35.8	

Figure 2.7: Results of numerical simulations for $T = 6$.

2.2.6 Appendix A - Derivation of Information State for Systems with Action Dependent Costs

In this appendix, we derive the information states for partially observed systems with control dependent costs as described in Subsection 2.2.3.4. We recall that for a general partially observed system, the information state at each $t = 0, \dots, T$ is given by the function $\Pi_t : \mathcal{X}_t \to \{-\infty\} \cup [-a_t^{\max}, 0]$. Given a realization $m_t \in \mathcal{M}_t$ of the memory M_t at any time t, it takes as its realization the functional form $p_t(X_t) = r_t(X_t|m_t)$. Next, we prove an important result to establish the information state.

Lemma 13. *Let the incurred cost at each $t = 0, \dots, T-1$ be $c_t(U_t) \in \mathbb{R}_{\geq 0}$. Then, for any $m_t \in \mathcal{M}_t$ and $x_t \in \mathcal{X}_t$, it holds that $r_t(x_t \mid m_t) = \mathbb{I}(x_t \mid m_t)$.*

Proof. Let $m_t \in \mathcal{M}_t$ and $x_t \in \mathcal{X}_t$ be realizations of the uncertain variables M_t and X_t, respectively, at each $t = 0, \dots, T$. Let $m_t = (y_{0:t}, u_{0:t-1})$ at time t. Then, we note that there exists a known function $\bar{c}_t : \prod_{\ell=0}^{t-1} \mathcal{U}_\ell \to \mathcal{A}_t$ such that $a_t = \bar{c}_t(u_{0:t-1})$. We use this property to write that

$$r_t(x_t|m_t) = \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(x_t, a_t|m_t)\big) - \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(a_t|m_t)\big)$$

$$
\begin{aligned}
&= \bar{c}_t(u_{0:t-1}) + \mathbb{I}\big(x_t | y_{0:t}, u_{0:t-1}\big) - \bar{c}_t(u_{0:t-1}) \\
&= \mathbb{I}\big(x_t | y_{0:t}, u_{0:t-1}\big) = \mathbb{I}(x_t | m_t), \qquad (2.141)
\end{aligned}
$$

where, in the second equality, we use the fact that $\max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(x_t, a_t | m_t)\big) = \max_{a_t \in [[A_t|m_t]]} \big(a_t + \mathbb{I}(x_t | a_t, m_t)\big)$ and $[[A_t|m_t]] = \{\bar{c}_t(u_{0:t-1})\}$. □

As a direct consequence of Lemma 13, for a given realization of the memory $m_t \in \mathcal{X}_t$ at time t, the realization of the information state for a perfectly observed system is the function form of the indicator function $\mathbb{I}(X_t|m_t)$, where for all $x_t \in \mathcal{X}$,

$$
\mathbb{I}(x_t|m_t) = \begin{cases} 0, & \text{if } x_t \in [[X_t|m_t]], \\ -\infty, & \text{if } x_t \notin [[X_t|m_t]]. \end{cases} \qquad (2.142)
$$

From (2.142), note that the functional form $\mathbb{I}(X_t|m_t) = \mathbb{I}(X_t|[[X_t|m_t]])$, and thus, at each time $t = 0, \dots, T$, given the realized memory $m_t \in \mathcal{M}_t$, it is sufficient to simply track the conditional range $[[X_t|m_t]]$ to derive the information state $r_t(X_t|m_t)$ for all $x_t \in \mathcal{X}$. This implies that $[[X_t|m_t]]$ satisfies all the properties of an information state.

2.2.7 Appendix B - Derivation of Approximate Information State for Perfectly Observed Systems

In this appendix, we derive the values of ϵ_t for all $t = 0, \dots, T$ while using state quantization for a perfectly observed system, as described in Subsection 2.2.4.1. Recall that the set of quantized states is $\hat{\mathcal{X}} \subseteq \mathcal{X}$ with a parameter γ and quantization function $\mu(x_t) = \arg\min_{\hat{x}_t \in \hat{\mathcal{X}}_t} d(x_t, \hat{x}_t)$ such that $d(x_t, \mu(x_t)) \leq \gamma$ for all $x_t \in \mathcal{X}_t$. Note that for any uncertain variable $Z \in \mathcal{Z}$, the conditional range given $\mu(x_t)$ at any t is $[[Z|\mu(x_t)]] = \cup_{\tilde{x}_t \in [[X_t|\mu(x_t)]]} [[Z|\tilde{x}_t]]$. Next, we prove the main result of this appendix.

Theorem 11. *For all $t = 0, \dots, T$, let $\mu : \mathcal{X}_t \to \hat{\mathcal{X}}_t$ such that $\max_{x_t \in \mathcal{X}_t} d(x_t, \mu(x_t)) \leq \gamma$ and let L_{f_t} be the Lipschitz constant for f_t. Then, for all t, $\hat{\Pi}_t = \mu(X_t)$ is an approximate information state which satisfies (2.122) with $\epsilon_t = 4\gamma + 2L_{f_t} \cdot \gamma$ and (2.123) with $\epsilon_T = 2\gamma$.*

Proof. For all $t = 0, \dots, T$, let $m_t = (x_{0:t}, u_{0:t-1})$ be the realization of M_t and let the approximate information state be $\hat{x}_t = \mu(x_t)$. We first derive the value of ϵ_T in the RHS of (2.123). Note that at time T, $[[X_T|m_T]] = [[X_T|x_T]] = \{x_T\}$ and $[[X_T|\hat{x}_T]] = \{x_T \in \mathcal{X} \mid d(x_T, \hat{x}_T) \leq \gamma\}$. Furthermore, $r_T(X_T|m_T) = r_T(X_T|x_T)$ because x_T is an information state. Then, we can expand the LHS in (2.123) using 2.121 as

$$\begin{aligned}&\mathcal{R}\big(r_T(X_T|x_T), r_T(X_T, \mu(x_T))\big)\\ &= \max\Big\{\mathcal{H}\big([[X_T|x_T]], [[X_T|\mu(x_T)]], \max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} |r_T(\psi^{x_T}(\tilde{x}_T)|x_T) - r_T(\tilde{x}_T|\mu(x_T))|\Big\}.\end{aligned} \tag{2.143}$$

We will analyse the RHS of (2.143) term by term. For the first term, we use the definition of the Hausdroff metric to state that

$$\begin{aligned}\mathcal{H}\big([[X_T|x_T]], [[X_T|\mu(x_T)]]\big) &= \max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} d(x_T, \tilde{x}_T)\\ &\leq \max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} \big(d(x_T, \mu(x_T)) + d(\mu(x_T), \tilde{x}_T)\big) \leq 2\gamma,\end{aligned} \tag{2.144}$$

where, in the first inequality we use the triangle inequality; and in the second inequality, we use the definition of the quantization function μ. For the second term in the RHS of (2.143), we use Definition 7 to write that

$$\begin{aligned}&\max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} |r_T(\psi^{x_T}(\tilde{x}_T)|x_T) - r_T(\tilde{x}_T|\mu(x_T))|\\ &\quad\leq \max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} \Big| \max_{a_T \in \mathcal{A}_T} (a_T + \mathbb{I}(\psi^{x_T}(\tilde{x}_T), a_T|x_T)) - \max_{a_T \in \mathcal{A}_T} (a_T + \mathbb{I}(\tilde{x}_T, a_T|\mu(x_T)))\\ &\qquad\qquad - \max_{a_T \in [[A_T|x_T]]} a_T + \max_{a_T \in [[A_T|\mu(x_T)]]} a_T\Big|.\end{aligned} \tag{2.145}$$

In the RHS of (2.145), note that $\mathbb{I}$ is a conditional cost distribution on the uncertain variables in the argument. Thus, we can use the property of conditional cost distributions from Definition 5 to state that

$$\begin{aligned}&\max_{a_T \in \mathcal{A}_T} (a_T + \mathbb{I}(\psi^{x_T}(\tilde{x}_T), a_T|x_T)) - \max_{a_T \in \mathcal{A}_T} (a_T + \mathbb{I}(\tilde{x}_T, a_T|\mu(x_T)))\\ &= \max_{a_T \in \mathcal{A}_T} \big(a_T + \mathbb{I}(\psi^{x_T}(\tilde{x}_T)|a_T, x_T) + \mathbb{I}(a_T|x_T)\big) - \max_{a_T \in \mathcal{A}_T} \big(a_T + \mathbb{I}(\tilde{x}_T|a_T, \mu(x_T)) + \mathbb{I}(a_T|\mu(x_T))\big)\end{aligned}$$

$$= \max_{a_T \in [[A_T|x_T]]} \big(a_T + \mathbb{I}(\psi^{x_T}(\tilde{x}_T)|a_T, x_T)\big) - \max_{a_T \in [[A_T|\mu(x_T)]]} \big(a_T + \mathbb{I}(\tilde{x}_T|a_T, \mu(x_T))\big),$$

$$= \max_{a_T \in [[A_T|x_T]]} a_T + \mathbb{I}(\psi^{x_T}(\tilde{x}_T)|x_T) - \max_{a_T \in [[A_T|\mu(x_T)]]} a_T - \mathbb{I}(\tilde{x}_T|\mu(x_T)), \quad (2.146)$$

where, in the second equality, we use (2.102); and in the third equality, because $\psi^{x_T}(\tilde{x}_T) = \min_{\bar{x}_T \in [[X_T|x_T]]} d(x_T, \bar{x}_T) = x_T$, it holds that

$$\mathbb{I}(\psi^{x_T}(\tilde{x}_T)|a_T, x_T) = \mathbb{I}(\psi^{x_T}(\tilde{x}_T)|x_T) = 0, \quad (2.147)$$

and because $a_T \in [[A_T|\bar{x}_T]]$ for all $\bar{x}_T \in [[X_T|\mu(x_T)]]$ when $a_T \in [[A_T|\mu(x_T)]] = \cup_{\bar{x}_T \in [[X_T|\mu(x_T)]]} [[A_T|\bar{x}_T]]$, it holds that

$$[[X_T \mid a_T, \mu(x_T)]] = \big\{\bar{x}_T \in \mathcal{X} \mid d(\bar{x}_T, \mu(x_T)) \leq \gamma, a_T \in [[A_T|\bar{x}_T]]\big\}$$
$$= \{\bar{x}_T \in \mathcal{X} \mid d(\bar{x}_T, \mu(x_T)) \leq \gamma\} = [[X_T \mid \mu(x_T)]]. \quad (2.148)$$

On substituting (2.146) into (2.145), it holds that

$$\max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} |r_T(\psi^{x_T}(\tilde{x}_T) \mid x_T) - r_T(\tilde{x}_T \mid \mu(x_T))|$$
$$= \max_{\tilde{x}_T \in [[X_T|\mu(x_T)]]} |\mathbb{I}(\psi^{x_T}(\tilde{x}_T) \mid x_T) - \mathbb{I}(\tilde{x}_T \mid \mu(x_T))| = 0. \quad (2.149)$$

Then, substituting both (2.144) and (2.149) into (2.143), we can conclude that $\epsilon_T = \max\{2\gamma, 0\} = 2\gamma$.

Next, to derive the value of ϵ_t for any $t = 0, \dots, T-1$, we expand the LHS of (2.122) using Definition 7 as

$$\mathcal{R}\big(r_t(X_t, \hat{X}_{t+1}|x_t, u_t), r_t(X_t, \hat{X}_{t+1}|\mu(x_t), u_t)\big)$$
$$= \max\Big\{\mathcal{H}\big([[X_t, \hat{X}_{t+1}|x_t, u_t]], [[X_t, \hat{X}_{t+1}|\mu(x_t), u_t]]\big),$$
$$\max_{(\tilde{x}_t, \tilde{x}_{t+1}) \in [[X_t, \hat{X}_{t+1}|\mu(x_t), u_t]]} \Big|r_t(\psi^{x_t}(\tilde{x}_t, \tilde{x}_{t+1})|x_t, u_t) - r_t(\tilde{x}_t, \tilde{x}_{t+1}|\mu(x_t), u_t)\Big|\Big\}. \quad (2.150)$$

Once again, we analyze the RHS of (2.150) term by term. For the first term, we use the definition of the Hausdorff metric and rearrange some terms to write that

$$\mathcal{H}\big([[X_t, \hat{X}_{t+1}|x_t, u_t]], [[X_t, \hat{X}_{t+1}|\mu(x_t), u_t]]\big)$$

$$\leq \max\Bigg\{ \min_{\tilde{x}_t \in [[X_t|\mu(x_t)]]} \Big(d(x_t, \tilde{x}_t) + \max_{\tilde{x}^1_{t+1} \in [[\hat{X}_{t+1}|x_t,u_t]]} \min_{\tilde{x}^2_{t+1} \in [[\hat{X}_{t+1}|\tilde{x}_t,u_t]]} d(\tilde{x}^1_{t+1}, \tilde{x}^2_{t+1}) \Big),$$

$$\max_{\tilde{x}_t \in [[X_t|\mu(x_t)]]} \Big(d(x_t, \tilde{x}_t) + \max_{\tilde{x}^2_{t+1} \in [[\hat{X}_{t+1}|\tilde{x}_t,u_t]]} \min_{\tilde{x}^1_{t+1} \in [[\hat{X}_{t+1}|x_t,u_t]]} d(\tilde{x}^1_{t+1}, \tilde{x}^2_{t+1}) \Big) \Bigg\}. \tag{2.151}$$

In the first term in the RHS of (2.151), note that

$$\max_{\tilde{x}^1_{t+1} \in [[\hat{X}_{t+1}|x_t,u_t]]} \min_{\tilde{x}^2_{t+1} \in [[\hat{X}_{t+1}|\tilde{x}_t,u_t]]} d(\tilde{x}^1_{t+1}, \tilde{x}^2_{t+1}) \leq \max_{w_t \in \mathcal{W}_t} d(\mu(f_t(x_t,u_t,w_t)), \mu(f_t(\tilde{x}_t,u_t,w_t)))$$

$$\leq \max_{w_t \in \mathcal{W}_t} \Big(d\big(\mu(f_t(x_t,u_t,w_t)), f_t(x_t,u_t,w_t)\big) + d\big(f_t(x_t,u_t,w_t), f_t(\tilde{x}_t,u_t,w_t)\big)$$

$$+ d\big(f_t(\tilde{x}_t,u_t,w_t), \mu(f_t(\tilde{x}_t,u_t,w_t))\big)\Big)$$

$$\leq \gamma + 2L_{f_t} \cdot \gamma + \gamma, \tag{2.152}$$

where, in the first inequality, we simply use the definition of $\tilde{x}^1_{t+1}$ and $\tilde{x}^2_{t+1}$; in the second inequality, we use the triangle inequality; and in the third inequality, we use the Lagrange continuity of f and the definition of the quanitzation function μ. This exact result also holds for the second term in the RHS of (2.151). Thus, we can substitute (2.152) in (2.151) to conclude that $\mathcal{H}\big([[X_t, \hat{X}_{t+1}|x_t,u_t]], [[X_t, \hat{X}_{t+1}|\mu(x_t),u_t]]\big) \leq \max_{\tilde{x}_t \in [[X_t|\mu(x_t)]]} d(x_t, \tilde{x}_t) + 2\gamma + 2L_{f_t} \cdot \gamma \leq 4\gamma + 2L_{f_t} \cdot \gamma$.

For the second term in the RHS of (2.150), we use Definition 7 to write that

$$\max_{(\tilde{x}_t, \hat{x}_{t+1}) \in [[X_t, \hat{X}_{t+1}|\mu(x_t),u_t]]} \Big| r_t(\psi^{x_t,u_t}(\tilde{x}_t, \hat{x}_{t+1})|x_t,u_t) - r_t(\tilde{x}_t, \hat{x}_{t+1}|\mu(x_t),u_t) \Big|$$

$$= \max_{(\tilde{x}_t, \hat{x}_{t+1}) \in [[X_t, \hat{X}_{t+1}|\mu(x_t),u_t]]} \Bigg| \max_{a_t \in \mathcal{A}_t} (a_t + \mathbb{I}(\psi^{x_t,u_t}(\tilde{x}_t, \hat{x}_{t+1}), a_t|x_t,u_t)) - \max_{a_t \in [[A_t|x_t,u_t]]} a_t$$

$$- \max_{a_t \in \mathcal{A}_t} (a_t + \mathbb{I}(\tilde{x}_t, \hat{x}_{t+1}, a_t|\mu(x_t),u_t)) + \max_{a_t \in [[A_t|\mu(x_t),u_t]]} a_t \Bigg|. \tag{2.153}$$

In the RHS of (2.153), the first two terms can be expanded using Definition 5 in a manner similar to (2.146), to write that

$$\max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(\psi^{x_t,u_t}(\tilde{x}_t, \hat{x}_{t+1}), a_t|x_t,u_t)\big) - \max_{a_t \in \mathcal{A}_t} \big(a_t + \mathbb{I}(\tilde{x}_t, \hat{x}_{t+1}, a_t|\mu(x_t),u_t)\big)$$

$$= \max_{a_t \in [[A_t|x_t,u_t]]} \big(a_t + \mathbb{I}(\psi^{x_t,u_t}(\tilde{x}_t, \hat{x}_{t+1})|a_t,x_t,u_t)\big) - \max_{a_t \in [[A_t|\mu(x_t),u_t]]} \big(a_t + \mathbb{I}(\tilde{x}_t, \hat{x}_{t+1}|a_t,\mu(x_t),u_t)\big)$$

$$= \max_{a_t \in [[A_t | x_t, u_t]]} a_t + \mathbb{I}(\psi^{x_t,u_t}(\tilde{x}_t, \hat{x}_{t+1}) | x_t, u_t) - \max_{a_t \in [[A_t | \mu(x_t), u_t]]} a_t - \mathbb{I}(\tilde{x}_t, \hat{x}_{t+1} | \mu(x_t), u_t). \tag{2.154}$$

where in the second inequality, because $\psi^{x_t,u_t}(\tilde{x}_t) = x_t$ and $\psi^{x_t,u_t}(\hat{x}_{t+1}) = \mu(f_t(x_t, u_t, \bar{w}_t))$, where $\tilde{w}_t = \arg\min_{\bar{w}_t \in \mathcal{W}} d(\hat{x}_{t+1}, f_t(x_t, u_t, \tilde{w}_t))$, it holds that $[[X_t, \hat{X}_{t+1} | a_t, x_t, u_t]] = [[X_t, \hat{X}_{t+1} | x_t, u_t]]$, and similarly, it holds that $[[X_t, \hat{X}_{t+1} | a_t, \mu(x_t), u_t]] = [[X_t, \hat{X}_{t+1} | \mu(x_t), u_t]]$. On substituting (2.154) into (2.153), it holds that

$$\begin{aligned} &\max_{(\tilde{x}_t, \tilde{x}_{t+1}) \in [[X_t, \hat{X}_{t+1} | \mu(x_t), u_t]]} \left| r_t(\psi^{x_t}(\tilde{x}_t, \hat{x}_{t+1}) | x_t, u_t) - r_t(\tilde{x}_t, \hat{x}_{t+1} | \mu(x_t), u_t) \right| \\ &\quad = \max_{(\tilde{x}_t, \hat{x}_{t+1}) \in [[X_t, \hat{X}_{t+1} | \mu(x_t), u_t]]} \left| \mathbb{I}(\psi^{x_t}(\tilde{x}_t, \hat{x}_{t+1}) | x_t, u_t) - \mathbb{I}(\tilde{x}_t, \hat{x}_{t+1} | a_t, \mu(x_t), u_t) \right| = 0, \end{aligned} \tag{2.155}$$

and subsequently, substituting into (2.150) yields that $\epsilon_t = \max\{4\gamma + 2L_{f_t} \cdot \gamma, 0\} = 4\gamma + 2L_{f_t} \cdot \gamma$. □

2.3 Worst-Case Control and Learning Using Partial Observations Over an Infinite Time Horizon

2.3.1 Notation and Preliminaries

In our exposition, we use the mathematical framework of uncertain variables [163] and cost distributions [53, 165]:

1) Cost Measures: Consider a sample space Ω with a sigma algebra $\mathcal{B}(\Omega)$. A cost measure is the non-stochastic analogue of a probability measure. Specifically, it is a function $q : \mathcal{B}(\Omega) \to \{-\infty\} \cup (-\infty, 0]$ satisfying the properties: (1) $q(\Omega) = 0$, (2) $q(\emptyset) = -\infty$, and (3) $q(B) = \sup_{\omega \in B} q(\omega)$ for all $B \in \mathcal{B}(\Omega)$, where $\sup_{\omega \in \emptyset} q(\omega) := -\infty$. Furthermore, for two sets $B^1, B^2 \in \mathcal{B}(\Omega)$ with $q(B^2) > -\infty$, the conditional cost measure of B^1 given B^2 is $q(B^1 \,|\, B^2) := q(B^1, B^2) - q(B^2)$, where $q(B^1, B^2) = \sup_{\omega \in B^1 \cap B^2} q(\omega)$.

2) Uncertain Variables: For a set $\mathscr{X}$, an uncertain variable is a mapping $\mathsf{X} : \Omega \to \mathscr{X}$ and is compactly denoted by $\mathsf{X} \in \mathscr{X}$. This is the non-stochastic equivalent of a random variable. For any $\omega \in \Omega$, its realization is $\mathsf{X}(\omega) = \mathsf{x} \in \mathscr{X}$. Its *marginal range* is the set of feasible realizations $[[\mathsf{X}]] := \{\mathsf{X}(\omega) \,|\, \omega \in \Omega\} \subseteq \mathscr{X}$. The *cost distribution* is $q(\mathsf{x}) := \sup_{\omega \in \{\Omega | \mathsf{X}(\omega) = \mathsf{x}\}} q(\omega)$ for all $\mathsf{x} \in [[\mathsf{X}]]$. The *joint range* of two uncertain variables $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ is the set of feasible simultaneous realizations $[[\mathsf{X}, \mathsf{Y}]] := \big\{\big(\mathsf{X}(\omega), \mathsf{Y}(\omega)\big) \,|\, \omega \in \Omega\big\} \subseteq \mathscr{X} \times \mathscr{Y}$. The two uncertain variables are *independent* if $[[\mathsf{X}, \mathsf{Y}]] = [[\mathsf{X}]] \times [[\mathsf{Y}]]$. The *conditional range* of X given a realization y of Y is the set $[[\mathsf{X}|\mathsf{y}]] := \big\{\mathsf{X}(\omega) \mid \mathsf{Y}(\omega) = \mathsf{y},\ \omega \in \Omega\big\}$. The cost distribution of any $\mathsf{x} \in [[\mathsf{X}]]$ given a realization $\mathsf{y} \in [[\mathsf{Y}]]$ with $q(\mathsf{y}) > -\infty$ is $q(\mathsf{x} \,|\, \mathsf{y}) = q(\mathsf{x}, \mathsf{y}) - q(\mathsf{y})$, where $q(x, y) = \sup_{\omega \in \{\Omega | X(\omega) = x, Y(\omega) = y\}} q(\omega)$.

3) Hausdorff Distance: Consider two bounded, non-empty subsets $\mathscr{X}, \mathscr{Y}$ of a metric space $(\mathscr{S}, \eta)$, where $\eta : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}_{\geq 0}$ is the metric. The Hausdorff distance between $\mathscr{X}$ and $\mathscr{Y}$ is the pseudo-metric [194, Chapter 1.12]:

$$\mathcal{H}(\mathscr{X}, \mathscr{Y}) := \max\big\{\sup_{\mathsf{x} \in \mathscr{X}} \inf_{\mathsf{y} \in \mathscr{Y}} \eta(\mathsf{x}, \mathsf{y}), \sup_{\mathsf{y} \in \mathscr{Y}} \inf_{\mathsf{x} \in \mathscr{X}} \eta(\mathsf{x}, \mathsf{y})\big\}. \tag{2.156}$$

Furthermore, if $f : \mathscr{S} \to \mathbb{R}$ is a Lipschitz continuous function with a constant $L_f \in \mathbb{R}_{\geq 0}$, then [166, Lemma 5]:

$$\Big| \sup_{x \in \mathscr{X}} f(x) - \sup_{y \in \mathscr{Y}} f(y) \Big| \leq L_f \cdot \mathcal{H}(\mathscr{X}, \mathscr{Y}). \tag{2.157}$$

2.3.2 Problem Formulation

We consider the control of an uncertain system which evolves in discrete time steps. At each time $t \in \mathbb{N} = \{0, 1, 2, \dots\}$ an agent collects an observation on the system as the uncertain variable $Y_t \in \mathcal{Y}$ and generates a control action denoted by the uncertain variable $U_t \in \mathcal{U}$. After generating the action at each t, the agent incurs a cost denoted by the uncertain variable $C_t \in \mathcal{C} \subset \mathbb{R}_{\geq 0}$. The set $\mathcal{C}$ is bounded by $\min\{\mathcal{C}\} = c^{\min}$ and $\max\{\mathcal{C}\} = c^{\max}$. We formulate our problem for a general case where the agent may not have knowledge of a state-space model for the system. Thus, we use an *input-output model* to describe the evolution of the system, as follows. At each $t \in \mathbb{N}$, the system receives two inputs: the action U_t, and an uncontrolled disturbance $W_t \in \mathcal{W}$. The disturbances $\{W_t \,|\, t \in \mathbb{N}\}$ constitute a sequence of independent uncertain variables. After receiving the inputs at each time $t \in \mathbb{N}$, the system generates two outputs: (1)

the observation $Y_{t+1} = h_{t+1}(W_{0:t}, U_{0:t})$, where $h_{t+1} : \mathcal{W}^t \times \mathcal{U}^t \to \mathcal{Y}$ is the observation function; and (2) the cost $C_t = d_t(W_{0:t}, U_{0:t})$, where $d_t : \mathcal{W}^t \times \mathcal{U}^t \to \mathcal{C}$ is the cost function. The initial observation is generated as $Y_0 = h_0(W_0)$.

The agent perfectly recalls all observations and control actions and at each $t \in \mathbb{N}$, the agent's memory is the uncertain variable $M_t := (Y_{0:t}, U_{0:t-1})$ taking values in $\mathcal{M}_t := \mathcal{Y}^t \times \mathcal{U}^{t-1}$. The agent uses a control law $g_t : \mathcal{M}_t \to \mathcal{U}$ to generate the action $U_t = g_t(M_t)$ as a function of the memory. The control strategy is the collection of control laws $\boldsymbol{g} := (g_0, g_1, \dots)$ with a feasible set $\mathcal{G}$. The performance of a strategy $\boldsymbol{g} \in \mathcal{G}$ is given by the *worst-case discounted cost*,

$$\mathcal{J}(\boldsymbol{g}) := \lim_{T \to \infty} \sup_{c_{0:T} \in [[C_{0:T}]]^{\boldsymbol{g}}} \sum_{t=0}^{T} \gamma^t \cdot c_t, \tag{2.158}$$

where $\gamma \in (0,1)$ is a discount parameter, the marginal range $[[C_{0:T}]]^{\boldsymbol{g}}$ is the set of all feasible costs consistent with the strategy $\boldsymbol{g}$ and with the set of feasible disturbances $\mathcal{W}$. The limit in (2.158) is well defined because $C_t \leq c^{\max}$ for all t. Next, we define the control problem with known dynamics.

Problem 3. The optimization problem is to derive the infimum value $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the feasible sets $\{\mathcal{U}, \mathcal{W}, \mathcal{Y}, \mathcal{C}\}$ and the functions $\{h_t, d_t \,|\, t \in \mathbb{N}\}$.

If the minimum value is achieved in Problem 3, the minimizing argument $\boldsymbol{g}^* = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$ is called an optimal control strategy. Our aim is to tractably compute the optimal value and an optimal strategy, if one exists. We impose the following assumption in our analysis.

Assumption 6. We consider that the sets $\{\mathcal{U}, \mathcal{W}, \mathcal{Y}\}$ are each bounded subsets of a metric space $(\mathscr{S}, \eta)$ and $\mathcal{C}$ is a bounded subset of $\mathbb{R}_{\geq 0}$.

Assumption 6 ensures that all uncertain variables take values in bounded sets and that we can use the Hausdorff pseudo-metric (2.156) as a distance measure between them.

Remark 15. We first derive results for Problem 3 with known dynamics. However, our main results in Section 2.3.6 are also suitable for reinforcement learning problems with unknown dynamics. We illustrate this with an example in Section 2.3.9.

2.3.3 Dynamic Program and Information States

In this section, we first present value functions to evaluate the performance of any strategy $\boldsymbol{g} \in \mathcal{G}$. Next, we present a memory-based DP decomposition of Problem 3 that approximately computes the value functions with arbitrary precision. However, because the memory grows in size with time, this DP suffers from exponentially increasing computation with an increase in precision. To alleviate this computational challenge, we present the notion of information states in Subsection 2.3.4. To construct value functions, we first define the *accrued cost* at each $t \in \mathbb{N}$ as the sum of past incurred costs

$$A_t := \sum_{\ell=0}^{t-1} \gamma^{\ell} \cdot C_{\ell}, \tag{2.159}$$

which satisfies $A_{t+1} = A_t + \gamma^t \cdot C_t$ with $A_0 := 0$. This is well defined in the limit $t \to \infty$ because $\lim_{t\to\infty} A_t \leq \lim_{t\to\infty} \sum_{\ell=0}^{t-1} \gamma^{\ell} \cdot c^{\max} = \frac{c^{\max}}{1-\gamma} =: a^{\max}$. Thus, $A_t \in [0, a^{\max}]$ for all $t \in \mathbb{N}$. Similarly, the *cost-to-go* at any $t \in \mathbb{N}$ is the sum of future all costs still to be incurred

$$C_t^{\infty} := \sum_{\ell=t}^{\infty} \gamma^{\ell-t} \cdot C_{\ell}. \tag{2.160}$$

Note that $C_t^{\infty} \in [0, a^{\max}]$ for all t and that $C_t^{\infty} = C_t + \gamma \cdot C_{t+1}^{\infty}$. Then, for all $t \in \mathbb{N}$, we can define a value function for any $\boldsymbol{g} \in \mathcal{G}$ as

$$V_t^{\boldsymbol{g}}(m_t) := \sup_{a_t, c_t^{\infty} \in [[A_t, C_t^{\infty} | m_t]]^{\boldsymbol{g}}} \left(a_t + \gamma^t \cdot c_t^{\infty} \right), \tag{2.161}$$

where $[[A_t, C_t^{\infty} | m_t]]^{\boldsymbol{g}}$ is the conditional range induced by the choice of strategy $\boldsymbol{g}$. From the definition of the value functions, at $t = 0$ it holds that $\sup_{y_0 \in \mathcal{Y}} V_0^{\boldsymbol{g}}(y_0) = \mathcal{J}(\boldsymbol{g})$, where $m_0 = y_0$. Thus, the value function $V_0^{\boldsymbol{g}}(y_0)$ evaluates the performance of any strategy $\boldsymbol{g}$ for an initial observation y_0. Similarly, the optimal value function at each $t \in \mathbb{N}$ is

$$V_t(m_t) := \inf_{\boldsymbol{g} \in \mathcal{G}} V_t^{\boldsymbol{g}}(m_t), \tag{2.162}$$

and the optimal value is $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g}) = \sup_{y_0 \in \mathcal{Y}} V_0(y_0)$.

Given the value functions in (2.161) and (2.162), we can evaluate the performance of a strategy and compare it with the optimal performance. However, there is no natural DP decomposition to compute these value functions in an infinite-horizon system with no terminal time. Thus, we construct a memory-based DP that assumes a finite horizon $T \in \mathbb{N}$ and use it to recursively compute approximations of the value functions. For any $\boldsymbol{g} \in \mathcal{G}$, we define finite-horizon evaluation functions for all $m_t \in \mathcal{M}_t$ and each $t = 0, \dots, T-1$ as

$$J_t^{\boldsymbol{g}}(m_t; T) := \sup_{m_{t+1} \in [[M_{t+1} | m_t]]^{\boldsymbol{g}}} J_{t+1}^{\boldsymbol{g}}(m_{t+1}; T), \tag{2.163}$$

where $J_T^{\boldsymbol{g}}(m_T; T) := \sup_{a_T, c_T \in [[A_T, C_T | m_T]]^{\boldsymbol{g}}} (a_T + \gamma^T \cdot c_T)$. Similarly, we define approximately optimal finite-horizon functions for all $m_t \in \mathcal{M}_t$ and each $t = 0, \dots, T-1$ as

$$J_t(m_t; T) := \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} J_{t+1}(m_{t+1}; T), \tag{2.164}$$

where, $J_T(m_T; T) := \inf_{u_T \in \mathcal{U}} \sup_{a_T, c_T \in [[A_T, C_T | m_T, u_T]]} (a_T + \gamma^T \cdot c_T)$. Note that the finite-horizon functions $J_t^{\boldsymbol{g}}(m_t; T)$ and $J_t(m_t; T)$ at any $t = 0, \dots, T$ are parameterized by the choice of horizon $T \in \mathbb{N}$. Next, we bound the approximation error between the value functions and their finite-horizon counterparts.

Lemma 14. *For any finite horizon $T \in \mathbb{N}$ and for all $m_t \in \mathcal{M}_t$ and each $t = 0, \dots, T$,*

$$\textbf{\textit{a)}}\ \frac{\gamma^{T+1} \cdot c^{\min}}{1-\gamma} + J_t^{\boldsymbol{g}}(m_t; T) \leq V_t^{\boldsymbol{g}}(m_t) \leq J_t^{\boldsymbol{g}}(m_t; T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}, \tag{2.165}$$

$$\textbf{\textit{b)}}\ \frac{\gamma^{T+1} \cdot c^{\min}}{1-\gamma} + J_t(m_t; T) \leq V_t(m_t) \ \leq J_t(m_t; T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}. \tag{2.166}$$

Proof. *a)* We prove each inequality in (2.165) using backward induction. For the upper bound at time T, we use the dynamics of the accrued cost and cost-to-go to write that

$$\begin{aligned} V_T^{\boldsymbol{g}}(m_T) &= \sup_{a_T,c_T,c^\infty_{T+1}\in[[A_T,C_T,C^\infty_{T+1}|m_T]]^{\boldsymbol{g}}} (a_T + \gamma^T \cdot c_T + \gamma^{T+1} \cdot c^\infty_{T+1}) \\ &\leq \sup_{a_T,c_T\in[[A_T,C_T|m_T]]^{\boldsymbol{g}}} \big(a_T + \gamma^T \cdot c_T\big) + \gamma^{T+1} \cdot a^{\max} \\ &\leq J_t^{\boldsymbol{g}}(m_t;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}. \end{aligned} \tag{2.167}$$

The lower bound at time T follows from $\frac{c^{\min}}{1-\gamma} \leq c^\infty_{T+1}$ using the same sequence of arguments as before. This forms the basis of our induction. Next, consider the hypothesis that (2.165) holds at time $t+1$. For the upper bound at time t, by definition

$$\begin{aligned} V_t^{\boldsymbol{g}}(m_t) &= \sup_{a_t,c_t,c^\infty_{t+1}\in[[A_t,C_t,C^\infty_{t+1}|m_t]]^{\boldsymbol{g}}} \big(a_t + \gamma^t \cdot c_t + \gamma^{t+1} \cdot c^\infty_{t+1}\big) \\ &= \sup_{a_{t+1},c^\infty_{t+1}\in[[A_{t+1},C^\infty_{t+1}|m_t]]^{\boldsymbol{g}}} (a_{t+1} + \gamma^{t+1} \cdot c^\infty_{t+1}) \\ &= \sup_{m_{t+1}\in[[M_{t+1}|m_t]]^{\boldsymbol{g}}} \sup_{a_{t+1},c^\infty_{t+1}\in[[A_{t+1},C^\infty_{t+1}|m_{t+1}]]^{\boldsymbol{g}}} (a_{t+1} + \gamma^{t+1} \cdot c^\infty_{t+1}) \\ &= \sup_{m_{t+1}\in[[M_{t+1}|m_t]]^{\boldsymbol{g}}} V_{t+1}^{\boldsymbol{g}}(m_{t+1}) \leq \sup_{m_{t+1}\in[[M_{t+1}|m_t]]^{\boldsymbol{g}}} J_{t+1}^{\boldsymbol{g}}(m_{t+1};T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma} \\ &= J_t^{\boldsymbol{g}}(m_t;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}, \end{aligned} \tag{2.168}$$

where, in the fourth equality, we use (2.161) for $V_{t+1}^{\boldsymbol{g}}(m_{t+1})$; and in the inequality, we use the hypothesis. The lower bound follows from the same sequence of arguments. Thus, (2.165) holds using induction.

b) We can prove the lower bound in (2.166) by taking the infimum on both sides of the lower bound in (2.165). To prove the upper bound in (2.166), we first note that $J_t(m_t;T) = \inf_{\boldsymbol{g}\in\mathcal{G}} J_t^{\boldsymbol{g}}(m_t;T)$ for all $t = 0,\dots,T$ using standard DP arguments for terminal-cost problems [53]. Then, at time T, by definition

$$\begin{aligned} V_T(m_T) &= \inf_{\boldsymbol{g}\in\mathcal{G}} V_T^{\boldsymbol{g}}(m_T) \\ &\leq \inf_{\boldsymbol{g}\in\mathcal{G}} J_T^{\boldsymbol{g}}(m_T;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma} = J_T(m_T;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}. \end{aligned} \tag{2.169}$$

Using this as the basis, the result follows for all $t = 0, \dots, T$ using the same induction arguments as in (2.165). □

Lemma 14 establishes that the approximation error between finite-horizon functions and corresponding value functions decreases as the horizon $T \in \mathbb{N}$ increases. A direct consequence of (2.166) is that $\lim_{T \to \infty} J_0(y_0; T) = V_0(y_0)$ for all $y_0 \in \mathcal{Y}$. Note, however, that the domain of $J_T(m_T; T)$ is $\mathcal{M}_T = \mathcal{Y}^T \times \mathcal{U}^{T-1}$ which grows with T, and in the limit $T \to \infty$, the set $\mathcal{M}_T$ is infinite-dimensional. Thus, it is computationally intractable to achieve close approximations of the optimal value using (2.164). We address this issue in the next subsection using *information states*, which take values in time-invariant spaces.

2.3.4 Information States

In this subsection, we present the notion of information states which take values in time-invariant spaces. Then, we use them to construct a time-invariant DP decomposition which converges to the optimal value of Problem 3. To begin, recall from Section 2.3.1 that a *cost distribution* is the non-stochastic equivalent of a probability distribution for uncertain variables. We use two specific cost distributions in our exposition, defined as follows.

Definition 11. Let $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ be two uncertain variables. The *indicator function* for $\mathbb{I} : \mathscr{X} \to \{\infty, 0\}$ for $\mathsf{x} \in \mathscr{X}$ is given by

$$\mathbb{I}(\mathsf{x}) := \begin{cases} 0, & \text{if } \mathsf{x} \in [[\mathsf{X}]], \\ -\infty, & \text{if } \mathsf{x} \notin [[\mathsf{X}]], \end{cases} \tag{2.170}$$

and the conditional indicator function $\mathbb{I} : \mathcal{X} \times \mathcal{Y} \to \{\infty, 0\}$ for any $\mathsf{x} \in \mathscr{X}$ given a realization $\mathsf{y} \in [[\mathsf{Y}]]$ is given by

$$\mathbb{I}(\mathsf{x} \,|\, \mathsf{y}) := \begin{cases} 0, & \text{if } \mathsf{x} \in [[\mathsf{X} \,|\, \mathsf{y}]], \\ -\infty, & \text{if } \mathsf{x} \notin [[\mathsf{X} \,|\, \mathsf{y}]]. \end{cases} \tag{2.171}$$

The indicator function verifies whether the input takes values within the conditional range of an uncertain variable and it satisfies the properties of a cost distribution from Subsection 2.3.1. Next, we use it to define the accrued distribution.

Definition 12. Let $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ be two uncertain variables and let $A_t \in \mathcal{A}$ be the accrued cost at any $t \in \mathbb{N}$. An *accrued distribution* for any $\mathsf{x} \in \mathscr{X}$ at any $t \in \mathbb{N}$ is a function $r_t : \mathscr{X} \to \{-\infty\} \cup [-a^{\max}, 0]$, given by

$$r_t(\mathsf{x}) := \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(\mathsf{x}, a_t)\big) - \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(a_t)\big), \tag{2.172}$$

and the conditional accrued distribution for $\mathsf{x} \in \mathscr{X}$ given a realization $\mathsf{y} \in [[\mathsf{Y}]]$ is a function $r_t : \mathscr{X} \times \mathscr{Y} \to \{-\infty\} \cup [-a^{\max}, 0]$, given by

$$r_t(\mathsf{x} \,|\, \mathsf{y}) := \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(\mathsf{x}, a_t \,|\, \mathsf{y})\big) - \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(a_t \,|\, \mathsf{y})\big). \tag{2.173}$$

The accrued distribution returns $-\infty$ when the input is not within the range of an uncertain variable and it returns an output from $[-a^{\max}, 0]$ otherwise. It also satisfies all properties of a cost distribution as defined in Subsection 2.3.1. Note that, at each $t \in \mathbb{N}$, given the realizations $m_t \in \mathcal{M}_t$, $u_t \in \mathcal{U}$ and the dynamics, we can compute the indicator $\mathbb{I}(c_t, m_{t+1} \,|\, m_t, u_t)$, and the accrued distribution $r_t(c_t, m_{t+1} \,|\, m_t, u_t)$ for all $c_t \in \mathcal{C}$ and $m_{t+1} \in \mathcal{M}_{t+1}$. We use these accrued distributions to define information states.

Definition 13. An *information state* at any $t \in \mathbb{N}$ is an uncertain variable $S_t = \sigma_t(M_t)$ taking values in a bounded, time-invariant subset $\mathcal{S}$ of a metric space $(\mathscr{S}, \eta)$. Furthermore, there exists a time-invariant cost distribution $\rho : \mathcal{C} \times \mathcal{S} \times \mathcal{S} \times \mathcal{U} \to \{-\infty\} \cup [-a^{\max}, 0]$ such that, for all $t \in \mathbb{N}$, for all $m_t \in \mathcal{M}_t$, $u_t \in \mathcal{U}$, $c_t \in \mathcal{C}$ and $s_{t+1} \in \mathcal{S}$, it satisfies

$$r_t(c_t, s_{t+1} \,|\, m_t, u_t) = \rho(c_t, s_{t+1} \,|\, \sigma_t(m_t), u_t), \tag{2.174}$$

where each conditional cost distribution in (2.174) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

Next, we use the information state to construct a time-invariant operator $\mathcal{T}$ that yields a fixed-point equation to recursively compute the optimal value in Problem 3. We first define the deterministic *cumulative discount* at any $t \in \mathbb{N}$ as $z_t := \gamma^t \in (0,1]$, where $z_0 = 1$ and $z_{t+1} = \gamma \cdot z_t$. Then, for any uniformly bounded function $\Lambda : \mathcal{S} \times (0,1] \to \mathbb{R}$ we define $\mathcal{T} : [\mathcal{S} \times (0,1] \to \mathbb{R}] \to [\mathcal{S} \times (0,1] \to \mathbb{R}]$, such that

$$[\mathcal{T}\Lambda](s,z) := \inf_{u \in \mathcal{U}} \sup_{c \in \mathcal{C},\, s' \in \mathcal{S}} \left(c + \gamma \cdot \Lambda(s', \gamma \cdot z) + \frac{\rho(c, s' \mid s, u)}{z} \right), \tag{2.175}$$

for all $s \in \mathcal{S}$ and $z \in (0,1]$. Note that we use time-invariant notation for all variables in (2.175) because the sets and functions in the RHS are time-invariant.

Remark 16. The RHS of (2.175) is finite in the limit $z \to 0$ because c has a finite upper bound, the function Λ is uniformly bounded, and $\sup_{c \in \mathcal{C}, s' \in \mathcal{S}} \rho(c, s' \mid s, u) = 0$ for all $s \in \mathcal{S}$ and $u \in \mathcal{U}$ from the definition of cost distributions.

Due to discounting, $\mathcal{T}$ is a contraction mapping (see proof in Appendix A) and therefore, using the Banach fixed point theorem, the equation $\Lambda = \mathcal{T}\Lambda$ admits a unique solution $\Lambda^\infty = \mathcal{T}\Lambda^\infty$. Starting at $\Lambda^0(s,z) := 0$, the fixed-point iteration around $\mathcal{T}$ generates a sequence of functions

$$\Lambda^{n+1}(s,z) = [\mathcal{T}\Lambda^n](s,z) = [\mathcal{T}^n \Lambda^0](s,z), \tag{2.176}$$

for all $n = 1, 2, \dots$, such that

$$\lim_{n \to \infty} \mathcal{T}^n V^0 = \Lambda^\infty. \tag{2.177}$$

The fixed-point iteration in (2.176) forms a time-invariant DP. Next, we establish that $\Lambda^n(\sigma_t(m_t), z_t)$, for any $n \in \mathbb{N}$, can be used to estimate the value function $V_t(m_t)$ of Problem 3 at any $t \in \mathbb{N}$, with estimation error that decreases in n.

Theorem 12. *Consider the function Λ^n, for any $n \in \mathbb{N}$, generated using (2.176). Then, for all $t \in \mathbb{N}$, it holds that*

$$\begin{aligned} \frac{\gamma^{n+t} \cdot c^{\min}}{1-\gamma} + \gamma^t \cdot \Lambda^n(\sigma_t(m_t), \gamma^t) + \sup_{a_t \in [[A_t | m_t]]} a_t \ &\leq \ V_t(m_t) \\ &\leq \sup_{a_t \in [[A_t | m_t]]} a_t + \gamma^t \cdot \Lambda^n(\sigma_t(m_t), \gamma^t) + \frac{\gamma^{n+t} \cdot c^{\max}}{1-\gamma}. \end{aligned} \tag{2.178}$$

Proof. We show (2.178) using (2.166) from Lemma 14. For this purpose, we first show that for any finite horizon $T \in \mathbb{N}$, the following property holds for each $t = 0, \dots, T$:

$$J_t(m_t; T) = \gamma^t \cdot \Lambda^{T-t+1}\big(\sigma_t(m_t), \gamma^t\big) + \sup_{a_t \in [[A_t | m_t]]} a_t, \tag{2.179}$$

where Λ^{T-t+1} is the $(T-t+1)$-th iterated function in (2.176). We prove (2.179) by induction. At time T, recall from (2.164) that $J_T(m_T; T) = \inf_{u_T \in \mathcal{U}} \sup_{a_T, c_T \in [[A_T, C_T | m_T, u_T]]}(a_T$ $+\gamma^T \cdot c_T)$. Using the indicator function in the RHS,

$$\sup_{a_T, c_T \in [[A_T, C_T | m_T, u_T]]} (a_T + \gamma^T \cdot c_T) = \sup_{a_T \in \mathcal{A}, c_T \in \mathcal{C}} (a_T + \gamma^T \cdot c_T + \mathbb{I}(a_T, c_T \,|\, m_T, u_T))$$
$$= \sup_{a_T \in \mathcal{A}, c_T \in \mathcal{C}} \Big(a_T + \gamma^T \cdot c_T + \mathbb{I}(a_T, c_T \,|\, m_T, u_T) - \sup_{a_T \in \mathcal{A}} (a_T + \mathbb{I}(a_T \,|\, m_T, u_T))\Big) + \sup_{a_T \in [[A_T | m_T]]} a_T$$
$$= \gamma^T \cdot \sup_{c_T \in \mathcal{C}} \big(c_T + \gamma^{-T} \cdot r_T(c_T |\, m_T, u_T)\big) + \sup_{a_T \in [[A_T | m_T]]} a_T, \tag{2.180}$$

where, in the second equality, we add and subtract the term $\sup_{a_T \in \mathcal{A}}(a_T + \mathbb{I}(a_T \,|\, m_T, u_T))$; and, in the last equality, we use the definition of the accrued cost. Consequently, we can write that

$$J_T(m_T; T) = \inf_{u_T \in \mathcal{U}} \sup_{c_T \in \mathcal{C}, s_{T+1} \in \mathcal{S}} \gamma^T \cdot \Big(c_T + \gamma \cdot \Lambda^0(s_{T+1}, \gamma^{T+1}) + \gamma^{-T} \cdot r_T(c_T, s_{T+1} \,|\, m_T, u_T)\Big)$$
$$+ \sup_{a_T \in [[A_T | m_T]]} a_T = \gamma^T \cdot \Lambda^1(\sigma_T(m_T), \gamma^T) + \sup_{a_T \in [[A_T | m_T]]} a_T, \tag{2.181}$$

where, in the first equality, recall that $\Lambda^0(s_{T+1}, \gamma^{T+1}) := 0$ identically; and in the second equality, we use $r_T(c_T, s_{T+1} | m_T, u_T) = \rho(c_T, s_{T+1} | m_T, u_T)$ and the definition of $\Lambda^1(\sigma_T(m_T), \gamma^T)$ in (2.176). This forms the basis of our induction. Next, consider the hypothesis that (2.179) holds for time $t+1$. Using the definition of the finite-horizon function at time t and the induction hypothesis,

$$J_t(m_t; T) = \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} J_{t+1}(m_{t+1}; T)$$
$$= \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \sup_{a_{t+1} \in [[A_{t+1} | m_{t+1}]]} \big(\gamma^{t+1} \cdot \Lambda^{T-t}(\sigma_{t+1}(m_{t+1}), \gamma^{t+1}) + a_{t+1}\big)$$
$$= \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1}, a_{t+1} \in [[M_{t+1}, A_{t+1} | m_t, u_t]]} (\gamma^{t+1} \cdot \Lambda^{T-t}(\sigma_{t+1}(m_{t+1}), \gamma^{t+1}) + \gamma^t \cdot c_t + a_t)$$

$$= \gamma^t \cdot V^{T-t+1}(\sigma_t(m_t), \gamma^t) + \sup_{a_t \in [[A_t | m_t]]} a_t, \quad (2.182)$$

where, in the third equality, we use $a_{t+1} = a_t + \gamma^t \cdot c_t$ and rearrange the terms, and the fourth equality follows from the same sequence of arguments as time step T. This proves (2.179) by induction. Then, (2.178) follows directly for all $t \in \mathbb{N}$ and all $n \in \mathbb{N}$ by substituting (2.179) into (2.166) and selecting the horizon $T = t + n - 1$. □

Theorem 12 allows us to characterize the error between the optimal value $V_0(y_0)$ and $\Lambda^n(\sigma_0(y_0), 1)$ for any $y_0 \in \mathcal{Y}$ by selecting $t = 0$ in (2.178), it follows that

$$\frac{\gamma^n \cdot c^{\min}}{1 - \gamma} + \Lambda^n(\sigma_0(y_0), 1) \leq V_0(y_0) \leq \Lambda^n(\sigma_0(y_0), 1) + \frac{\gamma^n \cdot c^{\max}}{1 - \gamma}, \quad (2.183)$$

where, recall that $a_0 = 0$ and $z_0 = 1$. These bounds imply that as the number of iterations $n \to \infty$, the fixed point Λ^∞ satisfies

$$\Lambda^\infty(\sigma_0(y_0), 1) = V_0(y_0). \quad (2.184)$$

Thus, the DP in (2.176) recursively computes the optimal value of Problem 3. Next, consider that the infimum is achieved in the RHS of (2.175) for each function Λ^n, $n \in \mathbb{N}$. Then, we define a time-invariant control strategy $\boldsymbol{\pi}^* := (\pi^*, \pi^*, \dots)$ where the control law at each $t \in \mathbb{N}$ is the minimizing argument for Λ^∞, i.e.,

$$\pi^*(s, z) := \arg\min_{u \in \mathcal{U}} \sup_{c \in \mathcal{C}, s' \in \mathcal{S}} \left(c + \gamma \cdot \Lambda^\infty(s', \gamma \cdot z) + \frac{\rho(c, s' \mid s, u)}{z} \right), \quad (2.185)$$

for all $s \in \mathcal{S}$ and $z \in (0, 1]$. A corresponding memory-based strategy is $\boldsymbol{g}^* = (g_0^*, g_1^*, \dots)$ with $g_t^*(m_t) := \pi^*(\sigma_t(m_t), \gamma^t)$ for all $m_t \in \mathcal{M}_t$ and all t. Then, $\boldsymbol{g}^*$ achieves the optimal value, i.e., $V_0^{\boldsymbol{g}^*}(y_0) = V_0(y_0)$ (see proof in Appendix B). Thus, the information-state gives an optimal solution to Problem 3.

Remark 17. Typically, the infinite-horizon worst-case problem is nonstationary due to the presence of discounted costs, a concern found also in the analysis of discounted risk-sensitive problems [200, 201]. To overcome this challenge, we take inspiration from recent developments in certainty-equivalent risk-sensitive problems from [16, 202].

Thus, when formulating out DP (2.175), we augment the information state $s \in \mathcal{S}$ with a deterministic term $z \in (0,1]$ and as a result, the control law π^* in (2.185) is a time-invariant function of the variables $(s,z) \in \mathcal{S} \times (0,1]$. During implementation, the control action at each $t \in \mathbb{N}$ is generated as $u_t = \pi^*(s_t, \gamma^t)$, where s_t the realization of the information state and $z_t = \gamma^t$ tracks the accumulated discount at time t. In contrast, the equivalent control laws using only the information state $p_t^*(s_t) := \pi^*(s_t, \gamma^t)$ and only the memory $g_t^*(m_t) = p_t^*(\sigma_t(m_t))$ are each not time-invariant because of the presence γ^t in their respective RHSs. Thus, explicitly tracking the accumulated discount $z_t = \gamma^t$ is sufficient to obtain a time-invariant control strategy when the information state takes values in a time-invariant space, a result consistent with [16, 202].

2.3.5 Examples of Information States

In this subsection, we consider a system with a known state-space model to present specific information states which satisfy Definition 13. Consider a system with a state $X_t \in \mathcal{X}$ which starts at X_0 and evolves as $X_{t+1} = f_t(X_t, U_t, W_t)$, the observation is $Y_t = h_t(X_t, W_t)$ and the incurred cost is $C_t = d_t(X_t, U_t)$ at each t, where $N_t \in \mathcal{N}$ is an uncontrolled disturbance. The uncontrolled variables $\{X_0, N_t, W_t \,|\, t \in \mathbb{N}\}$ take realizations independently. Next, we give examples of information states for different cases:

1) Perfectly observed systems: For all $t \in \mathbb{N}$, let $Y_t = X_t$, an information state is $S_t = X_t \in \mathcal{X}$.

2) Perfectly observed systems with deep dynamics: For all $t \in \mathbb{N}$, let $Y_t = X_t$ and $X_{t+1} = f(X_{t:t-k}, U_t, W_t)$, an information state is $S_t = (X_{t-k}, \dots, X_t) \in \mathcal{X}^{k+1}$.

3) Partially observed systems: Consider a generic partially observed system. An information state at each $t \in \mathbb{N}$ is the function-valued variable $S_t : \mathcal{X} \to \{-\infty\} \cup [-a^{\max}, 0]$. At each t, for a given $m_t \in \mathcal{M}_t$, its realization is a function $s_t(x_t) := r_t(x_t|m_t)$, where $r_t(\cdot)$ is an accrued distribution. This is a normalization of the results in [35, 53].

4) Systems with action dependent costs: For all $t \in \mathbb{N}$, let the cost of a partially observed system be $d_t(U_t) \in \mathbb{R}_{\geq 0}$. Then, an information state is the conditional range $S_t = [[X_t \mid M_t]] \in \mathcal{B}(\mathcal{X})$, where $\mathcal{B}(\mathcal{X})$ is the set of all subsets of $\mathcal{X}$.

Remark 18. Definition 13 helps us identify information states when system dynamics are known. However, such a representation often needs to be learned purely from observation and cost data, without knowledge of dynamics. Thus, in the next section, we specialize the notion of information states to systems with observable costs and define approximate information states that can be learned from output data.

2.3.6 Systems with Observable Costs

In this section, we analyze Problem 3 in the case where the agent observes the incurred cost at each instance of time. Thus, at each $t \in \mathbb{N}$, the agent receives a realization of (Y_t, C_t) and the memory is $M_t = (Y_{0:t}, C_{0:t-1}, U_{0:t-1})$. We first prove that for such a system, the accrued distribution in Definition 13 at each $t \in \mathbb{N}$ reduces to simply an indicator.

Lemma 15. *Consider Problem 3 with observable costs. At each $t \in \mathbb{N}$, for any given realizations $c_t \in \mathcal{C}$, $m_{t+1} \in \mathcal{M}_{t+1}$, $m_t \in \mathcal{M}_t$, and $u_t \in \mathcal{U}$, it holds that*

$$r_t(c_t, m_{t+1} \,|\, m_t, u_t) = \mathbb{I}(c_t, m_{t+1} \,|\, m_t, u_t). \tag{2.186}$$

Proof. Let the given realization of the memory at time t be $m_t = (\tilde{y}_{0:t}, \tilde{c}_{0:t-1}, \tilde{u}_{0:t-1})$. Then, we expand the accrued distribution at any t as

$$\begin{aligned} r_t(c_t, m_{t+1} \,|\, m_t, u_t) &= \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(a_t, c_t, m_{t+1} \,|\, m_t, u_t)\big) - \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(a_t \,|\, m_t)\big) \\ &= \sup_{a_t \in \mathcal{A}} \big(a_t + \mathbb{I}(a_t \,|\, m_t, u_t, c_t, m_{t+1}) + \mathbb{I}(c_t, m_{t+1} \,|\, m_t, u_t)\big) - \sum_{\ell=0}^{t-1} \gamma^\ell \cdot \tilde{c}_\ell \\ &= \sum_{\ell=0}^{t-1} \gamma^\ell \cdot \tilde{c}_\ell + \mathbb{I}(c_t, m_{t+1} \,|\, m_t, u_t) - \sum_{\ell=0}^{t-1} \gamma^\ell \cdot \tilde{c}_\ell = \mathbb{I}(c_t, m_{t+1} \,|\, m_t, u_t), \end{aligned} \tag{2.187}$$

where, in the second equality, the realization of A_t is determined as $\tilde{a}_t = \sum_{\ell=0}^{t-1} \gamma^\ell \cdot \tilde{c}_\ell$ given m_t; and in the third equality, $\mathbb{I}(a_t \,|\, m_t, u_t, c_t, m_{t+1}) = 0$ only if $a_t = \tilde{a}_t$. □

Motivated by the result of Lemma 15, we present a simpler notion of information states.

Definition 14. An *information state* for Problem 3 with observable costs at any $t \in \mathbb{N}$ is an uncertain variable $\bar{S}_t = \bar{\sigma}_t(M_t)$ taking values in a bounded, time-invariant set $\bar{\mathcal{S}}$. For all $t \in \mathbb{N}$, for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}_t$, it satisfies that

$$[[C_t, \bar{S}_{t+1} \,|\, m_t, u_t]] = [[C_t, \bar{S}_{t+1} \,|\, \bar{\sigma}_t(m_t), u_t]], \tag{2.188}$$

where each conditional range in (2.188) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

Next, we use the information state from Definition 14 to construct a time-invariant operator $\bar{\mathcal{T}} : [\bar{\mathcal{S}} \to \mathbb{R}] \to [\bar{\mathcal{S}} \to \mathbb{R}]$, such that, for any uniformly bounded function $\bar{\Lambda} : \bar{\mathcal{S}} \to \mathbb{R}$,

$$[\bar{\mathcal{T}}\bar{\Lambda}](\bar{s}) := \inf_{u \in \mathcal{U}} \sup_{c, \bar{s}' \in [[C, \bar{S}' | \bar{s}, u]]} \big(c + \gamma \cdot \bar{\Lambda}(\bar{s}')\big). \tag{2.189}$$

Note that (2.189) is simpler than (2.175) and $\bar{\mathcal{T}}$ is also a contraction mapping. Thus, $\bar{\Lambda} = \bar{\mathcal{T}}\bar{\Lambda}$ admits a unique solution $\bar{\Lambda}^\infty = \bar{\mathcal{T}}\bar{\Lambda}^\infty$. Starting with $\bar{\Lambda}^0(\bar{s}) := 0$, the fixed-point iteration around $\bar{\mathcal{T}}$ generates a sequence of functions

$$\bar{\Lambda}^{n+1}(\bar{s}) = [\bar{\mathcal{T}}\bar{\Lambda}^n](\bar{s}) = [\bar{\mathcal{T}}^n\bar{\Lambda}^0](\bar{s}), \tag{2.190}$$

for all $n = 1, 2, \dots$, such that $\lim_{n \to \infty} \bar{\mathcal{T}}^n \bar{\Lambda}^0 = \bar{\Lambda}^\infty$. Next, we establish error bounds for using $\bar{\Lambda}^n(\bar{\sigma}_t(m_t))$, $n \in \mathbb{N}$, to estimate $V_t(m_t)$ for all t in Problem 3 with observable costs.

Theorem 13. *Consider the function $\bar{\Lambda}^n$ generated using (2.190) for any $n \in \mathbb{N}$. Then, for all $t \in \mathbb{N}$, it holds that*

$$\begin{aligned} \frac{\gamma^{n+t} \cdot c^{\min}}{1 - \gamma} + \gamma^t \cdot \bar{\Lambda}^n(\bar{\sigma}_t(m_t)) + \sup_{a_t \in [[A_t | m_t]]} a_t \; &\leq \; V_t(m_t) \\ &\leq \sup_{a_t \in [[A_t | m_t]]} a_t + \gamma^t \cdot \bar{\Lambda}^n(\bar{\sigma}_t(m_t)) + \frac{\gamma^{n+t} \cdot c^{\max}}{1 - \gamma}. \end{aligned} \tag{2.191}$$

Proof. We show (2.191) by combining arguments in Theorem 12 with (2.186) from Lemma 15. Thus, we first show that for any horizon $T \in \mathbb{N}$, the following holds for each $t = 0, \dots, T$:

$$J_t(m_t; T) = \gamma^t \cdot \bar{\Lambda}^{T-t+1}\big(\bar{\sigma}_t(m_t)\big) + \sup_{a_t \in [[A_t | m_t]]} a_t. \tag{2.192}$$

We can prove (2.192) by induction. At time T, using the definition of the finite-horizon function

$$\begin{aligned}
&J_T(m_T; T) \\
&\quad = \inf_{u_T \in \mathcal{U}} \sup_{a_T, c_T \in [[A_T, C_T | m_T, u_T]]} (a_T + \gamma^T \cdot c_T) \\
&\quad = \inf_{u_T \in \mathcal{U}} \sup_{c_T \in [[C_T | m_T, u_T]]} c_T + \sup_{a_T \in [[A_T | m_T]]} a_T \\
&\quad = \inf_{u_T \in \mathcal{U}} \sup_{c_T, \bar{\sigma}_{T+1}(m_{T+1}) \in [[C_T, S_{T+1} | m_T, u_T]]} (c_T + \gamma^T \cdot \bar{\Lambda}^0(\bar{\sigma}_{T+1}(m_{T+1}))) + \sup_{a_T \in [[A_T | m_T]]} a_T \\
&\quad = \bar{\Lambda}^1(\sigma_T(m_T)) + \sup_{a_T \in [[A_T | m_T]]} a_T,
\end{aligned} \tag{2.193}$$

where, in the second equality, note that A_T is completely determined given M_T as in Lemma 15; and in the third equality, we note that $\bar{\Lambda}^0(\bar{\sigma}_{T+1}(m_{T+1})) = 0$. This forms the basis of our induction. Next, consider as a hypothesis that (2.192) holds at time $t + 1$. Using the definition of the finite-horizon function at time t,

$$\begin{aligned}
&J_t(m_t; T) \\
&\quad = \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} \sup_{c_t, a_t \in [[C_t, A_t | m_{t+1}]]} \big(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t + a_t\big) \\
&\quad = \inf_{u_t \in \mathcal{U}} \sup_{c_t, a_t, m_{t+1} \in [[C_t, A_t, M_{t+1} | m_t, u_t]]} \big(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t + a_t\big) \\
&\quad = \inf_{u_t \in \mathcal{U}} \sup_{c_t, \bar{\sigma}_{t+1}(m_{t+1}) \in [[C_t, \bar{S}_{t+1} | m_t, u_t]]} \big(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t\big) + \sup_{a_t \in [[A_t | m_t]]} a_t \\
&\quad = \inf_{u_t \in \mathcal{U}} \sup_{c_t, \bar{\sigma}_{t+1}(m_{t+1}) \in [[C_t, \bar{S}_{t+1} | \bar{\sigma}_t(m_t), u_t]]} \big(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t\big) + \sup_{a_t \in [[A_t | m_t]]} a_t \\
&\quad = \gamma^t \cdot \bar{\Lambda}^{T-t+1}(\bar{\sigma}_t(m_t)) + \sup_{a_t \in [[A_t | m_t]]} a_t
\end{aligned} \tag{2.194}$$

, where, in the third equality, we use the same arguments as in Lemma 15; in the fourth equality, we use (2.188) from Definition 14; and in the last equality, we use the definition of $\bar{\Lambda}^{T-t+1}$ from (2.190). This proves (2.192) using induction. Then, (2.191) follows directly for all $t \in \mathbb{N}$ and all $n \in \mathbb{N}$ by substituting (2.192) into (2.166) and selecting a horizon $T = t + n - 1$. □

In (2.191), we select $t = 0$ and let $n \to \infty$ to establish that $\bar{\Lambda}^{\infty}(\bar{\sigma}_0(y_0)) = V_0(y_0)$. Thus, when Problem 3 has observable costs, the fixed point $\bar{\Lambda}^{\infty}$ computes the optimal value function V_0 as a direct consequence of Theorem 13. Next, consider that the infimum is achieved in the RHS of $[\bar{\mathcal{T}}\bar{\Lambda}^{n}](\bar{s})$ for all $\bar{s} \in \bar{\mathcal{S}}$ and $n \in \mathbb{N}$. We define a strategy $\boldsymbol{\pi}^* = (\pi^*, \pi^*, \dots)$, where $\bar{\pi}^* : \bar{\mathcal{S}} \to \mathcal{U}$ is the minimizing argument in RHS of (2.189) for $\Lambda = \bar{\Lambda}^{\infty}$. Then, from the same arguments as in Subsection 2.3.4, it holds that the memory-based strategy $\bar{\boldsymbol{g}}^* = (\bar{g}_0^*, \bar{g}_1^*, \dots)$, where $\bar{g}_t^* := \bar{\pi}^*(\sigma_t(m_t))$, gives an optimal solution to Problem 3 with observable costs.

Remark 19. If a partially observed system with observable costs has the state-space model from Subsection 2.3.5, an information state at each t is $\bar{S}_t = [[X_t | M_t]] \in \mathcal{B}(\mathcal{X})$. This is simpler than the accrued cost function in Subsection 2.3.5.

Remark 20. When attempting to learn an information state that satisfies Definition 14 using only output data, we may not be able to satisfy (2.188) exactly. Thus, in Subsection 2.3.7, we relax this definition for approximate information states.

2.3.7 Approximate Information States

In this subsection, we define approximate information states that approximately satisfy (2.188), and construct a time-invariant approximate DP of Problem 3 using them. Then, we bound the resulting error estimating the optimal value and the performance loss of the resulting approximate strategy.

Definition 15. An *approximate information state* for Problem 3 with observable costs at any $t \in \mathbb{N}$ is an uncertain variable $\hat{S}_t = \hat{\sigma}_t(M_t)$ taking values in a bounded, time-invariant set $\hat{\mathcal{S}}$. Furthermore, there exists a parameter $\epsilon \in \mathbb{R}_{\geq 0}$ such that for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$ and $t \in \mathbb{N}$, it satisfies

$$\mathcal{H}\big([[C_t, \hat{S}_{t+1}|m_t, u_t]], [[C_t, \hat{S}_{t+1}|\hat{\sigma}_t(m_t), u_t]]\big) \leq \epsilon, \tag{2.195}$$

where each conditional range in (2.195) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

To compute an approximate value and an approximate control strategy, we proceed with approximate information states just as we did with information states. First, we construct a time-invariant operator $\hat{\mathcal{T}} : [\hat{\mathcal{S}} \to \mathbb{R}] \to [\hat{\mathcal{S}} \to \mathbb{R}]$, such that for any uniformly bounded function $\hat{\Lambda} : \hat{\mathcal{S}} \to \mathbb{R}$,

$$[\hat{\mathcal{T}}\hat{\Lambda}](\hat{s}) := \inf_{u \in \mathcal{U}} \sup_{c, \hat{s}' \in [[C, \hat{S}'|\hat{s}, u]]} \big(c + \gamma \cdot \hat{\Lambda}(\hat{s}')\big). \tag{2.196}$$

Note that $\hat{\mathcal{T}}$ is a contraction mapping and thus, the equation $\hat{\Lambda} = \hat{\mathcal{T}}\hat{\Lambda}$ admits a unique solution $\hat{\Lambda}^\infty = \hat{\mathcal{T}}\hat{\Lambda}^\infty$. Then, starting with $\hat{\Lambda}^0(\hat{s}) := 0$ the fixed-point iteration around $\hat{\mathcal{T}}$ recursively generates the functions

$$\hat{\Lambda}^{n+1}(\hat{s}) = [\hat{\mathcal{T}}\hat{\Lambda}^n](\hat{s}) = [\hat{\mathcal{T}}^n\hat{\Lambda}^0](\hat{s}), \tag{2.197}$$

for all $n = 1, 2, \dots$, such that $\lim_{n \to \infty} \hat{\mathcal{T}}^n \hat{\Lambda}^0 = \hat{\Lambda}^\infty$. This forms our approximate DP decomposition. Next, consider that the infimum is achieved in the RHS of $[\hat{\mathcal{T}}\hat{\Lambda}^n](\hat{s})$ for all $\hat{s} \in \hat{\mathcal{S}}$ and all $n \in \mathbb{N}$. We an approximate strategy $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}^*, \hat{\pi}^*, \dots)$, where $\hat{\pi}^* : \hat{S} \to \mathcal{U}$ is the minimizing argument in the RHS of (2.196) for $\hat{\Lambda} = \hat{\Lambda}^\infty$. Then, a corresponding memory-based strategy is $\hat{\boldsymbol{g}}^* := (\hat{g}_0^*, \hat{g}_1^*, \dots)$ with $\hat{g}_t^* := \pi^*(\sigma_t(m_t))$ for all $t \in \mathbb{N}$. Next, we bound both the approximation error between the optimal value $V_0(y_0)$ and $\hat{\Lambda}^\infty(\hat{\sigma}_0(y_0))$, and the performance loss when implementing $\hat{\boldsymbol{g}}^*$ to generate the control actions.

Theorem 14. *Let the functions $\hat{\Lambda}^n$ be Lipschitz continuous with a constant $L_{\hat{\Lambda}} \in \mathbb{R}_{\geq 0}$ for all $n \in \mathbb{N}$. Then, we have that*

$$\textbf{\textit{a)}} \quad |V_0(y_0) - \hat{\Lambda}^{\infty}(\hat{\sigma}_0(y_0))| \leq \hat{L} \cdot \epsilon \cdot (1-\gamma)^{-1}, \tag{2.198}$$

$$\textbf{\textit{b)}} \quad |V_0(y_0) - V_0^{\hat{\boldsymbol{g}}^*}(y_0)| \leq 2 \cdot \hat{L} \cdot \epsilon \cdot (1-\gamma)^{-1}, \tag{2.199}$$

where $\hat{L} = \max\{\gamma \cdot L_{\hat{\Lambda}}, 1\}$.

Proof. We show (2.198) using (2.166) from Lemma 14. Thus, we first show that for any $T \in \mathbb{N}$, it holds for all $t = 0, \dots, T$:

$$|J_t(m_t;T) - \gamma^t \cdot \hat{\Lambda}^{T-t+1}\big(\hat{\sigma}_t(m_t)\big) - \sup_{a_t \in [[A_t|m_t]]} a_t| \leq \beta_t(T), \tag{2.200}$$

where $\beta_t(T) = \beta_{t+1}(T) + \gamma^t \cdot \hat{L} \cdot \epsilon$ and $\beta_T(T) = \gamma^T \cdot \hat{L} \cdot \epsilon$. We prove (2.200) by induction. At time T, recall that $J_T(m_T;T) = \inf_{u_T \in \mathcal{U}} \sup_{c_T \in [[C_T|m_T,u_T]]} \gamma^T \cdot c_T + \sup_{a_T \in [[A_T|m_T]]} a_T$ using the arguments in Lemma 15. This implies that

$$\begin{aligned}
&\Big|J_T(m_T;T) - \gamma^T \cdot \hat{\Lambda}^1\big(\hat{\sigma}_T(m_T)\big) - \sup_{a_T \in [[A_T|m_T]]} a_T\Big| \\
&\qquad = \Big| \inf_{u_T \in \mathcal{U}} \sup_{c_T \in [[C_T|m_T,u_T]]} \gamma^T \cdot c_T - \gamma^T \cdot \hat{\Lambda}^1\big(\hat{\sigma}_T(m_T)\big)\Big| \\
&\qquad = \gamma^T \cdot \Big| \inf_{u_T \in \mathcal{U}} \sup_{c_T, \hat{s}_{T+1} \in [[C_T, \hat{S}_{T+1}|m_T,u_T]]} (c_T + \gamma \cdot \hat{\Lambda}^0(\hat{s}_{T+1})) \\
&\qquad\qquad - \inf_{u_T \in \mathcal{U}} \sup_{c_T, \hat{s}_{T+1} \in [[C_T, \hat{S}_{T+1}|\hat{\sigma}_T(m_T),u_T]]} (c_T + \gamma \cdot \hat{\Lambda}^0(\hat{s}_{T+1}))\Big| \\
&\qquad \leq \gamma^T \cdot \hat{L} \cdot \sup_{u_T \in \mathcal{U}} \mathcal{H}\big([[C_T, \hat{S}_{T+1}|m_T,u_T]], [[C_T, \hat{S}_{T+1}|\hat{\sigma}_T(m_T),u_T]]\big) \leq \gamma^T \cdot \hat{L} \cdot \epsilon,
\end{aligned} \tag{2.201}$$

where, in the second equality, we note that $\Lambda^0(\hat{s}_{T+1}) = 0$ identically; in the first inequality, we note that $\hat{L} = \max\{\gamma \cdot L_{\hat{\Lambda}}, 1\}$ is the Lipschitz constant of $(c_T + \gamma \cdot \hat{\Lambda}^0(\hat{s}_{T+1}))$ with respect to $(c_T, \hat{s}_{T+1})$ and use (2.157); and in the second inequality, we use (2.195). This forms the basis of our induction. Next, we consider the hypothesis that (2.198) holds at time $t+1$. Using the hypothesis and rearranging terms, $J_{t+1}(m_{t+1};T) \leq \beta_{t+1}(T) + \gamma^{t+1} \cdot \hat{\Lambda}^{T-t}\ (\hat{\sigma}_{t+1}(m_{t+1})) + \sup_{a_{t+1} \in [[A_{t+1}|m_{t+1}]]} a_{t+1}$. Then, at time t,

$$\begin{aligned}
&\Big|J_t(m_t;T) - \gamma^t\cdot\hat{\Lambda}^{T-t+1}\big(\hat{\sigma}_t(m_t)\big) - \sup_{a_t\in[[A_t|m_t]]} a_t\Big| \\
&\quad\leq \beta_{t+1}(T) + \Big|\inf_{u_t\in\mathcal{U}} \sup_{m_{t+1}\in[[M_{t+1}|m_t,u_t]]} \Big(\gamma^{t+1}\cdot\hat{\Lambda}^{T-t}(\hat{\sigma}_{t+1}(m_{t+1})) \\
&\qquad\qquad + \sup_{a_t,c_t\in[[A_t,C_t|m_{t+1}]]} (a_t + \gamma^t\cdot c_t)\Big) - \gamma^t\cdot\hat{\Lambda}^{T-t+1}(\hat{\sigma}_t(m_t)) - \sup_{a_t\in[[A_t|m_t]]} a_t\Big| \\
&\quad\leq \beta_{t+1}(T) + \gamma^t\cdot\sup_{u_t\in\mathcal{U}}\Big|\sup_{c_t,\hat{\sigma}_{t+1}(m_t)\in[[C_t,\hat{S}_{t+1}|m_t,u_t]]} \big(c_t + \gamma\cdot\hat{\Lambda}^{T-t}(\hat{\sigma}_{t+1}(m_{t+1}))\big) \\
&\qquad\qquad - \sup_{c_t,\hat{s}_{t+1}\in[[C_t,\hat{S}_{t+1}|\hat{\sigma}_t(m_t),u_t]]} (c_t + \gamma\cdot\hat{\Lambda}^{T-t}(\hat{s}_{t+1}))\Big| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \leq \beta_{t+1}(T) + \gamma^t\cdot\hat{L}\cdot\epsilon,
\end{aligned} \tag{2.202}$$

where, in the second inequality, we use arguments in Lemma 15 for $[[A_t, C_t|m_{t+1}]] = [[A_t|m_{t+1}]] \times [[C_t|m_{t+1}]]$; and, in the third inequality we use (2.157) and (2.195). This proves (2.200) for all t using induction.

Next, for the iterated function $\hat{\Lambda}^n$, we select a horizon $T = n-1$ and set $t = 0$ in (2.200), to write that $|J_0(y_0;T) - \hat{\Lambda}^n(\hat{\sigma}_0(y_0))| \leq \beta_0(T)$, where $\beta_0(T) = \sum_{\ell=0}^{n-1}\gamma^\ell\cdot\hat{L}\cdot\epsilon$. As $n\to\infty$ with $T = n-1$, note that $\lim_{T\to\infty} J_0(y_0;T) = V_0(y_0)$, $\lim_{n\to\infty}\hat{\Lambda}^n(\hat{\sigma}_0(y_0)) = \hat{\Lambda}^\infty(\hat{\sigma}_0(y_0))$, and $\lim_{T\to\infty}\beta_0(T) = \frac{\hat{L}\cdot\epsilon}{1-\gamma}$. The proof for (2.199) follows from a similar series of arguments. □

2.3.8 Alternate Characterization

When exploring whether an uncertain variable is a valid candidate to be considered for an approximate information state, it may be difficult to verify (2.195). Thus, we present two *stronger* conditions that are easier to verify. To establish that $\hat{S}_t = \hat{\sigma}_t(M_t)$, $t \in \mathbb{N}$, satisfies (2.195), the following two conditions should hold (see proof in Appendix C:

1) State-like evolution: There exists a Lipschitz continuous function $\psi : \hat{\mathcal{S}} \times \mathcal{U} \times \mathcal{Y} \to \mathcal{S}$ independent of the strategy $\boldsymbol{g}$. , such that

$$\hat{\sigma}_{t+1}(M_{t+1}) = \psi(\hat{\sigma}_t(M_t), U_t, Y_{t+1}). \tag{2.203}$$

2) Sufficient to approximate outputs: For all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$, there exists a constant $\delta \in \mathbb{R}_{\geq 0}$ such that

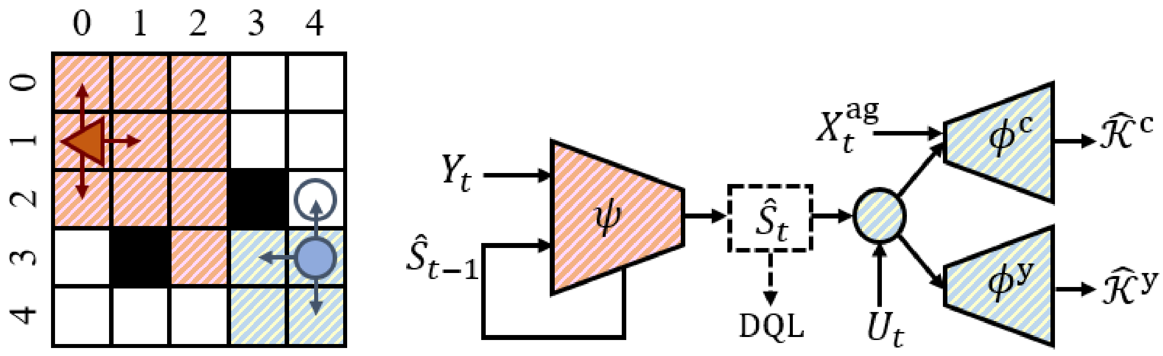
$$\mathcal{H}([[C_t, Y_{t+1} | m_t, u_t]]], [[C_t, Y_{t+1} | \hat{\sigma}_t(m_t), u_t]]) \leq \delta, \tag{2.204}$$

where each conditional range in (2.204) can be evaluated independent of the choice of strategy $\boldsymbol{g}$.

2.3.9 Numerical Example

We consider an agent pursuing a target across a 5 × 5 grid with obstacles. At each $t \in \mathbb{N}$, the agent's position X_t^{ag} and the target's position X_t^{ta} each take values in the set of grid cells $\mathcal{X} = \big\{(0,0), (0,1), \ldots, (4,4)\big\} \setminus \mathcal{O}$, where $\mathcal{O} \subset \mathcal{X}$ is the set of obstacles. Let $\mathcal{W} = \{(-1,0), (1,0), (0,0), (0,1), (0,-1)\}$, $\mathcal{N} = \{(0,-1), (0,0), (0,1)\}$, and $\mathcal{U} = \mathcal{W} \times \{\xi\}$, where ξ denotes a "stop" action. Starting at $X_0^{\text{ta}} \in \mathcal{X}$, the target's position evolves as $X_{t+1}^{\text{ta}} = \delta(X_t^{\text{ta}} + W_t \in \mathcal{X}) \cdot (X_t^{\text{ta}} + W_t) + (1 - \delta(X_t^{\text{ta}} + W_t \in \mathcal{X})) \cdot X_t^{\text{ta}}$, where $W_t \in \mathcal{W}$ and δ is returns 1 or 0 after checking the argument. At each t, the agent observes their own position perfectly and the target's position as $Y_t = \delta(X_t^{\text{ta}} + N_t \in \mathcal{X}) \cdot (X_t^{\text{ta}} + N_t) + (1 - \delta(X_t^{\text{ta}} + N_t \in \mathcal{X})) \cdot X_t^{\text{ta}}$, where $N_t \in \mathcal{N}$. Then, the agent selects an action $U_t \in \mathcal{U}$ which is either to move or to stop. If the agent moves, i.e., $U_t \neq \xi$, then $X_{t+1}^{\text{ag}} = \delta(X_t^{\text{ag}} + U_t \in \mathcal{X}) \cdot (X_t^{\text{ag}} + U_t) + (1 - \delta(X_t^{\text{ag}} + U_t \in \mathcal{X})) \cdot X_t^{\text{ag}}$. The agent incurs a cost $C_t = 2$. If the agent stops, i.e., $U_t = \xi$, they incur a terminal cost $10 \cdot \eta(X_T^{\text{ta}}, X_T^{\text{ag}})$ for the L1 distance from the target. We illustrate this pursuit problem in Fig. 2.8(a), where the black cells are obstacles, the red triangle is the agent, the blue circle is the observation, and the blue disk is the target.

We consider the pursuit problem when the agent is aware of their own dynamics, but unaware of the observation model and target's dynamics. Thus, we train an approximate information state (AIS) model to learn a representation of the target's dynamics using observations, actions, and incurred costs to enforce (2.203) and (2.204). The AIS is generated by a neural networks in an encoder-decoder architecture, as shown in Fig. 2.8(b). At each $t \in \mathbb{N}$, the *encoder* ψ receives as an input the observation Y_t



(a) The grid

(b) Encoder-decoder architecture

Figure 2.8: The pursuit problem with $x_0^{\text{ag}} = (0, 1)$, $x_0^{\text{ta}} = (4, 3)$ and $y_0 = (4, 2)$ is in (a). The neural network architecture for the AIS is in (b).

and previous AIS $\hat{S}_{t-1}$ and generates $\hat{S}_t$. It consists of a linear layer of size $(2, 4)$ with ReLU activation, followed by a gated recurrent unit (GRU) with a hidden state size of 4. The hidden state of the GRU constitutes the AIS $\hat{S}_t$ updated recurrently as $\hat{S}_t = \psi(\hat{S}_{t-1}, Y_t)$, thus enforcing (2.203). Note that our AIS is independent from the agent's position and action because the target moves independent from the agent. The decoder is comprised of two separate units, each of which is selected according to the action U_t. If $U_t = \xi$, we use the network ϕ^{c} which takes as an input the agent's position X_t^{ag} and the AIS $\hat{S}_t$ and generates a set of possible terminal costs $\hat{\mathcal{K}}^{\text{c}} := [[C_t | X_t^{\text{ag}}, \hat{S}_t]]$. This network comprises of two linear layers with dimensions $(6, 16)$ and $(16, 9)$, where the first layer has ReLU activation and the second has sigmoid activation. If $U_t \neq \xi$, we use the network ϕ^{y} which takes the AIS $\hat{S}_t$ as an input and generates the conditional range $\hat{\mathcal{K}}^{\text{y}} := [[Y_{t+1} | \hat{S}_t]]$. This network comprises of two linear layers with dimensions $(6, 16)$ and $(16, 23)$, where the first layer has ReLU activation and the second has sigmoid activation.

We train the entire model simultaneously using the outputs of the decoder. At each $t \in \mathbb{N}$, the training loss is given by the Hausdorff distance between the one-hot encoded incoming data point, either C_t or Y_{t+1}, and the current predicted set. Since the Hausdorff distance is not differentiable, we adapt the distance-transform-based surrogate loss proposed in [198]. Note that we cannot observe the true underlying set and thus train the predictions against sampled data points to eventually learn the feasible sets. We train the network for 3×10^6 instances with a learning rate of 0.0003. In each instance, we randomly initialize the agent and target's positions from the pink and blue hatched cells in Fig. 2.8(a) and randomize all subsequent noises, disturbances and actions.

Next, we utilize the trained encoder's output AIS and the agent's position as a state input to a deep Q-learning network (DQN) with two layers of $(6, 3)$ and $(3, 6)$ and a LeakyReLU activation each. We train this AIS-DQN using an exploratory policy for 3×10^6 instances with a learning rate of 0.0005 using a maximally risk-averse approach from [28] with high risk-aversion 0.9, to learn to minimize the worst-case discounted cost with $\gamma = 0.97$. We compare the worst-case performance of the greedy strategy of the trained AIS-DQN with the worst-case performance of a trained stochastic-DQN, which uses the observation and position as the state and has the same hyperparameters with no risk-aversion. In Fig. 2.9, we present the improvement in worst-case cost achieved by AIS-DQN over stochastic-DQN in 10^4 simulations each for different initial positions. Note that AIS-DQN outperforms stochastic-DQN for most cases.

Appendix A - Proof that the Dynamic Programming Operator is a Contraction Mapping

In this appendix, we prove that the operator $\mathcal{T}$ defined in (2.175) is a contraction mapping.

Lemma 16. *Consider the operator $\mathcal{T}$ defined in (2.175). There exists a constant $\alpha \in [0, 1)$ such that for two functions $\Lambda : \mathcal{S} \times (0, 1] \to \mathbb{R}$ and $\tilde{\Lambda} : \mathcal{S} \times (0, 1] \to \mathbb{R}$:*

$$||\mathcal{T}\Lambda - \mathcal{T}\tilde{\Lambda}||_\infty \leq \alpha \cdot ||\Lambda - \tilde{\Lambda}||_\infty. \tag{2.205}$$

Proof. Using the definition of $\mathcal{T}$, we expand the left hand side (LHS) of (2.205) as

$$||\mathcal{T}\Lambda - \mathcal{T}\tilde{\Lambda}||_\infty$$

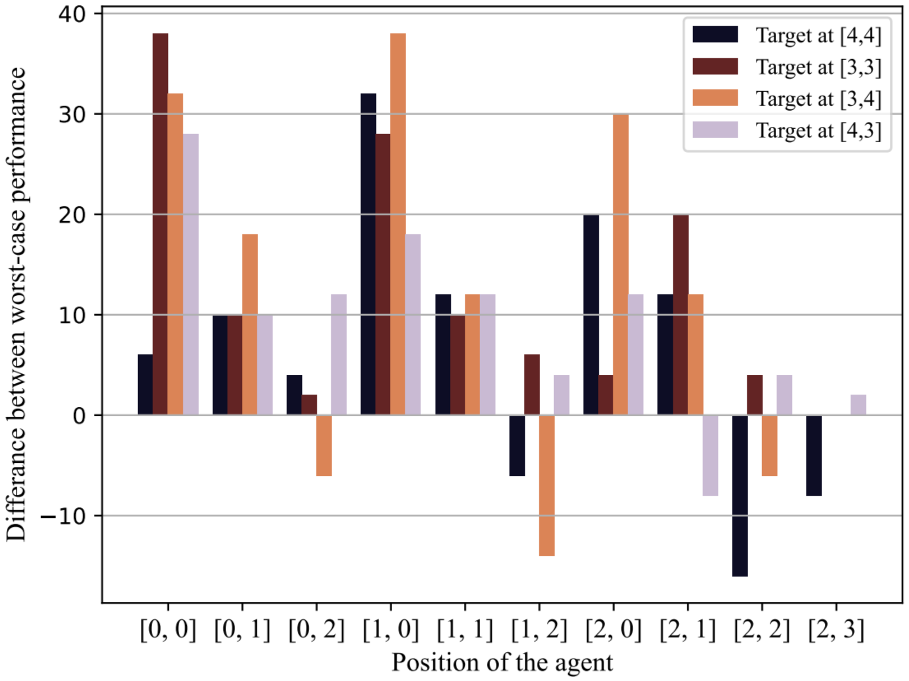


Figure 2.9: The improvement in worst-case performance using AIS-DQL over stochastic-DQL.

$$
\begin{aligned}
&= \sup_{s\in\mathcal{S}, z\in(0,1]} \Big| \inf_{u\in\mathcal{U}} \sup_{s'\in\mathcal{S}, c\in\mathcal{C}} \Big(c + \gamma\cdot\Lambda(s', \gamma\cdot z) + \frac{\rho(c, s' \mid s, u)}{z} \Big) \\
&\qquad\qquad - \inf_{u\in\mathcal{U}} \sup_{s'\in\mathcal{S}, c\in\mathcal{C}} \Big(c + \gamma\cdot\tilde{\Lambda}(s', \gamma\cdot z) + \frac{\rho(c, s' \mid s, u)}{z} \Big) \Big| \\
&\leq \gamma\cdot \sup_{s'\in\mathcal{S}, \gamma\cdot z\in(0,\gamma]} |\Lambda(s', \gamma\cdot z) - \tilde{\Lambda}(s', \gamma\cdot z)| \\
&\leq \gamma\cdot \sup_{s\in\mathcal{S}, z\in(0,1]} |\Lambda(s, z) - \tilde{\Lambda}(s, z)| \\
&\qquad\qquad\qquad\qquad\qquad\qquad = \gamma\cdot||\Lambda - \tilde{\Lambda}||_\infty, \quad (2.206)
\end{aligned}
$$

where, in the first inequality, we upper bound the difference between supremum values of two functions by the supremum difference between the two functions; and in the second inequality, we use $(0, \gamma] \subset (0, 1]$ in the argument of the supremum. This proves that the operator $\mathcal{T}$ is a contraction mapping by setting $\alpha = \gamma$. □

Appendix B - Proof that Information States Yield an Optimal Control Strategy

In this appendix, we prove that the information-state based control strategy $\boldsymbol{\pi}^* = (\pi^*, \pi^*, \dots)$ and corresponding memory-based control strategy $\boldsymbol{g}^* = (g_0^*, g_1^*, \dots)$ defined in Subsection 2.3.4 are optimal solutions to Problem 3. Recall from Subsection 2.3.4 that the information-based control law is $\pi^*(s, z) = \arg\min_{u \in \mathcal{U}} \sup_{c \in \mathcal{C}, s' \in \mathcal{S}} (c + \gamma \cdot \Lambda^\infty(s', \gamma \cdot z) + \frac{\rho(c, s'|s, u)}{z})$ for all $s \in \mathcal{S}$ and $z \in (0, 1]$. Furthermore, recall that the memory-based control law is $g_t^*(m_t) = \pi^*(\sigma_t(m_t), \gamma^t)$ for all $m_t \in \mathcal{M}_t$ and $t \in \mathbb{N}$. To begin, for any time-invariant control law $\pi : \mathcal{S} \times (0, 1] \to \mathcal{U}$, we define a law-dependent operator $\mathcal{T}(\pi) : [\mathcal{S} \times (0, 1] \to \mathbb{R}] \to [\mathcal{S} \times (0, 1] \to \mathbb{R}]$, such that for any uniformly bounded function $\Lambda : \mathcal{S} \times (0, 1] \to \mathbb{R}$, we have:

$$[\mathcal{T}(\pi)\Lambda](s, z) := \sup_{c \in \mathcal{C}, s' \in \mathcal{S}} \left(c + \gamma \cdot \Lambda(s', \gamma \cdot z) + \frac{\rho(c, s'|s, \pi(s, z))}{z} \right). \tag{2.207}$$

Note that the control action in the RHS of (2.207) is selected using as $u = \pi(s, z)$. Furthermore, by definition of π^* it holds that for all $s \in \mathcal{S}$ and $z \in (0, 1]$:

$$[\mathcal{T}(\pi^*)\Lambda^\infty](s, z) = [\mathcal{T}\Lambda^\infty](s, z) = \Lambda^\infty(s, z). \tag{2.208}$$

Then, we can construct a corresponding memory-based control law at each $t \in \mathbb{N}$ as $g_t(m_t) := \pi(\sigma_t(m_t), \gamma^t)$ and a memory-based control strategy $\boldsymbol{g} = (g_0, g_1, \dots)$. Next, we establish that we can use $[\mathcal{T}(\pi)^n \Lambda^0](\sigma_t(m_t), \gamma^t)$ for any $n \in \mathbb{N}$ to estimate the strategy-dependent value $V_t^{\boldsymbol{g}}(m_t)$ at each $t \in \mathbb{N}$.

Lemma 17. *For all $t \in \mathbb{N}$ and all $n \in \mathbb{N}$, it holds that:*

$$\begin{aligned} \frac{\gamma^{n+t} \cdot c^{\min}}{1 - \gamma} + \gamma^t \cdot [\mathcal{T}(\pi)^n \Lambda^0](\sigma_t(m_t), \gamma^t) + \sup_{a_t \in [[A_t | m_t]]} a_t & \\ \leq V_t^{\boldsymbol{g}}(m_t) & \\ \leq \sup_{a_t \in [[A_t | m_t]]} a_t + \gamma^t \cdot [\mathcal{T}(\pi)^n \Lambda^0](\sigma_t(m_t), \gamma^t) + \frac{\gamma^{n+t} \cdot c^{\max}}{1 - \gamma}&. \end{aligned} \tag{2.209}$$

Proof. The proof follows using the same sequence of arguments as the proof for Theorem 12, but by using (2.165) from Lemma 14 along with $u_t = g_t(m_t) := \pi(\sigma_t(m_t), \gamma^t)$. □

We can set $t = 0$ in (2.209) and note that

$$\frac{\gamma^n \cdot c^{\min}}{1-\gamma} + [\mathcal{T}(\pi)^n \Lambda^0](\sigma_0(y_0), 1) \ \leq \ V_t^{\boldsymbol{g}}(m_t) \leq [\mathcal{T}(\pi)^n \Lambda^0](\sigma_t(m_t), 1) + \frac{\gamma^n \cdot c^{\max}}{1-\gamma}, \quad (2.210)$$

where recall that $A_0 = 0$ always. Thus, as a direct consequence of Lemma 17, it holds that $[\lim_{n\to\infty} \mathcal{T}(\pi)^n]\Lambda^0(\sigma_0(y_0), 1) = V_0^{\boldsymbol{g}}(y_0)$. Next, we prove that $\lim_{n\to\infty} \mathcal{T}(\pi^*)^n = \Lambda^\infty$.

Lemma 18. *For all $s \in \mathcal{S}$ and $z \in (0,1]$, it holds that*

$$[\lim_{n\to\infty} \mathcal{T}(\pi^*)^n](s,z) = \Lambda^\infty(s,z). \quad (2.211)$$

Proof. We begin by showing that the LHS of (2.211) forms an upper bound on the RHS. By definition, note that $[\mathcal{T}(\pi)\Lambda](s,z) \geq [\mathcal{T}\Lambda](s,z)$ for any control law π and for all $s \in \mathcal{S}$, $z \in (0,1]$. Taking the limit on both sides with $\pi = \pi^*$, this implies for all $s \in \mathcal{S}$, $z \in (0,1]$ that

$$[\lim_{n\to\infty} \mathcal{T}(\pi^*)^n \Lambda^0](s,z) \geq [\lim_{n\to\infty} \mathcal{T}^n \Lambda^0](s,z) = \Lambda^\infty(s,z). \quad (2.212)$$

Next, we prove that the LHS of (2.211) also forms a lower bound on the RHS. From (2.208), it holds that $\Lambda^\infty = \mathcal{T}(\pi^*)\Lambda^\infty = \lim_{n\to\infty} \mathcal{T}(\pi^*)^n \Lambda^\infty$. Then, using $\Lambda^\infty \geq \Lambda^0 = 0$, we write for all $s \in \mathcal{S}$ and $z \in (0,1]$ that

$$\Lambda^\infty(s,z) = \lim_{n\to\infty} \mathcal{T}(\pi^*)^n \Lambda^\infty \geq [\lim_{n\to\infty} \mathcal{T}(\pi^*)^n \Lambda^0](s,z). \quad (2.213)$$

Using (2.212) and (2.213) simultaneously establishes (2.211). □

Then, as a direct consequence of both Lemmas 17 and 18, we can conclude that

$$V_t^{\boldsymbol{g}^*}(y_0) = [\lim_{n\to\infty} \mathcal{T}(\pi)^n]\Lambda^0(\sigma_0(y_0), 1) = \Lambda^\infty(\sigma_0(y_0), 1) = V_0(y_0), \quad (2.214)$$

where in the last equality, we use (2.184). This proves that the control strategies $\boldsymbol{g}^*$ and $\boldsymbol{\pi}^*$ are optimal solutions to Problem 3.

Appendix C - Proof that the Alternate Characterization Defines an Approximate Information State

In this appendix, we prove that the alternate characterization presented in Subsection 2.3.8 using the properties (2.203) and (2.204) is sufficient to establish (2.195) in Definition 15.

Lemma 19. *For all $t \in \mathbb{N}$, if an uncertain variable $\hat{S}_t = \hat{\sigma}_t(M_t)$ satisfies (2.203) - (2.204), it also satisfies (2.195).*

Proof. Let $m_t \in \mathcal{M}_t$ be a given realization of M_t and let $\hat{s}_t = \hat{\sigma}_t(s_t)$ satisfy (2.203) - (2.204), for all t. Let $\mathcal{K}_t^{\text{ob}} := [[C_t, Y_{t+1} | m_t, u_t]]$ and $\hat{\mathcal{K}}_t^{\text{ob}} := [[C_t, Y_{t+1} | \hat{\sigma}_t(m_t), u_t]]$. Then, using (2.203), we can write the LHS in (2.195) as

$$\begin{aligned}
&\mathcal{H}\big([[C_t, \psi(\hat{\sigma}_t(m_t), u_t, Y_{t+1}) | m_t, u_t]], [[C_t, \psi(\hat{\sigma}_t(m_t), u_t, Y_{t+1}) | \hat{\sigma}_t(m_t), u_t]]\big) \\
&= \max\Big\{ \sup_{(c_t, y_{t+1}) \in \mathcal{K}_t^{\text{ob}}} \inf_{(\hat{c}_t, \hat{y}_{t+1}) \in \hat{\mathcal{K}}_t^{\text{ob}}} \big(\eta(c_t, \hat{c}_t) + \eta\big(\psi(\hat{\sigma}_t(m_t), u_t, y_{t+1}), \psi(\hat{\sigma}_t(m_t), u_t, \hat{y}_{t+1})\big)\big), \\
&\sup_{(\hat{c}_t, \hat{y}_{t+1}) \in \hat{\mathcal{K}}_t^{\text{ob}}} \inf_{(c_t, y_{t+1}) \in \mathcal{K}_t^{\text{ob}}} \big(\eta(c_t, \hat{c}_t) + \eta\big(\psi(\hat{\sigma}_t(m_t), u_t, y_{t+1}), \psi(\hat{\sigma}_t(m_t), u_t, \hat{y}_{t+1})\big)\big)\Big\}, \qquad (2.215)
\end{aligned}$$

where, in the second equality, we use the definition of the Hausdorff distance from (2.156). Note that ψ is globally Lipschitz from the alternate characterization of the approximate information state. This implies that

$$\eta\big(\psi(\hat{\sigma}_t(m_t), u_t, y_{t+1}), \psi(\hat{\sigma}_t(m_t), u_t, \hat{y}_{t+1})\big) \leq L_\psi \cdot \eta(y_{t+1}, \hat{y}_{t+1}), \qquad (2.216)$$

and thus

$$\begin{aligned}
&\mathcal{H}\big([[C_t, \psi(\hat{\sigma}_t(m_t), u_t, Y_{t+1}) | m_t, u_t]], [[C_t, \psi(\hat{\sigma}_t(m_t), u_t, Y_{t+1}) | \hat{\sigma}_t(m_t), u_t]]\big) \\
&\leq L_\psi \cdot \max\Big\{ \sup_{(c_t, y_{t+1}) \in \mathcal{K}_t^{\text{ob}}} \inf_{(\hat{c}_t, \hat{y}_{t+1}) \in \hat{\mathcal{K}}_t^{\text{ob}}} \big(\eta(c_t, \hat{c}_t) + \eta(y_{t+1}, \hat{y}_{t+1})\big), \\
&\sup_{(\hat{c}_t, \hat{y}_{t+1}) \in \hat{\mathcal{K}}_t^{\text{ob}}} \inf_{(c_t, y_{t+1}) \in \mathcal{K}_t^{\text{ob}}} \big(\eta(c_t, \hat{c}_t) + \eta(y_{t+1}, \hat{y}_{t+1})\big)\Big\} \\
&= L_\psi \cdot \mathcal{H}(\mathcal{K}_t^{\text{ob}}, \hat{\mathcal{K}}_t^{\text{ob}}) \leq L_\psi \cdot \delta. \qquad (2.217)
\end{aligned}$$

Thus, (2.195) is satisfied by selecting $\epsilon = L_\psi \cdot \delta$. □

Chapter 3

DECENTRALIZED CONTROL FOR TEAMS WITH INFORMATION ASYMMETRY

3.1 Decentralized Control of Two Agents with Nested Accessible Information

3.1.1 Notation and Preliminaries

In our exposition, random variables are denoted with upper case letters and their realization by the corresponding lower case letters. For integers $a < b$, $X_{a:b}$ is shorthand for the vector $(X_a, X_{a+1}, \cdots, X_b)$ and $X^{a:b}$ is shorthand for the vector $(X^a, X^{a+1}, \cdots, X^b)$. When $a > b$, the dimension of $X^{a:b}$ is 0. The combined notation with $c < d$ and $a < b$, $X_{a:b}^{c:d}$ is short for the vector $(X_i^j : i = a, a+1, \cdots, b, j = c, c+1, \cdots, d)$.

3.1.2 Problem Formulation

We consider a team of two agents who take actions over $T \in \mathbb{N}$ discrete time steps. For each $t = 0, \ldots, T$, the state of the team is denoted by the random variable X_t which takes values in a finite set $\mathcal{X}_t$. The action of an agent $k = 1, 2$ at time t is U_t^k, which takes values in a finite set $\mathcal{U}_t^k$. We denote the tuple (U_t^1, U_t^2) by $U_t^{1:2}$. Starting at the initial state X_0, the system evolves as

$$X_{t+1} = f_t\left(X_t, U_t^{1:2}, W_t\right), \quad t = 0, \ldots, T-1, \tag{3.1}$$

where W_t is an uncontrolled disturbance which takes values in a finite set $\mathcal{W}_t$. At each $t = 0, \ldots, T$, each agent $k = 1, 2$ makes an observation

$$Y_t^k := h_t^k(X_t, V_t^k), \tag{3.2}$$

which takes values in a finite set $\mathcal{Y}_t^k$. Here, V_t^k is a measurement noise which takes values in a finite set $\mathcal{V}_t^k$. The external disturbances $\{W_t : t = 0, \dots, T\}$, measurement noises $\{V_t^1, V_t^2 : t = 0, \dots, T\}$, and initial state X_0 are collectively called the *primitive random variables* of the team and their probability distributions are known a priori. We assume that each primitive random variable is independent of all other primitive random variables to ensure that the system's evolution is Markovian [12].

Definition 16. For all $t = 0, \dots, T$, the *memory* of an agent $k = 1, 2$ is a set of random variables $M_t^k \subseteq \{Y_{0:t}^{1:2}, U_{0:t-1}^{1:2}\}$, which takes values in a finite collection of sets $\mathcal{M}_t^k$ and satisfies *perfect recall*, i.e, $M_{t-1}^k \subseteq M_t^k$, with $M_{-1}^k := \emptyset$.

We partition the memory M_t^2 of agent 2 into two components, the *common information* A_t^2 and *private information* L_t^2, which are described next:

1) The common information is a subset of the memory of agent 2 which is also available to agent 1. For all $t = 0, \dots, T$, we define the common information as a set of random variables $A_t^2 \subseteq M_t^2$ which takes values in a finite collection of sets $\mathcal{A}_t^2$ and satisfies the properties: (1) accessibility to agent 1, i.e., $A_t^2 \subseteq M_t^1$, and (2) perfect recall, i.e., $A_{t-1}^2 \subseteq A_t^2$, with $A_{-1}^2 := \emptyset$.

2) The private information of agent 2 is a subset of their memory which is unavailable to agent 1. For all $t = 0, \dots, T$, we define the private information as the set of random variables $L_t^2 := M_t^2 \setminus A_t^2$ which takes values in a finite collection of sets $\mathcal{L}_t^2$. We impose the condition $L_t^2 \cap M_t^1 = \emptyset$ to specify that agent 1 can not access the private information of agent 2 at each t.

The second property of the common information of agent 2 motivates us to define the *new information* added to A_t^2, for all $t = 0, \dots, T$, as the set of random variables $Z_t^2 := A_t^2 \setminus A_{t-1}^2$ which takes values in a finite collection of sets $\mathcal{Z}_t^2$. Note that $Z_0^2 := A_0^2$. Analogously, for all $t = 0, \dots, T$, we define the new information added to the memory of agent 1 as the set of random variables $Z_t^1 := M_t^1 \setminus M_{t-1}^1$ which takes values in a finite collection of sets $\mathcal{Z}_t^1$, where $Z_0^1 := M_0^1$. In our information structure, we enforce that for all t, the new information of agent 2 must satisfy $Z_t^2 \subseteq L_t^2 \cup \{Y_t^{1:2}, U_{t-1}^{1:2}\}$. This ensures that $Z_t^2 \not\subseteq M_{t-1}^1$ and $Z_t^2 \subseteq Z_t^1$, i.e., Z_t^2 is not accessible to agent 1 prior to time t and becomes accessible to agent 1 at time t.

Remark 21. The common information A_t^2 of agent 2 has a restriction in teams with nested accessible information imposed by the property $Z_t^2 \not\subseteq M_{t-1}^1$. The presence of this restriction allows us to specialize our results to systems where the information available to agent 1 but unavailable to agent 2, i.e., $M_t^1 \setminus A_t^2$, may grow in size with time.

Remark 22. As an example of an information structure which satisfies $Z_t^2 \not\subseteq M_{t-1}^1$, consider one-directional communication from 2 to 1 with a delay of $d \in \mathbb{N}$ time steps. In such a system, $M_t^1 = \{Y_{0:t}^1, U_{0:t-1}^1, Y_{0:t-d}^2, U_{0:t-d}^2\}$ and $M_t^2 = \{Y_{0:t}^2, U_{0:t-1}^2\}$. Then, $A_t^2 = \{Y_{0:t-d}^2, U_{0:t-d}^2\}$, $L_t^2 = \{Y_{t-d+1:t}^2, U_{t-d+1:t-1}^2\}$, and the set $M_t^1 \setminus A_t^2 = \{Y_{0:t}^1, U_{0:t-1}^1\}$ grows in size with time.

For all $t = 0, \dots, T$, each agent $k = 1, 2$ uses a control law $g_t^k : \mathcal{M}_t^k \to \mathcal{U}_t^k$ to select their action

$$U_t^k = g_t^k(M_t^k), \tag{3.3}$$

where $M_t^2 = \{L_t^2, A_t^2\}$. We define the control strategy of agent k as $\boldsymbol{g}^k := (g_t^k : t = 0, \dots, T)$ and the control strategy of the team as $\boldsymbol{g} := (\boldsymbol{g}^1, \boldsymbol{g}^2)$. The set of all feasible control strategies is $\mathcal{G}$. After each agent $k = 1, 2$ selects their action U_t^k at time t, the team incurs a cost $c_t(X_t, U_t^{1:2}) \in \mathbb{R}_{\geq 0}$. The performance criterion over the finite horizon T is

$$\mathcal{J}(\boldsymbol{g}) = \mathbb{E}^{\boldsymbol{g}}\left[\sum_{t=0}^{T} c_t\big(X_t, U_t^{1:2}\big)\right], \tag{3.4}$$

where the expectation is with respect to the joint probability distribution on all random variables. Next, we state the optimization problem for the team.

Problem 4. The optimization problem for the team is $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the distributions of the primitive random variables $\{X_0, W_{0:t}, V_{0:t}^{1:2}\}$, and the dynamics $\{c_t, f_t, h_t^{1:2} : t = 0, \dots, T\}$.

Problem 4 is guaranteed to have a solution because all variables take values in finite sets. Our goal is to derive a structural form for an optimal strategy $\boldsymbol{g}^* \in \mathcal{G}$ in Problem 4 which can be computed using a DP decomposition.

3.1.3 Analysis Using Prescriptions

3.1.3.1 Analysis for Agent 1

In this subsection, we derive a structural form for an optimal control strategy of agent 1. We first note that given a strategy $\boldsymbol{g}^2$, agent 1 cannot generate the action U_t^2 for each t because they cannot access the complete memory $M_t^2 = \{L_t^2, A_t^2\}$. However, they can access the component A_t^2. This motivates us to consider a two stage process for the generation of the action of agent 2: (1) agent 1 generates a prescription for agent 2 using only A_t^2, and (2) agent 2 computes U_t^2 using this prescription and their private information L_t^2.

Definition 17. For all $t = 0, \dots, T$, a *prescription* for agent 2 is a mapping $\Gamma_t^2 : \mathcal{L}_t^2 \to \mathcal{U}_t^2$ which takes values in a finite set $\mathcal{F}_t^2$.

At each t, the prescription for agent 2 is generated using a prescription law $\psi_t^2 : \mathcal{A}_t^2 \to \mathcal{F}_t^2$, which yields $\Gamma_t^2 = \psi_t^2(A_t^2)$. We call $\boldsymbol{\psi}^2 := (\psi_t^2 : t = 0, \dots, T)$ the prescription strategy for agent 2. Given a prescription Γ_t^2, the action of agent 2 is computed as $U_t^2 = \Gamma_t^2\big(L_t^2\big)$. Next, we use the person-by-person approach to set up a "new" centralized problem for agent 1. We proceed by arbitrarily fixing the prescription strategy $\boldsymbol{\psi}^2$ for agent 2. Since the prescription Γ_t^2 is generated using only the common information $A_t^2 \subseteq M_t^1$, agent 1 can derive the prescription using the fixed strategy as $\Gamma_t^2 = \psi_t^2(A_t^2)$. Then, we define a new state for agent 1 as $S_t^1 := \{X_t, L_t^2, A_t^2\}$ for all t, which takes values in a finite collection of sets $\mathcal{S}_t^1$. Given a prescription strategy $\boldsymbol{\psi}^2$, we can construct a state evolution function $\bar{f}_t^1(\cdot)$, such that $S_{t+1}^1 = \bar{f}_t^1(S_t^1, U_t^1, W_t, V_{t+1}^{1:2})$ and an observation rule $\bar{h}_t^1(\cdot)$ which yields $Z_{t+1}^1 = \bar{h}_t^1(S_t^1, U_t^1, W_t, V_{t+1}^{1:2})$ for all $t = 0, \dots, T-1$. The existence of these functions can be verified using the dynamics and information structure of the system to write the LHS in terms of the variables in the RHS. Similarly, we can construct a cost function $\bar{c}_t^1(\cdot)$ which yields the cost $\bar{c}_t^1(S_t^1, U_t^1) := c_t(X_t, U_t^1, \psi_t^2(A_t^2)(L_t^2))$ for all t. Then, for a given prescription strategy $\boldsymbol{\psi}^2$, the new centralized problem for agent 1 has state S_t^1, control action U_t^1, observation Z_t^1, and cost $\bar{c}_t^1(S_t^1, U_t^1)$ at time t. Furthermore, the performance criterion is $\mathcal{J}^1(\boldsymbol{g}^1) := \mathbb{E}^{\boldsymbol{g}^1}[\sum_{t=0}^T \bar{c}_t^1(S_t^1, U_t^1)]$.

Problem 5. The problem for agent 1 is $\inf_{\boldsymbol{g}^1} \mathcal{J}^1(\boldsymbol{g}^1)$, given a prescription strategy $\boldsymbol{\psi}^2$, the probability distributions of the primitive random variables $\{X_0, W_{0:t}, V_{0:t}^{1:2}\}$, and the dynamics $\{\bar{c}_t^1, \bar{f}_t^1, \bar{h}_t^1 : t = 0, \dots, T\}$.

Lemma 20. *For a given control strategy $\boldsymbol{g}^2$, consider a prescription strategy $\boldsymbol{\psi}^2$ such that*

$$\psi_t^2(A_t^2)(\cdot) := g_t^2(\cdot, A_t^2), \quad t = 0, \dots, T. \tag{3.5}$$

Then, $\mathcal{J}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \mathcal{J}^1(\boldsymbol{g}^1)$ for the fixed prescription strategy $\boldsymbol{\psi}^2$. Moreover, for any given prescription strategy $\boldsymbol{\psi}^2$, consider a control strategy $\boldsymbol{g}^2$ constructed as

$$g_t^2(\cdot, A_t^2) := \psi_t^2(A_t^2)(\cdot), \quad t = 0, \dots, T. \tag{3.6}$$

Then, $\mathcal{J}^1(\boldsymbol{g}^1)$ after fixing $\boldsymbol{\psi}^2$ is equal to $\mathcal{J}(\boldsymbol{g}^1, \boldsymbol{g}^2)$.

Proof. For the first part, given a control strategy $\boldsymbol{g}$ and prescription strategy $\boldsymbol{\psi}^2$, note that $U_t^2 = g_t^2(L_t^2, A_t^2) = \psi_t^2(A_t^2)(L_t^2)$, i.e., the control law and prescription law result in the same control action U_t^2 for a given memory $M_t^2 = \{L_t^2, A_t^2\}$, for all $t = 0, \dots, T$. Thus, after fixing $\boldsymbol{\psi}^2$, we can write the expected cost at each t as $\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2})] = \mathbb{E}^{\boldsymbol{g}^1}[c_t(X_t, U_t^1, \psi_t^2(A_t^2)(L_t^2))] = \mathbb{E}^{\boldsymbol{g}^1}[\bar{c}_t^1(S_t^1, U_t^1)]$, where the second equality holds using the construction of $\bar{c}_t^1(\cdot)$. The proof is complete by summing the cost over all time steps. For the second part, the proof follows from similar arguments as in the first part. □

Remark 23. We consider that a control strategy $\boldsymbol{g}^2$ and a prescription strategy $\boldsymbol{\psi}^2$ are always selected to satisfy (3.5) and (3.6) simultaneously. Thus, fixing $\boldsymbol{\psi}^2$ in Problem

5 also fixes $\boldsymbol{g}^2$, and vice versa. Next, consider a control strategy $(\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ which is an optimal solution to Problem 5. We construct a prescription strategy for agent 2 as $\psi_t^{*2}(A_t^2)(\cdot) := g_t^{*2}(\cdot, A_t^2)$, for all $t = 0, \ldots, T$, and use the first part of Lemma 20 to conclude that $\boldsymbol{g}^{*1}$ must an optimal solution for Problem 5 after fixing $\boldsymbol{\psi}^{*2}$. Thus, every optimal solution to Problem 4 yields a corresponding solution to Problem 5.

Problem 5 is a centralized stochastic control problem for agent 1, with a perfectly observed component A_t^2 of the state S_t^1 and a partially observed component $\{X_t, L_t^2\}$, which must be estimated using the memory M_t^1. For such an estimation problem, it is known [12, page 79] that agent 1 can use the probability distribution

$$\Pi_t^1 := \mathbb{P}^{\boldsymbol{g}}\big(X_t, L_t^2 \mid M_t^1, \Gamma_{0:t-1}^2\big), \quad t = 0, \ldots, T, \tag{3.7}$$

which takes values in the set of feasible distributions $\mathcal{P}_t^1 := \Delta(\mathcal{X}_t \times \mathcal{L}_t^2)$, where $\Gamma_{0:t-1}^2$ are known given $\boldsymbol{\psi}^2$ and M_t^1. Next, we show that the information state Π_t^1 evolves independent of the choice of strategies $\boldsymbol{g}^1$ and $\boldsymbol{\psi}^2$.

Lemma 21. *For all $t = 0, \ldots, T-1$, there exists a function $\tilde{f}_t^1(\cdot)$ independent of control strategy $\boldsymbol{g}^1$ and prescription strategy $\boldsymbol{\psi}^2$, such that $\Pi_{t+1}^1 = \tilde{f}_t^1(\Pi_t^1, U_t^1, \Gamma_t^2, Z_{t+1}^1)$, and subsequently, for any Borel subset $P^1 \subseteq \mathcal{P}_{t+1}^1$, $\mathbb{P}(\Pi_{t+1}^1 \in P^1 | M_t^1, U_{0:t}^1, \Gamma_{0:t}^2) = \mathbb{P}(\Pi_{t+1}^1 \in P^1 | \Pi_t^1, U_t^1, \Gamma_t^2)$.*

Proof. The proof follows the same arguments as the ones of Lemma 23 in Section 3.1.3.2. □

Lemma 22. *For any given prescription strategy $\boldsymbol{\psi}^2$ of agent 2, there exists a function $\tilde{c}_t^1(\cdot)$ for all $t = 0, \ldots, T$, such that*

$$\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2}) \mid M_t^1, U_t^1, \Gamma_t^2] = \tilde{c}_t^1(\Pi_t^1, A_t^2, U_t^1). \tag{3.8}$$

Proof. The proof follows the same arguments as the ones of Lemma 24 in Section 3.1.3.2. □

The distribution Π_t^1 is called an *information state* of agent 1 at time t. As a consequence of Lemmas 21 and 22, the information state yields the following result for Problem 5.

Theorem 15. *For any given prescription strategy $\boldsymbol{\psi}^2$ of agent 2 in Problem 5, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^{*1}$ with the structural form*

$$U_t^1 = g_t^{*1}(A_t^2, \Pi_t^1), \quad t = 0, \dots, T. \tag{3.9}$$

Proof. This proof follows standard arguments for centralized stochastic control problems in [12, page 79], and thus, it is omitted. □

Theorem 15 establishes a structural form for an optimal control strategy $\boldsymbol{g}^{*1}$ in Problem 5, which holds for all $\boldsymbol{\psi}^2$, and subsequently, for all $\boldsymbol{g}^2$. From Remark 23, we note that any optimal control strategy $(\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ for Problem 4 must yield a corresponding prescription strategy $\boldsymbol{\psi}^{*2}$ such that after fixing $\boldsymbol{\psi}^{*2}$, the control strategy $\boldsymbol{g}^{*1}$ is the optimal solution to Problem 5. Thus, there exists an optimal control strategy $(\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ for Problem 4 where $\boldsymbol{g}^{*1}$ takes the structural form in (3.9).

Remark 24. Consider that $|\mathcal{X}_t \times \mathcal{L}_t^2| = m \in \mathbb{N}$. Then, the information state Π_t^1 takes values in the continuous space $\mathcal{P}_t^1 = \big\{\big(p_t(1), \dots, p_t(m)\big) \in [0,1]^m : \sum_{i=1}^m p_t(i) = 1\big\}$. However, for all $t = 0, \dots, T$, the information state can only take *countably* many realizations because all random variables take values in finite sets. For example, at $t = 0$, for each $x_0 \in \mathcal{X}_0$ and $l_0^2 \in \mathcal{L}_0^2$, the probability $\mathbb{P}^{\boldsymbol{g}}(x_0, l_0^2 \mid z_0^1)$ can take only finitely many values, i.e., one value for each $z_0^1 \in \mathcal{Z}_0^1$. Similarly, at any finite t, the memory M_t^1 can take finitely many realizations and thus, there are finitely many realizations for Π_t^1. As the horizon $T \to \infty$, the information state may take at most countably infinite realizations.

3.1.3.2 Analysis for Agent 2

In this subsection, we restrict agent 1 to control strategies $\boldsymbol{g}^1$ which satisfy (3.9), and derive a structural form for the optimal prescription strategy of agent 2. Given

$\boldsymbol{g}^1$, agent 2 cannot generate the action U_t^1 at each t because they cannot access Π_t^1. Thus, we consider a two stage process to generate the action of agent 1: (1) agent 2 generates a prescription for agent 1 using only A_t^2, and (2) agent 1 computes U_t^1 using this prescription along with Π_t^1.

Definition 18. For all $t = 0, \dots, T$, a *prescription* for agent 1 is a function $\Gamma_t^1 : \mathcal{P}_t^1 \to \mathcal{U}_t^1$ which takes values in a finite set $\mathcal{F}_t^1$.

At each t, the prescription for agent 1 is generated using a prescription law $\psi_t^1 : \mathcal{A}_t^2 \to \mathcal{F}_t^1$, which yields $\Gamma_t^1 = \psi_t^1(A_t^2)$. We call $\boldsymbol{\psi}^1 := \big(\psi_t^1 : t = 0, \dots, T\big)$ the prescription strategy of agent 1 and $\boldsymbol{\psi} := (\boldsymbol{\psi}^1, \boldsymbol{\psi}^2)$ the prescription strategy of the system. For a given prescription Γ_t^1, agent 1 computes their action as $U_t^1 = \Gamma_t^1(\Pi_t^1)$. Next, we set up a new centralized problem from the perspective of agent 2 with a state $S_t^2 := \{X_t, L_t^2, \Pi_t^1\}$ for all t, which takes values in the finite collection of sets $\mathcal{S}_t^2$. Moreover, we can construct a state evolution function $\bar{f}_t^2(\cdot)$ such that $S_{t+1}^2 = \bar{f}_t^2(S_t^2, \Gamma_t^{1:2}, W_t, V_{t+1}^{1:2})$ and an observation rule $\bar{h}_t^2(\cdot)$ which yields $Z_{t+1}^2 = \bar{h}_t^2(S_t^2, \Gamma_t^{1:2}, W_t, V_{t+1}^{1:2})$ for all $t = 0, \dots, T-1$. Similarly, we can construct a cost function $\bar{c}_t^2(\cdot)$ such that $\bar{c}_t^2(S_t^2, \Gamma_t^{1:2}) := c_t(X_t, \Gamma_t^1(\Pi_t^1), \Gamma_t^2(L_t^2))$ for all t. Thus, the new centralized problem for agent 2 has the state S_t^2, observation Z_t^2 and action (Γ_t^1, Γ_t^2) at each t. The corresponding performance criterion is $\mathcal{J}^2(\boldsymbol{\psi}) = \mathbb{E}^{\boldsymbol{\psi}}[\sum_{t=0}^T \bar{c}_t^2(S_t^2, \Gamma_t^{1:2})]$.

Problem 6. The optimization problem for agent 2 is $\inf_{\boldsymbol{\psi}} \mathcal{J}^2(\boldsymbol{\psi})$, given the probability distributions of the primitive random variables $\{X_0, W_{0:t}, V_{0:t}^{1:2}\}$, and the dynamics $\{\bar{c}_t^2, \bar{f}_t^2, \bar{h}_t^2 : t = 0, \dots, T\}$.

Remark 25. Using the same sequence of arguments as Lemma 20, for each control strategy $\boldsymbol{g}$, we can construct an equivalent prescription strategy $\boldsymbol{\psi}$ such that $\mathcal{J}(\boldsymbol{g}) = \mathcal{J}(\boldsymbol{\psi})$ and vice versa. Thus, we always ensure that $\boldsymbol{\psi}$ is consistent with $\boldsymbol{g}$, which implies that for all t, $\Pi_t^1 = \mathbb{P}^{\boldsymbol{g}}(X_t \mid M_t^1, \Gamma_{0:t-1}^2) = \mathbb{P}^{\boldsymbol{\psi}}(X_t \mid M_t^1, \Gamma_{0:t-1}^2) = \mathbb{P}^{\boldsymbol{\psi}}(X_t \mid M_t^1, \Gamma_{0:t-1}^1, \Gamma_{0:t-1}^2)$, where we can add $\Gamma_{0:t-1}^1$ to the conditioning because they are functions of $A_t^2 \subseteq M_t^1$ and $\boldsymbol{\psi}^1$. Because of this property, we can equivalently write the dependence of a probability distribution on either $\boldsymbol{g}$ or $\boldsymbol{\psi}$.

Problem 6 is a partially observed centralized stochastic control problem and thus, agent 2 must estimate the state S^2_t at each time t. For this purpose, agent 2 can use the distribution

$$\Pi^2_t := \mathbb{P}^{\boldsymbol{\psi}}(X_t, L^2_t, \Pi^1_t \mid A^2_t, \Gamma^{1:2}_{0:t-1}), \quad t = 0, \dots, T, \tag{3.10}$$

which takes values in the set of feasible distributions $\mathcal{P}^2_t := \Delta(\mathcal{X}_t \times \mathcal{L}^2_t \times \mathcal{P}^1_t)$. Recall that at each t, the information state of agent 1, Π^1_t, can take at most countably infinitely many realizations in the space $\mathcal{P}^1_t$. Thus, the information state Π^2_t can be represented using a tuple of probability mass functions $\big(p_t(x_t, \ell^2_t, \cdot \,|a^2_t, \gamma^{1:2}_{0:t-1}) : x_t \in \mathcal{X}_t, \ell^2_t \in \mathcal{L}^2_t\big)$, where $p_t(x_t, \ell^2_t, \cdot \,|a^2_t, \gamma^{1:2}_{0:t-1}) : \mathcal{P}^1_t \to [0,1]$ for each $x_t \in \mathcal{X}_t$ and $\ell^2_t \in \mathcal{L}^2_t$. Next, we show that the evolution of Π^2_t is Markovian and independent of the prescription strategy $\boldsymbol{\psi}$.

Lemma 23. *For all $t = 0, \dots, T-1$, there exists a function $\tilde{f}^2_t(\cdot)$ independent of the prescription strategy $\boldsymbol{\psi}$, such that $\Pi^2_{t+1} = \tilde{f}^2_t(\Pi^2_t, \Gamma^1_t, \Gamma^2_t, Z^2_{t+1})$, and subsequently, for any Borel subset $P^2 \subseteq \mathcal{P}^2_{t+1}$, $\mathbb{P}(\Pi^2_{t+1} \in P^2 \mid A^2_t, \Gamma^{1:2}_{0:t}) = \mathbb{P}(\Pi^2_{t+1} \in P^2 \mid \Pi^2_t, \Gamma^{1:2}_t)$.*

Proof. Let x_t, γ^1_t, γ^2_t, a^2_t, and π^1_t be realizations of X_t, Γ^1_t, Γ^2_t, A^2_t, and the distribution Π^1_t, respectively, for all t. Then, using Bayes' rule

$$\begin{aligned}
&\mathbb{P}^{\boldsymbol{\psi}}(X_{t+1} = x_{t+1}, L^2_{t+1} = \ell^2_{t+1}, \Pi^1_{t+1} = \pi^1_{t+1} \mid a^2_{t+1}, \gamma^{1:2}_{0:t}) \\
&\qquad = \frac{\mathbb{P}^{\boldsymbol{\psi}}\big(X_{t+1} = x_{t+1}, L^2_{t+1} = \ell^2_{t+1}, \Pi^2_{t+1} = \pi^1_{t+1}, Z^2_{t+1} = z^2_{t+1} \mid a^2_t, \gamma^{1:2}_{0:t}\big)}{\mathbb{P}^{\boldsymbol{\psi}}\big(Z^2_{t+1} = z^2_{t+1} \mid a^2_t, \gamma^{1:2}_{0:t}\big)},
\end{aligned} \tag{3.11}$$

where $a^2_{t+1} = a^2_t \cup z^2_{t+1}$. Using the dynamics $\{\bar{f}^2_t, \bar{h}^2_t, \bar{c}^2_t\}$, we write that (x_{t+1}, ℓ^2_{t+1}) $= \eta^2_t(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1})$, $\pi^1_{t+1} = \xi^2_t(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1})$, $z^2_{t+1} = \bar{h}^2_t(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1})$, for some appropriate functions $\eta^2_t(\cdot)$ and $\xi^2_t(\cdot)$, where $s^2_t = \{x_t, \ell^2_t, \pi^1_t\}$. Substituting these relationships into the numerator in the RHS of (3.11) yields that

$$\begin{aligned}
&\mathbb{P}^{\boldsymbol{\psi}}\big(X_{t+1} = x_{t+1}, L^2_{t+1} = \ell^2_{t+1}, \Pi^1_{t+1} = \pi^1_{t+1}, Z^2_{t+1} = z^2_{t+1} \mid a^2_t, \gamma^{1:2}_{0:t}\big) \\
&\qquad = \sum_{s^2_t, w_t, v^{1:2}_{t+1}} \mathbb{I}[\eta^2_t(s^2_t, \gamma^{1:2}_t, w_t) = (x_{t+1}, \ell^2_{t+1})] \cdot \mathbb{P}(W_t = w_t, V^{1:2}_{t+1} = v^{1:2}_{t+1}) \\
&\qquad\qquad \cdot \mathbb{I}[\xi^2_t(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1}) = \pi^1_{t+1}] \cdot \mathbb{P}^{\boldsymbol{\psi}}\big(S^2_t = s^2_t \mid a^2_t, \gamma^{1:2}_{0:t-1}\big)
\end{aligned}$$

$$\cdot\, \mathbb{I}[\bar{h}^2_{t+1}(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1}) = z^2_{t+1}], \quad (3.12)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and where we can drop the prescriptions $\gamma^{1:2}_t$ from the conditioning in the last term because they are completely determined given $\boldsymbol{\psi}$ and a^2_t. Note that in (3.12), $\mathbb{P}^{\boldsymbol{\psi}}\big(S^2_t = s^2_t \mid a^2_t, \gamma^{1:2}_{0:t-1}\big) = \pi^2_t(s^2_t)$. Next, we expand the denominator in (3.11) as

$$\mathbb{P}^{\boldsymbol{\psi}}\big(Z^2_{t+1} = z^2_{t+1} \mid a^2_t, \gamma^{1:2}_{0:t}\big) \\ = \sum_{s^2_t, w_t, v^{1:2}_{t+1}} \mathbb{P}(W_t = w_t, V^{1:2}_{t+1} = v^{1:2}_{t+1}) \cdot \mathbb{I}[\bar{h}^2_t(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1}) = z^2_{t+1}] \cdot \pi^2_t(s^2_t). \quad (3.13)$$

Then, the first result holds by constructing an appropriate function $\tilde{f}^2_t(\cdot)$ using (3.11) - (3.13). To prove the second result, for any Borel subset $P^2 \subseteq \mathcal{P}^2_{t+1}$, we write that

$$\mathbb{P}(\Pi^2_{t+1} \in P^2 \mid a^2_t, \gamma^{1:2}_{0:t}, \pi^2_{0:t}) \\ = \sum_{z^2_{t+1}} \mathbb{I}[\tilde{f}^2_t(\pi^2_t, \gamma^{1:2}_t, z^2_{t+1}) \in P^2] \cdot \mathbb{P}(Z^2_{t+1} = z^2_{t+1} \mid a^2_t, \gamma^{1:2}_{0:t}, \pi^2_{0:t}). \quad (3.14)$$

The second term in (3.14) can be expanded as

$$\mathbb{P}(Z^2_{t+1} = z^2_{t+1} \mid a^2_t, \gamma^{1:2}_{0:t}, \pi^2_{0:t}) \\ = \sum_{s^2_t, w_t, v^{1:2}_{t+1}} \mathbb{I}[\bar{h}^2_t(s^2_t, \gamma^{1:2}_t, w_t, v^{1:2}_{t+1}) = z^2_{t+1}] \cdot \mathbb{P}(W_t = w_t, V^{1:2}_{t+1} = v^{1:2}_{t+1}) \cdot \pi^2_t(s^2_t). \quad (3.15)$$

The proof is complete by substituting this equation into (3.14). □

Lemma 24. *There exists a function $\tilde{c}^2_t(\cdot)$ for all t, such that*

$$\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U^{1:2}_t) \mid A^2_t, \Gamma^{1:2}_t] = \tilde{c}^2_t(\Pi^2_t, \Gamma^{1:2}_t). \quad (3.16)$$

Proof. Let a^3_t, $\gamma^{1:2}_t$, and π^2_t be realizations of the random variables A^3_t, $\Gamma^{1:2}_t$, and the conditional distribution Π^2_t, respectively, for all $t = 0, \dots, T$. To prove the result, we expand the expectation as

$$\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U^{1:2}_t) \mid a^2_t, \gamma^{1:2}_t] = \mathbb{E}^{\boldsymbol{\psi}}[\bar{c}^2_t(S^2_t, \Gamma^{1:2}_t) \mid a^2_t, \gamma^{1:2}_t]$$

$$= \sum_{s_t^2} \bar{c}_t^2(s_t^2, \gamma_t^{1:2}) \cdot \mathbb{P}^{\boldsymbol{\psi}}(S_t^2 = s_t^2 \mid a_t^2, \gamma_t^{1:2})$$

$$= \sum_{s_t^2} \bar{c}_t^2(s_t^2, \gamma_t^{1:2}) \cdot \pi_t^2(s_t^2) =: \tilde{c}_t^2(\pi_t^2, \gamma_t^{1:2}), \quad (3.17)$$

where we can drop the prescriptions $\gamma_t^{1:2}$ from the conditioning because they known given $\boldsymbol{\psi}$ and a_t^2. □

We call Π_t^2 the information state of agent 2 at time t. As a consequence of Lemmas 21 and 22, the information state yields the following result for Problem 6.

Theorem 16. *In Problem 6, without loss of optimality, we can restrict our attention to prescription strategies $\boldsymbol{\psi}^*$ with the structural form*

$$\Gamma_t^k = \psi_t^{*k}\big(\Pi_t^2\big), \quad k = 1, 2, \quad t = 0, \dots, T. \quad (3.18)$$

Proof. This proof follows similar arguments for centralized stochastic control problems in [12, page 79], and thus, it is omitted. □

Consider a prescription strategy $\boldsymbol{\psi}^* = (\boldsymbol{\psi}^{*1}, \boldsymbol{\psi}^{*2})$ which is an optimal solution to Problem 6, and a control strategy $\boldsymbol{g}^* = (\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ given by $g_t^{*1}(\Pi_t^{1:2}) := \psi_t^{*1}(\Pi_t^2)(\Pi_t^1)$ and $g_t^{*2}(L_t^2, \Pi_t^2) := \psi_t^{*2}(\Pi_t^2)(L_t^2)$ for each $k = 1, 2$ and $t = 0, \dots, T$. Using the same arguments as in Lemma 20, we conclude that $\mathcal{J}(\boldsymbol{g}^*) = \mathcal{J}^2(\boldsymbol{\psi}^*)$ and subsequently, that $\boldsymbol{g}^*$ is the optimal solution to Problem 4. Thus, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^*$ with the structural form $U_t^1 = g_t^{*1}(\Pi_t^1, \Pi_t^2)$ and $U_t^2 = g_t^{*2}(L_t^2, \Pi_t^2)$ for all $t = 0, \dots, T$.

Remark 26. Consider a system where, the feasible sets of system variables are time invariant, i.e., $\mathcal{X}_t = \mathcal{X}$, $\mathcal{W}_t = \mathcal{W}$, $\mathcal{V}_t^k = \mathcal{V}^k$, $\mathcal{Y}_t^k = \mathcal{Y}^k$ for each $k = 1, 2$ and $t = 0, \dots, T$, and the information structure satisfies $\mathcal{L}_t^2 = \mathcal{L}^2$, $\mathcal{Z}_t^1 = \mathcal{Z}^1$, $\mathcal{Z}_t^2 = \mathcal{Z}^2$ for all t. Note that the set $\mathcal{M}_t^1$ still grows in size with time. However, the spaces $\mathcal{P}^1 = \Delta(\mathcal{X} \times \mathcal{L}^2)$ and $\mathcal{P}^2 = \Delta(\mathcal{X} \times \mathcal{L}^2 \times \mathcal{P}^1)$ are time invariant and subequently, our optimal control strategies have time-invariant domains for both agents. This is a useful property to derive and implement optimal control strategies for long time horizons.

3.1.3.3 Dynamic Programming Decomposition

In this subsection, we construct the value functions and corresponding control laws to form a DP decomposition which can derive the optimal prescription strategies. Let γ_t^k and π_t^k be the realizations of the prescription Γ_t^k and information state Π_t^k, respectively, for each $k = 1, 2$ and $t = 0, \dots, T$. Then, we recursively define the value functions

$$J_t(\pi_t^2) \ := \ \inf_{\gamma_t^{1:2} \in \mathcal{F}_t^1 \times \mathcal{F}_t^2} \tilde{c}_t^2\big(\pi_t^2, \gamma_t^{1:2}\big) \ + \ \mathbb{E}^{\psi}\Big[J_{t+1}\big(\tilde{f}_t^2(\pi_t^2, \gamma_t^{1:2}, Z_{t+1}^2)\big) \ \mid \ \pi_t^2, \gamma_t^{1:2}\Big], \quad (3.19)$$

for all $t = 0, \dots, T$ and define $J_{T+1}(\pi_{T+1}^2) := 0$ identically. For each agent $k = 1, 2$, the prescription law at time t is $\gamma_t^{*k} = \psi_t^{*k}(\pi_t^2)$, i.e., the $\arg\inf$ in the RHS of (3.19). The prescription strategy $\boldsymbol{\psi}^*$ derived using this DP decomposition can be shown to be the optimal solution to Problem 5 using standard arguments [45, 168]. Recall that given an optimal strategy $\boldsymbol{\psi}^*$ derived using this DP decomposition, we can also derive the optimal control strategy $\boldsymbol{g}^*$ for Problem 4.

Remark 27. At each $t = 0, \dots, T$, our DP decomposition requires solving an optimization problem for each realization π_t^2 of the information state Π_t^2, which is a tuple of probability mass functions. Optimizing over probability mass functions is a computationally challenging problem. Next, we present two different approaches to alleviate the computational implications. In Section 3.1.4, we show how we can simplify our results when the system dynamics and information structure have additional favorable properties. In Section 3.1.5, we present an approximation for the information states which can reduce the number of computations required to derive an approximately optimal strategy.

3.1.4 Simplification for Decoupled Dynamics

In this subsection, we show how our results can be simplified when both agents have decoupled state and observation dynamics. We denote the state of each agent $k = 1, 2$ at time t by $X_t^k \in \mathcal{X}_t^k$. Starting at X_0^k, each state evolves as

$$X_{t+1}^k = f_t^k(X_t^k, U_t^k, W_t^k), \quad t = 0, \dots, T-1, \quad (3.20)$$

for $k = 1, 2$, where $W_t^k \in \mathcal{W}_t^k$ is a disturbance acting only on X_t^k. The observation of agent k at time t is $Y_t^k = h_t^k(X_t^k, V_t^k)$. We assume that all primitive random variables $\{X_0^k, W_t^k, V_t^k : k = 1, 2,\ t = 0, \dots, T\}$ are independent of each other and that the cost to the system at each $t = 0, \dots, T$ is $c_t(X_t^{1:2}, U_t^{1:2}) \in \mathbb{R}_{\geq 0}$. Without loss of optimality, we restrict attention to control strategies where $\boldsymbol{g}^1$ takes the form $U_t^1 = g_t^1(\Pi_t^1, \Pi_t^2)$ and where $\boldsymbol{g}^2$ takes the form $U_t^2 = g_t^2(L_t^2, \Pi_t^2)$, for all $t = 0, \dots, T$. Here, recall that $\Pi_t^1 = \mathbb{P}^{\boldsymbol{g}}\big(X_t^{1:2}, L_t^2 | M_t^1, \Gamma_{0:t-1}^2\big)$ and $\Pi_t^2 = \mathbb{P}^{\boldsymbol{g}}\big(X_t^{1:2}, L_t^2, \Pi_t^1 | A_t^2, \Gamma_{0:t-1}^{1:2}\big)$. Next, we show that the information state Π_t^1 can be simplified using the decoupled dynamics.

Lemma 25. *For each $k = 1, 2$ and $t = 0, \dots, T$, let x_t^k, m_t^k, l_t^2, and a_t^2 be realizations of the random variables X_t^k, M_t^k, L_t^2, and A_t^2, respectively. Then,*

$$\mathbb{P}^{\boldsymbol{g}}(X_t^{1:2} = x_t^{1:2}, L_t^2 = l_t^2 \mid m_t^1) = \mathbb{P}^{\boldsymbol{g}}(X_t^1 = x_t^1 \mid m_t^1) \cdot \mathbb{P}^{\boldsymbol{g}}(X_t^2 = x_t^2, L_t^2 = l_t^2 \mid a_t^2). \quad (3.21)$$

Proof. Given the realizations x_t^k, y_t^k, u_t^k, γ_t^k and l_t^2 of X_t^k, Y_t^k, U_t^k, Γ_t^k, and L_t^2, respectively, for each $k = 1, 2$ and $t = 0, \dots, T$, we prove (3.21) by mathematical induction. At $t = 0$, depending on the information sharing pattern of the system, there are two possible realizations of the memory of agent 1, either $m_0^1 = \{y_0^1\}$ or $m_0^1 = \{y_0^1, y_0^2\}$. For the first realization of the memory of agent 1, the private information of agent 2 is $l_0^2 = \{y_0^2\}$, and thus, we can expand the LHS of (3.21) as

$$\begin{aligned}\mathbb{P}^{\boldsymbol{g}}(X_0^{1:2} = x_0^{1:2}, Y_0^2 = y_0^2 \mid m_0^1) &= \mathbb{P}^{\boldsymbol{g}}(X_0^{1:2} = x_0^{1:2}, Y_0^2 = y_0^2 \mid y_0^1) \\ &= \mathbb{P}^{\boldsymbol{g}}(X_0^2 = x_0^1 \mid y_0^1) \cdot \mathbb{P}^{\boldsymbol{g}}(X_0^2 = x_0^2, Y_0^2 = y_0^2), \quad (3.22)\end{aligned}$$

where recall that the observation y_0^k depends only on x_0^k for each k, and the primitive random variables are independent of each other. For the second realization of the memory of agent 1, note that $l_t^2 = \emptyset$ because $l_t^2 \cap m_t^1 = \emptyset$, and thus, we can expand the LHS as $\mathbb{P}^{\boldsymbol{g}}(X_0^{1:2} = x_0^{1:2} | m_0^1) = \mathbb{P}^{\boldsymbol{g}}(X_0^1 = x_0^1 | y_0^1) \cdot \mathbb{P}^{\boldsymbol{g}}(X_0^2 = x_0^2 | y_0^2)$. For both cases at $t = 0$, we have shown the LHS is equal to the RHS in (3.21). This forms the basis of our induction. Next, we consider the induction hypothesis that (3.21) holds at each $0, \dots, t$, and expand the LHS at $t + 1$ as

$$
\begin{aligned}
&\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_{t+1} = l^2_{t+1} \mid m^1_{t+1}) \\
&\qquad = \frac{\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_{t+1} = l^2_{t+1}, z^1_{t+1} \mid m^1_t)}{\mathbb{P}^{\boldsymbol{g}}(Z^1_{t+1} = z^1_{t+1} \mid m^1_t)} \\
&\qquad = \frac{\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_{t+1} = l^2_{t+1}, Z^1_{t+1} = z^1_{t+1} \mid m^1_t)}{\sum_{x^{1:2}_{t+1}, l^2_{t+1}} \mathbb{P}^{\boldsymbol{g}}(X_{t+1} = x^{1:2}_{t+1}, L^2_{t+1} = l^2_{t+1}, Z^1_{t+1} = z^1_{t+1} \mid m^1_t)}. \qquad (3.23)
\end{aligned}
$$

Note that in the partially common information structure, $l^2_{t+1} \cup z^1_{t+1} = l^2_t \cup \{y^{1:2}_{t+1}, u^{1:2}_t\}$. Thus, we can write that

$$
\begin{aligned}
&\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_{t+1} = l^2_{t+1}, Z_{t+1} = z^1_{t+1} \mid m^1_t) \\
&\qquad = \mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, Y_{t+1} = y^{1:2}_{t+1}, U^{1:2}_t = u^{1:2}_t, L^2_t = l^2_t \mid m^1_t) \\
&\qquad = \mathbb{P}^{\boldsymbol{g}}(Y^1_{t+1} = y^1_{t+1} | x^1_{t+1}) \cdot \mathbb{P}^{\boldsymbol{g}}(Y^2_{t+1} = y^2_{t+1} | x^2_{t+1}) \cdot \mathbb{I}[g^1_t(m^1_t) = u^1_t] \\
&\qquad\qquad \cdot \mathbb{I}[\gamma^2_t(l^2_t) = u^2_t] \cdot \mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_t = l^2_t \mid m^1_t), \qquad (3.24)
\end{aligned}
$$

where $\mathbb{I}(\cdot)$ is the indicator function, and where γ^2_t and u^1_t are completely determined given m^1_t and $\boldsymbol{g}$. Furthermore, we expand the last term as $\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_t = l^2_t | m^1_t, u^1_t, \gamma^2_t) = \sum_{x^{1:2}_t, w^{1:2}_t} \mathbb{I}[f^1_t(x^1_t, u^1_t, w^1_t) = x^1_{t+1}] \cdot \mathbb{I}[f^2_t(x^2_t, \gamma^2_t(l^2_t), w^2_t) = x^2_{t+1}] \cdot \mathbb{P}(W^{1:2}_t = w^{1:2}_t) \cdot \mathbb{P}^{\boldsymbol{g}}(X^{1:2}_t = x^{1:2}_t, L^2_t = l^2_t | m^1_t)$, where we can use the induction hypothesis to obtain $\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_t = x^{1:2}_t, L^2_t = l^2_t | m^1_t) = \mathbb{P}^{\boldsymbol{g}}(x^1_t | m^1_t) \cdot \mathbb{P}^{\boldsymbol{g}}(X^{1:2}_t = x^{1:2}_t, L^2_t = l^2_t | a^2_t)$. Substituting these results into (3.23), and rearranging the terms yields

$$
\begin{aligned}
&\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_{t+1} = x^{1:2}_{t+1}, L^2_{t+1} = l^2_{t+1} | m^1_{t+1}) \\
&= \frac{\mathbb{P}^{\boldsymbol{g}}(X_{t+1} = x^1_{t+1}, Y^1_{t+1} = y^1_{t+1}, U^1_t = u^1_t | m^1_t)}{\mathbb{P}^{\boldsymbol{g}}(Y^1_{t+1} = y^1_{t+1}, U^1_t = u^1_t | m^1_t)} \cdot \mathbb{P}^{\boldsymbol{g}}(X^2_{t+1} = x^2_{t+1}, L^2_{t+1} = l^2_{t+1} | a^2_t, z^2_{t+1}) \\
&\qquad = \mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1} | m^1_t, y^1_{t+1}, u^1_t) \cdot \mathbb{P}^{\boldsymbol{g}}(X^2_{t+1} = x^2_{t+1}, Y^2_{t+1} = y^2_{t+1} | a^2_{t+1}). \qquad (3.25)
\end{aligned}
$$

To complete the proof by mathematical induction, we need to show that the first term in the RHS of the previous equation is equal to the first term in the RHS of (3.21). We achieve this by expanding

$$
\mathbb{P}^{\boldsymbol{g}}(X_{t+1} = x^1_{t+1} | m^1_t, y^{1:2}_{t+1}, u^{1:2}_t, l^2_t)
$$

$$\begin{aligned}
&= \frac{\mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1}, Y^1_{t+1} = y^1_{t+1} | m^1_t, l^2_t, y^2_{t+1})}{\sum_{x^1_{t+1}} \mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1}, Y^1_{t+1} = y^1_{t+1} | m^1_t, l^2_t, y^2_{t+1})} \\
&= \frac{\sum_{x^1_t} \mathbb{P}^{\boldsymbol{g}}(Y_{t+1} = y^1_{t+1} | x^1_{t+1}) \cdot \mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1} | x^1_t, u^1_t) \cdot \mathbb{P}^{\boldsymbol{g}}(X^1_t = x^1_t | m^1_t)}{\sum_{x^1_{t:t+1}} \mathbb{P}^{\boldsymbol{g}}(Y^1_{t+1} = y^1_{t+1} | x^1_{t+1}) \cdot \mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1} | x^1_t, u^1_t) \cdot \mathbb{P}^{\boldsymbol{g}}(X^1_t = x^1_t | m^1_t)},
\end{aligned} \tag{3.26}$$

where, in the last equality, we use Bayes' rule and the induction hypothesis. Recall that $z^1_{t+1} \subseteq l^2_t \cup \{y^{1:2}_{t+1}, u^{1:2}_t\}$. This implies that $\mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1} | m^1_t, y^{1:2}_{t+1}, u^{1:2}_t, l^2_t) = \mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1} | m^1_t, z^1_{t+1}) = \mathbb{P}^{\boldsymbol{g}}(X^1_{t+1} = x^1_{t+1} | m^1_t, y^1_{t+1}, u^1_t)$, which complete the proof. □

Motivated by Lemma 25, we define the distributions $\Theta^1_t := \mathbb{P}^{\boldsymbol{g}}(X^1_t | M^1_t)$ and $\Theta^2_t := \mathbb{P}^{\boldsymbol{g}}(X^2_t, L^2_t | A^2_t)$ and note that the information state Π^1_t at each $t = 0, \dots, T$ can be written as a function of (Θ^1_t, Θ^2_t). Thus, at time t, agent 1 can track the distributions (Θ^1_t, Θ^2_t) instead of Π^1_t to compute their optimal control action U^1_t. Next, we show that the evolution of Θ^k_t, for each $k = 1, 2$, is Markovian, strategy independent and decoupled from the dynamics of the other agent.

Lemma 26. *At each time t, there exists a function $\tilde{e}^k_t(\cdot)$, independent of the strategy $\boldsymbol{g}$, for all $k = 1, 2$ such that*

$$\Theta^k_{t+1} = \tilde{e}^k_t(\Theta^k_t, U^k_t, Y^k_{t+1}). \tag{3.27}$$

Proof. The proof follows the same arguments as the ones in Lemma 23 and thus, due to space limitations, it is omitted. □

Note that the distribution Θ^2_t is also available to agent 2 at each $t = 0, \dots, T$, because it depends only on the common information A^2_t. Subsequently, using the same sequence of arguments as the ones in Theorem 16, we conclude that, without loss of optimality, agent 2 can restrict attention to prescription strategies with the structural form $\Gamma^k_t = \psi^k_t\big(\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_t, L^2_t, \Theta^1_t | A^2_t), \Theta^2_t\big)$, for each $k = 1, 2$ and $t = 0, \dots, T$. Next, we show that the term $\mathbb{P}^{\boldsymbol{g}}(X^{1:2}_t, L^2_t, \Theta^1_t | A^2_t)$ in the argument of the prescription law for each k can also be simplified using the decoupled dynamics of the system.

Lemma 27. *For each $k = 1, 2$ and $t = 0, \dots, T$, let x_t^k, l_t^2, a_t^2, and θ_t^k be realizations of the random variables X_t^k, L_t^2, A_t^2, and the probability distribution Θ_t^k, respectively. Then,*

$$\mathbb{P}^{\boldsymbol{g}}(X_t^{1:2} = x_t^{1:2}, L_t^2 = l_t^2, \Theta_t^1 = \theta_t^1 | a_t^2) \\ = \mathbb{P}^{\boldsymbol{g}}(X_t^1 = x_t^1, \Theta_t^1 = \theta_t^1 | a_t^2) \cdot \mathbb{P}^{\boldsymbol{g}}(X_t^2 = x_t^2, L_t^2 = l_t^2 | a_t^2). \quad (3.28)$$

Proof. The proof follows by mathematical induction using the same arguments as the ones in Lemma 25, and thus, due to space limitations, it is omitted. □

Starting with the structural form of optimal prescription strategies in Theorem 16, we can use Lemmas 25 and 27, to conclude that in systems with decoupled dynamics, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^*$ with the structural form

$$U_t^1 = g_t^{*1}\big[\Theta_t^1, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1, \Theta_t^1 \mid A_t^2)\big], \quad (3.29)$$

$$U_t^2 = g_t^{*2}\big[L_t^2, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1, \Theta_t^1 \mid A_t^2)\big], \; t = 0, \dots, T. \quad (3.30)$$

Remark 28. The control strategy $\boldsymbol{g}^1$ yielded a control law for each $t = 0, \dots, T$ for agent 1 with the form $U_t^1 = g_t^1(\Pi_t^1, \Pi_t^2)$, which has the domain $\Delta(\mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \mathcal{L}_t^2 \times \Delta(\mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \mathcal{L}_t^2))$. In contrast, the domain of the control law g_t^{*1} in (3.29) is $\Delta(\mathcal{X}_t^1) \times \Delta(\mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1 \times \Delta(\mathcal{X}_t^1))$, which is a space with a smaller dimension than the one before. Similarly, the control laws of agent 2 have a domain with a smaller dimension in (3.30) than the control laws derived using Theorem 16. Thus, we have obtained a simpler form for an optimal control strategy in systems with decoupled dynamics.

We can further simplify the structural form of the optimal control strategies when agent 1 can perfectly observe the state X_t^1, i.e, $Y_t^1 = X_t$ and subsequently, $X_t^1 \subseteq M_t^1$ at each $t = 0, \dots, T$. Then, for a given realization m_t^1 of the memory M_t^1, the probability distribution Θ_t^1 at each t is simply given by $\Theta_t^1 = \mathbb{I}[X_t^1 = x_t^1]$ for the realization $x_t^1 \in m_t^1$ of X_t^1, where $\mathbb{I}$ is the indicator function. Using this result in (3.29) and (3.30), we conclude that, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^*$ with the form

$$U_t^1 = g_t^1\big[X_t^1, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1 \mid A_t^2)\big], \tag{3.31}$$

$$U_t^2 = g_t^2\big[L_t^2, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1 \mid A_t^2)\big], \quad t = 0, \dots, T. \tag{3.32}$$

Remark 29. When agent 1 can perfectly observe their own state, at each t, the domains of the optimal control laws g_t^{*1} and g_t^{*2} are $\mathcal{X}_t^1 \times \Delta(\mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1)$ and $\mathcal{L}_t^2 \times \Delta(\mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1)$, respectively. These domains are small enough that the optimal control laws at each t are functions of distributions over finite sets instead of probability mass functions. Thus, the resulting DP can be solved using standard techniques for centralized problems.

3.1.5 Implementation

In this subsection, we present an approach to approximate the information state Π_t^1 for all $t = 0, \dots, T$ which ensures that the approximation can only take finitely many values. To simplify the notation, we restrict our attention to systems where $|\mathcal{X}_t \times \mathcal{L}_t^2| = m$, $m \in \mathbb{N}$ for all $t = 0, \dots, T$. Furthermore, we consider that the maximum cost at each t is bounded above by $||c||_\infty < \infty$. Recall that the space of feasible values for Π_t^1 is the simplex $\mathcal{P}^1 = \big\{\big(p(1), \dots, p(m)\big) \in [0,1]^m : \sum_{i=1}^m p(i) = 1\big\}$. We use the procedure in [203] to generate a set of equally distributed points in $\mathcal{P}^1$. Specifically, we select a number $n \in \mathbb{N}$ and define a set $\mathcal{Q}_n := \big\{\big(q(1), \dots, q(m)\big) \in \mathcal{P}^1 : n{\cdot}q(i) \in \mathbb{N}_{\geq 0}, i = 1, \dots, m\big\}$. The set $\mathcal{Q}_n$ forms a lattice containing $|\mathcal{Q}_n| = \binom{m+n-1}{m-1}$ points in the simplex $\mathcal{P}^1$. For example, let $\mathcal{X}_t = \{0,1\}$ and $\mathcal{L}_t^2 = \emptyset$, which implies that $m = 2$. Then, by selecting $n = 2$ we construct the set $\mathcal{Q}_2 = \big\{(0,1), (1/2,1/2), (1,0)\big\}$. Similarly, if $m = 3$ and we select $n = 2$, we construct the set $\mathcal{Q}_2 = \big\{(1,0,0), (1/2,1/2,0), (0,1,0), (0,1/2,1/2), (0,0,1), (1/2,0,1/2)\big\}$. Next, we define the total variation distance between any point in $\mathcal{P}^1$ and $\mathcal{Q}_n$, and then, we use this metric to define an approximate information state.

Definition 19. The total variation distance between any $\pi_t^1 = (p(1), \ldots, p(m)) \in \mathcal{P}^1$ and any $q_t = (q(1), \ldots, q(m)) \in \mathcal{Q}_n$ is $|\pi_t^1 - q|_{TV} = \sum_{i=1}^m |p(i) - q(i)|$.

Definition 20. The approximate information state for agent 1 at each $t = 0, \ldots, T$ is a random variable $\hat{\Pi}_t^1$ which takes values in the finite set $\mathcal{Q}_n$, and which is given by

$$\hat{\Pi}_t^1 = \sigma(\Pi_t^1) := \arg\min_{q \in \mathcal{Q}_n} |\Pi_t^1 - q|_{TV}. \tag{3.33}$$

Given any distribution $\pi_t^1 \in \mathcal{P}^1$, the corresponding realization of the approximate information state $\hat{\pi}_t^1 = \sigma(\pi_t^1)$ can be efficiently computed using the algorithm in [203]. Next, we present an upper bound in the total variation distance between any information state and its approximation.

Lemma 28. *For all $t = 0, \ldots, T$, for any realization π_t^1 of the information state Π_t^1, it holds that*

$$|\pi_t^1 - \sigma(\pi_t^1)|_{TV} \leq \frac{2a \cdot (1 + a)}{m \cdot n}, \tag{3.34}$$

where $a = \lfloor m/2 \rfloor \in \mathbb{N}$ and $\lfloor \cdot \rfloor$ is the floor function.

Proof. The proof follows from [203, Proposition 2]. □

Given any upper bound $\epsilon \in \mathbb{R}_{>0}$, we can use Lemma 28 to construct a set $\mathcal{Q}_n$ which satisfies $|\pi_t^1 - \sigma(\pi_t^1)| \leq \epsilon$ for all $\pi_t^1 \in \mathcal{P}^1$, by selecting $n \geq \frac{2a(1+a)}{m \cdot \epsilon}$. Furthermore, the resulting approximate information state $\hat{\Pi}_t^1$ can be updated in a Markovian and strategy independent manner as $\hat{\Pi}_{t+1}^1 = \sigma[\tilde{f}_t^1(\hat{\Pi}_t^1, U_t^1, \Gamma_t^2, Z_{t+1}^1)]$, for all $t = 0, \ldots, T-1$.

Our aim is to solve the centralized Problem 5 for agent 1 using the approximate information state $\hat{\Pi}_t^1$, which takes only finitely many values for all $t = 0, \ldots, T$, instead of the information state Π_t^1, which can take countably infinitely many values. For a fixed prescription strategy $\boldsymbol{\psi}^2$, recall from Lemma 22 that the expected cost at time t can be written as $\tilde{c}_t^1(\Pi_t^1, A_t^2, U_t^1)$. Then, in Problem 5, we can optimize the performance criterion $\mathcal{J}^1(\boldsymbol{g}^1)$ using a centralized DP as follows. Let u_t^1, a_t^2, and π_t^1 be the realizations of U_t^1, A_t^2 and Π_t^1, respectively. Then, we define the value functions for all $t = 0, \ldots, T$ as

$$J_t^1(\pi_t^1, a_t^2) := \inf_{u_t^1 \in \mathcal{U}_t^1} \tilde{c}_t^1(\pi_t^1, a_t^2, u_t^1) + \mathbb{E}[J_{t+1}^1(\Pi_{t+1}^1, A_{t+1}^2) \mid \pi_t^1, a_t^2, u_t^1], \tag{3.35}$$

and $J^1_{T+1}(\pi_{T+1}, a_{T+1}) := 0$ identically. The person-by-person optimal control law at time t is $u^{*1}_t = g^{*1}_t(\pi^1_t, a^2_t)$, i.e., the arg inf in the RHS of (3.35), and the performance of the system is $\mathcal{J}^1(\boldsymbol{g}^{*1}) = \mathbb{E}[J^1_0(\Pi^1_0, A^2_0)]$.

However, we seek the best control strategy $\hat{\boldsymbol{g}}^{*1}$ for Problem 5 which takes the structural form $u^1_t = \hat{g}^1_t(\hat{\pi}^1_t, a^2_t)$ for all $t = 0, \dots, T$. Thus, we define the modified value functions for all $t = 0, \dots, T$ as

$$\hat{J}^1_t(\hat{\pi}^1_t, a^2_t) := \inf_{u^1_t \in \mathcal{U}^1} \tilde{c}^1_t(\hat{\pi}^1_t, a^2_t, u^1_t) + \mathbb{E}[\hat{J}^1_{t+1}(\hat{\Pi}^1_{t+1}, A^2_{t+1}) \mid \hat{\pi}^1_t, a^2_t, u^1_t], \tag{3.36}$$

and $\hat{J}^1_{t+1}(\hat{\pi}^1_t, a^2_t) := 0$ identically, where $\hat{\pi}^1_t = \sigma(\pi^1_t)$. For a fixed $\boldsymbol{\psi}^2$, the best control law using the approximate information state at each t is $u^{*1}_t = \hat{g}^{*1}_t(\hat{\pi}^1_t, a^2_t)$, i.e., the arg inf in the RHS of (3.36), and the performance of the system is $\mathcal{J}^1(\hat{\boldsymbol{g}}^{*1}) = \mathbb{E}[\hat{J}^1_0(\hat{\Pi}^1_0, A^2_0)]$. The *loss in person-by-person performance* which arises from using the approximate information state is measured by the difference $|\mathcal{J}^1(\boldsymbol{g}^{*1}) - \mathcal{J}^1(\hat{\boldsymbol{g}}^{*1})|$. Next, we present a result for this performance loss.

Lemma 29. *For any given prescription strategy $\boldsymbol{\psi}^2$,*

$$\lim_{n \to \infty} |\mathcal{J}^1(\boldsymbol{g}^{*1}) - \mathcal{J}^1(\hat{\boldsymbol{g}}^{*1})| = 0. \tag{3.37}$$

Proof. The proof follows directly from [204, Theorem 3]. □

Lemma 29 establishes the asymptotic convergence of the optimal performance by using the approximate information state towards the exact person-by-person optimal performance. Furthermore, it implies that for any desired upper bound on loss $\alpha_0 \in \mathbb{R}_{\geq 0}$, there exists a number $n \in \mathbb{N}$ and set $\mathcal{Q}_n$, such that $|\mathcal{J}^1(\boldsymbol{g}^{*1}) - \mathcal{J}^1(\hat{\boldsymbol{g}}^{*1})| < \alpha_0$. An explicit relationship between the upper bound α_0 and the upper bound on total variation distance, ϵ can be obtained using Theorem 9 and Proposition 46 of [100]. This is given by recursively defining

$$\alpha_t = 2(\epsilon \cdot ||c||_\infty + 3\epsilon \cdot ||\hat{J}^1_{t+1}||_\infty + 3\epsilon \cdot \hat{J}^1_L + \alpha_{t+1}), \tag{3.38}$$

where $||\hat{J}^1_{t+1}||_\infty := \sup_{\hat{\pi}^1_t, a^2_t} \hat{J}^1_{t+1}(\hat{\pi}^1_{t+1}, a^2_{t+1})$ and $\hat{J}^1_L$ is a finite upper bound on the Lipschitz constant of $\hat{J}^1_t$ for all $t = 0, \dots, T$. Note that an upper bound on the value of

$\hat{J}_t$ exists for all $t = 0, \ldots, T$ because cost is upper bounded. Furthermore, the Lipschitz continuity of $\hat{J}_t^1$ arises naturally from the fact that it is piece-wise linear and concave with respect to $\hat{\pi}_t^1$ for all $t = 0, \ldots, T$ [205].

The maximum loss in person-by-person performance from using an approximate information state in $\mathcal{Q}_n$ is $||\alpha_0||_\infty := \sup_{\boldsymbol{\psi}^2} \alpha_0$. Furthermore, we define an approximate information state for agent 2 as $\hat{\Pi}_t^2 := \mathbb{P}^{\boldsymbol{\psi}}(X_t^1, L_t^2, \hat{\Pi}_t^1 \mid M_t^2, \Gamma_{0:t-1}^{1:2})$. In a manner similar to Lemma 23, we can show that at each $t = 0, \ldots, T-1$, there exists a function $\hat{f}_t^2$ such that $\hat{\Pi}_{t+1}^2 = \hat{f}_t^2(\hat{\Pi}_t^2, \Gamma_t^{1:2}, Z_{t+1}^2)$. Thus, using the same sequence of arguments as Theorem 16, we conclude that if we restrict our attention to control strategies with the structural form $U_t^1 = \hat{g}_t^1(\hat{\Pi}_t^1, \hat{\Pi}_t^2)$, and $U_t^2 = \hat{g}_t^2(L_t^2, \hat{\Pi}_t^2)$ for all $t = 0, \ldots, T$, the maximum loss in optimal performance in Problem 4, $|\mathcal{J}(\boldsymbol{g}^*) - \mathcal{J}(\hat{\boldsymbol{g}}^{*1}, \hat{\boldsymbol{g}}^{*1})|$, is also $||\alpha_0||_\infty$.

Remark 30. In this approximation technique, the set of feasible values of $\hat{\Pi}_t^1$, $\mathcal{Q}_n$, is finite and does not grow in size with time. Thus, $\hat{\Pi}_t^2$ is a simple probability distribution with a finite support, which, in turn, simplifies the implementation of our DP. However, it is still challenging to compute globally optimal prescription strategies for moderate and large values of the parameter $n \in \mathbb{N}$ because the number of possible prescriptions of agent 1, $|\mathcal{U}_t^1|^{|\mathcal{Q}_n|}$, grows exponentially with n. Instead, this approach may be utilized when only person-by-person optimal strategies are required.

3.2 Decentralized Worst-Case Control with Nested Subsystems

3.2.1 Notation and Preliminaries

We utilize the mathematical framework for *uncertain variables* from [163, 192] which was introduced for non-stochastic information theory. An uncertain variable is a non-stochastic analogue of a random variable with set-valued uncertainty. For a sample space Ω and a set $\mathcal{X}$, an uncertain variable is a mapping $X : \Omega \to \mathcal{X}$. For any $\omega \in \Omega$, it has the realization $X(\omega) = x \in \mathcal{X}$. The *marginal range* of X is the set $[[X]] := \{X(\omega) \mid \omega \in \Omega\}$. For two uncertain variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, their *joint range* is $[[X, Y]] := \{\big(X(\omega), Y(\omega)\big) \mid \omega \in \Omega\}$. For a given realization y of Y, the *conditional range* of X is $[[X|y]] := \{X(\omega) \mid Y(\omega) = y,\ \omega \in \Omega\}$ and, generally, $[[X|Y]] := \{[[X|y]] \mid y \in [[Y]]\}$.

3.2.2 Problem Formulation

We consider a decentralized system with a set of subsystems $\mathcal{N} = \{1, \dots, N\}$, $N \in \mathbb{N}$. Each subsystem $n \in \mathcal{N}$ contains $K^n \in \mathbb{N}$ agents in a set $\mathcal{K}^n = \{1, \dots, K^n\}$, and each agent acts across $T \in \mathbb{N}$ discrete time steps. For each $t = 0, \dots, T$, the state of the system is denoted by an uncertain variable X_t which takes values in a finite set $\mathcal{X}_t$ and the control action of each agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, is denoted by an uncertain variable $U_t^{k,n}$ which takes values in a finite set $\mathcal{U}_t^{k,n}$. Recall that uncertain variables are non-stochastic analogues of random variables with known feasible sets but unknown distributions [163]. We denote all actions in a subsystem $n \in \mathcal{N}$ collectively by $U_t^n := (U_t^{k,n} : k \in \mathcal{K}^n)$ and all actions in the system by $U_t^{1:N} := (U_t^1, \dots, U_t^N)$. Starting at X_0, the state of the system evolves as

$$X_{t+1} = f_t\left(X_t, U_t^{1:N}, W_t\right), \quad t = 0, \dots, T-1, \tag{3.39}$$

where W_t is the disturbance of the system at time t which takes values in a finite set $\mathcal{W}_t$. At each $t = 0, \dots, T$, each agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, has an observation $Y_t^{k,n} := h_t^{k,n}(X_t, V_t^{k,n})$, which takes values in a finite set $\mathcal{Y}_t^{k,n}$. Here, $V_t^{k,n}$ is an observation noise which takes values in a finite set $\mathcal{V}_t^{k,n}$. We denote all noises in subsystem $n \in \mathbb{N}$ collectively by $V_t^n := (V_t^{k,n} : k \in \mathcal{K}^n)$ with a feasible set $\mathcal{V}_t^n := \prod_{k \in \mathcal{K}^n} \mathcal{V}_t^{k,n}$. The disturbances $\{W_t : t = 0, \dots, T\}$, measurement noises $\left\{V_t^{k,n} : k \in \mathcal{K}^n, n \in \mathcal{N}, t = 0, \dots, T\right\}$, and initial state X_0 are collectively called the *primitive variables.* Each primitive variable is independent from all other primitive variables. This ensures that the system is Markovian in a non-stochastic sense [49, 163]. Next, we describe the information structure of the system.

Definition 21. The *memory* of agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, at each t, is a set $M_t^{k,n} \subseteq \{Y_{0:t-1}^{i,m}, U_{0:t-1}^{i,m} : i \in \mathcal{K}^m, m \geq n\}$, which takes values in a finite collection of sets $\mathcal{M}_t^{k,n}$ and satisfies *perfect recall*, i.e., $M_t^{k,n} \subseteq M_{t+1}^{k,n}$ for all $t = 0, \dots, T-1$.

Remark 31. To be consistent with the exposition in the literature [50,142], we consider that for all $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, and $t = 0, \ldots, T$, the memory $M_t^{k,n}$ is updated before the observation $Y_t^{k,n}$ is realized.

In an information structure with nested subsystems, we partition the memory $M_t^{k,n}$ of each agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, into two components, C_t^n and $L_t^{k,n}$, with the following properties:

1) The common information of all agents in a subsystem $n \in \mathcal{N}$ at each $t = 0, \ldots, T$ is the information available to all agents in $\mathcal{K}^n$. We define it as the set $C_t^n := \bigcap_{k \in \mathcal{K}^n} M_t^{k,n}$ which takes values in a finite collection of sets $\mathcal{C}_t^n$, and satisfies the following properties: (1) *nestedness*, i.e., $C_t^m \subseteq C_t^n$ for all $m \in \mathcal{N}$ with $m > n$ and $t = 0, \ldots, T$, and (2) *perfect recall*, i.e., $C_t^n \subseteq C_{t+1}^n$ for all t. The property of nestedness implies that the common information of a subsystem is also available to all preceding subsystems.

2) The private information of an agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, at each $t = 0, \ldots, T$ is the information available only to agent k. We define it as the set $L_t^{k,n} := M_t^{k,n} \backslash C_t^n$ which takes values in a finite collection of sets $\mathcal{L}_t^{k,n}$. For any pair of subsystems $n, m \in \mathcal{N}$ with $m > n$, we impose the condition that $L_t^{k,m} \cap C_t^n = \emptyset$ for all $k \in \mathcal{K}^m$, i.e., no element in the private information of an agent can be available to *all* agents of a preceding subsystem.

For each $n \in \mathcal{N}$, we also define the *new information* added to C_t^n at each $t = 0, \ldots, T$ as a set $Z_t^n := C_t^n \setminus C_{t-1}^n$, which takes values in a finite collection of sets $\mathcal{Z}_t^n$, where $C_{-1}^n := \emptyset$. The new information is also nested, i.e., $Z_t^m \subseteq Z_t^n$, for all $m, n \in \mathcal{N}$, $m > n$, and t. Thus, the new information available to any subsystem is a subset of the new information available to any preceding subsystem at each t. As an illustration, consider a system with three agents divided into two subsystems in Fig. 3.1. Subsystem 1 contains two agents and subsystem 2 contains one agent. The red dashed arrows indicate the communication between the agents at time t. The one-directional, dashed arrow indicates sharing of common information from subsystem 2 to all agents in subsystem 1 at time t, which ensures that $C_t^2 \subseteq C_t^1$.

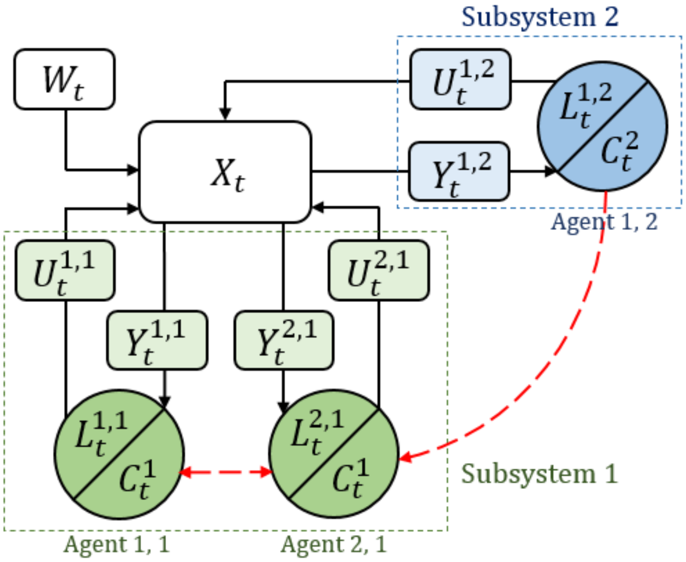


Figure 3.1: A system with three agents in two nested subsystems at time t

After updating their memory and realizing their observations at each $t = 0, \dots, T$, each agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, selects an action

$$U_t^{k,n} = g_t^{k,n}(Y_t^{k,n}, M_t^{k,n}) = g_t^{k,n}(Y_t^{k,n}, L_t^{k,n}, C_t^n), \tag{3.40}$$

where $M_t^{k,n} = \{L_t^{k,n}, C_t^n\}$, and $g_t^{k,n} : \mathcal{Y}_t^{k,n} \times \mathcal{M}_t^{k,n} \to \mathcal{U}_t^{k,n}$ is the *control law* of agent k in subsystem n at time t. The *control strategy* of subsystem $n \in \mathcal{N}$ is $\boldsymbol{g}^n := (g_{0:T}^{k,n} : k \in \mathcal{K}^n)$ and the control strategy of the system is $\boldsymbol{g} := \big(\boldsymbol{g}^1, \dots, \boldsymbol{g}^N\big)$. The set of all feasible strategy profiles is $\mathcal{G}$. All agents collectively incur a shared terminal cost $d_T(X_T) \in \mathbb{R}_{\geq 0}$ at the time horizon T. The system performance is measured by the worst-case terminal cost

$$\mathcal{J}(\boldsymbol{g}) := \max_{X_0, W_{0:T}, V_{0:T}^{1:N}} d_T(X_T). \tag{3.41}$$

Problem 7. The optimization problem is $\min_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the feasible sets $\{\mathcal{X}_0, \mathcal{W}_t, \mathcal{V}_t^n \mid n \in \mathcal{N}, t = 0, \dots, T\}$, the cost function d_T, and the system dynamics $\{f_t, h_t^{k,n} \mid k \in \mathcal{K}^n, n \in \mathcal{N}, t = 0, \dots, T\}$.

Our aim is to develop a DP which can derive an optimal strategy profile $\boldsymbol{g}^* \in \mathcal{G}$ for Problem 7. Note that an optimal strategy is guaranteed to exist because all variables take values in finite sets.

Remark 32. Variations of many existing information structures are special cases of our model. For example, a system with partial history sharing [45, 50], emerges from our model when all agents belong to a single subsystem, i.e., we set $N = 1$. Similarly, a system with one-directional delayed communication among a chain of $K \in \mathbb{N}$ agents is the special case where we have one agent per subsystem, i.e., $N = K$ and $|\mathcal{K}^n| = 1$ for all $n \in \mathcal{N}$. Simpler cases of instantaneous one-directional communication among just two agents were considered in [154, 174] for stochastic control.

3.2.3 Equivalent System with Nested Information

In this section, we construct an equivalent system with a simpler information structure, where each subsystem acts as the decision maker. For example, consider two distinct agents $k, i \in \mathcal{K}^n$ in a subsystem $n \in \mathcal{N}$. In general, given a strategy $\boldsymbol{g}$, agent k cannot generate the action $U_t^{i,n}$ of agent i since they cannot access the private information $L_t^{i,n}$. However, all agents in subsystem $n \in \mathcal{N}$ can access the common information C_t^n. This motivates us to consider a two-stage mechanism for action generation: (1) each subsystem n generates a *partial action* for each agent in $\mathcal{K}^n$ using only the common information C_t^n, and then (2) each agent $k \in \mathcal{K}^n$ uses the partial action along with their private information $L_t^{k,n}$ to generate their action U_t^k.

Definition 22. A *partial action* for agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, at each $t = 0, \dots, T$ is a mapping $\hat{U}_t^{k,n} : \mathcal{Y}_t^{k,n} \times \mathcal{L}_t^{k,n} \to \mathcal{U}_t^{k,n}$ where $\hat{\mathcal{U}}_t^{k,n}$ is a finite set.

At each $t = 0, \dots, T$, a subsystem $n \in \mathcal{N}$ generates a partial action for an agent $k \in \mathcal{K}^n$ using a partial control law $\hat{g}_t^{k,n} : \mathcal{C}_t^n \to \hat{\mathcal{U}}_t^{k,n}$, which yields $\hat{U}_t^{k,n} = \hat{g}_t^{k,n}(C_t^n)$. We define the partial control law for subsystem n at time t as $\hat{g}_t^n := (g_t^{k,n} : k \in \mathcal{K}^n)$ and the partial control strategy for the system as $\hat{\boldsymbol{g}} := (\hat{g}_{0:T}^1, \dots, \hat{g}_{0:T}^N)$. To simplify the notation, we also define $\hat{U}_t^n := \big(\hat{U}_t^{k,n} : k \in \mathcal{K}^n\big)$ and compactly write $\hat{U}_t^n = \hat{g}_t^n(C_t^n)$ for each $t = 0, \dots, T$. The partial action $\hat{U}_t^n$ is generated using only the common information in subsystem n and thus is available to all agents in $\mathcal{K}^n$. Furthermore, given a partial action $\hat{U}_t^n$, each agent $k \in \mathcal{K}^n$ must generate their action as $U_t^{k,n} = \hat{U}_t^{k,n}(Y_t^{k,n}, L_t^{k,n})$.

Next, we construct a new state which is controlled only using the partial actions of all subsystems. For each $n \in \mathcal{N}$, let $L_t^n := \{Y_t^{k,n}, L_t^{k,n} : k \in \mathcal{K}^n\}$. Then, the new state is

$$\hat{X}_t := \left\{X_t, L_t^1, \dots, L_t^N\right\}, \quad t = 0, \dots, T, \tag{3.42}$$

and it takes values in a finite collection of sets $\hat{\mathcal{X}}_t$.

Lemma 30. *For each $t = 0, \dots, T-1$, we can construct a state evolution function $\hat{f}_t(\cdot)$ and an observation rule $\hat{h}_t^n(\cdot)$, $n \in \mathcal{N}$, such that $\hat{X}_{t+1} = \hat{f}_t(\hat{X}_t, \hat{U}_t^{1:N}, W_t, V_{t+1}^{1:N})$ and $Z_{t+1}^n = \hat{h}_t^n(\hat{X}_t, \hat{U}_t^{1:N})$, respectively. In addition, we can construct a terminal cost function $\hat{d}_T(\cdot)$ which yields a shared cost $\hat{d}_T(\hat{X}_T) = d_T(X_T)$.*

Proof. We first show how $\hat{f}_t(\cdot)$ can be constructed at each $t = 0, \dots, T$ by establishing that each component in $\hat{X}_{t+1} = \left\{X_{t+1}, L_{t+1}^1, \dots, L_{t+1}^N\right\}$ can be written in terms of the variables in the RHS. Note that $X_{t+1} = f_t(X_t, U_t^{1:N}, W_t)$, where $U_t^{k,n} = \hat{U}_t^{k,n}(Y_t^{k,n}, L_t^{k,n})$ for all $k \in \mathcal{K}^n$, $n \in \mathcal{N}$. For each $n \in \mathcal{N}$, $L_{t+1}^n \subseteq \{L_t^m, Y_{t+1}^{k,m}, U_t^{k,m} \mid k \in \mathcal{K}^m,\ m \geq n\}$ and for each $k \in \mathcal{K}^n$, $Y_{t+1}^{k,n} = h_{t+1}^{k,n}(X_{t+1}, V_{t+1}^{k,n})$. Using these equations, each term in $\hat{X}_{t+1}$ can be written using the variables in the RHS. The proof is complete by defining an appropriate function $\hat{f}_t(\cdot)$. The other functions can also be constructed in a similar manner. □

Lemma 30 yields a "new" decentralized system with a shared state $\hat{X}_t$, action $\hat{U}_t^n$, and observation Z_t^n for each subsystem $n \in \mathcal{N}$ and $t = 0, \dots, T$. The cost is $\hat{d}_T(\hat{X}_T)$ and the performance criterion is $\hat{\mathcal{J}}(\hat{\boldsymbol{g}}) := \max_{X_0, W_{0:T}, V_{0:T}^{1:N}} \hat{d}_T(\hat{X}_T)$. This new decentralized system has a nested information structure [174] since at each t, a partial action $\hat{U}_t^n$ is generated using only $C_t^n = Z_{0:t}^n$, while $Z_t^m \subseteq Z_t^n$ for all $n, m \in \mathcal{N}$ with $m > n$. Next, we define a new optimization problem and show that it is equivalent to Problem 7.

Problem 8. The new problem is $\min_{\hat{\boldsymbol{g}} \in \hat{\mathcal{G}}} \hat{\mathcal{J}}(\hat{\boldsymbol{g}})$, given the sets $\{\mathcal{X}_0, \mathcal{W}_t, \mathcal{V}_t^n | n \in \mathcal{N}, t = 0, \dots, T\}$, the cost function $\hat{d}_T$, and the dynamics $\{\hat{f}_t, \hat{h}_t^n | k \in \mathcal{K}^n, n \in \mathcal{N}, t = 0, \dots, T\}$.

Lemma 31. *For any given control strategy $\boldsymbol{g}$, consider a partial control strategy $\hat{\boldsymbol{g}}$*

$$\hat{g}_t^{k,n}(C_t^n)(\cdot,\cdot) := g_t^{k,n}(\cdot,\cdot,C_t^n),\ t=0,\dots,T, \tag{3.43}$$

for all $k \in \mathcal{K}^n$, $n \in \mathcal{N}$. Then, $\mathcal{J}(\boldsymbol{g}) = \hat{\mathcal{J}}(\hat{\boldsymbol{g}})$. Moreover, for any given partial control strategy $\hat{\boldsymbol{g}}$, we can construct a control strategy $\boldsymbol{g}$ such that

$$g_t^{k,n}(\cdot,\cdot,C_t^n) := \hat{g}_t^{k,n}(C_t^n)(\cdot,\cdot),\ t=0,\dots,T, \tag{3.44}$$

for all $k \in \mathcal{K}^n$, $n \in \mathcal{N}$. Then, $\hat{\mathcal{J}}(\hat{\boldsymbol{g}}) = \mathcal{J}(\boldsymbol{g})$.

Proof. For the first part, given a control strategy $\boldsymbol{g}$ and the partial control strategy $\hat{\boldsymbol{g}}$, the action at each t for any agent $k \in \mathcal{K}^n$, $n \in \mathcal{N}$, is $U_t^{k,n} = g_t^{k,n}(Y_t^{k,n}, L_t^{k,n}, C_t^n) = \hat{g}_t^{k,n}(C_t^n)(Y_t^{k,n}, L_t^{k,n})$. Thus, the control law and partial control law result in the same action $U_t^{k,n}$ for any given $\{Y_t^{k,n}, L_t^{k,n}, C_t^n\}$. Subsequently, for any given realizations of all primitive variables $\{X_0, W_{0:T}, V_{0:T}^{1:3}\}$, the two strategies yield the same terminal state X_T and cost $d_T(X_T)$. The proof of the second part follows using arguments similar to the first part. □

Lemma 31 implies that given an optimal partial strategy $\hat{\boldsymbol{g}}^*$ for Problem 8, a control strategy $\boldsymbol{g}^*$ constructed by (3.44) is an optimal solution to Problem 7, with performance $\mathcal{J}(\boldsymbol{g}^*)$. Similarly, an optimal strategy for Problem 7 yields an optimal solution to Problem 8. Thus, the two problems are equivalent.

Remark 33. It is easier to analyze the equivalent system with a nested information structure than the original system. In the next section, we use this new information structure to derive results for optimal partial strategies, and then use Lemma 31 to characterize the optimal control strategies.

3.2.4 Main Results

In this section, we derive a structural form for optimal partial strategies in Problem 8. To simplify the exposition, we carry out the detailed derivation for a system with only 3 subsystems. This illustrates all the key arguments required to prove the results for different cases. Later, in subsection IV-D, we present the results for N subsystems.

3.2.4.1 Analysis for Subsystem 1

In this subsection, we analyze the optimal partial strategy of subsystem 1 in a system with 3 subsystems, i.e., $\mathcal{N} = \{1, 2, 3\}$. We use the person-by-person approach and thus, arbitrarily fix the partial control strategies $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$, such that $\hat{U}_t^n = \hat{g}_t^n(C_t^n)$, for all $n = 2, 3$ and $t = 0, \dots, T$, and set up a centralized problem for subsystem 1. Recall that $C_t^n \subseteq C_t^1$, and thus, any agent in subsystem 1 can derive the partial action $\hat{U}_t^n$ for all $n = 2, 3$ and $t = 0, \dots, T$. Next, we construct a centralized problem for subsystem 1 with a new state $S_t^1 := \left\{\hat{X}_t, C_t^2\right\}$ for each $t = 0, \dots, T$, which takes values in a finite collection of sets $\mathcal{S}_t^1$.

Lemma 32. *For any given partial strategies $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$, we can construct a state evolution function $\bar{f}_t^1(\cdot)$ and an observation rule $\bar{h}_t^1(\cdot)$, such that $S_{t+1}^1 = \bar{f}_t^1(S_t^1, \hat{U}_t^1, W_t, V_{t+1}^{1:3})$, and $Z_{t+1}^1 = \bar{h}_t^1(S_t^1, \hat{U}_t^1)$, respectively, for all $t = 0, \dots, T-1$. In addition, we can construct a terminal cost function $\bar{d}_T^1(\cdot)$, which yields a shared cost $\bar{d}_T^1(S_T^1) = \hat{d}_T(\hat{X}_T)$.*

Proof. We first show how $\bar{f}_t^1(\cdot)$ can be constructed at each $t = 0, \dots, T$ by establishing that each component in $S_{t+1}^1 = \left\{\hat{X}_{t+1}, C_{t+1}^2\right\}$ can be written in terms of the variables in the RHS. Note that $\hat{X}_{t+1} = \hat{f}_t(\hat{X}_t, \hat{U}_t^1, \hat{g}_t^2(C_t^2), \hat{g}_t^3(C_t^3),\ W_t, V_{t+1}^{1:3})$, where $\hat{g}_t^2(\cdot)$ and $\hat{g}_t^3(\cdot)$ are given and $C_t^3 \subseteq C_t^2$. Furthermore, $C_{t+1}^2 = C_t^2 \cup Z_{t+1}^2$, where $Z_{t+1}^2 = \hat{h}_t^2(\hat{X}_t, \hat{U}_t^1, \hat{g}_t^2(C_t^2), \hat{g}_t^3(C_t^3))$. Thus, given $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$, each term in S_{t+1}^1 can be written as a function of the variables in the argument of the RHS. The proof is complete by defining an appropriate function $\bar{f}_t^1(\cdot)$. The other functions can be constructed similarly by rewriting the LHS as a function of the variables in the argument in the RHS. □

Given the partial control strategies $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$, Lemma 32 yields a new *centralized* system with state S_t^1, action $\hat{U}_t^1$, and observation Z_t^1 for all $t = 0, \dots, T$. The terminal cost incurred by the system is $\bar{d}_T^1(S_T^1)$ and the worst-case performance criterion is $\mathcal{J}^1(\hat{\boldsymbol{g}}^1) := \max_{X_0, W_{0:T}, V_{0:T}^{1:3}} \bar{d}_T^1(S_T^1)$.

Problem 9. The problem for subsystem 1 is $\min_{\hat{\boldsymbol{g}}^1} \mathcal{J}^1(\hat{\boldsymbol{g}}^1)$, given the control partial strategies $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$, the feasible sets $\{\mathcal{X}_0, \mathcal{W}_t, \mathcal{V}_t^n : n = 1, 2, 3,\ t = 0, \dots, T\}$, the cost function $\bar{d}_T^1$, and the dynamics $\{\bar{f}_t^1, \bar{h}_t^1 : t = 0, \dots, T\}$.

Remark 34. Using Lemma 32, we construct a cost $\bar{d}_T^1(S_T^1)$ for the centralized problem such that $\bar{d}_T^1(S_T^1) = \hat{d}_T(\hat{X}_T)$ for all realizations of the primitive variables $\{X_0, W_{0:T}, V_{0:T}^{1:3}\}$. Thus, the performance $\mathcal{J}^1(\hat{\boldsymbol{g}}^1)$ for any given $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$ is equal to $\mathcal{J}(\hat{\boldsymbol{g}}^1, \hat{\boldsymbol{g}}^2, \hat{\boldsymbol{g}}^3)$.

Note that Problem 9 is a partially observed centralized problem with subsystem 1 as the sole decision maker. Specifically, at each $t = 0, \dots, T$, the component C_t^2 of the state S_t^1 is completely observed by subsystem 1, whereas, the component $\hat{X}_t$ must be estimated using the common information C_t^1 and history of partial actions $\hat{U}_{0:t-1}^{1:3}$. For partially observed centralized problems, e.g., Problem 9, a DP was provided in [49] and its complexity was reduced using a *sufficiently informative function.* We simply call this an *information state* to be consistent with its stochastic counterpart. In minimax problems, an information state at time t is a set of feasible values for the partially observed state which are compatible with the information available to a subsystem at time t. Since subsystem 1 completely observes C_t^2, we need to define an information state only for the component $\hat{X}_t$. For all $t = 0, \dots, T$, given the realizations c_t^1 and $\hat{u}_{0:t-1}^{1:3}$ of C_t^1 and $\hat{U}_{0:t-1}^{1:3}$, respectively, the realized information state of subsystem 1 is the set

$$\begin{aligned} P_t^1 := \Big\{ \hat{x}_t \in \hat{\mathcal{X}}_t \mid \exists \big(x_0 \in \mathcal{X}_0, w_{0:t-1} \in \prod_{\ell=0}^{t-1} \mathcal{W}_\ell, v_{0:t}^n \in \prod_{\ell=0}^{t} \mathcal{V}_\ell^n,\ n = 1, 2, 3 \big) \\ \text{such that } s_{\ell+1}^1 = \bar{f}_\ell^1\big(s_\ell^1, \hat{u}_\ell^1, w_\ell, v_{\ell+1}^{1:3}\big), z_{\ell+1}^1 = \bar{h}_\ell^1\big(s_\ell^1, \hat{u}_\ell^1\big),\ \ell = 0, \dots, t-1 \Big\}. \end{aligned} \tag{3.45}$$

The information state each t is a function of uncertain variables and thus, in general, we denote it with the set-valued uncertain variable Π_t^1 which takes values in a finite collection of feasible sets $\mathcal{P}_t^1 \subseteq 2^{\hat{\mathcal{X}}_t}$, where $2^{\hat{\mathcal{X}}_t}$ denotes the power set of $\hat{\mathcal{X}}_t$. Next, we show that the information state Π_t^1 does not depend on the choice of partial strategy $\hat{\boldsymbol{g}}$.

Lemma 33. *At each $t = 0, \ldots, T-1$, there exists a function $\tilde{f}_t^1(\cdot)$ independent of $\hat{\boldsymbol{g}}$, such that*

$$\Pi_{t+1}^1 = \tilde{f}_t^1(\Pi_t^1, \hat{U}_t^{1:3}, Z_{t+1}^1). \tag{3.46}$$

Proof. The proof follows the same arguments as the ones of Lemma 34 in Section IV-B. □

Lemma 33 establishes the Markovian and strategy independent evolution of Π_t^1 at each t. Thus, we state the following result for Problem 9 using the centralized DP in [49].

Theorem 17. *For any given partial strategies $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$ of subsystems 2 and 3, respectively, in Problem 9, without loss of optimality, we can restrict attention to partial strategies $\hat{\boldsymbol{g}}^{*1}$ with the structural form*

$$\hat{U}_t^1 = \hat{g}_t^{*1}\big(\Pi_t^1, C_t^2\big), \quad t = 0, \ldots, T. \tag{3.47}$$

Proof. This result follows from standard arguments for centralized minimax control problems in [49]. □

Theorem 17 establishes a structural form for an optimal partial strategy $\hat{\boldsymbol{g}}^{*1}$ of subsystem 1 in Problem 3, which holds for all $\hat{\boldsymbol{g}}^2$ and $\hat{\boldsymbol{g}}^3$. From Remark 34, we note that any globally optimal partial strategy $(\hat{\boldsymbol{g}}^{*1}, \hat{\boldsymbol{g}}^{*2}, \hat{\boldsymbol{g}}^{*3})$ for Problem 8, must necessarily be a person-by-person optimal solution of Problem 9, i.e., after fixing $\hat{\boldsymbol{g}}^{*2}$ and $\hat{\boldsymbol{g}}^{*3}$ for subsystems 2 and 3, respectively, the strategy $\hat{\boldsymbol{g}}^{*1}$ is optimal for Problem 9. Thus, there exists a globally optimal partial strategy $(\hat{\boldsymbol{g}}^{*1}, \hat{\boldsymbol{g}}^{*2}, \hat{\boldsymbol{g}}^{*3})$ where $\hat{\boldsymbol{g}}^{*1}$ takes the structural form in (3.47).

3.2.4.2 Analysis for Subsystem 2

In this subsection, we restrict attention to partial control strategies $\hat{\boldsymbol{g}}^1$ of agent 1 which satisfy (3.47), and derive a structural form for the partial strategy of subsystem 2. Note that given $\hat{\boldsymbol{g}}^1$, agents in subsystem 2 cannot generate $\hat{U}_t^1$ at any t because they cannot access Π_t^1. Thus, before using the person-by-person approach for subsystem 2, we consider that $\hat{U}_t^1$ is generated in two stages for all $t = 0, \dots, T$: (1) subsystem 2 generates a *prescription* for subsystem 1 using only C_t^2, and then (2) subsystem 1 uses the prescription and its information state Π_t^1 to generate its partial action $\hat{U}_t^1$.

Definition 23. A *prescription* by subsystem 2 for subsystem 1 at any $t = 0, \dots, T$ is a mapping $\Gamma_t^{[2,1]} : \mathcal{P}_t^1 \to \mathcal{U}_t^1$ which takes values in a finite set $\mathcal{F}_t^{[2,1]}$.

At each $t = 0, \dots, T$, subsystem 2 generates a prescription for subsystem 1 using a prescription law $\psi_t^{[2,1]} : \mathcal{C}_t^2 \to \mathcal{F}_t^{[2,1]}$, which yields $\Gamma_t^{[2,1]} = \psi_t^{[2,1]}(C_t^2)$. We define the prescription strategy for subsystem 2 as $\boldsymbol{\psi}^2 := (\psi_t^{[2,1]} : t = 0, \dots, T)$. Since the prescription $\Gamma_t^{[2,1]}$ is generated only using $C_t^2 \subseteq C_t^1$, it is available to both subsystems 1 and 2. Then, subsystem 1 must generate its partial action as $\hat{U}_t^1 = \Gamma_t^{[2,1]}\big(\Pi_t^1\big)$. Note that in this formulation, the prescription $\Gamma_t^{[2,1]}$ is generated by subsystem 2 and subsystem 1 simply utilizes the given prescription to obtain $\hat{U}_t^1$. Thus, we can consider that at each $t = 0, \dots, T$, subsystem 2 selects a *complete action* $\Theta_t^2 := (\Gamma_t^{[2,1]}, U_t^2)$. This motivates us to simultaneously derive a structural form for the optimal prescription and partial strategies of subsystem 2. To this end, we use the person-by-person approach by fixing the partial control strategy $\hat{\boldsymbol{g}}^3$ of subsystem 3 and noting that subsystem 2 can derive $\hat{U}_t^3$ for all t because $C_t^3 \subseteq C_t^2$. Next, we construct a centralized problem for subsystem 2 with a state $S_t^2 := \{\hat{X}_t, \Pi_t^1, C_t^3\}$ at each t, which takes values in a finite collection of sets $\mathcal{S}_t^2$. Using the same arguments as Lemma 32, we can construct a state evolution function $\bar{f}_t^2(\cdot)$ such that $S_{t+1}^2 = \bar{f}_t^2(S_t^2, \Theta_t^2, W_t, V_{t+1}^{1:3})$, and an observation rule $\bar{h}_t^2(\cdot)$ which yields $Z_{t+1}^2 = \bar{h}_t^2(S_t^2, \Theta_t^2)$ to obtain a centralized problem with state S_t^2, observation Z_t^2 and complete action $\Theta_t^2 = (\Gamma_t^{[2,1]}, U_t^2)$ for all t. Furthermore, we can construct a cost function $\bar{d}_T^2(\cdot)$ which yields the terminal cost $\bar{d}_T^2(S_T^2) = \hat{d}_T(\hat{X}_T)$ and a corresponding performance criterion $\mathcal{J}^2(\boldsymbol{\psi}^2, \hat{\boldsymbol{g}}^2) := \max_{X_0, W_{0:T}, V_{0:T}^{1:3}} \bar{d}_T^2(S_T^2)$.

Problem 10. The problem for subsystem 2 is $\inf_{\boldsymbol{\psi}^2, \hat{\boldsymbol{g}}^2} \mathcal{J}^2(\boldsymbol{\psi}^2, \hat{\boldsymbol{g}}^2)$, given a partial control strategy $\hat{\boldsymbol{g}}^3$, the feasible sets $\{\mathcal{X}_0, \mathcal{W}_t, \mathcal{V}_t^n : n = 1, 2, 3, t = 0, \dots, T\}$, the cost function $\bar{d}_T^2$, and the dynamics $\{\bar{f}_t^2, \bar{h}_t^2 : t = 0, \dots, T\}$.

Remark 35. Consider a fixed partial control strategy $\hat{\boldsymbol{g}}^3$ for agent 3. Using the same sequence of arguments as Lemma 31, for each prescription strategy $\boldsymbol{\psi}^2$, we can construct a partial control strategy $\hat{\boldsymbol{g}}^1$ such that the performance criterion $\mathcal{J}(\hat{\boldsymbol{g}}^1, \hat{\boldsymbol{g}}^2, \hat{\boldsymbol{g}}^3)$ for Problem 8 is equal to $\mathcal{J}^2(\boldsymbol{\psi}^2, \hat{\boldsymbol{g}}^2)$. Similarly, for each partial control strategy $\hat{\boldsymbol{g}}^1$ we can construct a prescription strategy $\boldsymbol{\psi}^2$ which yields the same performance. Thus, it is equivalent to select either a partial control strategy $\hat{\boldsymbol{g}}^1$ or a prescription strategy $\boldsymbol{\psi}^2$.

Problem 10 is a partially observed centralized problem with subsystem 2 as the sole decision maker. Specifically, at each $t = 0, \dots, T$, the component C_t^3 in state S_t^2 is perfectly observed by subsystem 2, whereas $\hat{X}_t$ and Π_t^1 must be estimated using an information state like (3.45). Given the realizations c_t^2, $\gamma_{0:t-1}^{[2,1]}$, and $\hat{u}_{0:t-1}^{2:3}$ of C_t^2, $\Gamma_{0:t-1}^{[2,1]}$, and $\hat{U}_{0:t-1}^{2:3}$, respectively, the realized information state of subsystem 2 is

$$P_t^2 := \Big\{\hat{x}_t \in \hat{\mathcal{X}}_t, P_t^1 \in \mathcal{P}_t^1 \mid \exists \big(x_0 \in \mathcal{X}_0, w_{0:t-1} \in \prod_{\ell=0}^{t-1} \mathcal{W}_\ell, v_{0:t}^n \in \prod_{\ell=0}^{t} \mathcal{V}_\ell^n,$$
$$n = 1, 2, 3\big) \text{ s.t. } s_{\ell+1}^2 = \bar{f}_\ell^2\big(s_\ell^2, \theta_\ell^2, w_\ell, v_{\ell+1}^{1:3}\big), z_{\ell+1}^2 = \bar{h}_\ell^2\big(s_\ell^2, \theta_\ell^2\big), \ell = 0, \dots, t-1\Big\}, \quad (3.48)$$

Thus, the information state of subsystem 2 is a set-valued uncertain variable Π_t^2 which takes values in a finite collection of feasible sets $\mathcal{P}_t^2 \subseteq 2^{\hat{\mathcal{X}}_t} \times 2^{\mathcal{P}_t^1}$. Next, we show that the information state Π_t^2 is independent of the choice of $\hat{\boldsymbol{g}}$.

Lemma 34. *At each $t = 0, \dots, T-1$, there exists a function $\tilde{f}_t^2(\cdot)$ independent from $\hat{\boldsymbol{g}}$, such that*

$$\Pi_{t+1}^2 = \tilde{f}_t^2(\Pi_t^2, \Gamma_t^{[2,1]}, \hat{U}_t^{2:3}, Z_{t+1}^2). \quad (3.49)$$

Proof. Let $P_t^2 \in \mathcal{P}_t^2$ be a given set-valued realization of the information state Π_t^2 at time t. Then, for the realizations $\gamma_t^{[2,1]}$, $\hat{u}_t^{2:3}$ and z_{t+1}^2 of $\Gamma_t^{[2,1]}$, $\hat{U}_t^{2:3}$ and Z_{t+1}^2, subsystem

2 can eliminate some possible elements in P_t^2 at the start of time $t+1$. Specifically, we define an interim set

$$Q_t^2 := \big\{(\hat{x}_t, P_t^1) \in P_t^2 \mid z_{t+1}^2 = \hat{h}_t^2(\hat{x}_t^2, \hat{u}_t^{1:3}), \hat{u}_t^1 = \gamma_t^{[2,1]}(\pi_t^1)\big\}, \tag{3.50}$$

which is completely determined given the set P_t^2 and the realizations $\{\gamma_t^{[2,1]}, \hat{u}_t^{2:3}, z_{t+1}^2\}$. Next, we derive the realization P_{t+1}^2 of the information state Π_{t+1}^2 using set Q_t^2 as

$$\begin{aligned} P_{t+1}^2 = \Big\{ &\hat{x}_{t+1} \in \hat{\mathcal{X}}_{t+1}, P_{t+1}^1 \in \mathcal{P}_{t+1}^1 \mid \hat{x}_{t+1} = \hat{f}_t\big(\hat{x}_t, \hat{u}_t^{1:3}, w_t, v_{t+1}^{1:3}\big), \\ &P_{t+1}^1 = \tilde{f}_t^1(P_t^1, \hat{u}_t^{1:3}, z_{t+1}^1), z_{t+1}^1 = \hat{h}_t^1(\hat{x}_t, \hat{u}_t^{1:3}),\ \hat{u}_t^1 = \gamma_t^{[2,1]}(P_t^1) \\ &\text{for all } (\hat{x}_t, P_t^1) \in Q_t^2, w_t \in \mathcal{W}_t, v_{t+1}^n \in \mathcal{V}_{t+1}^n, n = 1,2,3\Big\}. \end{aligned} \tag{3.51}$$

Thus, we can define an appropriate function $\tilde{f}_t^2(\cdot)$ using (3.51) such that that $P_{t+1}^2 = \tilde{f}_t^2(P_t^2, \gamma_t^{[2,1]}, \hat{u}_t^{2:3}, z_{t+1}^2)$. □

Lemma 34 establishes that the evolution of Π_t^2 is Markovian. Thus, we can state the following result for Problem 10 using the standard centralized DP in [49].

Theorem 18. *For a given partial strategy $\hat{\boldsymbol{g}}^3$ of subsystem 3 in Problem 10, without loss of optimality, we can restrict attention to prescription strategies $\boldsymbol{\psi}^{*2}$ and partial strategies $\hat{\boldsymbol{g}}^{*2}$ with the structural form*

$$\Gamma_t^{[2,1]} = \psi_t^{*[2,1]}\big(\Pi_t^2, C_t^3\big), \tag{3.52}$$

$$\hat{U}_t^2 = \hat{g}_t^{*2}\big(\Pi_t^2, C_t^3\big), \quad t = 0, \dots, T. \tag{3.53}$$

Proof. This result follows from standard arguments for centralized minimax control problems [49]. □

Using an optimal prescription strategy which satisfies (3.52), we construct an optimal partial strategy $\hat{\boldsymbol{g}}^{*1}$ for Problem 8 such as $\hat{g}_t^{*1}(\cdot, \Pi_t^2, C_t^3) := \psi_t^{*[2,1]}(\Pi_t^2, C_t^3)(\cdot)$ and recall from Remark 35 that they yield the same performance. Thus, we can derive an optimal partial strategy $\hat{\boldsymbol{g}}^{*1}$ in the form

$$\hat{U}_t^1 = \hat{g}_t^{*1}(\Pi_t^{1:2}, C_t^3), \quad t = 0, \dots, T. \tag{3.54}$$

3.2.4.3 Analysis for Subsystem 3

In this subsection, we restrict attention to partial strategies $\hat{\boldsymbol{g}}^1$ and $\hat{\boldsymbol{g}}^2$ of the form in (3.54) and (3.53), respectively, and derive a structural form for the partial strategy of subsystem 3. Given $\hat{\boldsymbol{g}}^1$ and $\hat{\boldsymbol{g}}^2$, subsystem 3 cannot generate the control actions $\hat{U}_t^{1:2}$ for any $t = 0, \dots, T$, because it cannot access the information states Π_t^1 and Π_t^2. Instead, subsystem 3 generates a prescription for both subsystems 1 and 2 in a manner similar to Section III-B. Specifically, for each t, a prescription of subsystem 3 for subsystem 1 is a mapping $\Gamma_t^{[3,1]} : \mathcal{P}_t^1 \times \mathcal{P}_t^2 \to \hat{\mathcal{U}}_t^1$ which takes values in a finite set $\mathcal{F}_t^{[3,1]}$ and that for subsystem 2 is a mapping $\Gamma_t^{[3,2]} : \mathcal{P}_t^2 \to \hat{\mathcal{U}}_t^2$ which takes values in a finite set $\mathcal{F}_t^{[3,2]}$. At each $t = 0, \dots, T$, subsystem 3 generates a prescription for each $n = 1, 2$ using a prescription law $\psi_t^{[3,n]} : \mathcal{C}_t^3 \to \mathcal{F}_t^{[3,n]}$, which yields $\Gamma_t^{[3,n]} = \psi_t^{[3,n]}(C_t^3)$. We define the prescription strategy of subsystem 3 as $\boldsymbol{\psi}^3 := (\psi_t^{[3,n]} : n = 1, 2,\ t = 0, \dots, T)$. In this formulation, after receiving the prescriptions of subsystem 3, subsystem 1 generates its partial action as $\hat{U}_t^1 = \Gamma_t^{[3,1]}(\Pi_t^{1:2})$ and subsystem 2 generates its partial action as $\hat{U}_t^2 = \Gamma_t^{[3,2]}(\Pi_t^2)$. Thus, subsystem 3 is the sole decision maker with a complete action $\Theta_t^3 := (\Gamma_t^{[3,1]}, \Gamma_t^{[3,2]}, \hat{U}_t^3)$. Next, we construct a new state for subsystem 3 as $S_t^3 := \{\hat{X}_t, \Pi_t^1, \Pi_t^2\}$ for all t, which takes values in a finite collection of sets $\mathcal{S}_t^3$. Using the same arguments as Lemma 32, we can construct a state evolution function $\bar{f}_t^3(\cdot)$ such that $S_{t+1}^3 = \bar{f}_t^3(S_t^3, \Theta_t^3, W_t, V_{t+1}^{1:3})$, and an observation rule $\bar{h}_t^3(\cdot)$ which yields $Z_{t+1}^3 = \bar{h}_t^3(S_t^3, \Theta_t^3)$ to obtain a centralized problem with state S_t^3, observation Z_t^3 and complete action Θ_t^3 for all t. Furthermore, we can construct a cost function $\bar{d}_T^3(\cdot)$ which yields a terminal cost $\bar{d}_T^3(S_T^3) := \hat{d}_T(\hat{X}_T)$ and a performance criterion $\mathcal{J}^3(\boldsymbol{\psi}^3, \hat{\boldsymbol{g}}^3) := \max_{X_0, W_{0:T}, V_{0:T}^{1:3}} \bar{d}_T^3(S_T^3)$.

Problem 11. The problem for subsystem 3 is $\inf_{\boldsymbol{\psi}^3, \hat{\boldsymbol{g}}^3} \mathcal{J}^3(\boldsymbol{\psi}^3, \hat{\boldsymbol{g}}^3)$, given the feasible sets $\big\{\mathcal{X}_0, \mathcal{W}_t, \mathcal{V}_t^n : n = 1, 2, 3,\ t = 0, \dots, T\big\}$, the cost function $\bar{d}_T^3$ and the dynamics $\big\{\bar{f}_t^3, \bar{h}_t^3 : t = 0, \dots, T\big\}$.

Using the same arguments as in Lemma 31 and Remark 35, we conclude that we can generate partial actions $\hat{U}_t^n$, for $n = 1, 2$ and $t = 0, \dots, T$, equivalently using either a prescription strategy $\boldsymbol{\psi}^3$ or an appropriate partial control strategy $\hat{\boldsymbol{g}}^n$. In Problem 11, at each t, subsystem 3 must estimate the entire state S_t^3. Thus, for the realizations c_t^3 and $\theta_{0:t-1}^3$ of C_t^3 and $\Theta_{0:t-1}^3$, respectively, the realized information state for subsystem 3 is

$$P_t^3 := \Big\{\hat{x}_t \in \hat{\mathcal{X}}_t, P_t^1 \in \mathcal{P}_t^1, P_t^2 \in \mathcal{P}_t^2 \mid \exists \big(x_0 \in \mathcal{X}_0, w_{0:t-1} \in \prod_{\ell=0}^{t-1} \mathcal{W}_\ell, v_{0:t}^n \in \prod_{\ell=0}^{t} \mathcal{V}_\ell^n,$$
$$n = 1,2,3\big) \text{ s.t. } s_{\ell+1}^3 = \bar{f}_\ell^3\big(s_\ell^3, \theta_\ell^3, w_\ell, v_{\ell+1}^{1:3}\big), z_{\ell+1}^3 = \bar{h}_\ell^3\big(s_\ell^3, \theta_\ell^3, \big), \ell = 0, \dots, t-1\Big\}. \quad (3.55)$$

We define the information state of subsystem 3 as a set-valued uncertain variable which takes values in a finite collection of feasible sets $\mathcal{P}_t^3 \subseteq 2^{\hat{\mathcal{X}}_t} \times 2^{\mathcal{P}_t^1} \times 2^{\mathcal{P}_t^2}$. Next, we show that Π_t^2 is independent of $\hat{\boldsymbol{g}}$.

Lemma 35. *At each $t = 0, \dots, T-1$, there exists a function $\tilde{f}_t^3(\cdot)$ independent from $\hat{\boldsymbol{g}}$, such that*

$$\Pi_{t+1}^3 = \tilde{f}_t^3(\Pi_t^3, \Gamma_t^{[3,1]}, \Gamma_t^{[3,2]}, \hat{U}_t^3, Z_{t+1}^3). \quad (3.56)$$

Proof. The proof is the same as the proof of Lemma 34. Due to space constraints, it is omitted. □

Lemma 3.56 establishes that the evolution of the Π_t^3 is Markovian. Thus, we can use the standard centralized DP decomposition from [49] to state the following result.

Theorem 19. *In Problem 11, without loss of optimality, we can restrict attention to prescription strategies $\boldsymbol{\psi}^{*3}$ and partial strategies $\hat{\boldsymbol{g}}^{*3}$ with the structural form*

$$\Gamma_t^{[3,n]} = \psi_t^{*[3,n]}\big(\Pi_t^3\big), \quad n = 1, 2, \quad (3.57)$$

$$\hat{U}_t^3 = \hat{g}_t^{*3}\big(\Pi_t^3\big), \quad t = 0, \dots, T. \quad (3.58)$$

Proof. This proof follows similar arguments for centralized minimax control problems [49]. Due to space constraints, it is omitted. □

As a result of Theorem 19, we conclude that we can derive an optimal partial strategy $\hat{\boldsymbol{g}}^*$ for Problem 8 with the form

$$\hat{U}_t^n = \hat{g}_t^{*n}(\Pi_t^{n:3}), \quad n = 1, 2, 3, \quad t = 0, \dots, T. \tag{3.59}$$

3.2.4.4 Result for N Subsystems

The results equivalent to Theorems 17 - 19 for $N \in \mathbb{N}$ subsystems can be proven using arguments similar to those for 3 subsystems and mathematical induction. The steps for the mathematical induction are detailed in Appendix A. Here, due to space constraints, we simply state the main result for Problem 8 and Problem 7.

Theorem 20. *In Problem 8, without loss of optimality, we can restrict attention to partial strategies $\hat{\boldsymbol{g}}^*$ with the form*

$$\hat{U}_t^n = \hat{g}_t^{*n}(\Pi_t^{n:N}), \quad t = 0, \dots, T. \tag{3.60}$$

Furthermore, in Problem 7, there exists an optimal control strategy $\boldsymbol{g}^{*k,n}$ for each $k \in \mathcal{K}^n$ and $n \in \mathcal{N}$, with the structural form

$$U_t^{k,n} = g_t^{*k,n}(Y_t^{k,n}, L_t^{k,n}, \Pi_t^{n:N}), \quad t = 0, \dots, T. \tag{3.61}$$

Proof. The proof follows from the arguments detailed in Appendix A. □

Remark 36. The structural form of optimal control strategies in (3.61) cannot be obtained by a direct application of the common information approach [50]. Note that the domains of our optimal strategies do not grow in size with time when the feasible sets $\mathcal{X}_t$ and $\mathcal{L}_t^{k,n}$ do not grow in size with time for all $k \in \mathcal{K}^n$ and $n \in \mathcal{N}$. This improves the computational tractability of our approach for larger values of horizon T. In contrast, the common information approach considers C_t^1 as a part of the private information of all agents in $\mathcal{K}^1$ and C_t^1 grows in size with time. Thus, the domains of optimal strategies increase in size with time for $N \geq 2$.

3.2.4.5 Dynamic Programming Decomposition

In this subsection, we construct a DP decomposition for the optimal prescription and partial strategy of subsystem N. Let $\theta_t^N = (\gamma_t^{[N,1]}, \dots, \gamma_t^{[N,N-1]}, \hat{u}_t^N)$ be the realization of Θ_t^N for all $t = 0, \dots, T$. Then, for each possible information state $P_T^N \in \mathcal{P}_T^N$, we define a value function at time T as $V_T(P_T^N) := \max_{s_T^N \in P_T^N} \bar{d}_T^N(s_T^N)$. Furthermore, for each possible information state $P_t^N \in \mathcal{P}_t^N$, for all $t = 0, \dots, T-1$, we iteratively define the value functions

$$V_t(P_t^N) := \min_{\theta_t^N} \max_{z_{t+1}^N \in \tilde{\mathcal{Z}}_{t+1}^N(P_t^N, \theta_t^N)} V_{t+1}[\tilde{f}_t^N(P_t^N, \theta_t^N, z_{t+1}^N)], \tag{3.62}$$

where the set of feasible values of z_{t+1}^N is given by

$$\tilde{\mathcal{Z}}_{t+1}^N(P_t^N, \theta_t^N) := \Big\{ z_{t+1}^N \in \mathcal{Z}_{t+1}^N \mid z_{t+1}^N = \bar{h}_t^N(s_t, \theta_t^N, w_t, v_{t+1}^{1:N}) \\ \text{for all } s_t^N \in P_t^N, w_t \in \mathcal{W}_t, v_{t+1}^n \in \mathcal{V}_{t+1}^n, n \in \mathcal{N} \Big\}. \tag{3.63}$$

The prescription law at time t for each $n < N$ is $\gamma_t^{*[N,n]} = \psi_t^{*[N,n]}(P_t^N)$ and the partial control law for subsystem N is $\hat{u}_t^{*N} = \hat{g}_t^{*N}(P_t^N)$, i.e., they are the arg inf in the RHS of (3.62). The corresponding prescription strategy $\boldsymbol{\psi}^{*N}$ and partial strategy $\hat{\boldsymbol{g}}^{*N}$ can be shown to be optimal for $\mathcal{J}^N(\boldsymbol{\psi}^N, \hat{\boldsymbol{g}}^N)$ using standard arguments [49, 50].

Remark 37. After solving the DP, we can construct optimal control strategies for Problem 7 in a manner similar to Lemma 31.

Remark 38. When applying our results to an arbitrary decentralized system, there may be multiple feasible ways to allocate the agents to subsystems. For example, consider a three agent system with one-directional communication from agent 3 to agent 2 and from agent 2 to agent 1. We can either allocate all agents to one subsystem or allocate each agent to a unique subsystem. Both of these are valid options because they both ensure that the common information of each subsystem is nested within the common information of all preceding subsystems. However, Theorem 20 leads to a different DP for different allocations. Thus, a system designer must decide on an allocation before deriving the optimal control strategies. In general, our DP performs better when more agents have private information which does not grow in size with time. Thus, in the three agent system with one-directional communication, allocating each agent to a unique subsystem is better than allocating all agents to one subsystem.

3.2.4.6 Extension to Additive Cost

Consider a variation of our problem where the system incurs a cost $d_t(X_t, U_t^{1:N})$ at each $t = 0, \dots, T$ and the performance is measured by the worst-case total cost

$$\Xi(\boldsymbol{g}) := \max_{X_0, W_{0:T}, V_{0:T}^{1:N}} \sum_{t=0}^{T-1} d_t(X_t, U_t^{1:N}) + d_T(X_T). \tag{3.64}$$

We can transform the additive cost in (3.64) into a terminal cost using a technique from [49, 50]. At each $t = 0, \dots, T$, we define an uncertain variable $A_t := \sum_{\ell=0}^{t-1} d_\ell(X_\ell, U_\ell^{1:N}) \in \mathcal{A}_t$ which tracks the cost incurred by the system up to time t. Note that $A_0 = 0$. Then, at each t, we consider an augmented state for the system, $\{X_t, A_t\}$ and note that the augmented state at time $t+1$, $\{X_{t+1}, A_{t+1}\}$, evolves as a function of $\{X_t, A_t\}$, the control actions $U_t^{1:N}$ and the disturbance W_t. Then the total cost is simply given by the terminal cost $A_T + d_T(X_T)$ and thus, we can apply our results to this equivalent terminal cost problem. This extension to total cost problems does present a challenge because the information state at time t takes values in a subset of the power set $2^{\mathcal{X}_t \times \mathcal{A}_t}$, which may grow in size with time as $\mathcal{A}_t$ generally grows in size with time. We plan to address this in the future by exploring alternative augmented states.

3.2.5 Numerical Example

In this subsection, we validate our results with a simple example. We consider two agents who seek to surround a target. The agents and target can each move along a linear grid with $\Lambda \in \mathbb{N}$ points. Starting at X_0^0, the position of the target is updated at each $t = 0, \dots, T$ as

$$X_{t+1}^0 = \begin{cases} X_t^0 + W_t^0, & \text{if } 1 \leq X_t^0 + W_t^0 \leq \Lambda, \\ \Lambda, & \text{if } X_t^0 + W_t^0 \geq \Lambda, \\ 1, & \text{if } X_t^0 + W_t^0 \leq 1, \end{cases} \tag{3.65}$$

with a disturbance $W_t^0 \in \{-1,0,1\}$. Each agent occupies a separate subsystem, and thus, we refer to an agent simply by its subsystem $n=1,2$. At each t, each agent n selects an action $U_t^n \in \{-1,0,1\}$, and updates its position as

$$X_{t+1}^n = \begin{cases} X_t^n + U_t^n, & \text{if } 1 \leq X_t^n + U_t^n \leq \Lambda, \\ \Lambda, & \text{if } X_t^n + U_t^n \geq \Lambda, \\ 1, & \text{if } X_t^n + U_t^n \leq 1. \end{cases} \tag{3.66}$$

The positions of both agents are observed perfectly by both the agents. Additionally, agent 2 perfectly observes the position of the target, but can only communicate this to agent 1 with a delay of 1 time step. Agent 1 also receives a noisy observation of the target's position $Y_t^1 = \max\{1, \min\{X_t^0 + V_t^1, \Lambda\}\}$, with a noise $V_t^1 \in \{-1,0\}$. Agent 1 has faulty equipment and cannot transmit their observations to agent 2. Thus, for all $t=0,\dots,T$, the information structure is given by $C_t^1 = \{X_{0:t-1}^0, Y_{0:t-1}^1, U_{0:t-1}^{1:2}, X_{0:t}^{1:2}\}$ and $L_t^1 = \{Y_t^1\}$ for agent 1, and $C_t^2 = \{X_{0:t-1}^0, U_{0:t-1}^2, X_{0:t}^{1:2}\}$ and $L_t^2 = \{X_t^2\}$ for agent 2. At the onset, both agents receive a common observation Y_0 implying that the target's starting position is in the set $\{Y_0 - 1, Y_0, Y_0 + 1\}$. The terminal cost is

$$d_T(X_T^{0:3}) = \begin{cases} \sum_{n=1}^2 |X_T^0 - X_T^n|, & \text{if } I^s = 1, \\ D + \sum_{n=1}^2 |X_T^0 - X_T^n|, & \text{if } I^s = 0, \end{cases} \tag{3.67}$$

where $I^s \in \{0,1\}$ indicates if the agents have successfully surrounded the target, $\sum_{n=1}^2 |X_T^0 - X_T^n|$ penalizes the distance of the agents from the target, and $D > 0$ is a penalty for failing to surround the target. We summarize in Table 3.1 the optimal worst-case performance achieved for different realizations of the initial conditions $\{x_0^1, x_0^2, y_0\}$, with and without using information states in the DP. We note that their values are the same, which validates our results.

3.2.6 Appendix A - N Subsystems

In this subsection, we describe the steps to prove our results using mathematical induction for $N \in \mathbb{N}$ subsystems. The analysis for subsystem 1 is the same in the presence of N subsystems as in Section 3.2.4.1. As in (3.45), we can define the information

Table 3.1: Optimal cost for $\Lambda = 8$, $T = 3$, and $D = 10$.

$\{x_0^1, x_0^2, y_0\}$	With Information States	Without Information States
$\{8, 8, 2\}$	18	18
$\{3, 6, 7\}$	4	4
$\{3, 3, 4\}$	14	14
$\{3, 5, 8\}$	4	4

state for subsystem 1 at any t as the set-valued uncertain variable Π_t^1 which contains feasible values of $\hat{X}_t \in \hat{\mathcal{X}}_t^1$ given $\{C_t^1, \hat{U}_{0:t-1}^{1:N}\}$, and takes values in $\mathcal{P}_t^1 \subseteq 2^{\hat{\mathcal{X}}_t}$. Using the same arguments as Theorem 17, in Problem 8, without loss of optimality, we can restrict attention to partial strategies $\hat{\boldsymbol{g}}^{*1}$ with the form

$$\hat{U}_t^1 = \hat{g}_t^{*1}(\Pi_t^1, C_t^2), \quad t = 0, \dots, T. \tag{3.68}$$

This forms the basis of our mathematical induction. Given (3.68), we can iteratively prove the results for each subsystem $n \in \{2, \dots, N\}$. For any $n \in \mathcal{N}$, we consider the induction hypotheses for all $m \in \mathcal{N}$ where $m < n$:

Hypothesis 1) The information state is well defined for all t as an appropriate set-valued uncertain variable Π_t^m which takes values in a space $\mathcal{P}_t^m \subseteq 2^{\hat{\mathcal{X}}_t} \times 2^{\mathcal{P}_t^1} \times \dots \times 2^{\mathcal{P}_t^{m-1}}$.

Hypothesis 2) Without loss of optimality in Problem 8, we can restrict attention to partial strategies $\hat{\boldsymbol{g}}^*$ with the form $\hat{U}_t^m = \hat{g}_t^{*m}(\Pi_t^{m:n-1}, C_t^n)$ for all $m < n$ and $t = 0, \dots, T$.

Remark 39. In hypothesis 2, we consider that a structural form using information states has already been derived for each subsystem $m < n$. Note that the form in (3.68) for $\hat{\boldsymbol{g}}^{*1}$ is consistent with this for $n = 2$. We later show that the structural form of strategies derived for each subsystem $n \in \mathcal{N}$ is also consistent with hypothesis 2.

Given the two hypotheses, the proof by mathematical induction follows from the following steps:

Step 1) For all $t = 0, \dots, T$, we define a prescription by subsystem n for all $m < n$ as a mapping $\Gamma_t^{[n,m]} : \mathcal{P}_t^m \times \dots \times \mathcal{P}_t^{n-1} \to \hat{\mathcal{U}}_t^m$ which takes values in a finite set $\mathcal{F}_t^{[n,m]}$.

Each prescription $\Gamma_t^{[n,m]}$ is generated using a prescription law $\psi_t^{[n,m]} : \mathcal{C}_t^n \to \mathcal{F}_t^{[n,m]}$ which yields $\Gamma_t^{[n,m]} = \psi_t^{[n,m]}(C_t^n)$. We define the prescription strategy of subsystem n as $\boldsymbol{\psi}^n := (\psi_t^{[n,m]} : m < n, t = 0, \ldots, T)$. Furthermore, we define the complete action of subsystem n as $\Theta_t^n := (\Gamma_t^{[n,1]}, \ldots, \Gamma_t^{[n,n-1]}, \hat{U}_t^n)$ for all $t = 0, \ldots, T$.

Step 2) We fix the partial control strategies $\hat{\boldsymbol{g}}^{n+1}, \ldots, \hat{\boldsymbol{g}}^N$ of all subsystems $n+1, \ldots, N$, and use the person-by-person approach to derive a structural form for $\hat{\boldsymbol{g}}^{*n}$. We construct a state for subsystem n as $S_t^n := \{\hat{X}_t, \Pi_t^1, \ldots, \Pi_t^{n-1}, C_t^{n+1}\}$, which takes values in a finite collection of sets $\mathcal{S}_t^n$. Note that $C_t^{N+1} := \emptyset$. As in Lemma 32, we can also construct a state evolution function $\bar{f}_t^n(\cdot)$ such that $S_{t+1}^n = \bar{f}_t^n(S_t^n, \Theta_t^n, W_t, V_{t+1}^{1:N})$, and an observation rule $\bar{h}_t^n(\cdot)$ which yields $Z_{t+1}^n = \bar{h}_t^n(S_t^n, \Theta_t^n)$ to obtain a centralized problem with state S_t^n, observation Z_t^n and complete action Θ_t^n for all t. Furthermore, we can construct a cost function $\bar{d}_T^n(\cdot)$ which yields a terminal cost $\bar{d}_T^n(S_T^n) := \hat{d}_T(\hat{X}_T)$ and a performance criterion $\mathcal{J}^n(\boldsymbol{\psi}^n, \hat{\boldsymbol{g}}^n) = \max_{X_0, W_{0:T}, V_{0:T}^{1:N}} \bar{d}_T^n(S_T^n)$.

Step 3) Using the same arguments as Lemma 31, we can prove that for each prescription strategy $\boldsymbol{\psi}^n$, we can construct partial strategies $(\hat{\boldsymbol{g}}^1, \ldots, \hat{\boldsymbol{g}}^{n-1})$ that lead to the same partial actions $(\hat{U}_t^1, \ldots, \hat{U}_t^{n-1})$ for all t, and vice versa. Furthermore, this construction ensures that after fixing strategies $\hat{\boldsymbol{g}}^{n+1}, \ldots, \hat{\boldsymbol{g}}^N$, it holds that $\mathcal{J}^n(\boldsymbol{\psi}^n, \hat{\boldsymbol{g}}^n) = \mathcal{J}(\hat{\boldsymbol{g}})$.

Step 4) In the constructed centralized problem of $\min_{\boldsymbol{\psi}^n, \hat{\boldsymbol{g}}^n} \mathcal{J}^n(\boldsymbol{\psi}^n, \hat{\boldsymbol{g}}^n)$, the unobserved components in S_t^n need to be estimated at each $t = 0, \ldots, T$ using an information state. For realizations c_t^n and $\theta_{0:t}^n$ of C_t^n and $\Theta_{0:t}^n$, respectively, the realization of the information state is

$$P_t^n := \Big\{\hat{x}_t \in \hat{\mathcal{X}}_t, P_t^1 \in \mathcal{P}_t^1, \ldots, P_t^{n-1} \in \mathcal{P}_t^{n-1} \mid \exists \big(x_0 \in \mathcal{X}_0, w_{0:t-1} \in \prod_{\ell=0}^{t-1} \mathcal{W}_\ell, v_{0:t}^n \in \prod_{\ell=0}^{t} \mathcal{V}_\ell^n, n \in \mathcal{N}\big) \text{ s.t. } s_{\ell+1}^n = \bar{f}_\ell^n\big(s_\ell^n, \theta_\ell^n, w_\ell, v_{\ell+1}^{1:N}\big), z_{\ell+1}^n = \bar{h}_\ell^n\big(s_\ell^n, \theta_\ell^n\big),\ \ell = 0, \ldots, t-1\Big\}. \tag{3.69}$$

Thus, the information state for subsystem n at time t is a set-valued uncertain variable Π_t^n which takes values in a finite set $\mathcal{P}_t^n \subseteq 2^{\hat{\mathcal{X}}_t} \times 2^{\mathcal{P}_t^1} \times \cdots \times 2^{\mathcal{P}_t^{n-1}}$. Using the same sequence of arguments as Lemma 34 we can show that at each time t, there exists a function $\tilde{f}_t^n(\cdot)$ independent of the partial control strategy $\hat{\boldsymbol{g}}$, such that

$\Pi_{t+1}^n = \tilde{f}_t^n(\Pi_t^n, \Theta_t^n, \hat{U}_t^{n+1:N}, Z_{t+1}^n)$. Furthermore, using the same arguments as Theorems 18 and 19, we conclude that without loss of optimality, we can restrict attention to prescription strategies $\boldsymbol{\psi}^{*n}$ and partial strategies $\hat{\boldsymbol{g}}^{*n}$ with the structural form

$$\Gamma_t^{[n,m]} = \psi_t^{*[n,m]}(\Pi_t^n, C_t^{n+1}),\ m < n,\ t = 0, \dots, T, \tag{3.70}$$

$$\hat{U}_t^n = g_t^{*n}(\Pi_t^n, C_t^{n+1}), \quad t = 0, \dots, T. \tag{3.71}$$

The results (3.70) and (3.71) are consistent with the induction hypotheses and complete the proof by mathematical induction. Then, the first result in Theorem 20 follows from (3.70) and (3.71) by constructing a partial control law $\hat{g}_t^{*n}(\Pi_t^{n:N}) := \psi_t^{*[N,n]}(\Pi_t^N)(\Pi_t^{n:N-1})$, for all $n \in \mathcal{N}$ and $t = 0, \dots, T$. The second result follows by constructing an appropriate control strategy $\boldsymbol{g}^*$ from $\hat{\boldsymbol{g}}^*$ using Lemma 31.

Chapter 4

MECHANISM DESIGN FOR COMPETING AGENTS WITH PRIVATE INFORMATION

4.1 Social Media and Misleading Information in a Democracy: A Mechanism Design Approach

4.1.1 Introduction

For the last few years, political commentators have been indicating that we live in a *post-truth* era [206], wherein the deluge of information available on the internet has made it extremely difficult to identify facts. As a result, individuals have developed a tendency to form their opinions based on the *believability* of presented information rather than its truthfulness [207]. This phenomenon is exacerbated by the business practices of social media platforms, which often seek to maximize the *engagement* of their users at all costs. In fact, the algorithms developed by platforms for this purpose often promote conspiracy theories among their users [208].

The sensitivity of users of social media platforms to conspiratorial ideas makes them an ideal terrain to conduct political misinformation campaigns [209, 210]. Such campaigns are especially effective tools to disrupt democratic institutions, because the functioning of stable democracies relies on *common knowledge* about the political actors and the processes they can use to gain public support [211]. The trust held by the citizens of a democracy on common knowledge includes: (1) trust that all political actors act in good faith when contesting for power, (2) trust that elections lead to a free and fair transfer of power between the political actors, and (3) trust that democratic institutions ensure that elected officials wield their power in the best interest of the citizens. In contrast, citizens of democracies often have a *contested knowledge* regarding who should hold power and how they should use it [211]. The introduction of *alternative* *facts* can reduce the trust on common knowledge about democracy, especially if they become accepted beliefs among the citizens and increase polarization of opinions [212]. Such disruptions on the trust on common knowledge can be found in the 2016 U.S. elections [213] and Brexit Campaign in 2016 [214], where the spread of misinformation through social media platforms resulted in a large number of citizens mistrusting the results of voting.

To tackle this growing phenomenon of misinformation, this chapter considers a finite group of social media platforms, whose users represent the citizens in a democracy, and a democratic government. Every post in the platforms is associated with a parameter that captures its informativeness, which can take values between two extremes: (1) completely factual and (2) complete misinformation. Posts that exhibit misinformation can lead to a decrease in trust on common knowledge among the users [185–188]. In addition, social media platforms are considered to have the technologies to *filter*, or label, posts that intend to sacrifice trust on common knowledge, but are reluctant to apply them because of potential losses in revenue. Thus, the government seeks to incentivize the social media platforms to use these technologies and filter any misinformation included in the posts.

A *misinformation filtering game* describes the interactions between the social media platforms and the government. In this game, each platform acts as strategic agent seeking to maximize their advertisement revenue from the engagement of their users [213, 215]. User engagement is a metric that can be used to quantify the interaction of users with a platform, and subsequently, how much time they spend on the platform. Recent efforts reported in the literature on misinformation in social media platforms have indicated that increasing filtering of misinformation leads to decreasing of user engagement [9]. There are many possible reasons for this phenomenon. First, filtering reduces the total number of posts propagating across the social network. Second, the users whose opinions are filtered may perceive this action as dictatorial censorship [216], and as a result, they may chose to express their opinions in other platforms. Finally, misinformation tends to elicit stronger reactions, e.g., surprise, joy, sadness, as compared to factual posts [217], which may increase user engagement. Thus, each platform is reluctant to filter misinformation.

In the misinformation filtering game, the government is also a strategic agent, whose utility increases as the trust of the users of social media platforms on common knowledge increases. Consequently, increasing filtering of misinformation by the social media platforms increases the utility of the government. Thus the government is willing to make an investment to incentivize the social media platforms to filter misinformation. In the proposed approach, mechanism design is utilized to distribute this investment among the platforms optimally, and in return, implement an optimal level of filtering.

4.1.2 Problem Formulation

We consider a democratic society with a nonempty set of social media platforms $\mathcal{I} = \{1, \ldots, I\}$, $I \in \mathbb{N}$, and a government. We refer to the social media platforms and the government collectively as the *agents*, and denote the set of all agents by $\mathcal{J} = \mathcal{I} \cup \{0\}$, where the index 0 corresponds to the government. The agents strategically take actions in a *misinformation filtering game*, as described next.

Let the informativeness of a post on platform $i \in \mathcal{I}$ be $x_i \in [0, 1]$, where $x_i = 0$ indicates that the post contains complete misinformation, and $x_i = 1$ indicates that the post is completely factual. Our hypothesis states that the emergence of posts with many falsehoods, i.e., $x_i \to 0$, decreases the trust of the users on common knowledge about democracy [186, 211]. Recall that the common knowledge refers to knowledge of the political actors and the processes to gain public support. Each platform $i \in \mathcal{I}$ has the technological means to detect and filter misinformation. In the misinformation filtering game, the action $a_i \in \mathcal{A} = [0, 1]$ of platform i represents the level of filtering imposed by i. Action a_i minimizes the spread of a post with informativeness $x_i < a_i$, while a post with $x_i \geq a_i$ is unaffected. In practice, misinformation filters can be implemented by either placing warnings on each post with $x_i < a_i$, or limiting the reach of such posts. Thus, the action a_i is a lower threshold on informativeness that is accepted by platform i. To this end, we call a_i the *filter* of platform i.

Each platform $i \in \mathcal{I}$ generates revenue by monetizing the *engagement* of their users with advertisements [215]. With an increase in filtering, there is a decrease in user engagement [9]. Users may perceive filters as censorship [216], and as a result, they may choose to express their opinions on other platforms. Consider, for example, platform $l \in \mathcal{I}$ with a filter $a_l > a_i$. Some users of l, whose posts have been marked up by the filter, may migrate to platform i and increase the engagement of i. This motivates us to define a set of *competing platforms*.

Definition 24. For each platform $i \in \mathcal{I}$, the set $\mathcal{C}_i \subset \mathcal{I}$, with $i \in \mathcal{C}_i$, is the set of *competing platforms* whose choice of filters has an impact on the engagement of platform $i \in \mathcal{I}$.

To simplify our exposition, we consider that for $i, k \in \mathcal{I}$, if $i \in \mathcal{C}_k$, then $k \in \mathcal{C}_i$. However, our mechanism can easily be extended to allow asymmetric competition among platforms.

Definition 25. The *valuation function* of a social media platform $i \in \mathcal{I}$ is $v_i\big(a_k : k \in \mathcal{C}_i\big) : \mathcal{A}^{|\mathcal{C}_i|} \to \mathbb{R}_{\geq 0}$. It is a decreasing function with respect to a_i and strictly increasing with respect to a_l for all $l \in \mathcal{C}_{-i}$, where $\mathcal{C}_{-i} = \mathcal{C}_i \setminus \{i\}$.

The valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ gives the revenue of platform i from user engagement after filtering by all platforms. A higher value of a_i will decrease the revenue of platform i. A higher value of a_l for another competing platform $l \in \mathcal{C}_{-i}$ will increase the revenue of platform i. Recall from the discussion in the previous section that filtering of misinformation in a platform increases the trust of their users on common knowledge about democracy. Thus, for each $i \in \mathcal{I}$, we define the *average trust function* on common knowledge.

Definition 26. The *average trust function* on common knowledge of the users of platform $i \in \mathcal{I}$ is $h_i(a_i) : \mathcal{A} \to [0, 1]$, and it is a strictly increasing function with respect to a_i.

The average trust function $h_i(a_i)$ captures the impact of filter a_i on the trust on common knowledge across the users of platform i. A low value of $h_i(a_i)$ implies that a_i leads to low trust on common knowledge for the users of platform i, and vice versa. In practice, platform i can measure the opinions of their users through surveys [218], and thus eventually estimate the impact of filter a_i using the average trust function $h_i(a_i)$.

Recall that, in our framework, the government is the strategic agent $0 \in \mathcal{J}$ who seeks to maximize the trust on common knowledge of the users of all social media platforms. Therefore, the government selects an action $a_0 \in \mathcal{A} = [0, 1]$ that designates a lower bound which must be satisfied by the aggregate average trust of all platforms in $\mathcal{I}$. To this end, we refer to the action a_0 as the government's lower bound on trust on common knowledge.

Let $N_i \in \mathbb{N}$ be the total number of users of the social media platform $i \in \mathcal{I}$. The fraction of the number of users of i with respect to the total number of users of all platforms is $n_i = \frac{N_i}{\sum_{l \in \mathcal{I}} N_l}$. The fraction n_i represents the contribution of users in platform i on the average trust on common knowledge. Since $\sum_{i \in \mathcal{I}} n_i = 1$, the *aggregate average trust* is $\sum_{i \in \mathcal{I}} n_i \cdot h_i(a_i)$. In our framework, the government's role is to select the lower-bound a_0 for the aggregate average trust. After the government decides on a_0, each platform $i \in \mathcal{I}$ that participates in the game must select a filter a_i that satisfies:

$$a_0 - \sum_{i \in \mathcal{I}} n_i \cdot h_i(a_i) \leq 0. \tag{4.1}$$

Next, we define the government's valuation function.

Definition 27. The *valuation function* of the government is $v_0(a_0) : [0, 1] \to \mathbb{R}_{\geq 0}$, and it is an increasing function with respect to the lower bound a_0.

The government's valuation function $v_0(a_0)$ assigns a monetary value to the lower bound a_0. Recall that the government seeks to increase the trust on common knowledge among the users of all platforms. Thus, the government's valuation increases as the lower bound on aggregate average trust increases. The government also has a fixed, finite budget $b_0 \in \mathbb{R}_{\geq 0}$, which denotes the maximum possible investment.

The private and public information structure corresponding to each agent is as follows: (i) *Public information:* The set of competing platforms $\mathcal{C}_i$, set of feasible actions $\mathcal{A}$, and fraction of users n_i of each platform $i \in \mathcal{I}$ are known to all agents in set $\mathcal{J}$. (ii) *Valuation functions:* The valuation function $v_i(\cdot)$ of each platform $i \in \mathcal{I}$ is known only to platform i. Similarly, the valuation function $v_0(\cdot)$ and the budget b_0 of the government are known only to the government. (iii) *Average trust functions:* The average trust function $h_i(\cdot)$ of platform $i \in \mathcal{I}$ is known only to platform i.

We impose the following assumptions in our game:

Assumption 7. For each platform $i \in \mathcal{I}$, $|\mathcal{C}_i| \geq 3$.

This assumption simplifies the exposition of our mechanism. Assumption 7 implies that each user frequents multiple social media platforms. For the case with $|\mathcal{C}_i| \geq 2$ and extended results see Appendix A.

Assumption 8. The valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big) : \mathcal{A}^{|\mathcal{C}_i|} \to \mathbb{R}_{\geq 0}$ of each social media platform $i \in \mathcal{I}$ is a concave and differentiable function with respect to a_k.

The concavity of $v_i\big(a_k : k \in \mathcal{C}_i\big)$ captures the diminishing marginal change in engagement due to additional filtering. The higher the value of a_i, the more users of platform i will perceive the filter as censorship. Thus, for platform i, increasing a low-value filter may lead to a smaller loss in engagement as compared to increasing a high-value filter.

Assumption 9. The average trust function $h_i(a_i) : \mathcal{A} \to [0, 1]$ of each social media platform $i \in \mathcal{I}$ is a concave and differentiable function with respect to a_i.

The concavity of $h_i(a_i)$ implies that, for large values of a_i, a small incremental change in a_i would not have a significant impact on the average trust of users. Practically, this implies low values of a_i will have a major impact on the average trust.

Assumption 10. The valuation function of the government $v_0(a_0) : [0, 1] \to \mathbb{R}_{\geq 0}$ is a concave and differentiable function with respect to the lower-bound a_0.

Practically, for high values of a_0, the government might not be interested in investing additional resources to increase a_0 even more, as the impact on average trust would not be significant. Nevertheless, we also analyze our system by relaxing Assumptions 8 - 10 in Subsection 4.1.4.1.

Assumption 11. The output of the function $h_i(a_i)$ can be monitored by any competing platform $l \in \mathcal{C}_{-i}$, and a violation of the condition (4.1) can be detected by the government.

Assumption 11 helps us enforce the mechanism, presented in Section 4.1.3, in a static environment. In the mechanism, each platform $i \in \mathcal{I}$ commits to a minimum value of their average trust function, to be achieved by choosing an appropriate a_i. Consider that a platform i selects a value a_i that fails to satisfy this commitment. The government can detect a violation of (4.1) by gauging public opinion on the internet. However, the government does not know the function $h_i(\cdot)$, and thus, would penalize each platform in $\mathcal{I}$ equally for the violation of (4.1). To avoid a penalty for the failure of platform i, a competing platform $l \in \mathcal{C}_{-i}$ can report the violation. Thus, it is reasonable to consider that each platform $i \in \mathcal{I}$ monitors the output $h_l(a_l)$ of each competing platform $l \in \mathcal{C}_{-i}$ to maximize their own utility. In future research, we could potentially relax Assumption 11 using a dynamic mechanism [183].

Assumption 12. The government ensures that any social media platform $i \in \mathcal{I}$ that does not participate in the mechanism receives no benefits from the filters of participating platforms.

In static mechanisms, the ability to exclude a agent from receiving benefits of some common resource is a necessary condition for voluntary participation of agents without any monetary investment [219]. This condition is often assumed implicitly in the literature [181]. In our mechanism, the government can make an investment up to the budget b_0. Thus, we assume *partial excludability* in Assumption 12, where a non-participating platform i still receives the maximum valuation for selecting filter

$a_i = 0$, but cannot receive benefits from the filters of any participating platforms. In practice, the government can publicize that platform i is non-cooperative in a collective endeavor to filter misinformation. The resulting loss in credibility among the users of participating platforms will minimize their migration to platform i. In future research, we can relax this assumption using a dynamic mechanism [220].

4.1.2.1 Problem Statement

To resolve the conflict of interest between the government and the platforms, the government hires a social planner to design a mechanism and impose the misinformation filtering game. The mechanism must: (i) incentivize all platforms to voluntarily participate in the game, and (ii) induce a selection of filters that maximizes the *social welfare*. The social welfare is the sum of utilities of all agents, defined next. To meet these objectives, the social planner asks each agent $i \in \mathcal{J}$ to send a message m_i from a set $\mathcal{M}_i$. Using the message profile $m = (m_0, \dots, m_{|\mathcal{I}|})$, the social planner assigns a tax $\tau_i(m) \in \mathbb{R}$ for each platform $i \in \mathcal{I}$, and an investment $\tau_0(m) \in \mathbb{R}_{\geq 0}$ for the government. The message and tax of each agent are defined in Subsection 4.1.3.2. By convention, a tax $\tau_i(m) > 0$ is a payment made by agent $i \in \mathcal{J}$, and a tax $\tau_i(m) < 0$ is a subsidy given to agent i. While the taxes of the platforms can be either payments or subsidies, the government may never collect a subsidy from any platform. Note that the social planner must not receive any profit, nor incur any losses, for designing and implementing the mechanism, i.e., the mechanism should be budget balanced with $\sum_{i\in\mathcal{J}} \tau_i(m) = 0$.

Definition 28. The *utility* of platform $i \in \mathcal{I}$ is $u_i\big(m, a_k : k \in \mathcal{C}_i\big) := v_i\big(a_k : k \in \mathcal{C}_i\big) - \tau_i(m)$, while government's utility is $u_0(m, a_0) := v_0(a_0) - \tau_0(m)$.

Then, the social welfare to be maximized by the social planner is $u_0(m, a_0) + \sum_{i\in\mathcal{I}} u_i(m, a_k : k \in \mathcal{C}_i)$.

Problem 12. The social planner's optimization problem is

$$\max_{a} \Bigg(v_0(a_0) - \tau_0(m) + \sum_{i\in\mathcal{I}} \Big(v_i\big(a_k : k \in \mathcal{C}_i\big) - \tau_i(m) \Big) \Bigg),$$

$$\text{subject to: } 0 \leq a_i \leq 1, \quad \forall i \in \mathcal{J}, \tag{4.2}$$

$$a_0 - \sum_{i \in \mathcal{I}} n_i \cdot h_i(a_i) \leq 0, \tag{4.3}$$

$$0 \leq \tau_0(m) \leq b_0, \tag{4.4}$$

$$\sum_{i \in \mathcal{J}} \tau_i(m) = 0, \tag{4.5}$$

where $a = (a_0, \dots, a_{|\mathcal{I}|})$ and $\tau(m) = (\tau_0(m), \dots, \tau_{|\mathcal{I}|}(m))$.

In Problem 12, (4.3) ensures that the aggregate average trust of all users satisfies the government's lower bound a_0, (4.4) restricts the government's investment $\tau_0(m)$ to be within the available budget, and (4.5) ensures budget balance. The objective function of Problem 12 is differentiable and concave, and the set of feasible solutions is non-empty, convex, and compact. Thus, Problem 12 has a unique solution. However, this solution cannot be computed directly by the social planner because she has no knowledge of the functional form of either the valuation function $v_i(\cdot)$ of any agent $i \in \mathcal{J}$, or the average trust function $h_i(\cdot)$ of any platform $i \in \mathcal{I}$. If the social planner simply asks the agents to report their private information, the agents may not be truthful. Thus, the social planner seeks to design the taxes $\tau_i(m)$ for each agent $i \in \mathcal{J}$ that will incentivize the agents to be truthful while also maximizing the social welfare.

Remark 40. The government has no compelling reason to misreport their budget b_0 to the social planner. Thus, we consider that the social planner has knowledge of b_0.

Remark 41. By maximizing the social welfare, the utility of each agent is maximized in Problem 12. Hence, agents have an incentive to participate in the mechanism. Note that the government can not design the mechanism because they would optimize only their own utility $u_0(m, a_0)$. Thus, the social planner is essential to design and implement our mechanism.

4.1.3 Mechanism Design Approach

In this section, we present a two-step mechanism to incentivize the filtering of misinformation among social media platforms. The aim of step one is to ensure the voluntary participation of all platforms. The aims of step two are to (i) gain truthful information from the platforms, (ii) derive the optimal investment, and (iii) maximize the social welfare.

4.1.3.1 Step One - The Participation Step

In step one of the mechanism, each social media platform $i \in \mathcal{I}$ must decide whether to participate in the mechanism. This decision is taken with complete knowledge of the rules of step two, described in the next subsection. Let platform $i \in \mathcal{I}$ choose to not participate. Platform i neither pays taxes nor receives subsidies, i.e., $\tau_i(m) = 0$, and they are free to select the filter $a_i = 0$ that maximizes $v_i(a_k : k \in \mathcal{C}_i)$. Meanwhile, another competing platform $l \in \mathcal{C}_{-i}$ may decide to participate in the mechanism and subsequently, implement a non-zero filter a_l. From Assumption 12, the government ensures that platform i receives no utility from the competing filter a_l. Thus, the utility of platform i is $v_i(a_k = 0 : k \in \mathcal{C}_i)$. We will use this utility in Theorem 24 of Section IV to establish the voluntary participation of all platforms in our mechanism.

4.1.3.2 Step Two - The Bargaining Step

In step two, the social planner asks each agent $i \in \mathcal{J}$ to broadcast a message $m_i \in \mathcal{M}_i$. For all $i \in \mathcal{I}$, let $\mathcal{D}_i = \mathcal{C}_i \cup \{0\}$, and $\mathcal{D}_{-i} = \mathcal{D}_i \setminus \{i\}$. The message of platform i is

$$m_i := (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i), \tag{4.6}$$

where $\tilde{h}_i \in \mathbb{R}_{\geq 0}$ is the minimum average trust that platform i proposes to achieve through filtering; $\tilde{p}_i := (\tilde{p}^i_l : l \in \mathcal{D}_{-i})$, $\tilde{p}_i \in \mathbb{R}^{|\mathcal{D}_{-i}|}_{\geq 0}$, is the collection of prices that platform i is willing to pay or receive per unit changes in the filters of other competing platforms (except i) and the government's lower bound; and $\tilde{a}_i = (\tilde{a}^i_k : k \in \mathcal{D}_i)$, $\tilde{a}_i \in \mathbb{R}^{|\mathcal{D}_i|}$, is the profile of filters proposed by platform i for all competing platforms (including i) and government's lower bound.

Remark 42. Note that each platform proposes a filter for themselves, denoted by $\tilde{a}_i^i$, in their message m_i. However, platform i does not propose a price for $\tilde{a}_i^i$. Thus, every platform can influence their filter, but not the associated price.

The message of the government is $m_0 := (\tilde{p}_0, \tilde{a}_0^0)$, where $\tilde{p}_0 \in \mathbb{R}_{\geq 0}$ is the price that the government is willing to pay or receive per unit change in the average trust, and $\tilde{a}_0^0 \in \mathbb{R}$ is the lower bound proposed by the government. Note that our social planner respects the privacy of each platform $i \in \mathcal{I}$ since the valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ or average trust function $h_i(a_i)$ are not requested. Similarly, the government is not forced to reveal the functional form of $v_0(a_0)$. Each agent $i \in \mathcal{J}$ is free to send any feasible value of the message m_i. Given the messages $m := (m_0, \dots, m_{|\mathcal{I}|})$, the social planner allocates the following parameters to the agents:

1) The social planner allocates a *filter* to each platform $i \in \mathcal{I}$ and a *lower bound* to the government such that the constraints of Problem 1 are satisfied. The filter allocated by the social planner to platform i is

$$\alpha_i(m) := \sum_{k \in \mathcal{C}_i} \frac{\tilde{a}_i^k}{|\mathcal{C}_i|}, \tag{4.7}$$

i.e., the average of the filters proposed by all competing platforms including i. The lower bound allocated by the social planner to the government is

$$\alpha_0(m) = \sum_{k \in \mathcal{J}} \frac{\tilde{a}_0^k}{|\mathcal{J}|}, \tag{4.8}$$

i.e., the average of the lower bounds proposed by all platforms and the government.

2) The social planner allocates to each platform $i \in \mathcal{I}$ a *minimum average trust*:

$$\eta_i(m) := \min\left\{ \frac{n_i \cdot \tilde{h}_i}{\sum_{k \in \mathcal{I}} n_k \cdot \tilde{h}_k} \cdot \alpha_0(m),\ 1 \right\}, \tag{4.9}$$

where $\eta_i(m) \in [0, 1]$, and where the social planner will not accept a message m_i that might lead to $\sum_{k \in \mathcal{I}} n_k \cdot \tilde{h}_k = 0$. The allocated minimum average trust, $\eta_i(m)$, is a lower bound on average trust that must be achieved by platform i. Let the filter implemented by platform i be a_i. Then, platform i must ensure that $n_i \cdot h_i(a_i) \geq \eta_i(m)$. Recall from the information structure that a violation of this condition cannot be detected by the social planner since she does not have explicit knowledge of the function $h_i(\cdot)$. However, by Assumption 11, the output of $h_i(a_i)$ can be monitored by any other competing platform $l \in \mathcal{C}_{-i}$. Platform l will then report any violation of $n_i \cdot h_i(a_i) \geq \eta_i(m)$ to ensure that platform i implements the largest filter a_i, and maximizes the utility $u_l(m, a_k : k \in \mathcal{C}_l)$. This prevents platforms from violating the constraint imposed by $\eta_i(m)$.

3) The social planner allocates a to each platform $i \in \mathcal{I}$ *price* $\pi_l^i \in \mathbb{R}_{\geq 0}$ corresponding to the allocated filter $\alpha_l(m)$ of every other competing platform $l \in \mathcal{C}_{-i}$, where

$$\pi_l^i := \sum_{k \in \mathcal{C}_{-l}: k \neq i} \frac{\tilde{p}_l^k}{|\mathcal{C}_l| - 2}. \tag{4.10}$$

This is the average of prices proposed for the allocated filter $\alpha_l(m)$ by all competing platforms in $\mathcal{C}_{-l}$ except i. Thus, the allocated price π_l^i is independent of the prices proposed by both platforms i and l. Similarly, the social planner allocates the price $\pi_0 \in \mathbb{R}_{\geq 0}$ to the government, where

$$\pi_0 = \sum_{i \in \mathcal{I}} \frac{p_0^i}{|\mathcal{I}|}. \tag{4.11}$$

We write the prices allocated to the agents without the argument m to simplify the notation.

4) The social planner allocates the following *tax* to each social media platform $i \in \mathcal{I}$,

$$\begin{aligned}\tau_i(m) := -\tilde{p}_0 \cdot \eta_i(m) - \sum_{l \in \mathcal{C}_{-i}} \pi_i^l \cdot \alpha_i(m) + \sum_{l \in \mathcal{C}_{-i}} \pi_l^i \cdot \alpha_l(m) \\ + \sum_{l \in \mathcal{D}_{-i}} \tilde{p}_l^i \cdot (\tilde{a}_l^i - \tilde{a}_l^{-i})^2,\end{aligned} \tag{4.12}$$

where, for all $l \in \mathcal{C}_{-i}$,

$$\tilde{a}_l^{-i} = \sum_{k \in \mathcal{C}_l: k \neq i} \frac{\tilde{a}_l^k}{|\mathcal{C}_l| - 1} \tag{4.13}$$

represents the average of the proposed filters for l by all competing platforms except $i \in \mathcal{I}$, and

$$\tilde{a}_0^{-i} = \sum_{k \in \mathcal{J}_{-i}} \frac{\tilde{a}_0^k}{|\mathcal{J}| - 1} \tag{4.14}$$

represents the average of lower bounds proposed by all agents except i. The tax $\tau_i(m)$ of platform i in (4.12) can be interpreted as follows: (i) the first term is a subsidy given by the government to platform i for the increase in average trust among the users of i; (ii) the second term is a collection of subsidies given by each competing platform $l \in \mathcal{C}_{-i}$ to platform i for the increase in valuation $v_l\big(a_k : k \in \mathcal{C}_l\big)$ due to the allocated filter α_i; (iii) the third term is a payment by platform i for the increase in valuation $v_i\big(a_k : k \in \mathcal{C}_i\big)$ due to the allocated filter α_l of each competing platform $l \in \mathcal{C}_{-i}$; and (iv) the fourth term is a collection of penalties to platform i if either the filter proposed in message m_i for any competing platform $l \in \mathcal{C}_{-i}$ is inconsistent with the filters proposed by other platforms, or if the proposed lower bound is inconsistent with that proposed by other agents. The fourth term also penalizes platform i for higher values of $\tilde{p}_l^i$, and thus, incentivizes proposing lower prices for other agents.

The social planner also allocates an investment to the government as $\tau_0(m) := \pi_0 \cdot \alpha_0(m) + (\tilde{p}_0 - \pi_0)^2$, where the first term is the total investment for the allocated lower bound $\alpha_0(m)$, and the second term is a penalty for any deviation between the proposed price $\tilde{p}_0$ and the allocated price π_0.

Remark 43. For some filter $a_i > 0$ of any platform i in (4.12), the social planner takes a payment from each competing platform $l \in \mathcal{C}_{-i}$ and allocates an equal subsidy to i. This subsidy incentivizes platform i to implement the filter a_i, and helps to fairly distribute the government's investment.

Remark 44. In the bargaining step, we have used *all* platforms in $\mathcal{I}$ when defining the allocations, for e.g., π_0. However, this does not cause any issues due to non-participating platforms because, as we prove in Theorem 24, all platforms eventually participate in the mechanism in the participation step.

Step two of the mechanism is characterized by the tuple $\langle \mathcal{M}, g(\cdot) \rangle$, where $\mathcal{M} = \mathcal{M}_0 \times \cdots \times \mathcal{M}_{|\mathcal{I}|}$ is the message space, and $g(m) : \mathcal{M} \to \mathcal{O}$ maps it to a set of outcomes $\mathcal{O} := \big\{ \big(\alpha_0(m), \dots, \alpha_{|\mathcal{I}|}(m)\big), \big(\tau_0(m), \dots,\ \tau_{|\mathcal{I}|}(m)\big) : \alpha_i(m) \in \mathcal{A},\ \tau_i(m) \in \mathbb{R},\ i \in \mathcal{J} \big\}$. The mechanism $\langle \mathcal{M}, g(\cdot) \rangle$ together with the utility functions $(u_i : i \in \mathcal{J})$ induces a game in which the social planner allocates the filters $(\alpha_1(m), \dots, \alpha_i(m))$ to the platforms and the lower bound $\alpha_0(m)$ to the government. Each platform $i \in \mathcal{I}$ that participates in the mechanism must implement the filter $a_i = \alpha_i(m)$, and the government must select the lower bound $a_0 = \alpha_0(m)$. Platform i can influence their allocated filter $\alpha_i(m)$ with their message m_i. Thus, the strategy of platform i is given by the message $m_i \in \mathcal{M}_i$, with the constraint $\alpha_i(m) \in \mathcal{S}_i(m)$, where $\mathcal{S}_i(m) = \{a_i \in \mathcal{A} : n_i \cdot h_i(a_i) \geq \eta_i(m)\}$. The set of feasible allocations $\mathcal{S}_i(m)$ for $i \in \mathcal{I}$ is a function of the messages of all agents. The government's strategy is the message $m_0 \in \mathcal{M}_0$. For such a game, we select the solution concept of the GNE [221]. Let $m_{-i} = (m_0, \dots, m_{i-1}, m_{i+1}, \dots, m_{|\mathcal{I}|})$. A message profile m^* is a GNE of the induced game, if

$$u_i\big((m_i^*, m_{-i}^*), \alpha_k(m_i^*, m_{-i}^*) : k \in \mathcal{C}_i\big) \geq u_i\big((m_i, m_{-i}^*), \alpha_k(m_i, m_{-i}^*) : k \in \mathcal{C}_i\big), \tag{4.15}$$

for all $m_i \in \mathcal{M}_i$ and $\alpha_i \in \mathcal{S}_i(m)$, for all $i \in \mathcal{I}$; and the message m_0^* of the government satisfies $u_0\big((m_0^*,\ m_{-0}^*), \alpha_0(m_0^*, m_{-0}^*)\big) \geq u_0\big((m_0, m_{-0}^*), \alpha_0(m_0,\ m_{-0}^*)\big)$, for all $m_0 \in \mathcal{M}_0$. In the rest of the paper, we denote the utility of agent $i \in \mathcal{J}$ by $u_i(m_i, m_{-i})$.

Remark 45. In general, the GNE solution concept is defined for a game with complete information. However, we adopt this solution in our induced game despite the fact that the valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ and the average trust function $h_i(a_i)$ are the private information of platform i. We resolve this discrepancy by considering that the induced game is played repeatedly over multiple iterations, and thus, the agents can iteratively converge to a GNE [179,181,183,184].

Remark 46. We have summarized our notation in Table 4.1.

Table 4.1: A summary of the key variables

Symbol	**Explanation**
m_i	The message broadcast by agent $i \in \mathcal{J}$
a_i	The filter of platform $i \in \mathcal{I}$
$\tilde{a}_k^i$	The filter proposed by platform $i \in \mathcal{I}$ for platform $k \in \mathcal{C}_i$
$\alpha_i(m)$	The filter allocated to platform $i \in \mathcal{I}$
a_0	The government's lower bound on trust
$\tilde{a}_0$	The lower bound proposed by the government
$\tilde{a}_0^i$	The lower bound proposed by $i \in \mathcal{I}$ for the government
$\alpha_0(m)$	The lower bound allocated to the government
$v_i(\cdot)$	The valuation function of agent $i \in \mathcal{J}$
$h_i(\cdot)$	The average trust function of platform $i \in \mathcal{I}$
$\tilde{h}_i$	The proposed minimum average trust of platform $i \in \mathcal{I}$
$\eta_i(m)$	The allocated minimum average trust for platform $i \in \mathcal{I}$
$\tilde{p}_l^i$	The price proposed by platform $i \in \mathcal{I}$ for agent $l \in \mathcal{D}_{-i}$
π_l^i	The price allocated to platform $i \in \mathcal{I}$ for agent $l \in \mathcal{D}_{-i}$
$\tilde{p}_0$	The price proposed by the government
π_0	The price allocated to the government
$\tau_i(m)$	The tax allocated to agent $i \in \mathcal{J}$

4.1.4 Properties of the Mechanism

In this section, we establish the properties of our mechanism. Recall that each social media platform $i \in \mathcal{I}$ is a strategic agent who seeks to maximize $u_i(m_i, m_{-i})$ through the choice of $m_i \in \mathcal{M}_i$. Thus, we can define the following optimization problem for platform $i \in \mathcal{I}$ in the induced game.

Problem 13. Platform i's optimization problem is

$$\max_{m_i \in \mathcal{M}_i} v_i\big(\alpha_k(m) : k \in \mathcal{C}_i\big) - \tau_i(m), \tag{4.16}$$

$$\text{subject to: } 0 \leq \alpha_i(m) \leq 1, \tag{4.17}$$

$$\eta_i(m) - n_i \cdot h_i\big(\alpha_i(m)\big) \leq 0, \tag{4.18}$$

where (4.16) is the utility $u_i(m_i, m_{-i})$ of platform i, (4.17) ensures the feasibility of the allocated filter $\alpha_i(m)$, and (4.18) ensures that the allocated minimum average trust is achieved.

Note that the social planner can ensure that (4.17) and (4.18) are hard constraints by imposing a tax $\tau_i(m) \to \infty$ when they are violated. Also recall that the government strategically selects message $m_0 \in \mathcal{M}_0$ to maximize their utility $u_0(m_0, m_{-0})$.

Problem 14. The government's optimization problem is

$$\max_{m_0 \in \mathcal{M}_0} v_0\big(\alpha_0(m)\big) - \tau_0(m), \tag{4.19}$$

$$\text{subject to: } 0 \leq \alpha_0(m) \leq 1, \tag{4.20}$$

$$\pi_0 \cdot \alpha_0(m) - b_0 \leq 0, \tag{4.21}$$

where the objective in (4.19) is the utility $u_0(m_0, m_{-0})$, (4.20) ensures that the government's lower bound a_0 is feasible, and (4.21) is the budgetary constraint on total investment.

Remark 47. Consider optimal solutions $m_i^* \in \mathcal{M}_i$ of Problem 13 for each platform $i \in \mathcal{I}$, and $m_0^* \in \mathcal{M}_0$ of Problem 14 for the government. Then, the profile $m^* = \big(m_0^*, \ldots, m_{|\mathcal{I}|}^*\big)$ satisfies (4.15), and thus, is a GNE of the induced game.

Next, we establish some basic properties of the mechanism in Lemmas 36 and 37 at any GNE. We prove later in Theorem 23 that a GNE for the induced game always exists. In Lemma 36, we show that the government's proposed price at any GNE is equal to the average price proposed by all platforms.

Lemma 36. *Let the message profile $m^* \in \mathcal{M}$ be a GNE of the induced game. Then, $\tilde{p}_0^* = \pi_0^*$ for the government.*

Proof. We note that (4.19) is concave with respect to $\tilde{p}_0$. At GNE, we have that $\frac{\partial u_0}{\partial \tilde{p}_0}\big|_{\tilde{p}_0^*} = 2 \cdot (\tilde{p}_0^* - \pi_0^*) = 0$, thus $\tilde{p}_0^* = \pi_{-0}^*$. □

Similarly, in Lemma 37, we show that, at any GNE, the filters proposed by all social media platforms in $\mathcal{C}_i$ for platform i are equal, and the lower bound proposed by all platforms is the same, unless the corresponding price proposal is 0.

Lemma 37. *Let the message profile $m^* \in \mathcal{M}$ be a GNE of the induced game. Then, for $\tilde{p}_k^i \neq 0$, we have $\tilde{a}_k^{i*} = \tilde{a}_k^{-i*}$ for each social media platform $i \in \mathcal{I}$, for each $k \in \mathcal{D}_{-i}$.*

Proof. The proof follows from the same sequence of arguments as Lemma 36. □

Next, we show that our proposed mechanism is budget balanced at any GNE, i.e., the social planner redistributes all the payments it collects from the agents as subsidies to them.

Theorem 21 (**Budget Balance**)**.** *Consider any GNE $m^* \in \mathcal{M}$ of the induced game. The proposed mechanism is budget balanced at GNE, i.e., $\sum_{i\in\mathcal{J}} \tau_i(m^*) = 0$.*

Proof. From Lemmas 36 and 37, the tax $\tau_i^* = \tau_i(m^*)$ for social media platform i at GNE is $\tau_i^* = -\tilde{p}_0^* \cdot \eta_i(m^*) - \sum_{l\in\mathcal{C}_{-i}} \pi_i^l \cdot \alpha_i(m^*) + \sum_{l\in\mathcal{C}_{-i}} \pi_l^i \cdot \alpha_l(m^*)$. The tax τ_0^* for the government at GNE is $\tau_0^* = \tilde{p}_0^* \cdot \alpha_0(m^*)$, where $\tilde{p}_0^*$ is the price per unit change on average trust at GNE. Since $\sum_{i\in\mathcal{I}} \eta_i(m) = \alpha_0(m)$ for all $m \in \mathcal{M}$, then at GNE we have $\sum_{i\in\mathcal{J}} \tau_i^* = \sum_{i\in\mathcal{I}}[-\sum_{l\in\mathcal{C}_{-i}} \pi_i^l \cdot \alpha_i(m^*) + \sum_{l\in\mathcal{C}_{-i}} \pi_l^i \cdot \alpha_l(m^*)] = 0$. □

In Lemma 38, we establish that every GNE of the induced game leads to an allocation of a filter profile and a lower bound such that all constraints of Problem 12 are satisfied.

Lemma 38 (**Feasibility**)**.** *Every GNE message profile $m^* \in \mathcal{M}$ leads to a filter profile $\big(\alpha_1(m^*), \dots, \alpha_{|\mathcal{I}|}(m^*)\big)$ and lower bound $\alpha_0(m^*)$, which is a feasible solution of Problem 12.*

Proof. Every GNE message profile m^* satisfies (4.17) - (4.18) and (4.20) - (4.21). From Theorem 21, $\sum_{i\in\mathcal{J}} \tau_i(m^*) = 0$. For each $i \in \mathcal{I}$, $\eta_i(m) \leq n_i \cdot h_i(\alpha_i(m))$, and $\sum_{i\in\mathcal{I}} \eta_i(m) = \alpha_0(m)$. Hence, $\sum_{i\in\mathcal{I}} h_i(\alpha_i(m)) \geq \alpha_0(m)$. □

Next, we establish that each platform $i \in \mathcal{I}$ can unilaterally deviate in the message $m_i \in \mathcal{M}_i$, to achieve any desired allocation of filter profile. This property ensures that platform i can always attain any filter $\hat{a}_i \in \mathcal{A}$ for themselves.

Lemma 39. *Given the message profile $m_{-i} \in \mathcal{M}_{-i}$, the social media platform $i \in \mathcal{I}$ can unilaterally deviate in their message $m_i \in \mathcal{M}_i$ to attain any filter $\hat{a}_k \in \mathcal{A}$ as the allocated filter $\alpha_k(m) \in \mathcal{S}_k(m)$, for all $k \in \mathcal{C}_i$.*

Proof. Let m_{-i} be the message profile of all agents in $\mathcal{J}_{-i}$. Then, platform i can propose a filter $\tilde{a}_k^i = \hat{a}_k - \sum_{l \in \mathcal{C}_k : l \neq i} \frac{\tilde{a}_k^l}{|\mathcal{C}_k|-1}$, to ensure that $\alpha_k(m) = \hat{a}_k$ for each $k \in \mathcal{C}_i$. Moreover, platform i can propose a lower bound $\tilde{a}_0^i = -\sum_{l \in \mathcal{J}_{-i}} \tilde{a}_0^l$ for the government, to ensure that $\alpha_0(m) = 0$, and subsequently, $\alpha_k(m) = \hat{a}_k \in \mathcal{S}_k(m)$ for all $k \in \mathcal{C}_i$. □

Next, we show that, at any GNE, the allocated filters for all platforms and the allocated lower bound for the government result in the optimal solution of Problem 12.

Theorem 22 (**Strong Implementation**)**.** *Consider any GNE $m^* \in \mathcal{M}$ of the induced game. The allocated filter profile $(\alpha_1(m^*), \ldots, \alpha_{|\mathcal{I}|}(m^*))$ and lower bound $\alpha_0(m^*)$ at equilibrium are equal to the optimal solution a^{*o} of Problem 12.*

Proof. Let $\alpha(m^*) = (\alpha_1(m^*), \ldots, \alpha_{|\mathcal{I}|}(m^*))$. Then, the GNE message profile m^* satisfies, for platform $i \in \mathcal{I}$, the following Kush-Kahn-Tucker (KKT) conditions for optimality: (i) $\frac{\partial v_i}{\partial \alpha_i}\big|_{\alpha(m^*)} + \sum_{l \in \mathcal{C}_{-i}} \pi_i^l - \lambda_i^i + \mu_i^i + \nu_i^i \cdot \frac{\partial h_i}{\partial \alpha_i}\big|_{\alpha(m^*)} = 0$, (ii) $\frac{\partial v_i}{\partial \alpha_l}\big|_{\alpha(m^*)} - \pi_l^i = 0$, for all $l \in \mathcal{C}_{-i}$, (iii) $\tilde{p}_0^* - \nu_i^i = 0$, (iv) $\lambda_i^i \cdot (\alpha_i(m^*) - 1) = 0$, (v) $\mu_i^i \cdot \alpha_i(m^*) = 0$, (vi) $\nu_i^i \cdot (\eta_i(m^*) - h_i(\alpha_i(m^*))) = 0$, (vii) $\lambda_i^i, \mu_i^i, \nu_i^i \geq 0$, where (i) - (iii) are the derivatives of the Lagrangian of Problem 13 for platform i with respect to $\alpha(m)$ and $\eta_i(m)$, and (iv) - (vii) are constraints on the Lagrange multipliers $(\lambda_i^i, \mu_i^i, \nu_i^i)$. Using (ii) and (iii), $\sum_{k \in \mathcal{C}_i} \frac{\partial v_k}{\partial \alpha_i}\big|_{\alpha(m^*)} - \lambda_i^i + \mu_i^i + \nu_i^i \cdot \frac{\partial h_i}{\partial \alpha_i}\big|_{\alpha(m^*)} = 0$, for all $i \in \mathcal{I}$. Similarly, we can write the KKT conditions for Problem 14 with the Lagrange multipliers $(\lambda_0^0, \mu_0^0, \omega_0^0)$. The optimal solution $a^{*o} = \left(a_0^{*o}, a_1^{*o}, \ldots, a_{|\mathcal{I}|}^{*o}\right)$ satisfies the KKT conditions of Problem 12 with the Lagrange multipliers $(\lambda_i, \mu_i, \omega, \nu : i \in \mathcal{J})$. We set $\pi_0 = \tilde{p}_0^*$, $\lambda_i = \lambda_i^i$, $\mu_i = \mu_i^i$, $\nu = \tilde{p}_0^*$, $\omega = \omega_0^0$, $a_i^{*o} = \alpha_i(m^*)$, which implies that the efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNEs. □

Next, we show that our mechanism guarantees the existence of at least one GNE for the induced game. This ensures that the results in this section are always valid for the mechanism.

Theorem 23 (**GNE existence**). *Let $a^{*o} = \left(a_0^{*o}, a_1^{*o}, \dots, a_{|\mathcal{I}|}^{*o}\right)$ be the optimal solution of Problem 12. There is a GNE message profile $m^* \in \mathcal{M}$ of the induced game that guarantees that the filter profile $\left(\alpha_1(m^*), \dots, \alpha_{|\mathcal{I}|}(m^*)\right)$ and lower bound $\alpha_0(m^*)$ at GNE satisfy $\alpha_i(m^*) = a_i^{*o}$, for all $i \in \mathcal{J}$.*

Proof. Consider that the optimal solution a^{*o} which satisfies the KKT conditions for Problem 12 with the corresponding Lagrange multipliers $(\lambda_i, \mu_i, \nu, \omega : i \in \mathcal{J})$. Taking similar steps to the proof of Theorem 22, we can show that for $\tilde{p}_0 = \pi_0 = \nu$, the Lagrange multipliers of Problems 13 and 14 are $\lambda_i^i = \lambda_i$, $\mu_i^i = \mu_i$, $\nu_i^i = \nu$, $\omega_0^0 = \omega$, $i \in \mathcal{J}$, and the allocated prices are $\pi_l^i = \frac{\partial v_i}{\partial \alpha_l}\Big|_{a^{*o}}$, for all $l \in \mathcal{C}_{-i}$. This implies that the allocated filters at GNE are $\alpha_i(m^*) = a_i^{*o}$ for each platform $i \in \mathcal{I}$, and the allocated lower bound is $\alpha_0(m^*) = a_0^{*o}$. □

Next, we consider the participation step from Subsection 4.1.3.1. The government always participates in the mechanism for the opportunity to incentivize misinformation filtering among the platforms. In the following result (Theorem 24), we invoke Assumption 12 and the properties of our mechanism to show that in step one, every social media platform voluntarily participates in the mechanism. Thus, with rational agents, the mechanism can be implemented without dictatorship.

Theorem 24 (**Individually Rational**). *Each platform $i \in \mathcal{I}$ prefers the outcome of every GNE of the induced game to the outcome of not participating in the mechanism.*

Proof. Let m^* be a GNE message profile. By Lemma 39, there exists a message $m_i \in \mathcal{M}_i$ for platform i such that $\alpha_0(m_i, m_{-i}^*) = 0$. Platform i can unilaterally deviate in their message m_i to ensure that for every $k \in \mathcal{C}_i$, the allocated filter is $\alpha_k(m_i, m_{-i}^*) = 0$. From Section 4.1.3.1, the utility of a non-participating platform $i \in \mathcal{I}$ is $v_i(0, \dots, 0)$. Consider the message $m_i = (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i)$ with $\tilde{p}_l^i = 0$, for all

$l \in \mathcal{D}_{-i}$, $\tilde{a}^i_k = -\sum_{l\in\mathcal{C}_{-i}} \tilde{a}^l_k$, for all $k \in \mathcal{C}_{-i}$, and $\tilde{a}^i_0 = -\sum_{l\in\mathcal{J}_{-i}} \tilde{a}^i_0$. Then, the allocation $\alpha_k(m_i, m^*_{-i}) = 0$ is feasible for every platform $k \in \mathcal{C}_i$. The tax for platform i is $\tau_i = 0$ and utility is $u_i(m_i, m^*_{-i}) = v_i(0, \dots, 0) - 0$. Using (4.15), $u_i(m^*) \geq u_i(m_i, m^*_{-i})$. Hence, $u_i(m^*) \geq v_i(0, \dots, 0)$. Thus, in the participation step, the weakly dominant action of every platform $i \in \mathcal{I}$ is to participate in the mechanism. □

4.1.4.1 Extension to Quasi-Concave Valuations

In this subsection, we replace Assumptions 8 - 10 with the following more general assumptions: (i) The valuation functions $v_i\big(a_k : k \in \mathcal{C}_i\big)$ and $v_0(a_0)$ of each platform $i \in \mathcal{I}$ and the government, respectively, are quasi-concave and differentiable. (ii) The average trust function $h_i(a_i)$ for all $i \in \mathcal{I}$ is differentiable. Thus, we cannot use the KKT conditions to prove the existence of an induced GNE and strong implementation. However, if a GNE exists, the proposed mechanism is still budget balanced, individually rational, and Lemmas 36 - 39 hold. In Theorem 25, we show that there still exists a GNE and it induces a Pareto efficient equilibrium in the game. Pareto efficiency is the condition where we cannot improve the utility of any agent without decreasing the utility of another agent [181]. This is weaker than Theorem 22.

Theorem 25. *Let the valuation function $v_i(a_k : k \in \mathcal{C}_i)$ be quasi-concave and differentiable for all $i \in \mathcal{J}$ in the game $\langle \mathcal{M}, g(\cdot), (u_i)_{i\in\mathcal{I}} \rangle$. Then, (i) there exists a GNE for the induced game, and (ii) every induced GNE is Pareto efficient.*

Proof. (i) Existence: By Lemma 2, $\mathcal{M}'_i := \{m_i \in \mathcal{M}_i : \tilde{p}^i_l \cdot (\tilde{a}^i_l - a^{-i}_l) = 0,\ \forall l \in \mathcal{D}_{-i}\}$ at GNE. For all $m_i \in \mathcal{M}'_i$, $u_i(m) = v_i(\alpha_k(m) : k \in \mathcal{C}_i) + \tilde{p}_0 \cdot \eta_i(m) + \sum_{l\in\mathcal{C}_{-i}} \pi^l_i \cdot \alpha_i(m) - \sum_{l\in\mathcal{C}_{-i}} \pi^i_l \cdot \alpha_l(m)$, where $\tilde{p}_0, \pi^l_i, \pi^i_l$ are independent of m_i for all $l \in \mathcal{C}_{-i}$, and $u_i(m) = u_i(\eta_i, \alpha_k : \alpha_k \in \mathcal{D}_i)$. By Lemma 4, platform i can unilaterally deviate in message $m_i \in \mathcal{M}_i$ to receive any allocation $\alpha_k(m) \in \mathcal{A}$, for all $k \in \mathcal{D}_i$. Thus, platform i's action is to select $\beta_i = (\eta_i, \alpha_k : k \in \mathcal{D}_i)$ from the set $\mathcal{B}_i = \{[0,1] \times \mathcal{A}^{|\mathcal{D}_i|} : n_i \cdot h_i(\alpha_i) - \eta_i \geq 0\}$, which is convex, compact, and independent of the message profile m_{-i}, while $\alpha_0 \in \mathcal{A}$, where $\mathcal{A}$ is compact, convex, and independent of m_{-0}, and $v_i(a_k : k \in \mathcal{C}_i)$,

$i \in \mathcal{I}$, is quasi-concave and differentiable. The utility $u_i(\beta)$, $\beta = (\beta_0, \ldots, \beta_{|\mathcal{I}|})$, is quasi-concave and differentiable with respect to $\beta_i \in \mathcal{B}_i$ for all $i \in \mathcal{I}$. Similarly, the government's utility $u_0(\alpha_0)$ is quasi-concave and differentiable with respect to α_0. It follows from Glicksberg's theorem that there exists a GNE for the induced game. *(ii) Pareto efficiency:* It is sufficient to show that the NE can be characterized by a Walrasian equilibrium, thus Pareto efficient. Consider any NE action profile $\beta^* = (\alpha_0^*, \beta_1^*, \ldots, \beta_{|\mathcal{I}|}^*) \in \mathcal{A} \times \mathcal{B}_1 \times \cdots \times \mathcal{B}_{|\mathcal{I}|}$. By the NE definition, for every platform $i \in \mathcal{I}$ it holds that $u_i(\beta^*) \geq u_i(\beta_i, \beta_{-i}^*)$, for all $\beta_i \in \mathcal{B}_i$. Then, we have $\beta_i^* = \arg\max_{\beta_i \in \mathcal{B}_i} \{v_i(\alpha_k : k \in \mathcal{C}_i) + \tilde{p}_0^* \cdot \eta_i + \sum_{l \in \mathcal{I}_{-i}} \pi_i^{*l} \cdot \alpha_i - \sum_{l \in \mathcal{I}_{-i}} \pi_l^{*i} \cdot \alpha_l\}$. Similarly, for the government, $\alpha_0^* = \arg\max_{\alpha_0 \in \mathcal{A}} \{v_0(\alpha_0) - \pi_0^* \cdot \alpha_0\}$. Therefore, the NE profile β^* is a Walrasian equilibrium. □

Remark 48. With quasi-concave valuations, the induced GNE may not lead to allocations that are optimal for Problem 1. However, Theorem 25 shows that there still exists a GNE and that it is a Pareto efficient allocation, and thus, our mechanism incentivizes filtering with suboptimal social welfare.

4.1.4.2 Example

In this subsection, we present an example of how our proposed mechanism may be executed. Consider three major social media platforms: Facebook, Twitter, and Reddit. These platforms allow users from different political backgrounds to obtain the latest news. Typically, users engage with these platforms by scrolling, liking, or sharing posts featuring news and personal opinions. The time spent by all users on a platform defines the total engagement in the platform [213].

Since user engagement is a primary driver of advertisement revenue, Facebook, Twitter, and Reddit regularly optimize their post recommendation algorithms to maximize user engagement without accounting for the impact on user opinions [208]. Thus, many users form echo chambers, where they repeatedly interact only with biased posts. The biases of many users expose them to misinformation. This might cause uncertainty regarding the integrity of democratic institutions [186], or the results of the elections [211]. In practice, each platform can filter misinformation by flagging inaccurate posts. However, filtering is expensive because of (i) the large cost to identify inaccurate posts [222], and (ii) the potential decrease in engagement of censored users [9]. Thus, the government allocates a budget for the problem and appoints an independent agency to design monetary incentives for the platforms. This agency seeks a mechanism that: (i) induces voluntary participation among all platforms and (ii) maximizes the social welfare. Such a mechanism incentivizes platforms to implement filters. A sufficiently high lower bound ensures that some platforms implement non-zero filters. Using the mechanism in Section 4.1.3, the agency must achieve (i) and (ii) without knowledge of how the engagement and average trust on common knowledge evolve.

In step one (the participation step), the agency asks each platform whether they wish to participate in the mechanism since the government is not dictatorial. However, the agency assures the three platforms that they need not reveal private information, and that they can avoid filtering misinformation even after participating in the mechanism (Lemma 39). The government announces that platforms that choose not to participate will be labeled as uncooperative. Then, the weakly dominant action of every platform in step one is to participate in the mechanism (Theorem 24), ensuring property (i).

In step two (the bargaining step) of the mechanism, the agency asks each platform for a message proposing filtering levels for competing agents and corresponding prices, a lower bound for the government, and a minimum level for the average trust of their own users. Similarly, the government also proposes a lower bound and a price associated with the lower bound. The agency then publicly reveals all proposals and uses the rules of the mechanism to assign a potential subsidy/payment and potential filtering level to each platform. Similarly, the agency assigns a potential amount of investment and minimum average to the government. Note that the subsidy given to any platform is proportional to their assigned minimum average trust and filtering level. These assignments become binding only if all stakeholders, Facebook, Twitter, Reddit, and the government, accept them. If any stakeholder is dissatisfied, then all stakeholders change their proposals and resubmit. This process is repeated until all stakeholders reach a consensus, known as a GNE. The mechanism ensures that a consensus exists (Theorem 23) and that it maximizes the social welfare among all stakeholders (Theorem 22), thus establishing property (ii). Furthermore, the mechanism also ensures that each stakeholder is consistent in their messages with respect to the messages of other stakeholders (Lemmas 36 - 37) and that the agency makes neither profit nor loss (Theorem 21). Therefore, as long as the government is committed to addressing the problem of misinformation, the mechanism ensures the platforms will eventually agree to implement filters. The allocations become binding on all stakeholders, and the agency collects the government's investment. This investment is distributed to Facebook, Twitter, and Reddit as a subsidy, only after they implement the assigned filters.

4.1.5 Conclusions and Future Work

In this paper, we designed a mechanism to induce a GNE solution in the misinformation filtering game, where (i) each platform agrees to participate voluntarily, and (ii) the collective utility of the government and the platforms is maximized. Our proposed mechanism also satisfies budget balance. We also analyzed our mechanism under relaxed assumptions. Our ongoing work focuses on improving estimates of the valuation and average trust functions of the platforms using data and explicitly considering modeling uncertainty. These refinements of the modeling framework will allow us to make our mechanism more practically useful. Future research should include extending the mechanism to a dynamic setting where the platforms react in real-time to the proposed taxes/subsidies.

4.1.6 Appendix A - Extension to 2 Agents

In this appendix, we present an extension of by relaxing Assumption 7 to a more general assumption that no platform has a monopoly on its users. The mechanism presented in this assumes that for any platform $i \in \mathcal{I}$ with the set of competing platforms $\mathcal{C}_i$, it holds that $|\mathcal{C}_i| \geq 2$.

We consider the same step one (the participation step) for the mechanism as before. Then in step two (the bargaining step), the message of platform i is defined as

$$m_i := (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i), \tag{4.22}$$

where $\tilde{h}_i \in \mathbb{R}_{\geq 0}$ is the minimum average trust that platform i proposes to achieve through filtering; $\tilde{p}_i$, is the collection of prices that platform i is willing to pay or receive per unit changes in the filters of other competing platforms (except i) and the government's lower bound, given by

$$\tilde{p}_i := \begin{cases} (\tilde{p_l}^i : l \in \mathcal{D}_i), & \text{if } |\mathcal{C}_i| = 2, \\ (\tilde{p_l}^i : l \in \mathcal{D}_{-i}), & \text{if } |\mathcal{C}_i| \geq 3, \end{cases} \tag{4.23}$$

where $\tilde{p_l}^i \in \mathbb{R}_{\geq 0}$ for all $i, l \in \mathcal{J}$; and $\tilde{a}_i := (\tilde{a}_k^i : k \in \mathcal{D}_i)$, with $\tilde{a}_i \in \mathbb{R}^{|\mathcal{D}_i|}$ is the profile of filters for all competing platforms (including i) and government's lower bound proposed by platform i.

The message of the government is $m_0 := (\tilde{p}_0, \tilde{a}_0^0)$, where $\tilde{p}_0 \in \mathbb{R}_{\geq 0}$ is the price that the government is willing to pay or receive per unit change of the average trust, and $\tilde{a}_0^0 \in \mathbb{R}$ is the lower bound proposed by the government.

Based on the message profile $m := (m_0, m_1, \ldots, m_{|\mathcal{I}|})$ that the social planner receives, she allocates the following parameters to the agents:

1) The social planner allocates a filter to each platform $i \in \mathcal{I}$ and a lower bound to the government such that the constraints of Problem 1 are satisfied. The filter allocated by the social planner to platform i is $\alpha_i(m) := \sum_{k \in \mathcal{C}_i} \frac{\tilde{a}_i^k}{|\mathcal{C}_i|}$. The lower bound allocated by the social planner to the government is $\alpha_0(m) := \sum_{k \in \mathcal{J}} \frac{\tilde{a}_0^k}{|\mathcal{J}|}$.

2) The social planner allocates a minimum average trust $\eta_i(m) \in [0, 1]$ to each platform $i \in \mathcal{I}$, given by

$$\eta_i(m) := \min \left\{ \frac{n_i \cdot \tilde{h}_i}{\sum_{k \in \mathcal{I}} n_k \cdot \tilde{h}_k} \cdot \alpha_0(m),\ 1 \right\}, \tag{4.24}$$

where the social planner will not accept a message m_i from a platform i that might lead to a situation where $\sum_{k \in \mathcal{I}} n_k \cdot \tilde{h}_k = 0$. The allocated minimum average trust, $\eta_i(m)$, is a lower bound on average trust that must be achieved by platform i. Let the filter implemented by platform i be a_i. Then, platform i must ensure that $n_i \cdot h_i(a_i) \geq \eta_i(m)$. Recall from Section III-B that, as a result of Assumption 11, the social planner can prevent the platforms from violating the constraint imposed by $\eta_i(m)$.

3) The social planner also allocates a payment price

$$ {\pi_l}^i := \begin{cases} \tilde{p}_l^l, & \text{if } |\mathcal{C}_l| = 2, \\ \sum_{k \in \mathcal{C}_{-l} : k \neq i} \dfrac{\tilde{p_l}^k}{|\mathcal{C}_l| - 2}, & \text{if } |\mathcal{C}_l| \geq 3, \end{cases} \tag{4.25} $$

where $\pi_l^i \in \mathbb{R}_{\geq 0}$, to be paid by platform $i \in \mathcal{I}$ for a unit change in allocated filter $\alpha_l(m)$ of every other competing platform $l \in \mathcal{C}_{-i}$. Furthermore, the social planner allocates a subsidy price

$$ \sigma_l^i := \begin{cases} \tilde{p}_i^l, & \text{if } |\mathcal{C}_l| = 2, \\ \sum_{k \in \mathcal{C}_{-l} : k \neq i} \dfrac{\tilde{p_l}^k}{|\mathcal{C}_l| - 2}, & \text{if } |\mathcal{C}_l| \geq 3, \end{cases} \tag{4.26} $$

where $\sigma_l^i \in \mathbb{R}_{\geq 0}$, to be received by platform $i \in \mathcal{I}$ from every other competing platform $l \in \mathcal{C}_{-i}$, for a unit change in allocated filter $\alpha_i(m)$. For the government, the social planner simply allocates a price $\pi_0 := \sum_{i \in \mathcal{I}} \frac{\tilde{p}_0^i}{|\mathcal{I}|}$ to be paid for a unit change in lower bound $\alpha_0(m)$.

Remark 49. Note that when $|\mathcal{C}|_i = 2$, platform $i \in \mathcal{I}$ proposes a price corresponding to their own proposed action $\tilde{a}_i(m)$. In contrast, when $|\mathcal{C}_i| \geq 3$, platform i does not proposes a price corresponding to their own filter. However, we have designed the payment price in (4.25) and subsidy price (4.26) so that platform i cannot affect either of these prices with their message m_i. Thus, each still platform behaves as a *price taker* when $|\mathcal{C}_i| = 2$.

4) The social planner allocates the following tax to each social media platform $i \in \mathcal{I}$,

$$\tau_i := -\tilde{p}_0 \cdot \eta_i(m) - \sum_{l \in \mathcal{C}_{-i}} \sigma_l^i \cdot \alpha_i(m) + \sum_{l \in \mathcal{C}_{-i}} \pi_l^i \cdot \alpha_l(m) + \sum_{l \in \mathcal{C}_{-i} \cup \{0\}} \tilde{p}_l{}^i \cdot (\tilde{a}_l^i - \tilde{a}_l{}^{-i})^2 + \sum_{l \in \mathcal{C}_{-i}} \Big(\mathbb{I}(|\mathcal{C}_i| = 2) \cdot (\tilde{p}_i^i - \tilde{p}_i^l)^2 + \mathbb{I}(|\mathcal{C}_l| = 2) \cdot (\tilde{p}_l^i - \tilde{p}_l^l)^2 \Big), \tag{4.27}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\tilde{a}_l{}^{-i} = \sum_{k \in \mathcal{C}_{-l}} \frac{\tilde{a}_l^k}{|\mathcal{C}_l|-1}$, for each $l \in \mathcal{C}_{-i}$, is the average of the proposed filters for l by all competing platforms except $i \in \mathcal{I}$, and and $\tilde{a}_0{}^{-i} = \sum_{k \in \mathcal{J}_{-i}} \frac{\tilde{a}_0^k}{|\mathcal{J}|-1}$ is the average of lower bounds proposed by all agents except i. The tax $\tau_i(m)$ of platform i in (4.27) can be interpreted as follows: (i) the first term in (4.27) represents a subsidy given by the government to platform i for the increase in average trust among the users of platform i; (ii) the second term in (4.27) is a collection of subsidies given by each competing platform $l \in \mathcal{C}_{-i}$ to platform i for the increase in valuation $v_l\big(\alpha_k(m) : k \in \mathcal{C}_l\big)$ due to the allocated filter $\alpha_i(m)$; (iii) the third term in (4.27) is a payment by platform i for the increase in valuation $v_i\big(\alpha_k(m) : k \in \mathcal{C}_i\big)$ due to the allocated filter $\alpha_l(m)$ of each competing platform $l \in \mathcal{C}_{-i}$; (iv) the fourth term in (4.27) is a collections of penalties to platform i if either the filter proposed in message m_i for any competing platform $l \in \mathcal{C}_{-i}$ is inconsistent the filters proposed by other platforms, or if the lower bound proposed in m_i is inconsistent with the lower bound proposed by other agents; and (v) the fifth term is a collection of penalties to social media i for inconsistency in the proposed price, only applicable if $|\mathcal{C}_i| = 2$, or if $|\mathcal{C}_l| = 2$ for some $l \in \mathcal{C}_{-i}$.

The social planner also proposes the following payment function to the government:

$$\tau_0 := \pi_0 \cdot \alpha_0(m) + (\tilde{p}_0 - \pi_0)^2, \tag{4.28}$$

where the first term is the total investment made by the government for the allocated lower bound $\alpha_0(m)$, and the second term is a penalty when the price proposed by the government deviates from the price allocated to the government.

Remark 50. Note that the presence of the indicator function $\mathbb{I}(\cdot)$ in (4.27) does not lead to discontinuities in the utility $u_i(m_i, m_{-i})$ with respect to the message profile m.

Remark 51. The extended mechanism induces a game where the strategy of platform $i \in \mathcal{I}$ is $m_i \in \mathcal{M}_i$, such that $\alpha_i(m) \in \mathcal{S}_i(m)$. The equilibrium for the induced game is given by the GNE, defined in (4.15).

Then, we note that the results of Lemmas 36 - 37 hold for the extended mechanism. In addition, we prove in Lemma 40 that the equilibrium price received by any platform $i \in \mathcal{I}$ for an allocated filter $\alpha_l(m)$, $l \in \mathcal{C}_{-i}$, is the same as the price paid by the competing platform l.

Lemma 40. *Let message profile $m^* \in \mathcal{M}$ be a GNE of the induced game. Then, for each platform $i \in \mathcal{I}$ and each competing platform $l \in \mathcal{C}_{-i}$, it holds that $\sigma_l^{*i} = {\pi_i}^{*l}$.*

Proof. Consider two social media platforms $i \in \mathcal{I}$ and $l \in \mathcal{C}_{-i}$. The result holds from the definition of σ_l^i and ${\pi_i}^l$ when both $|\mathcal{C}_i| \geq 3$ and $|\mathcal{C}_l| \geq 3$.

Let $|\mathcal{C}_i| = 2$, with $\mathcal{C}_i = \{i, l\}$, and let m_{-i}^* be the message profile at GNE of all agents except i. In order to maximize their utility $u_i(m_i, m_{-i}^*)$, platform i must select a price $\tilde{p}_i^{*i}$ that minimizes the tax τ_i in (4.27). Thus, $\frac{\partial u_i}{\partial \tilde{p}_i^i}\Big|_{\tilde{p}_i^{*i}} = 2 \cdot (\tilde{p}_i^{*i} - \tilde{p}_i^{*l}) = 0$, which yields $\tilde{p}_i^{*i} = \tilde{p}_i^{*l}$. Then, the result holds using the definitions of σ_l^i and ${\pi_i}^l$ in (4.26) and (4.25), respectively. Through a similar analysis, we can prove the result when $|\mathcal{C}_l| = 2$. □

An additional implication of Lemma 40 is that the fifth term in (4.27) is 0 at any GNE. Thus, it can be verified that the results of Lemmas 38 - 39 and Theorems 21 - 25 hold for the extended mechanism.

Chapter 5

CONCLUSIONS

5.1 Summary of the Contributions

In this dissertation, we analyzed various decision-making problems relevant to both control and learning in cyber-physical systems. Chapter 1 completed a review of the literature pertaining to centralized worst-case learning and control, as well as to the decentralized control of teams. We also explored open questions in the literature related to both these areas of research and highlighted specific cyber-physical applications that pertain to these open questions. Finally, Chapter 1 introduced the theory of mechanism design to coordinate the actions of competing agents which formed the basis of an application to incentivize misinformation filtering in social media platforms.

Chapter 2 developed a principled approach to worst-case control and learning in partially observed systems using non-stochastic approximate information states. In the context of worst-case control problems, this chapter provides a unifying theory to rigorously derive performance bounds for strategies computed using various approximation schemes. In the context of worst-case reinforcement learning problems, the proposed approach introduces an approximate information state model that can be learned from output data without prior knowledge of system dynamics. Thus, this theoretical framework can facilitate data-driven worst-case control and reinforcement learning, which makes it particularly well suited for safety-critical applications in cyber-physical systems such as connected and automated vehicles, power grids, cyber-security and healthcare. Specifically, the major technical contributions of this chapter can be summarized as follows:

1) The introduction of a general notion of information states which yields an optimal DP decomposition for worst-case control problems.

2) The introduction of the notion of approximate information states that can either be constructed from output variables or learned from output data.

3) The formulation of an approximate DP and the derivation of explicit performance bounds for the resulting approximate control strategy.

4) An extension of the notions of information states and approximate information states to infinite horizon problems with the derivation of performance bounds.

5) The exposition of examples of information states and approximate information states along with theoretical approximation bounds for control problems.

6) The illustration of the performance of this approach in various worst-case control and reinforcement learning problems using numerical examples.

Chapter 3 derived structural forms for optimal control strategies in decentralized teams with one-directional information sharing amongst agents in both stochastic and non-stochastic settings. A salient feature of the structural forms is that they yield control strategies with time-invariant domains and thus, they facilitate the use of DP for their computation across long time horizons. In the stochastic setting, the information structure of “nested accessible information” was introduced for a team of two agents. This information structure unifies many previously studied information structures with one-directional communication and consequently, it gives a general property of the information sharing pattern within a team that can simplify the search for optimal strategies. In the non-stochastic setting, the ideas behind the nested accessible information structure were extended to teams of multiple agents with “nested subsystems.” The results obtained for decentralized teams with nested subsystems generalize the standard common information approach for non-stochastic problems. Collectively, the theoretical developments of this chapter advance the state-of-the-art on decision-making in complex cyber-physical systems consisting of multiple decision-makers with one-directional communication, such as vehicle platoons, large organizations and robot swarms. In summary, the major technical contributions of this chapter are as follows:

1) The establishment of a structural form of optimal control strategies in decentralized problems with one-directional communication which restricts their domains to spaces which do not grow in size with time.

2) A DP decomposition utilizing these results in both stochastic and non-stochastic formulations.

3) The extension of these results to special cases with decoupled dynamics for the stochastic formulation and their validation with numerical simulations in the non-stochastic formulation.

Chapter 4 introduced a model for the interactions between different social media platforms that care about maximizing user engagement and a democratic government that cares about the user's trust in democratic institutions. The proposed model captured the relationship between user engagement on social media platforms, misinformation filtering, and trust of users in democratic institutions by drawing upon theoretical and empirical conclusions in political science and sociology. Then, this chapter presented a mechanism to create incentives from a fixed monetary budget of the government, and to distribute these incentives amongst social media platforms in a manner that maximizes the filtering of misinformation. Finally, various desirable properties of the proposed mechanism were proved, including voluntary participation of social media platforms, no reliance on additional governmental spending, and the induction of a socially optimal equilibrium that balances the utilities of all platforms and the government. In summary, the major contributions of this chapter to can be summarized as follows:

1) The introduction of a mechanism to incentivize social media platforms to filter misleading information.

2) Derivations of key properties of the proposed mechanism and a case study to show how these properties help the implementation of such a mechanism.

In conclusion, this dissertation presents contributions towards multiple problems of decision-making in cyber-physical systems with partial information. The theoretical results presented in each of the three chapters have promising implications for both worst-case reinforcement learning and control in a range of applications including connected and automated vehicles, medicine and healthcare, and robotic swarms. For worst-case control with partial observations, approximate information state framework facilitates the computation of approximate solutions with known error bounds. This has the potential to make robust control computationally viable for a larger spectrum of safety-critical applications [33, 34], as well as applications vulnerable to adversarial attacks [31, 32]. This framework also facilitates the use of robust reinforcement learning approaches [92–94] in partially observed systems. Such robust strategies perform better than stochastic optimal strategies when they face unanticipated disturbances during real-life implementation [223–225], thus making them appropriate for use in cyber-physical systems. Furthermore, the data-driven nature of this framework also provides a principled approach to learn a model for human decision-making purely from observations during human-robot interactions. For example, [190] learned a stochastic approximate information state model for the behavior of human-driven vehicles and subsequently, used this model to safely control the merging of connected and automated vehicles in mixed traffic. Additionally, the results of the dissertation's decentralized control section can extend these advancements to problems consisting of multiple agents, such as connected and automated vehicles acting cooperatively to minimize energy consumption, maximize traveler safety, and improve traffic flow. Overall, this dissertation's contributions have the potential to drive significant progress towards safer and more computationally efficient decision-making in cyber-physical systems.

5.2 Directions for Future Research

While the ideas explored in this dissertation advance the state-of-the-art for both centralized and decentralized decision-making in cyber-physical systems, they also raise a number of compelling questions. We explore some of these questions in this section.

A well known drawback of the worst-case decision making approach considered in the first part of this dissertation is that in its effort to be robust against *every* possible disturbance, it often yields control actions whose performance is overly conservative for the application at-hand [226]. This drawback can be addressed by considering the risk-sensitive [28,35,65] and robust stochastic [26,27] formulations in both control and reinforcement learning problems. Both of these formulations allow for a better regulated trade off between the best-case performance of a stochastically optimal strategy and the worst-case performance of a non-stochastic strategy. It is my belief that the mathematical frameworks presented in the first two chapters of this dissertation can both be extended to both of these formulations. Such an extension would have two contributions, namely a generalization of the proposed frameworks towards a wider range of applications and an illumination on the connections between different models of uncertainty in decision-making [227].

The approach for decision-making in teams presented in the second part of this dissertation also yields many avenues of future research. One exciting direction is to consider the extension of these ideas to teams consisting of both humans and robotic agents. Human-robot teams are characterized by one-directional communication from the robot towards the human but are likely to feature imperfect communication from the human to the robot [228]. Thus, the information structure of these teams is likely to be consistent with those analyzed in this dissertation, and thus, it is my belief that these ideas can be used to improve the challenging problem of coordinating the actions of the agents to achieve a shared goal. This constitutes an important open problem because human-robot teams are expected to feature in a number of applications including connected and automated vehicles in mixed traffic [229,230], hospitals [231,232] and manufacturing [233, 234].

Finally, consider the mechanism presented in the third part of this dissertation, which incentivizes misinformation filtering in social media platforms. This mechanism relies upon a static formulation for all the interactions between the competing agents. To make this framework more practically useful, I believe that the model and the mechanism both need to be generalized to dynamic settings, where the effect of misinformation filtering on engagement and trust in democratic institutions is allowed to evolve with time.

BIBLIOGRAPHY

[1] K.-D. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.

[2] A. A. Malikopoulos, L. E. Beaver, and I. V. Chremos, "Optimal time trajectory and coordination for connected and automated vehicles," *Automatica*, vol. 125, no. 109469, 2021.

[3] B. Chalaki and A. A. Malikopoulos, "Robust learning-based trajectory planning for emerging mobility systems," in *2022 American Control Conference (ACC)*, 2022, pp. 2154–2159.

[4] M. Fatemi, T. W. Killian, J. Subramanian, and M. Ghassemi, "Medical dead-ends and learning to identify high-risk states and treatments," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4856–4870, 2021.

[5] A. Boloori, S. Saghafian, H. A. Chakkera, and C. B. Cook, "Data-driven management of post-transplant medications: An ambiguous partially observable markov decision process approach," *Manufacturing & Service Operations Management*, vol. 22, no. 5, pp. 1066–1087, 2020.

[6] J. A. Ansere, G. Han, L. Liu, Y. Peng, and M. Kamal, "Optimal resource allocation in energy-efficient internet-of-things networks with imperfect csi," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5401–5411, 2020.

[7] Z. Li, R. Chen, L. Liu, and G. Min, "Dynamic resource discovery based on preference and movement pattern similarity for large-scale social internet of things," *IEEE Internet of Things Journal*, vol. 3, no. 4, pp. 581–589, 2015.

[8] Y. Papanastasiou, "Fake news propagation and detection: A sequential model," *Management Science*, vol. 66, no. 5, pp. 1826–1846, 2020.

[9] O. Candogan and K. Drakopoulos, "Optimal signaling of content accuracy: Engagement vs. misinformation," *Operations Research*, vol. 68, no. 2, pp. 497–515, 2020.

[10] L. E. Beaver and A. A. Malikopoulos, "An Overview on Optimal Flocking," *Annual Reviews in Control*, vol. 51, pp. 88–99, 2021.

[11] ——, “Beyond Reynolds: A Constraint-Driven Approach to Cluster Flocking,” in *IEEE 59th Conference on Decision and Control*, 2020, pp. 208–213.

[12] P. R. Kumar and P. P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control.* Englewood Cliffs, NJ: Prentice-Hall, 1986.

[13] D. Burago, M. De Rougemont, and A. Slissenko, “On the complexity of partially observed markov decision processes,” *Theoretical Computer Science*, vol. 157, no. 2, pp. 161–183, 1996.

[14] A. A. Malikopoulos, “Separation of learning and control for cyber-physical systems,” *Automatica*, vol. 151, no. 110912, 2023.

[15] I. V. Chremos and A. A. Malikopoulos, “An analytical study of a two-sided mobility game,” in *2022 American Control Conference (ACC)*, 2022, pp. 1254–1259.

[16] N. Bäauerle and U. Rieder, “Partially observable risk-sensitive markov decision processes,” *Mathematics of Operations Research*, vol. 42, no. 4, pp. 1180–1196, 2017.

[17] A. A. Malikopoulos, “A duality framework for stochastic optimal control of complex systems,” *IEEE Transactions on Automatic Control*, vol. 61, no. 10, pp. 2756–2765, 2015.

[18] M. Ahmadi, N. Jansen, B. Wu, and U. Topcu, “Control theory meets pomdps: A hybrid systems approach,” *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5191–5204, 2020.

[19] A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yüksel, “Information structures in optimal decentralized control,” in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC).* IEEE, 2012, pp. 1291–1306.

[20] A. A. Malikopoulos, “On team decision problems with nonclassical information structures,” *IEEE Transactions on Automatic Control*, 2023.

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

[22] R. K. Mishra, D. Vasal, and S. Vishwanath, “Decentralized multi-agent reinforcement learning with shared actions,” in *2021 55th Annual Conference on Information Sciences and Systems (CISS).* IEEE, 2021, pp. 1–6.

[23] K. Zhang, Z. Yang, and T. Başar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

[24] H. Kao and V. Subramanian, “Common information based approximate state representations in multi-agent reinforcement learning,” in *International Conference on Artificial Intelligence and Statistics.* PMLR, 2022, pp. 6947–6967.

[25] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, “Bias and variance approximation in value function estimates,” *Management Science*, vol. 53, no. 2, pp. 308–322, 2007.

[26] W. Wiesemann, D. Kuhn, and B. Rustem, “Robust markov decision processes,” *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[27] G. N. Iyengar, “Robust dynamic programming,” *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[28] O. Mihatsch and R. Neuneier, “Risk-sensitive reinforcement learning,” *Machine learning*, vol. 49, pp. 267–290, 2002.

[29] H. S. Witsenhausen, “Minimax controls of uncertain systems,” , 1966.

[30] T. Başar and P. Bernhard, *H-infinity optimal control and related minimax design problems: a dynamic game approach.* Springer Science & Business Media, 2008.

[31] M. Rasouli, E. Miehling, and D. Teneketzis, “A scalable decomposition method for the dynamic defense of cyber networks,” in *Game Theory for Security and Risk Management.* Springer, 2018, pp. 75–98.

[32] Y. Shoukry, J. Araujo, P. Tabuada, M. Srivastava, and K. H. Johansson, “Minimax control for cyber-physical systems under network packet scheduling attacks,” in *Proceedings of the 2nd ACM international conference on High confidence networked systems*, 2013, pp. 93–100.

[33] M. Giuliani, J. Lamontagne, P. Reed, and A. Castelletti, “A state-of-the-art review of optimal reservoir control for managing conflicting demands in a changing world,” *Water Resources Research*, vol. 57, no. 12, p. e2021WR029927, 2021.

[34] Q. Zhu and T. Başar, “Robust and resilient control design for cyber-physical systems with an application to power systems,” in *2011 50th IEEE Conference on Decision and Control and European Control Conference.* IEEE, 2011, pp. 4066–4071.

[35] M. R. James, J. S. Baras, and R. J. Elliott, “Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems,” *IEEE transactions on automatic control*, vol. 39, no. 4, pp. 780–792, 1994.

[36] S. P. Coraluppi and S. I. Marcus, “Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes,” *Automatica*, vol. 35, no. 2, pp. 301–309, 1999.

[37] D. P. Bertsekas, "Control of uncertain systems with a set-membership description of the uncertainty." Ph.D. dissertation, Massachusetts Institute of Technology, 1971.

[38] D. Bertsekas, "Distributed asynchronous policy iteration for sequential zero-sum games and minimax control," *arXiv preprint arXiv:2107.10406*, 2021.

[39] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria.* Springer Science & Business Media, 2012, vol. 30.

[40] J. Moon and T. Başar, "Minimax control over unreliable communication channels," *Automatica*, vol. 59, pp. 182–193, 2015.

[41] D. Bertsekas, *Dynamic programming and optimal control: Volume I.* Athena scientific, 2012, vol. 1.

[42] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of markov decision processes," *Mathematics of operations research*, vol. 12, no. 3, pp. 441–450, 1987.

[43] P. Bernhard, "A separation theorem for expected value and feared value discrete time control," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 1, pp. 191–206, 1996.

[44] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[45] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1644–1658, 2013.

[46] J. Subramanian and A. Mahajan, "Approximate information state for partially observed systems," in *2019 IEEE 58th Conference on Decision and Control (CDC).* IEEE, 2019, pp. 1629–1636.

[47] Y. Cong, X. Wang, and X. Zhou, "Rethinking the mathematical framework and optimality of set-membership filtering," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2544–2551, 2021.

[48] D. P. Bertsekas and I. B. Rhodes, "On the minimax reachability of target sets and target tubes," *Automatica*, vol. 7, no. 2, pp. 233–247, 1971.

[49] D. Bertsekas and I. Rhodes, "Sufficiently informative functions and the minimax feedback control of uncertain dynamic systems," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 117–124, 1973.

[50] M. Gagrani and A. Nayyar, "Decentralized minimax control problems with partial history sharing," in *2017 American Control Conference (ACC).* IEEE, 2017, pp. 3373–3379.

[51] R. R. Moitié, M. Quincampoix, and V. M. Veliov, "Optimal control of discrete-time uncertain systems with imperfect measurement," *IEEE transactions on automatic control*, vol. 47, no. 11, pp. 1909–1914, 2002.

[52] C. Piccardi, "Infinite-horizon minimax control with pointwise cost functional," *Journal of Optimization Theory and Applications*, vol. 78, no. 2, pp. 317–336, 1993.

[53] P. Bernhard, "Minimax - or feared value - L_1 / L_∞ control," *Theoretical computer science*, vol. 293, no. 1, pp. 25–44, 2003.

[54] M. Gagrani, Y. Ouyang, M. Rasouli, and A. Nayyar, "Worst-case guarantees for remote estimation of an uncertain source," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1794–1801, 2020.

[55] P. Bernhard, "Max-plus algebra and mathematical fear in dynamic optimization," *Set-Valued Analysis*, vol. 8, no. 1, pp. 71–84, 2000.

[56] J. S. Baras and N. S. Patel, "Robust control of set-valued discrete-time dynamical systems," *IEEE Transactions on Automatic Control*, vol. 43, no. 1, pp. 61–75, 1998.

[57] S. P. Coraluppi and S. I. Marcus, "Mixed risk-neutral/minimax control of discrete-time, finite-state markov decision processes," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 528–532, 2000.

[58] A. Nilim and L. El Ghaoui, "Robust control of markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[59] E. M. Wolff, U. Topcu, and R. M. Murray, "Robust control of uncertain markov decision processes with temporal logic specifications," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012, pp. 3372–3379.

[60] T. Osogami, "Robust partially observable markov decision process," in *International Conference on Machine Learning*. PMLR, 2015, pp. 106–115.

[61] M. Rasouli and S. Saghafian, "Robust partially observable markov decision processes," in *HKS Working Paper No. RWP18-027*, 2018.

[62] S. Saghafian, "Ambiguous partially observable markov decision processes: Structural results and applications," *Journal of Economic Theory*, vol. 178, pp. 1–35, 2018.

[63] M. Cubuktepe, N. Jansen, S. Junges, A. Marandi, M. Suilen, and U. Topcu, "Robust finite-state controllers for uncertain pomdps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 11 792–11 800.

[64] M. Ahmadi, U. Rosolia, M. D. Ingham, R. M. Murray, and A. D. Ames, "Constrained risk-averse markov decision processes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 718–11 725.

[65] M. Ahmadi, X. Xiong, and A. D. Ames, "Risk-averse control via cvar barrier functions: Application to bipedal robot locomotion," *IEEE Control Systems Letters*, vol. 6, pp. 878–883, 2021.

[66] R. Laraki and S. Sorin, "Advances in zero-sum dynamic games," in *Handbook of game theory with economic applications.* Elsevier, 2015, vol. 4, pp. 27–93.

[67] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning.* PMLR, 2017, pp. 2817–2826.

[68] F. Dufour and T. Prieto-Rumeau, "Approximation of discounted minimax markov control problems and zero-sum markov games using hausdorff and wasserstein distances," *Dynamic Games and Applications*, vol. 9, no. 1, pp. 68–102, 2019.

[69] S. Meyn, *Control Systems and Reinforcement Learning.* Cambridge University Press, 2022.

[70] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[71] A. A. Malikopoulos, P. Y. Papalambros, and D. N. Assanis, "A Real-Time Computational Learning Model for Sequential Decision-Making Problems Under Uncertainty," *Journal of Dynamic Systems, Measurement, and Control*, vol. 131, no. 4, 05 2009, 041010. [Online]. Available: https://doi.org/10.1115/1.3117200

[72] T. M. Moerland, J. Broekens, and C. M. Jonker, "Model-based reinforcement learning: A survey," *arXiv preprint arXiv:2006.16712*, 2020.

[73] J. Ramírez, W. Yu, and A. Perrusquía, "Model-free reinforcement learning from expert demonstrations: a survey," *Artificial Intelligence Review*, vol. 55, no. 4, pp. 3213–3241, 2022.

[74] J. Clifton and E. Laber, "Q-learning: Theory and applications," *Annual Review of Statistics and Its Application*, vol. 7, pp. 279–301, 2020.

[75] H. Hasselt, "Double q-learning," *Advances in neural information processing systems*, vol. 23, 2010.

[76] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep q-learning," in *Learning for Dynamics and Control.* PMLR, 2020, pp. 486–489.

[77] 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.

[78] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[79] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine learning*, vol. 38, pp. 287–308, 2000.

[80] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[81] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang, "Variational policy gradient method for reinforcement learning with general utilities," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4572–4583, 2020.

[82] J. Zhang, J. Kim, B. O'Donoghue, and S. Boyd, "Sample efficient reinforcement learning with reinforce," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 887–10 895.

[83] M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski *et al.*, "What matters for on-policy deep actor-critic methods? a large-scale study," in *International conference on learning representations*, 2021.

[84] J. Morimoto and K. Doya, "Robust reinforcement learning," *Neural computation*, vol. 17, no. 2, pp. 335–359, 2005.

[85] M. Heger, "Consideration of risk in reinforcement learning," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 105–111.

[86] G. Jiang, C.-P. Wu, and G. Cybenko, "Minimax-based reinforcement learning with state aggregation," in *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, vol. 2. IEEE, 1998, pp. 1236–1241.

[87] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free q-learning designs for linear discrete-time zero-sum games with application to h-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[88] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[89] A. P. Valadbeigi, A. K. Sedigh, and F. L. Lewis, “H_∞ static output-feedback control design for discrete-time systems using reinforcement learning,” *IEEE transactions on neural networks and learning systems*, vol. 31, no. 2, pp. 396–406, 2019.

[90] S. Chakravorty and D. Hyland, “Minimax reinforcement learning,” in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2003, p. 5718.

[91] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, “H_∞ control of linear discrete-time systems: Off-policy reinforcement learning,” *Automatica*, vol. 78, pp. 144–152, 2017.

[92] Y. Wang and S. Zou, “Online robust reinforcement learning with model uncertainty,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 7193–7206, 2021.

[93] N. Kumar, K. Levy, K. Wang, and S. Mannor, “An efficient solution to s-rectangular robust markov decision processes,” *arXiv preprint arXiv:2301.13642*, 2023.

[94] Y. Wang and S. Zou, “Policy gradient method for robust reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 484–23 526.

[95] P. Clavier, E. L. Pennec, and M. Geist, “Towards minimax optimality of model-based robust reinforcement learning,” *arXiv preprint arXiv:2302.05372*, 2023.

[96] A. Rantzer, “Minimax adaptive control for a finite set of linear systems,” in *Learning for Dynamics and Control*. PMLR, 2021, pp. 893–904.

[97] N. Agarwal, B. Bullins, E. Hazan, S. Kakade, and K. Singh, “Online control with adversarial disturbances,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 111–119.

[98] P. Gradu, J. Hallman, and E. Hazan, “Non-stochastic control with bandit feedback,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 764–10 774, 2020.

[99] M. Littman and R. S. Sutton, “Predictive representations of state,” *Advances in neural information processing systems*, vol. 14, 2001.

[100] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, “Approximate information state for approximate planning and reinforcement learning in partially observed systems,” *Journal of Machine Learning Research*, vol. 23, no. 12, pp. 1–83, 2022.

[101] G. Patil, A. Mahajan, and D. Precup, “On learning history based policies for controlling markov decision processes,” *arXiv preprint arXiv:2211.03011*, 2022.

[102] A. D. Kara and S. Yüksel, "Near optimality of finite memory feedback policies in partially observed markov decision processes." *Journal of Machine Learning Research*, vol. 23, pp. 11–1, 2022.

[103] L. Yang, K. Zhang, A. Amice, Y. Li, and R. Tedrake, "Discrete approximate information states in partially observable environments," in *2022 American Control Conference (ACC).* IEEE, 2022, pp. 1406–1413.

[104] T. W. Killian, H. Zhang, J. Subramanian, M. Fatemi, and M. Ghassemi, "An empirical study of representation learning for reinforcement learning in healthcare," in *Machine Learning for Health.* PMLR, 2020, pp. 139–160.

[105] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," *Advances in neural information processing systems*, vol. 31, 2018.

[106] A. Hefny, Z. Marinho, W. Sun, S. Srinivasa, and G. Gordon, "Recurrent predictive state policy networks," in *International Conference on Machine Learning.* PMLR, 2018, pp. 1949–1958.

[107] A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello, "Learning causal state representations of partially observable environments," *arXiv preprint arXiv:1906.10437*, 2019.

[108] M. Okada and T. Taniguchi, "Dreaming: Model-based reinforcement learning by latent imagination without reconstruction," in *2021 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 2021, pp. 4209–4215.

[109] J. Marschak, "Elements for a theory of teams," *Management science*, vol. 1, no. 2, pp. 127–137, 1955.

[110] R. Radner, "Team decision problems," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 857–881, 1962.

[111] J. Marschak and R. Radner, *Economic Theory of Teams.* Yale University Press, 1972.

[112] J. Zabojnik, "Centralized and decentralized decision making in organizations," *Journal of Labor Economics*, vol. 20, no. 1, pp. 1–22, 2002.

[113] A. M. I. Mahbub and A. A. Malikopoulos, "A Platoon Formation Framework in a Mixed Traffic Environment," *IEEE Control Systems Letters (LCSS)*, vol. 6, pp. 1370–1375, 2021.

[114] A. A. Malikopoulos, C. Charalambous, and I. Tzortzis, "The average cost of markov chains subject to total variation distance uncertainty," in *Systems & Control Letters*, vol. 120, 2018, pp. 29–35.

[115] J. Krainak, J. Speyer, and S. Marcus, “Static team problems–part i: Sufficient conditions and the exponential cost criterion,” *IEEE Transactions on Automatic Control*, vol. 27, no. 4, pp. 839–848, 1982.

[116] ——, “Static team problems–part ii: Affine control laws, projections, algorithms, and the legt problem,” *IEEE Transactions on Automatic Control*, vol. 27, no. 4, pp. 848–859, 1982.

[117] H. S. Witsenhausen, “On information structures, feedback and causality,” *SIAM Journal on Control*, vol. 9, no. 2, pp. 149–160, 1971.

[118] S. Yuksel, “Stochastic nestedness and the belief sharing information pattern,” *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2773–2786, 2009.

[119] N. Saldi and S. Yüksel, “Geometry of information structures, strategic measures and associated stochastic control topologies,” *Probability Surveys*, vol. 19, pp. 450–532, 2022.

[120] H. S. Witsenhausen, “Separation of estimation and control for discrete time systems,” *Proceedings of the IEEE*, vol. 59, no. 11, pp. 1557–1566, 1971.

[121] J. Tsitsiklis and M. Athans, “On the complexity of decentralized decision making and detection problems,” *IEEE Transactions on Automatic Control*, vol. 30, no. 5, pp. 440–446, 1985.

[122] A. A. Malikopoulos, “Centralized stochastic optimal control of complex systems,” in *2015 European Control Conference (ECC)*. IEEE, 2015, pp. 721–726.

[123] ——, “Equilibrium control policies for markov chains,” in *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, 2011, pp. 7093–7098.

[124] L. Lessard and A. Nayyar, “Structural results and explicit solution for two-player lqg systems on a finite time horizon,” in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 6542–6549.

[125] A. Mahajan, “Optimal decentralized control of coupled subsystems with control sharing,” *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2377–2382, 2013.

[126] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, “The complexity of decentralized control of markov decision processes,” *Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, 2002.

[127] Y.-C. Ho and K.-C. Chu, “Team decision theory and information structures in optimal control problems–Part I,” *IEEE Transactions on Automatic Control*, vol. 17, no. 1, pp. 15–22, 1972.

[128] ——, "Team decision theory and information structures in optimal control problems-part II," *IEEE Trans. Autom. Control*, vol. 17, pp. 22–28, 1972.

[129] C. D. Charalambous and N. U. Ahmed, "Centralized versus decentralized optimization of distributed stochastic differential decision systems with different information structures-part i: a general theory," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1194–1209, 2016.

[130] C. D. Charalambous, "Decentralized optimality conditions of stochastic differential decision problems via girsanov's measure transformation," *Mathematics of Control, Signals, and Systems*, vol. 28, no. 3, pp. 1–55, 2016.

[131] H. Witsenhausen, "On the structure of real time source coders," *Bell Syst. Tech. J*, vol. 58, no. 6, pp. 1437–1451, 1979.

[132] P. Varaiya and J. Walrand, "Optimal Causal Coding-Decoding Problems," *Math. Oper. Res. Bell Syst. Tech. J. Bell Syst. Tech. J*, vol. 59, no. 8, pp. 814–820, 1983.

[133] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4017–4035, 2006.

[134] A. Nayyar and D. Teneketzis, "On jointly optimal real-time encoding and decoding strategies in multi-terminal communication systems," in *Proceedings of the IEEE Conference on Decision and Control*, 2008, pp. 1620–1627.

[135] Y. Kaspi and N. Merhav, "Structure theorem for real-time variable-rate lossy source encoders and memory-limited decoders with side information," in *IEEE International Symposium on Information Theory - Proceedings*, 2010, pp. 86–90.

[136] D. Teneketzis and P. Varaiya, "The Decentralized Quickest Detection Problem," *IEEE Transactions on Automatic Control*, vol. 29, no. 7, pp. 641–644, 1984.

[137] V. V. Veeravalli, T. Başar, and H. V. Poor, "Decentralized Sequential Detection with a Fusion Center Performing the Sequential Test," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 433–442, 1993.

[138] J. Wu and S. Lall, "A dynamic programming algorithm for decentralized markov decision processes with a broadcast structure," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 6143–6148.

[139] P. Varaiya and J. Walrand, "Causal coding and control for Markov chains," *Systems and Control Letters*, vol. 3, no. 4, pp. 189–192, 1983.

[140] L. Lessard and S. Lall, "Optimal controller synthesis for the decentralized two-player problem with output feedback," in *2012 American Control Conference (ACC)*. IEEE, 2012, pp. 6314–6321.

[141] L. Lessard and A. Nayyar, “Structural results and explicit solution for two-player lqg systems on a finite time horizon,” in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 6542–6549.

[142] A. Nayyar and L. Lessard, “Structural results for partially nested lqg systems over graphs,” in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 5457–5464.

[143] H. S. Witsenhausen, “A standard form for sequential stochastic control,” *Mathematical Systems Theory*, vol. 7, no. 1, pp. 5–11, 1973.

[144] A. Mahajan, “Sequential decomposition of sequential dynamic teams: Applications to real-time communication and networked control systems.” Ph.D. dissertation, University of Michigan, 2008.

[145] S. Yuksel, “A universal dynamic program and refined existence results for decentralized stochastic control,” *SIAM Journal on Control and Optimization*, vol. 58, no. 5, pp. 2711–2739, 2020.

[146] A. Mahajan and D. Teneketzis, “On the design of globally optimal communication strategies for real-time noisy communication systems with noisy feedback,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 580–595, 2008.

[147] ——, “Optimal design of sequential real-time communication systems,” *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5317–5338, 2009.

[148] A. Nayyar, T. Başar, D. Teneketzis, and V. V. Veeravalli, “Optimal Strategies for Communication and Remote Estimation With an Energy Harvesting Sensor,” *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2246–2260, 2013.

[149] A. Nayyar, A. Mahajan, and D. Teneketzis, “Optimal control strategies in delayed sharing information structures,” *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1606–1620, 2011.

[150] J. Arabneydi and A. Mahajan, “Team optimal control of coupled subsystems with mean-field sharing,” in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 1669–1674.

[151] T. Yoshikawa, “Decomposition of dynamic team decision problems,” *IEEE Transactions on Automatic Control*, vol. 23, no. 4, pp. 627–632, 1978.

[152] M. Aicardi, F. Davoli, and R. Minciardi, “Decentralized optimal control of markov chains with a common past information set,” *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 1028–1031, 1987.

[153] G. Casalino, F. Davoli, R. Minciardi, P. Puliafito, and R. Zoppoli, “Partially nested information structures with a common past,” *IEEE transactions on automatic control*, vol. 29, no. 9, pp. 846–850, 1984.

[154] Y. Xie, J. Dibangoye, and O. Buffet, “Optimally solving two-agent decentralized pomdps under one-sided information sharing,” in *International Conference on Machine Learning.* PMLR, 2020, pp. 10 473–10 482.

[155] A. Nayyar and D. Teneketzis, “On the structure of real-time encoding and decoding functions in a multiterminal communication system,” *IEEE transactions on information theory*, vol. 57, no. 9, pp. 6196–6214, 2011.

[156] S. M. Asghari, Y. Ouyang, and A. Nayyar, “Optimal local and remote controllers with unreliable uplink channels,” *IEEE Transactions on Automatic Control*, vol. 64, no. 5, pp. 1816–1831, 2018.

[157] Y. Ouyang, S. M. Asghari, and A. Nayyar, “Optimal local and remote controllers with unreliable communication,” in *2016 IEEE 55th Conference on Decision and Control (CDC).* IEEE, 2016, pp. 6024–6029.

[158] H. Tavafoghi, Y. Ouyang, and D. Teneketzis, “A unified approach to dynamic decision problems with asymmetric information: Nonstrategic agents,” *IEEE Transactions on Automatic Control*, vol. 67, no. 3, pp. 1105–1119, 2021.

[159] M. Rotkowitz and S. Lall, “A characterization of convex problems in decentralized control,” *IEEE transactions on Automatic Control*, vol. 50, no. 12, pp. 1984–1996, 2005.

[160] P. Shah and P. A. Parrilo, “$\mathcal{H}_2$-optimal decentralized control over posets: A state-space solution for state-feedback,” *IEEE Transactions on Automatic Control*, vol. 58, no. 12, pp. 3084–3096, 2013.

[161] E. Gallestey, M. James, and W. McEneaney, “Max-plus methods in partially observed h/sub/spl infin//control,” in *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, vol. 3. IEEE, 1999, pp. 3011–3016.

[162] N. Saldi, T. Linder, and S. Yüksel, “Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control,” *IEEE Transactions on Automatic Control*, vol. 60, no. 2, pp. 553–558, 2014.

[163] G. N. Nair, “A nonstochastic information theory for communication and state estimation,” *IEEE Transactions on automatic control*, vol. 58, no. 6, pp. 1497–1510, 2013.

[164] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "Approximate information states for worst-case control of uncertain systems," in *Proceedings of the 61th IEEE Conference on Decision and Control (CDC)*, 2022, pp. 4945–4950.

[165] ——, "On robust control of partially observed uncertain systems with additive costs," in *Proceedings of the 2023 American Control Conference (ACC)*, 2023 (to appear).

[166] ——, "Approximate Information States for Worst-Case Control and Learning in Uncertain Systems," *arXiv:2301.05089 (in review)*, 2023.

[167] A. Dave, I. Faros, N. Venkatesh, and A. A. Malikopoulos, "Worst-Case Control and Learning Using Partial Observations Over an Infinite Time Horizon," *arXiv:2303.16321 (in review)*, 2023.

[168] A. Nayyar and D. Teneketzis, "Common knowledge and sequential team problems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 5108–5115, 2019.

[169] S. Sudhakara, D. Kartik, R. Jain, and A. Nayyar, "Optimal communication and control strategies in a multi-agent mdp problem," *arXiv preprint arXiv:2104.10923*, 2021.

[170] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "On decentralized control of two agents with nested accessible information," in *2022 American Control Conference (ACC)*. IEEE, 2022, pp. 3423–3430.

[171] ——, "On decentralized minimax control with nested subsystems," in *2022 American Control Conference (ACC)*. IEEE, 2022, pp. 3437–3444.

[172] A. Dave and A. A. Malikopoulos, "Decentralized stochastic control in partially nested information structures," *IFAC-PapersOnLine*, vol. 52, no. 20, pp. 97–102, 2019.

[173] ——, "Structural results for decentralized stochastic control with a word-of-mouth communication," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 2796–2801.

[174] ——, "A dynamic program for a team of two agents with nested information," in *2021 IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 3768–3773.

[175] ——, "The prescription approach for decentralized stochastic control with word-of-mouth communication," *arXiv preprint, arXiv:1907.12125*, 2021.

[176] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic theory.* Oxford University Press, 1995.

[177] P. N. Brown and J. R. Marden, "Optimal mechanisms for robust coordination in congestion games," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2437–2448, 2017.

[178] I. V. Chremos and A. A. Malikopoulos, "The design and analysis of a mobility game," *arXiv preprint arXiv:2202.07691*, 2022.

[179] S. Sharma and D. Teneketzis, "Local public good provisioning in networks: A Nash implementation mechanism," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 11, pp. 2105–2116, 2012.

[180] A. Sinha and A. Anastasopoulos, "Generalized proportional allocation mechanism design for multi-rate multicast service on the internet," *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 146–153, 2013.

[181] A. Kakhbod and D. Teneketzis, "An efficient game form for unicast service provisioning," *IEEE Transactions on Automatic Control*, vol. 57, no. 2, pp. 392–404, 2011.

[182] R. Jain and J. Walrand, "An efficient nash-implementation mechanism for network resource allocation," *Automatica*, vol. 46(8), pp. 1276–1283, 2010.

[183] M. Zhang and J. Huang, "Efficient network sharing with asymmetric constraint information," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1898–1910, 2019.

[184] I. V. Chremos and A. A. Malikopoulos, "A socially-efficient emerging mobility market," *arXiv preprint arXiv:2011.14399*, 2020.

[185] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Science vs conspiracy: Collective narratives in the age of misinformation," *PloS one*, vol. 10, no. 2, p. e0118093, 2015.

[186] E. Brown, "Propaganda, misinformation, and the epistemic value of democracy," *Critical Review*, vol. 30(3–4), pp. 194–218, 2018.

[187] J. A. Tucker, Y. Theocharis, M. E. Roberts, and P. Barberá, "From liberation to turmoil: Social media and democracy," *Journal of Democracy*, vol. 28(4), pp. 46–59, 2017.

[188] A. Sternisko, A. Cichocka, and J. J. Van Bavel, "The dark side of social movements: Social identity, non-conformity, and the lure of conspiracy theories," *Current opinion in psychology*, vol. 35, pp. 1–6, 2020.

[189] A. Dave, I. V. Chremos, and A. A. Malikopoulos, "Social media and misleading information in a democracy: A mechanism design approach," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2633–2639, 2022.

[190] N. Venkatesh, V.-A. Le, A. Dave, and A. A. Malikopoulos, "Connected and automated vehicles in mixed-traffic: Learning human driver behavior for effective on-ramp merging," *arxiv: 2304.00397*, 2023 (in review).

[191] I. V. Chremos, H. Bang, A. Dave, V.-A. Le, and A. A. Malikopoulos, "Modeling travel behavior in mobility systems with an atomic routing game and prospect theory," *arXiv:2303.17790*, 2023.

[192] A. Rangi and M. Franceschetti, "Towards a non-stochastic information theory," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 997–1001.

[193] A. Girard and G. J. Pappas, "Approximation metrics for discrete and continuous systems," *IEEE Transactions on Automatic Control*, vol. 52, no. 5, pp. 782–798, 2007.

[194] M. F. Barnsley, *Superfractals.* Cambridge University Press, 2006.

[195] G. Didinsky, *Design of minimax controllers for nonlinear systems using cost-to-come methods.* University of Illinois at Urbana-Champaign, 1995.

[196] P. Bernhard, "Sketch of a theory of nonlinear partial information min-max control," Ph.D. dissertation, INRIA, 1993.

[197] D. Bertsekas, "Convergence of discretization procedures in dynamic programming," *IEEE Transactions on Automatic Control*, vol. 20, no. 3, pp. 415–419, 1975.

[198] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[199] V. Kolokoltsov and V. P. Maslov, *Idempotent analysis and its applications.* Springer Science & Business Media, 1997, vol. 401.

[200] K.-J. Chung and M. J. Sobel, "Discounted mdp's: Distribution functions and exponential utility maximization," *SIAM journal on control and optimization*, vol. 25, no. 1, pp. 49–62, 1987.

[201] S. I. Marcus, E. Fernández-Gaucherand, D. Hernández-Hernandez, S. Coraluppi, and P. Fard, "Risk sensitive markov decision processes," in *Systems and control in the twenty-first century.* Springer, 1997, pp. 263–279.

[202] N. Bäuerle and U. Rieder, "More risk-sensitive markov decision processes," *Mathematics of Operations Research*, vol. 39, no. 1, pp. 105–120, 2014.

[203] Y. A. Reznik, "An algorithm for quantization of discrete probability distributions," in *2011 Data Compression Conference.* IEEE, 2011, pp. 333–342.

[204] N. Saldi, S. Yüksel, and T. Linder, "Asymptotic optimality of finite model approximations for partially observed markov decision processes with discounted cost," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 130–142, 2019.

[205] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[206] W. Davies, "The age of post-truth politics," *The New York Times*, vol. 24, p. 2016, 2016.

[207] J. Cone, K. Flaharty, and M. J. Ferguson, "Believability of evidence matters for correcting social impressions," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9802–9807, 2019.

[208] Z. Tufekci, "Youtube, the great radicalizer," *The New York Times*, vol. 10, 2018.

[209] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[210] J. Weedon, W. Nuland, and A. Stamos, "Information operations and facebook," *Retrieved from: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf*, 2017.

[211] H. Farrell and B. Schneier, "Common-knowledge attacks on democracy," *Berkman Klein Center Research Publication*, vol. 7, no. 2018, 2018.

[212] E. Klein and J. Robison, "Like, post, and distrust? how social media use affects trust in government," *Political Communication*, vol. 37, no. 1, pp. 46–64, 2020.

[213] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[214] O. Analytica, "Russia will deny cyberattacks despite more us evidence," *Emerald Expert Briefings*, 2018.

[215] R. Jaakonmäki, O. Müller, and J. Vom Brocke, "The impact of content, context, and creator on user engagement in social media marketing," *Proceedings of the 50th Hawaii international conference on system sciences*, 2017.

[216] E. A. Vogels, A. Perrin, and M. Anderson, *Most Americans Think Social Media Sites Censor Political Viewpoints.* Pew Research Center, Washington D.C., 2020. [Online]. Available: https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints

[217] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[218] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, “Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France,” *New media & society*, vol. 16, no. 2, pp. 340–358, 2014.

[219] T. Saijo and T. Yamato, “Fundamental impossibility theorems on voluntary participation in the provision of non-excludable public goods,” *Review of Economic Design*, vol. 14, no. 1-2, pp. 51–73, 2010.

[220] F. Farhadi, H. Tavafoghi, D. Teneketzis, and S. J. Golestani, “An efficient dynamic allocation mechanism for security in networks of interdependent strategic agents,” *Dynamic Games and Applications*, vol. 9(4), pp. 914–941, 2019.

[221] F. Facchinei and C. Kanzow, “Generalized Nash equilibrium problems,” *Annals of Operations Research*, vol. 175, no. 1, pp. 177–211, 2010.

[222] D. Graves, “Understanding the promise and limits of automated fact-checking,” *Technical Report from Reuters Institute for the Study of Journalism*, 2018.

[223] A. Kumar, A. Zhou, G. Tucker, and S. Levine, “Conservative q-learning for offline reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[224] J. Moos, K. Hansel, H. Abdulsamad, S. Stark, D. Clever, and J. Peters, “Robust reinforcement learning: A review of foundations and recent advances,” *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 276–315, 2022.

[225] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, “Safe learning in robotics: From learning-based control to safe reinforcement learning,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[226] Y. Wang and M. P. Chapman, “Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control,” *Artificial Intelligence*, p. 103743, 2022.

[227] S. Samson, J. A. Reneke, and M. M. Wiecek, “A review of different perspectives on uncertainty and risk and an alternative modeling paradigm,” *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 558–567, 2009.

[228] D. Malik, M. Palaniappan, J. Fisac, D. Hadfield-Menell, S. Russell, and A. Dragan, “An efficient, generalized bellman update for cooperative inverse reinforcement learning,” in *International Conference on Machine Learning.* PMLR, 2018, pp. 3394–3402.

[229] V.-A. Le and A. A. Malikopoulos, “A Cooperative Optimal Control Framework for Connected and Automated Vehicles in Mixed Traffic Using Social Value Orientation,” in *2022 61th IEEE Conference on Decision and Control (CDC)*, 2022, pp. 6272–6277.

[230] W. Wang, L. Wang, C. Zhang, C. Liu, L. Sun *et al.*, “Social interactions for autonomous driving: A review and perspectives,” *Foundations and Trends® in Robotics*, vol. 10, no. 3-4, pp. 198–376, 2022.

[231] G. Hoffman and C. Breazeal, “Collaboration in human-robot teams,” in *AIAA 1st intelligent systems technical conference*, 2004, p. 6434.

[232] H. S. Koppula and A. Saxena, “Anticipating human activities for reactive robotic response.” in *IROS*. Tokyo, 2013, p. 2071.

[233] S. Nikolaidis and J. Shah, “Human-robot teaming using shared mental models,” *ACM/IEEE HRI*, 2012.

[234] W. P. Chan, G. Hanks, M. Sakr, H. Zhang, T. Zuo, H. M. Van der Loos, and E. Croft, “Design and evaluation of an augmented reality head-mounted display interface for human robot teams collaborating in physically shared manufacturing tasks,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 3, pp. 1–19, 2022.

Appendix

REPUBLICATION PERMISSIONS

This appendix lists the articles that appear in this dissertation.

- [164]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "Approximate information states for worst-case control of uncertain systems," in *Proceedings of the 61th IEEE Conference on Decision and Control (CDC)*, pages 4945–4950, 2022.

- [165]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "On robust control of partially observed uncertain systems with additive costs," in *Proceedings of the 2023 American Control Conference (ACC)*, pp. 4639-4644, 2023.

- [170]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "On decentralized control of two agents with nested accessible information," in *Proceedings of the 2022 American Control Conference (ACC)*, pages 3423–3430, 2022.

- [171]: **Aditya Dave**, Nishanth Venkatesh, and Andreas A. Malikopoulos, "On decentralized minimax control with nested subsystems," in *Proceedings of the 2022 American Control Conference (ACC)*, pages 3437–3444. 2022.

- [189]: **Aditya Dave**, Ioannis Vasileios Chremos, and Andreas A. Malikopoulos, "Social media and misleading information in a democracy: A mechanism design approach," *IEEE Transactions on Automatic Control*, 67(5), pp. 2633-2639, 2022.

The permission grant to reuse this material in the dissertation is attached in the next page.

CCC RightsLink

Home | Help | Live Chat | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

Approximate Information States for Worst-case Control of Uncertain Systems

Conference Proceedings: 2022 IEEE 61st Conference on Decision and Control (CDC)

Author: Aditya Dave

Publisher: IEEE

Date: 06 December 2022

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW