New Concepts in Team Theory: Mean Field Teams and Reinforcement Learning

Jalal Arabneydi

Doctor of Philosophy
Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

December 2016

A thesis submitted to McGill University in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

To my father and my mother
Mandani Arabneydi and Zinat Maroofi

To my older and younger brothers
Hossein and Reza

To my best friend
Hadi Amirheidari

ABSTRACT

This thesis consists of two parts wherein each part introduces a new concept in team theory.

In the first part, we introduce systems with partially exchangeable agents. A system is called partially exchangeable if it can be partitioned into sub-populations where agents are exchangeable. A sub-population of agents is called exchangeable if the manner in which agents are indexed does not affect the dynamics and cost. In practice, this insensitivity to the index naturally emerges in many applications. For example, in power systems, the system dynamics and cost would not change if the houses in a residential neighborhood were numbered differently; in swarm robotics, the dynamics and cost depend on the position of the robots, not on how the agents are indexed. We first show that a system with partially exchangeable agents is equivalent to a system where agents are coupled in the dynamics and cost through the aggregate behavior of agents (called mean-field). Then, we investigate and identify the optimal strategy—under mean-field sharing information structure which is non-classical—for two different models: linear quadratic and controlled Markov chain. We show that the optimal strategies, unlike the existing results in team theory, are scalable to large scale systems. We use the theory to solve idealized models of demand response in power systems and resource allocation in networks.

In the second part, we study systems with partial history sharing information structure—which encompasses a large class of team problems including mean-field teams—when agents do not know the complete model of the system. The agents must learn the optimal strategies by interacting with their environment using reinforcement learning. We develop a reinforcement learning algorithm that guarantees ε-team-optimal performance. As an intermediate step of this development, we revisit the well-known partially observable Markov decision process and propose a novel approach to find an ε-optimal solution. The novelty of this approach is to identify the planning space based on the structure of the model. To illustrate the algorithm, we develop a reinforcement learning algorithm for the benchmark example of two-user multi access broadcast channel and present numerical results.

ABRÉGÉ

Cette thèse se compose de deux parties dans lesquels chaque partie introduit un nouveau concept dans la théorie de l'équipe.

Dans la première partie, nous introduisons des systèmes avec des agents patiellements échangeables. Un système est appelé partiellement échangeable si elle peut être divisée dans des sous-populations où les agents sont échangeables. Une sous-population d'agents est appelé échangeables si la manière dont les agents sont indexés n'a aucune incidence sur la dynamique et le coût de celle-ci. En pratique, cette insensibilité à l'indexation émerge dans de nombreuses applications naturelles. Par exemple, dans les systèmes de puissance, la dynamique du système et le coût ne changerait pas si les maisons dans un quartier résidentiel ont été numérotés différemment; en robotique en essaim, la dynamique et le coût dépend de la position des robots, et non de la manière dont les agents sont indexés. Premièrement, nous montrons qu'un système avec des agents partiellement échangeables est équivalent à un système où la dynamique et le coût sont couplés à travers le comportement global des agents (appelés champ moyen). Ensuite, nous enquêtons et identifions la solution optimale —sous la structure du partage des information du champ moyen qui n'est pas classique—de deux modèles différents: la linéaire quadratique ainsi que la chaîne contrôlée de Markov. Nous démontrons que les solutions optimales, contrairement aux résultats existants dans la théorie de l'équipe, sont évolutives aux systèmes à grande échelle. Nous utilisons cette théorie pour résoudre des modèles idéalisés de réponse à la demande dans les systèmes d'alimentation et de l'allocation des ressources dans les réseaux.

Dans la deuxième partie, nous étudions les systèmes avec une structure de partage d'information historique partielle—qui englobe une grande classe de problèmes de l'équipe, y compris les équipes à champ moyens—où les agents ne connaissent pas le modèle complet du système. Ces agents doivent apprendre les stratégies optimales en interagissant avec leur environnement en utilisant l'apprentissage par renforcement. Nous développons un algorithme de renforcement d'apprentissage qui garantit solution ε-optimale performance. Comme une étape intermédiaire de ce développement, nous revissions le processus de décision partiellement observable bien connu de Markov et nous proposons une nouvelle approche pour apprendre solution ε-optimale. La nouveauté de cette approche est d'identifier l'espace de planification basé sur la structure du modèle. Pour illustrer l'algorithme, nous développons un algorithme d'apprentissage de renforcement pour deux utilisateurs d'un accès multi canal

de diffusion, qui est un exemple de référence.

ACKNOWLEDGMENTS

First, I would like to thank my Ph.D. supervisor, Professor Aditya Mahajan, for providing me with the opportunity to do my Ph.D. thesis at McGill University. I want to thank him for his support and guidance. I am grateful for his motivation, enthusiasm, patience, and immense knowledge in the control theory. I also appreciate the opportunities he has provided me to attend conferences and workshops and to interact with leading researchers around the world.

I would also like to acknowledge my committee members: Professors Peter E. Caines and Ioannis Psaromiligkos from the department of Electrical and Computer Engineering of McGill University, who graciously agreed to serve on my committee,

I am grateful for the financial support throughout my study provided by McGill Engineering Doctoral Award (MEDA), the Natural Sciences and Engineering Research Council of Canada (NSERC). I am also thankful for doctoral fellowship from GERAD[1], Greville Smith and Geoff Hyland Fellowships in Engineering, and other awards from McGill University.

I want to thank the staff and members of the McGill Centre for Intelligent Machines (CIM) and acknowledge the professional services of CIM, GERAD and REPARTI[2].

I would like to thank my fellow labmates: Hossein Alizadeh, Ali Pakniyat, Prokopis Prokopiou, Jhelum Chakravorty, Mohammad Afshari, and Jayakumar Subramanian for their positive energies and sincere suggestions throughout these five years.

Finally, I would like to acknowledge my family who supported me during these five years. I would like to thank my father, Mandani Arabneydi, my mother, Zinat Maroofi, my older brother Hossein Arabneydi, my younger brother Reza Arabneydi, and my best friend Hadi Amirheidari for their constant love and support.

Jalal Arabneydi
McGill University, Montreal, Quebec
September, 2016.

[1]Groupe d' Études et de Recherche en Analyse des Décisions (GERAD) Institute (Group for Study and Research in Decision Analysis)

[2]Regroupement pour l' étude des Environnements PARTagés Intelligents (Centre for the Study of Distributed Intelligent Shared Environments)

CLAIMS OF ORIGINALITY AND PUBLISHED WORK

Claims of originality

The following original contributions are presented in this thesis:

Chapter 2

We introduce the notion of systems with partially exchangeable agents.

1. We show that any linear quadratic system with partially exchangeable agents—irrespective of the information structure—is equivalent to a mean-field coupled system with the same information structure. In this case, the mean-field is empirical average.

2. We show that any controlled Markov chain system with partially exchangeable agents—irrespective of the information structure—is equivalent to a mean-field coupled system with the same information structure. In this case, the mean-field is empirical distribution.

Chapter 3

We study linear quadratic mean-field teams. The related publications are [5,6].

1. The linear quadratic mean-field team under mean-field sharing information structure (MFS-IS) is a decentralized system with non-classical information structure that is neither partially nested nor quadratic invariant; yet we show linear control laws are optimal.

2. Let K denote the number of sub-populations; then, we obtain the corresponding optimal gains by $K + 1$ *decoupled* Riccati equations: one for each sub-population and one for the mean-field term. In fact, each agent simply needs to solve two Riccati equations: one corresponding to its own sub-population and one to the mean-field (i.e., distributed and decentralized implementation). The dimensions of these Riccati equations do not depend on the number of agents in each sub-population. Thus, the solution complexity does not depend on the number of agents in the system (i.e., the solution is scalable for large sub-populations).

3. In general, the matrices in the dynamics and cost are allowed to depend on $\mathcal{N}$ (the number of agents) although their size only depends on K (the number of sub-populations). However, if these matrices do not depend on the number of agents, then neither do the optimal gains. Consequently, the agents *need not even be aware of the number of agents.*

4. We show that the centralized performance is achieved by sharing only the mean-field (which can be shared using distributed algorithms such as consensus [Xiao and Boyd, 2004]). Thus, instead of requiring that agents have the capability or the energy to communicate to a centralized controller, agents only require the capacity or the energy to communicate to their neighbours.

5. For partial mean-field sharing information structure (PMFS-IS), we propose a linear strategy that is approximately optimal where the approximation error is inversely proportional to the size of sub-populations whose mean-fields are not observed. The proposed strategy is a certainty equivalence strategy in which all agents generate an estimate of the unobserved components of the mean-field using the observed components of the mean-field. This estimate is used in the optimal strategy identified for MFS-IS. We show that the approximation error between the proposed strategy for PMFS-IS and the optimal strategy for MFS-IS is given by terms of the weighted cost of a linear system, which can be computed by a Lyapunov equation.

6. As a consequence of the PMFS-IS result, when a sub-population is large, sharing its mean-field has a vanishingly small advantage. In the extreme case when all sub-populations are large, the approximately optimal solution can be implemented under a completely decentralized information structure.

7. We show that our results generalize to several variations of the basic mean-field model including: (i) systems where a major agent interacts with a collection of minor agents in Section 3.5.1, (ii) systems where agents have individual tracking cost in Section 3.5.2, and (iii) systems where agents have individual weights in Section 3.5.3. In Section 3.6, we show our results generalize to infinite horizon setup using standard arguments.

Chapter 4

We study controlled Markov chain mean-field teams. The related publications are [1,2].

1. We identify a dynamic program to obtain globally optimum control strategies.

2. We show that the mean-field is an information state for the dynamic program. Since the mean-field is observed by each agent, each agent can independently solve the dynamic program by agreeing upon a deterministic rule to break ties, while using arg min, which ensures that all agents compute the same optimal strategy (i.e., decentralized implementation). In addition, agents may compute the mean-field in a distributed manner using methods such as consensus-based algorithms [Olfati-Saber et al., 2006, Bishop and Doucet, 2014].

3. The size of the mean field does not increase with time. Thus, our results extend naturally to infinite horizon setups.

4. The size of the mean field increases polynomially with the number of agents rather than exponentially. This allows us to solve problems with moderate number of agents. (In Sections 4.7 and 4.8, we give two examples with $n = 100$ agents).

5. The solution methodology and dynamic programming decomposition extend to the scenario where all agents observe a noisy version of the mean-field.

6. We generalize our results to arbitrary coupled per-step cost, heterogeneous population, system with a major agent and a sub-populations of minor agents, and randomized strategy.

7. We present salient features of Markov chains with exchangeable transition probability.

Chapter 5

We propose a novel approach to compute an approximate solution of centralized POMDP. The key feature of this approach is the ability to use MDP solvers for finite state MDPs to find an $\varepsilon-$optimal solution when the model is known, partly known, or not known.

1. We define a new notion, that we call Incrementally Expanding Representation (IER), based on the notion of information state (not necessarily belief state) that allows us to exploit the structure of the model. Unlike many existing approaches that treat all POMDPs similarly, this approach is able to exploit the structure of the model and use

that to identify an efficient planning space. See the example of machine maintenance in Section 5.4 where this feature contributes to an efficient solution.

2. The proposed approach provides a framework for reinforcement learning in POMDP. In particular, this framework allows us to utilize any MDP reinforcement learning algorithm in order to to learn an $\varepsilon-$optimal solution of POMDP.

Chapter 6

We propose a decentralized reinforcement learning algorithm for systems with partial history sharing information structure and we show that the proposed algorithm learns $\varepsilon-$optimal solution. The related publications are [3,4].

1. We propose a novel approach to perform reinforcement learning in a large class of decentralized stochastic control systems with partial history sharing (PHS) information structure that guarantees ϵ-team-optimal solution. To the best of the author's knowledge, none of the existing RL approaches guarantee team-optimal (or ϵ-team-optimal) solution.

2. We illustrate the proposed approach and verify it numerically by designing a decentralized Q-learning algorithm for two-user multi-access broadcast channel, which is a benchmark example for decentralized control systems.

Appendix C

We present an alternative proof for common information approach [Nayyar et al., 2013] that sheds new light on this general approach.

Appendix D

Since we use Q-learning algorithm for our numerical examples, we present a preliminary document on Q-learning algorithm and its proof based on [Tsistsiklis, 1994, Bertsekas, 1998].

Publications

1. Jalal Arabneydi and Aditya Mahajan, "Team Optimal Control of Coupled Subsystems with Mean-Field Sharing", IEEE Conference on Decision and Control (CDC), pp.

1669-1674, Dec., 2014.

2. Jalal Arabneydi and Aditya Mahajan,"Team Optimal Control of Coupled Major-Minor Subsystems with Mean-Field Sharing", IEEE Indian Control Conference (ICC), pp. 95-100, Jan., 2015.

3. Jalal Arabneydi and Aditya Mahajan, "Reinforcement Learning in Multi-Agent Systems with Partial History Sharing", The 2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM), June, 2015.

4. Jalal Arabneydi and Aditya Mahajan, "Reinforcement Learning in Decentralized Stochastic Control Systems with Partial History Sharing", IEEE American Control Conference (ACC), pp. 5449–5456, Jul., 2015.

5. Jalal Arabneydi and Aditya Mahajan, "Team Optimal Solution of Finite Number of Mean-Field Coupled LQG Subsystems", IEEE Conference on Decision and Control (CDC), pp. 5308 - 5313, Dec., 2015.

6. Jalal Arabneydi and Aditya Mahajan, "Team Optimal Decentralized Control of System with Partially Exchangeable Agents—Part 1: Linear Quadratic Mean Field Teams", Submitted to IEEE Transactions on Automatic Control, August 2016.

Contribution of co-authors

Professor Aditya Mahajan contributed in the problem formulations and their analyses. These contributions amounted to 25% of the papers cited above.

Contents

List of Figures

List of Tables

LIST OF ACRONYMS

FPK:	Fokker-Planck-Kolmogorov
HJB:	Hamilton-Jacobi-Bellman
IER:	Incrementally Expanding Representation
LQ:	Linear Quadratic
MDP:	Markov Decision Process
MABC:	Multi Access Broadcast Channel
MFG:	Mean Field Game
MFS-IS:	Mean Field Sharing Information Structure
PDE:	Partial Differential Equation
PHS:	Partial History Sharing
PMFS-IS:	Partial Mean Field Sharing Information Structure
POMDP:	Partially Observable Markov Decision Process
NMFS-IS:	Noisy Mean Field Sharing Information Structure
ODE:	Ordinary Differential Equation
RL:	Reinforcement Learning

CHAPTER 1

Introduction

1.1 Motivation

Team theory studies multiple decision makers that wish to collaborate in order to accomplish a common task. The salient features of teams are as follows.

- *Multiple decision makers:* There are more than one decision maker (agent).
- *Decentralized information:* The decision makers are allowed to have different information (perspective) about the other decision makers and the environment.
- *Common objective:* The objective of all the decision makers is the same.

In practice, team theory arises in various applications ranging from smart grids, social networks, robotics, network controlled systems, transportation networks, communication networks, sensor networks, and economics.

There is a long and rich history of research on team theory, starting from the work of Radner [Radner, 1962, Marschack and Radner, 1972], Witsenhausen [Witsenhausen, 1968, Witsenhausen, 1971, Witsenhausen, 1973] and others; and continuing to various solution approaches that have been proposed in recent years. The reader is referred to [Yüksel and Başar, 2013, Mahajan et al., 2012] for detailed overviews. However, many of the initial research results were negative and showed that even simple dynamical systems with two agents can be difficult to optimally control [Witsenhausen, 1968, Whittle and Rudge, 1974]. Since then, various solution methodologies for team optimal control have been proposed and

Fig. 1.1 Teams are almost everywhere.

there has been considerable progress in understanding the nature of system dynamics and the information structure under which these methodologies work. See [Mahajan et al., 2012] and references therein for an overview.

In spite of this progress, there is a big gap between the team theory and applications. On the one hand, the envisioned applications—which include networked control systems, swarm robotics, and modern power systems—often consist of multiple interconnected dynamical systems and agents. On the other hand, *explicit* optimal solutions are available for systems with only a few (often two or three) agents [Ouyang and Teneketzis, 2015,Lipsa and Martins, 2011b, Lessard and Lall, 2015]. In addition, most of the literature on team theory assumes that the agents know the system model *completely*; however, in practice, this might not be the case. Therefore, it is important for agents to be able to *learn* the optimal solution. In the literature, learning in centralized stochastic control is well studied and there exist many approaches such as model-predictive control, adaptive control, and reinforcement learning. This is in contrast to the learning in decentralized stochastic control; it is not immediately clear on how centralized learning approaches would work for decentralized systems. To the best of author's knowledge, there is no algorithm in the literature that guarantees team-optimal solution.

The first part of this thesis attempts to reduce the gap between the theory and applications by introducing mean-field teams whose optimal solution can be computed explicitly for the moderate and large scale systems. The second part of this thesis introduces a novel decentralized reinforcement learning algorithm that guarantees $\varepsilon-$optimal solution for a large class of team problems.

1.2 Scope of this thesis

This thesis has two folds. In the first fold, we study optimal control of stochastic decentralized systems in which the dynamics and cost satisfy a property that we call *exchangeability*. In a dynamical system, we say agents i and j are *exchangeable* if exchanging (or interchanging) agents i and j does not affect the dynamics or the cost. Or, equivalently, the dynamics and the cost do not depend on the index assigned to the two agents. In many applications of decentralized systems, the system may be partitioned into sub-populations where all agents within a sub-population are exchangeable. In Chapter 2, we show such systems are equivalent to systems where the agents are coupled in both dynamics and cost only through the aggregate behavior of agents (called mean-field). We develop a framework for the design of optimal decentralized control of two models: linear quadratic and controlled Markov chain.

In the second fold, we study stochastic decentralized systems in which agents do not know the complete model of the system. The agents must learn the optimal strategies by interacting with their environment, i.e., by decentralized Reinforcement Learning (RL). In particular, we develop a decentralized reinforcement learning algorithm that learns ε-team-optimal solution for partial history sharing information structure [Nayyar et al., 2013], which encompasses a large class of decentralized control systems including delayed sharing [Nayyar et al., 2011], control sharing [Mahajan, 2013], mean-field sharing [Arabneydi and Mahajan, 2014], etc.

In principle, the proposed decentralized RL algorithm can be used for mean-field teams so that the agents can learn ϵ-team-optimal strategy when the complete model is not available. In this thesis, however, we do not study this combination and leave it as future work.

1.3 Main challenges, ideas, and results of this thesis

In this section, we briefly describe mean-field teams and the proposed decentralized reinforcement learning and present the main challenges, ideas, and results.

Part 1: Mean-field teams

We first introduce the notion of systems with partially exchangeable agents and show that they are equivalent to systems in which agents are coupled through the mean-field (aggregate behavior). For this reason, we call these systems mean-field teams. Then, we study and find the optimal solution of two different models of mean-field teams: linear quadratic and controlled Markov chain.

Systems with partially exchangeable agents

There are many natural applications where the manner in which agents are numbered does not matter. For example, in power systems, the system dynamics and cost would not change if the houses in a residential neighborhood were numbered differently; in swam robotics, the dynamics and cost depend on the position of the robots, not on how we index them. When agents are modeled as linear quadratic (or controlled Markov chain), we show that

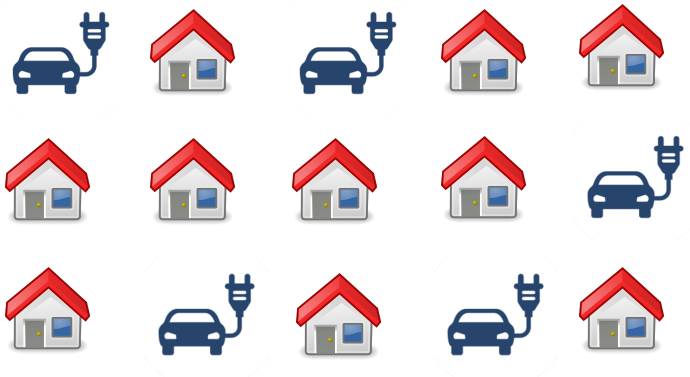

Fig. 1.2 The notion of partially exchangeable agents in smart grids. The manner in which residential houses in a neighborhood or electric vehicles of the same size are numbered does not matter.

the mean-field is the empirical average (or the empirical distribution), respectively. Notice that the notion of exchangeable agents is different from the notion of exchangeable random variables. For more details see Remark 2.1.

Remark 1.1 In mean-field games [Huang et al., 2003, Lasry and Lions, 2006a, Lasry and Lions, 2006b, Huang et al., 2007, Li and Zhang, 2008, Caines, 2013], the term *mean-field* refers to the expectation and the probability distribution of infinite population while in mean-field teams, it refers to the empirical average and the empirical distribution of finite population[1].

Linear quadratic mean-field teams

Two information structures are investigated: mean-field sharing and partial mean-field sharing. In the former, all agents observe their local state and the mean-field of all sub-populations; in the latter, all agents observe their local state but the mean-field of only a subset of the sub-populations. Both information structures are non-classical. In was first shown by the celebrated Witsenhausen's counterexample[2] that Linear Quadratic Gaussian (LQG) systems with non-classical information structure are difficult. Since then, other counterexamples are introduced [Whittle and Rudge, 1974, Lipsa and Martins, 2011a, Yuksel and Tatikonda, 2009] that show the optimal strategies are not necessarily linear. The only known information structure for which the linear strategies are optimal is partially nested. However, the mean-field sharing is not partially nested and the noises are allowed to be non-Gaussian.

For the mean-field sharing, we first construct an auxiliary centralized system; second, we use a linear transformation, parallel axis theorem, and certainty equivalence theorem to decouple the optimal design into the local design and global design without loss of optimality[3]; third, we show the centralized solution is implementable under the mean-field sharing. We show the optimal strategy is unique, identical across the sub-populations, and linear. For the partial mean-field sharing, we present an $\varepsilon-$optimal strategy. In particular, we propose a certainty equivalence strategy that estimates the unobserved components of the mean-field using the observed components of the mean-field. We characterize the approximation error by a Lyapunov equation and show that the error converges to zero as the sub-populations, whose mean-fields are not observed, grow to infinity. The computational complexity of our

[1] Notice that the empirical average and distribution converge to the expected value and the probability distribution in infinite population due to the law of large numbers.

[2] It is still an open problem after 48 years.

[3] This decoupling holds for any arbitrary number of agents (not necessarily large population)

solution does not depend on the size of population and it depends only on the number of sub-populations.

Controlled Markov chain mean-field teams

Two information structures are investigated: mean-field sharing and noisy mean-field sharing. In the former, all agents observe their local state and the mean-field of all sub-populations; in the latter, all agents observe their local state but the noisy version of mean-field. Both information structures are non-classical. This problem is conceptually challenging because, in general, team optimal control problems with non-classical information structure belong to NEXP[4] complexity class [Bernstein et al., 2002]. Although it is possible to get a dynamic programming decomposition for problems with non-classical information structure [Witsenhausen, 1973], the size of the corresponding information state increases with time. For some information structures, we can find information states that do not increase with time [Mahajan et al., 2012], but even for these models the size of the information state increases exponentially with the number of agents.

For the both information structures, we restrict attention to exchangeable control laws in each sub-population. Let us call the mapping from the local state to the local action *local rule*. Then, we show that the mean-field is a sufficient statistic and it evolves in Markovian manner under the local rules (not under actions). Based on this result, we develop a non-standard dynamic program where the minimization is over the local rules (i.e., function minimization). This dynamic program identifies a global optimal strategy. The computational complexity of solving this dynamic program is polynomial in the number of agents and linear in time.

Part 2: Decentralized reinforcement learning

The presence of multiple agents with different information makes decentralized reinforcement learning conceptually more difficult than centralized reinforcement learning. Most of the existing RL algorithms [Sutton and Barto, 1998] are designed for Markov Decision Processes (MDPs); however, a decentralized control system is not MDP in general. To overcome this difficulty, we use the existing approach called common information approach for the decentralized control systems with partial history sharing (PHS) [Nayyar et al., 2013]. Following

[4]If the horizon is larger than the number of agents, then it is NEXP-hard and if the horizon is limited to be less than the number of agents it is NEXP-complete [Bernstein et al., 2002].

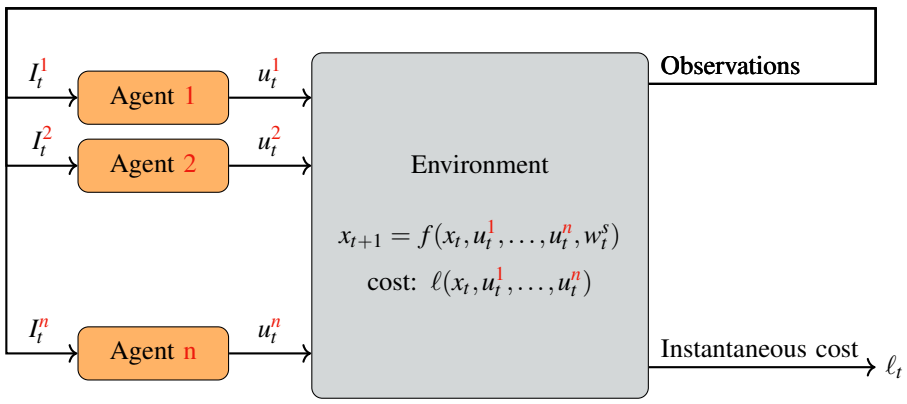

Fig. 1.3 General decentralized reinforcement learning problem.

[Nayyar et al., 2013] the decentralized stochastic control problem is transformed into an equivalent centralized Partially Observed Markov Decision Process (POMDP). For more information on the common information approach and PHS information structure, we refer the reader to Appendix C. In principle, any RL of POMDP may be used to learn the resultant centralized POMDP. However, reinforcement learning in POMDP is difficult because

- when the model is known, the finite-horizon POMDP is PSpace-complete [Papadimitriou and Tsitsiklis, 1987] and the infinite-horizon POMDP is undecidable [Madani et al., 1999]; hence, finding an optimal solution is more difficult when the model in not completely known.

- most of the existing RL algorithms [Sutton and Barto, 1998] are designed for the finite state-action MDPs and do not directly work for POMDPs. For example, the standard Q-learning algorithm may not be directly used in the belief-state MDP because: (a) the belief space is uncountable and (b) knowing the belief state itself requires the knowledge of the model.

Our approach consists of two main steps. In the first step, we convert the decentralized control system to an equivalent centralized POMDP using the common information approach. This conversion is performed from the point of view of a virtual coordinator that observes the common information among all decision makers and chooses partially evaluated functions that map the local information to the actions. The coordinator's problem is a centralized POMDP. Therefore, any learning algorithm for centralized (partially observed) systems may

be used for the coordinated system. However, since the reinforcement learning in POMDP is difficult, we develop a novel approach to solve centralized POMDPs. In particular, our main focus is on the ability of the approach to be applicable for the RL purposes. To do this, we revisit the centralized POMDP and introduce a notion that we call Incrementally Expanding Representation (IER) that is based on the notion of information state. The key feature of the IER is that it allows us to take the model structure into account and come up with an efficient planning space. To illustrate the importance of the planning space, consider the belief space which is widely used as the planning space for many existing POMDP solvers because the value function is piece wise linear and convex in the belief state. For the reinforcement learning purposes, however, the belief space may not be efficient because (i) knowing the belief space requires the knowledge of the dynamics and (2) knowing the value function requires the knowledge of the cost. Hence, the main challenge in designing an RL algorithm of POMDP is to find an efficient planning space.

1.4 Organization of this thesis

Part 1: mean-field teams

In Chapter 2, we introduce systems with partially exchangeable agents. We define the notion of partial exchangeable agents and show that the linear quadratic and controlled Markov chain systems are equivalent to mean-field coupled systems.

In Chapter 3, we study linear quadratic systems with partially exchangeable agents under two information structures: Mean-field Sharing Information Structure (MFS-IS) and Partial Mean-Field Sharing Information Structure (PMFS-IS). In MFS-IS, all agents observe their local state and the mean-field of all sub-populations; in PMFS-IS, all agents observe their local state but the mean-field of only a subset of the sub-populations. We generalize our results to major-minor agents, tracking cost, weighted mean-field, and infinite horizon setups. We illustrate our results using an example of demand response in smart grids.

In Chapter 4, we study a class of controlled Markov chain systems with partially exchangeable agents in which agents of the same sub-population have identical dynamics. Two information structures are investigated: Mean-Field Sharing Information Structure (MFS-IS) and Noisy Mean-Field Sharing Information Structure (NMFS-IS). In MFS-IS, all agents observe their local state and the mean-field of all sub-populations; in NMFS-IS, all agents

observe their local state but a noisy version of the mean-field. We generalize our results to arbitrary coupled cost, heterogeneous population, and major-minor. We illustrate our approach by two examples: the first is motivated by smart grids and the second by economics.

Part 2: Decentralized reinforcement learning

In Chapter 5, we present a novel approach for an approximate solution of centralized Partially Observable Markov Decision Process (POMDP) that guarantees $\varepsilon-$optimal performance when the model is known, partly known, or not known.

In Chapter 6, we study systems with multiple agents that wish to collaborate in order to accomplish a common task while a) the agents have different information (decentralized information) and b) the agents do not know the model of the system completely. We develop a decentralized reinforcement learning algorithm that learns ϵ-team-optimal solution for the partial history sharing information structure. We illustrate the proposed approach and verify it numerically by obtaining a decentralized Q-learning algorithm for two-user Multi Access Broadcast Channel (MABC) which is a benchmark example for decentralized control systems.

CHAPTER 2

Systems with partially exchangeable agents

Consider a multi-agent dynamical system where $\mathcal{N}$ denotes the set of agents. The state, action, and noise of agent $i, i \in \mathcal{N}$, at time t are denoted by x_t^i, u_t^i, and w_t^i where $x_t^i \in \mathcal{X}^i$, $u_t^i \in \mathcal{U}^i$, and $w_t^i \in \mathcal{W}^i$. Let $\mathbf{x}_t = (x_t^i)_{i\in\mathcal{N}}$, $\mathbf{u}_t = (u_t^i)_{i\in\mathcal{N}}$, and $\mathbf{w}_t = (w_t^i)_{i\in\mathcal{N}}$ denote the state, action, noise of the entire system. The dynamics are given by

$$\mathbf{x}_{t+1} = f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t), \tag{2.1}$$

where f_t is system dynamics and the primitive random variables $\{\mathbf{x}_1, \{\mathbf{w}_t\}_{t\geq 1}\}$ are defined on a common probability space. A per-step cost $c_t(\mathbf{x}_t, \mathbf{u}_t)$ is incurred at each time t.

For now, we do not specify the information structure as we want to identify the system properties that do not depend on the information structure.

For any state $\mathbf{x}$ and agents $i, j \in \mathcal{N}$, let $\sigma_{i,j}\mathbf{x}$ denote the state when agents i and j are exchanged. For example, if $\mathbf{x} = (x^1, x^2, x^3, x^4, x^5)$, then $\sigma_{2,4}\mathbf{x} = (x^1, x^4, x^3, x^2, x^5)$. Similar interpretation holds for $\sigma_{i,j}\mathbf{u}$ and $\sigma_{i,j}\mathbf{w}$.

Definition 2.1 (Exchangeable agents) A pair (i, j) of agents is exchangeable if the following conditions hold:

1) For any t, and any $\mathbf{x}$, $\mathbf{u}$, and $\mathbf{w}$,

$$\sigma_{i,j}\big(f_t(\mathbf{x}, \mathbf{u}, \mathbf{w})\big) = f_t(\sigma_{i,j}\mathbf{x}, \sigma_{i,j}\mathbf{u}, \sigma_{i,j}\mathbf{w}),$$

i.e., exchanging agents i and j does not affect the system dynamics.

2) For any t, and any $\mathbf{x}$ and $\mathbf{u}$,

$$c_t(\mathbf{x}, \mathbf{u}) = c_t(\sigma_{i,j}\mathbf{x}, \sigma_{i,j}\mathbf{u}),$$

i.e., exchanging agents i and j does not affect the cost.

Definition 2.2 (Exchangeable set of agents) A set $\mathcal{S}$ of agents, $\mathcal{S} \subseteq \mathcal{N}$, is exchangeable if every pair of agents in $\mathcal{S}$ is exchangeable.

Definition 2.3 (System with partially exchangeable agents) The multi-agent described above is called a *system with partially exchangeable agents* if the set $\mathcal{N}$ of agents can be partitioned into K disjoint subsets $\mathcal{N}^k$, $k \in \mathcal{K} \coloneqq \{1, \dots, K\}$, such that for each $k \in \mathcal{K}$, the set $\mathcal{N}^k$ of agents is exchangeable.

Remark 2.1 Note that the notion of exchangeability defined above is in the sense of model structure and it is different from the notion of exchanegability of random variables (in the sense of probability). To elaborate on this difference, consider two decoupled agents with dynamics $x^i_{t+1} = f^i_t(x^i_t, u^i_t, w^i_t), i \in \{1, 2\}$ and the per-step cost $c^1_t(x^1_t, u^1_t) + c^2_t(x^2_t, u^2_t)$.

- Let $f^1_t = f^2_t$ and $c^1_t = c^2_t$; hence, the agents are exchangeable. However, the primitive random variables $\{x^1_1, x^2_1\}$ and $\{w^1_t, w^2_t\}$ need not be exchangeable (in probability).
- Let $x^1_1 = x^2_1$ and $w^1_t = w^2_t$; hence, the primitive random variables are exchangeable (in probability). However, agents need not be exchangeable (when $f^1_t \neq f^2_t$ and $c^1_t \neq c^2_t$).

2.1 Linear quadratic systems

Suppose the dynamics (2.1) are linear, i.e.,

$$\mathbf{x}_{t+1} = A_t\mathbf{x}_t + B_t\mathbf{u}_t + \mathbf{w}_t, \tag{2.2}$$

where A_t and B_t are matrices of appropriate dimensions. The cost is quadratic, i.e., for $t \in \{1, \dots, T-1\}$,

$$c_t(\mathbf{x}_t, \mathbf{u}_t) = \mathbf{x}_t{}^\intercal Q_t \mathbf{x}_t + \mathbf{u}_t{}^\intercal R_t \mathbf{u}_t, \tag{2.3}$$

and $t = T$,

$$c_T(\mathbf{x}_T) = \mathbf{x}_T{}^\intercal Q_T \mathbf{x}_T, \tag{2.4}$$

where Q_t and R_t are matrices of appropriate dimensions. Furthermore, assume that the above system is partially exchangeable, i.e., agents $\mathcal{N}$ can be partitioned into K disjoint sub-populations $\mathcal{N}^k$, $k \in \mathcal{K} \coloneqq \{1, \dots, K\}$, such that for each $k \in \mathcal{K}$, the agents $\mathcal{N}^k$ are exchangeable. Moreover, for any sub-population $k \in \mathcal{K}$ and agent $i \in \mathcal{N}^k$, state x_t^i takes values in $\mathbb{R}^{d_x^k}$ and action u_t^i takes values in $\mathbb{R}^{d_u^k}$.

The *mean-field* of states $\bar{x}_t^k$ of sub-population k, $k \in \mathcal{K}$, is defined as the empirical mean of the states of all agents in that sub-population, i.e.,

$$\bar{x}_t^k \coloneqq \langle (x_t^i)_{i \in \mathcal{N}^k} \rangle = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} x_t^i, \quad k \in \mathcal{K}.$$

Similarly, the mean-field of the actions $\bar{u}_t^k$ of sub-population k, $k \in \mathcal{K}$, is defined as the empirical mean of the actions of all agents in that sub-population, i.e.,

$$\bar{u}_t^k \coloneqq \langle (u_t^i)_{i \in \mathcal{N}^k} \rangle = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} u_t^i, \quad k \in \mathcal{K}.$$

The mean-field of states and actions of the entire population are denoted by $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{u}}_t$ respectively, i.e.,

$$\bar{\mathbf{x}}_t = \operatorname{vec}(\bar{x}_t^1, \dots, \bar{x}_t^K), \quad \bar{\mathbf{u}}_t = \operatorname{vec}(\bar{u}_t^1, \dots, \bar{u}_t^K).$$

Proposition 2.1 *In the linear quadratic system with partially exchangeable agents described above, there exist matrices* $\{A_t^k, B_t^k, D_t^k, E_t^k, Q_t^k, R_t^k\}_{k \in \mathcal{K}}$ *and* P_t^x *and* P_t^u *such that the dynamics of agent* $i \in \mathcal{N}^k$ *of sub-population* $k, k \in \mathcal{K}$*, may be written as*

$$x_{t+1}^i = A_t^k x_t^i + B_t^k u_t^i + D_t^k \bar{\mathbf{x}}_t + E_t^k \bar{\mathbf{u}}_t + w_t^i; \tag{2.5}$$

the per-step cost at time $t \in \{1, \dots, T-1\}$*, may be written as*

$$c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \bar{\mathbf{x}}_t^\intercal P_t^x \bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^\intercal P_t^u \bar{\mathbf{u}}_t + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[(x_t^i)^\intercal Q_t^k x_t^i + (u_t^i)^\intercal R_t^k u_t^i \right]; \tag{2.6}$$

and the per-step cost at time $t = T$*, may be written as*

$$c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T) = \bar{\mathbf{x}}_T^\intercal P_T^x \bar{\mathbf{x}}_T + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} (x_T^i)^\intercal Q_T^k x_T^i. \tag{2.7}$$

Thus, any linear quadratic system with partial exchangeable agents—irrespective of the in-

formation structure—is equivalent to a mean-field coupled system with the same information structure.

Proof of Proposition 2.1

Let $A_t^{i,j}$ denote the (i,j)-th element of matrix A_t. We use a similar notation for other matrices as well. Fix a sub-population $k, k \in \mathcal{K}$. If we exchange agents $i, j \in \mathcal{N}^k$, then property 1 of Definition 2.1 implies that $A_t^{i,i} = A_t^{j,j}$ and for any other agent $n \in \mathcal{N}$, $A_t^{i,n} = A_t^{j,n}$ and $A_t^{n,i} = A_t^{n,j}$. (Similar relationships hold for B_t as well). Property 2 of Definition 2.1 implies that $Q_t^{i,i} = Q_t^{j,j}$, $Q_t^{i,n} = Q_t^{j,n}$, and $Q_t^{n,i} = Q_t^{n,j}$. (Similar relationships hold for R_t as well). Define the following:

- For $i, j \in \mathcal{N}^k, A_t^{i,i} = A_t^{j,j}$ and $B_t^{i,i} = B_t^{j,j}$. Denote these by a_t^k and b_t^k, respectively.
- For $i, j \in \mathcal{N}^k$ and $n, m \in \mathcal{N}^l, l \neq k, A_t^{i,n} = A_t^{j,m}$ and $B_t^{i,n} = B_t^{j,m}$. Denote these by $d_t^{k,l}$ and $e_t^{k,l}$, respectively.
- For $i, j \in \mathcal{N}^k, Q_t^{i,i} = Q_t^{j,j}$ and $R_t^{i,i} = R_t^{j,j}$. Denote these by q_t^k and r_t^k, respectively.
- For $i, j \in \mathcal{N}^k$ and $n, m \in \mathcal{N}^l, l \neq k, Q_t^{i,n} = Q_t^{j,m}$ and $R_t^{i,n} = R_t^{j,m}$. Denote these by $p_t^{x,k,l}$ and $p_t^{u,k,l}$, respectively.

Now, consider the dynamics according to (2.2), the dynamics of agent i of sub-population k can be written as

$$x_{t+1}^i = A^{i\bullet}\mathbf{x}_t + B^{i\bullet}\mathbf{u}_t + w_t^i, \tag{2.8}$$

where $A^{i\bullet}$ and $B^{i\bullet}$ denote the ith row of A_t and B_t. Note that

$$\begin{aligned} A^{i\bullet}\mathbf{x}_t &= A_t^{i,i}x_t^i + \sum_{j\in\mathcal{N}^k, j\neq i} A_t^{i,j}x_t^j + \sum_{l\in\mathcal{K}, l\neq k}\sum_{n\in\mathcal{N}^l} A^{i,n}x_t^n \\ &= a_t^k x_t^i + d_t^{k,k}\sum_{j\in\mathcal{N}^k, j\neq i} x_t^j + \sum_{l\in\mathcal{K}, l\neq k} d_t^{k,l}\sum_{n\in\mathcal{N}^l} x_t^n \\ &= a_t^k x_t^i + d_t^{k,k}(|\mathcal{N}^k|\bar{x}_t^k - x_t^i) + \sum_{l\in\mathcal{K}, l\neq k} d_t^{k,l}|\mathcal{N}^l|\bar{x}_t^l \\ &=: A_t^k x_t^i + \sum_{l\in\mathcal{K}} D_t^{k,l}\bar{x}_t^l, \end{aligned} \tag{2.9}$$

where $A_t^k = a_t^k - d_t^{k,k}$ and $D_t^{k,l} = |\mathcal{N}^l| d_t^{k,l}$. By a similar algebra, we can define B_t^k and $E_t^{k,l}$ such that

$$B_t^{i\bullet}\mathbf{u}_t = B_t^k u_t^i + \sum_{l\in\mathcal{K}} E_t^{k,l}\bar{u}_t^l, \tag{2.10}$$

where $B_t^k = b_t^k - e_t^{k,k}$ and $E_t^{k,l} = |\mathcal{N}^l| e_t^{k,l}$. Substituting (2.9) and (2.10) in (2.8), we get (3.1). Now consider the per-step cost given by (2.3). Note that

$$\begin{aligned}
\mathbf{x}_t^\intercal Q_t \mathbf{x}_t &= \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}\sum_{j\in\mathcal{N}^l}(x_t^i)^\intercal Q_t^{i,j} x_t^j = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{K},l\neq k}\sum_{i\in\mathcal{N}^k}\sum_{j\in\mathcal{N}^l}(x_t^i)^\intercal p_t^{x,k,l} x_t^j \\
&\quad + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}\sum_{j\in\mathcal{N}^k,j\neq i}(x_t^i)^\intercal p_t^{x,k,k} x_t^j + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}(x_t^i)^\intercal q_t^k x_t^i \\
&= \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{K},l\neq k}|\mathcal{N}^k||\mathcal{N}^l|(\bar{x}_t^k)^\intercal p_t^{x,k,l}\bar{x}_t^l \quad + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}\sum_{j\in\mathcal{N}^k}|\mathcal{N}^k|^2(\bar{x}_t^k)^\intercal p_t^{x,k,k}\bar{x}_t^k \\
&\quad - \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}(x_t^i)^\intercal p_t^{x,k,k} x_t^i + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}(x_t^i)^\intercal q_t^k x_t^i = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{K}}|\mathcal{N}^k||\mathcal{N}^l|(\bar{x}_t^k)^\intercal p_t^{x,k,l}\bar{x}_t^l \\
&\quad + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}(x_t^i)^\intercal (q_t^k - p_t^{x,k,k}) x_t^i \\
&=: \mathbf{x}_t^\intercal P_t^x \mathbf{x}_t + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}\frac{1}{|\mathcal{N}^k|}(x_t^i)^\intercal Q_t^k x_t^i,
\end{aligned} \tag{2.11}$$

where $P_t^{xk,l} = |\mathcal{N}^k||\mathcal{N}^l| p^{x,k,l}$ and $Q_t^k = |\mathcal{N}^k|(q_t^k - p_t^{x,k,k})$. By similar algebraic manipulation, we can show

$$\mathbf{u}_t^\intercal R_t \mathbf{u}_t = \mathbf{u}_t^\intercal P_t^u \mathbf{u}_t + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}\frac{1}{|\mathcal{N}^k|}(u_t^i)^\intercal R_t^k u_t^i, \tag{2.12}$$

where $P_t^{uk,l} = |\mathcal{N}^k||\mathcal{N}^l| p^{u,k,l}$ and $R_t^k = |\mathcal{N}^k|(r_t^k - p_t^{u,k,k})$. Substituting (2.11) and (2.12) in (2.3) and (2.4), we get (3.2) and (3.3).

2.2 Controlled Markov chain systems

Suppose the dynamics (2.1) are controlled Markov chain. i.e.,

$$\mathbf{x}_{t+1} = f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t).$$

At time $t \in \{1, \dots, T\}$, the per step cost is

$$c_t(\mathbf{x}_t, \mathbf{u}_t).$$

Assume that the above system is partially exchangeable, i.e., agents $\mathcal{N}$ can be partitioned into K disjoint sub-populations $\mathcal{N}^k$, $k \in \mathcal{K} \coloneqq \{1, \dots, K\}$, such that for each $k \in \mathcal{K}$, the agents $\mathcal{N}^k$ are exchangeable. Moreover, for any sub-population $k \in \mathcal{K}$ and agent $i \in \mathcal{N}^k$, state x_t^i takes values in $\mathcal{X}^k$ and action u_t^i takes values in $\mathcal{U}^k$.

The *mean-field* of joint states and actions ξ_t^k of sub-population k, $k \in \mathcal{K}$, is defined as the empirical distribution of the states and actions of all agents in that sub-population, i.e.,

$$\xi_t^k(x, u) = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \mathbb{1}(x_t^i = x, u_t^i = u), \quad x \in \mathcal{X}^k, u \in \mathcal{U}^k,$$

The mean-field of states and actions of the entire population is denoted by $\boldsymbol{\xi}_t$, i.e.,

$$\boldsymbol{\xi}_t \coloneqq \operatorname{vec}(\xi_t^1, \dots, \xi_t^K).$$

Proposition 2.2 *In the controlled Markov chain system with partially exchangeable agents described above, there exist functions $\{\{f_t^k\}_{k \in \mathcal{K}}, \ell_t\}$ such that the dynamics of agent $i \in \mathcal{N}^k, k \in \mathcal{K}$, may be written as*

$$x_{t+1}^i = f_t^k(x_t^i, u_t^i, \boldsymbol{\xi}_t, w_t^i),$$

and the per-step cost at time t, may be written as $\ell_t(\boldsymbol{\xi}_t)$.

Thus, any controlled Markov system with partial exchangeable agents—irrespective of the information structure—is equivalent to a mean-field coupled system with the same information structure.

Proof of Proposition 2.2

Prior to the proof, we present the following Lemma that simplifies the proof.

Lemma 2.1 *Let $\mathbf{x} = (x^1, \dots, x^n)$ denote a vector of n deterministic variables where $x^i \in \mathcal{X}, i \in \{1, \dots, n\}$, and $f(\mathbf{x})$ denote any arbitrary function over the product space $\prod_{i=1}^n \mathcal{X}$. If*

for any arbitrary permutation σ, $f(\mathbf{x}) = f(\sigma\mathbf{x})$; then, there exists a function $\bar{f}$ such that

$$f(\mathbf{x}) = \bar{f}(\xi),$$

where ξ denotes the empirical distribution of $\mathbf{x}$.

Let $\mathbf{x}_t^{-i}, \mathbf{u}_t^{-i}$, and $\mathbf{w}_t^{-i}$ denote the state $\mathbf{x}_t$, action $\mathbf{u}_t$, and noise $\mathbf{w}_t$ of all agents except agent i, respectively. Consider agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$. In general, the dynamics may be written as follows.

$$x_{t+1}^i = f_t^i(x_t^i, u_t^i, \mathbf{x}_t^{-i}, \mathbf{u}_t^{-i}, w_t^i).$$

According to property 1, arbitrarily permuting (exchanging) the state and the action of the other agents must not change the dynamics of agent $i \in \mathcal{N}^k$. Let $\sigma(k')$, $k' \in \mathcal{K}$, denote any arbitrary permutation of the other agents of sub-population $k' \in \mathcal{K}$. Then, we have

$$f_t^i(x_t^i, u_t^i, \mathbf{x}_t^{-i}, \mathbf{u}_t^{-i}, w_t^i) = f_t^i(x_t^i, u_t^i, \sigma(k')\mathbf{x}_t^{-i}, \sigma(k')\mathbf{u}_t^{-i}, w_t^i).$$

By the recursive application of Lemma 2.1, there exists a function $\bar{f}^i$ such that

$$f_t^i(x_t^i, u_t^i, \mathbf{x}_t^{-i}, \mathbf{u}_t^{-i}, w_t^i) = \bar{f}_t^i(x_t^i, u_t^i, \boldsymbol{\xi}_t, w_t^i).$$

Note that the empirical distribution of the other agents of sub-population k, $(x_t^j, u_t^j)_{j \neq i \in \mathcal{N}^k}$, can be characterized by (x_t^i, u_t^i) and ξ_t^k. Now, consider two arbitrary agents i and j of sub-population k. From property 1, we have

$$x_t^i = \bar{f}_t^i(x_t^i, u_t^i, \boldsymbol{\xi}_t, w_t^i) = \bar{f}_t^j(x_t^i, u_t^i, \boldsymbol{\xi}_t, w_t^i),$$

and

$$x_t^j = \bar{f}_t^j(x_t^j, u_t^j, \boldsymbol{\xi}_t, w_t^j) = \bar{f}_t^i(x_t^j, u_t^j, \boldsymbol{\xi}_t, w_t^j).$$

Hence, $\bar{f}_t^i = \bar{f}_t^j = \bar{f}_t^k, \forall i, j \in \mathcal{N}^k$.

The similar argument holds for the per-step cost. In particular, from the property 2 of Definition 2.1, arbitrarily permuting agents of sub-population $k \in \mathcal{K}$ does not change the cost. Thus, by the recursive application of Lemma 2.1, there exist a function ℓ_t such that $c_t(\mathbf{x}_t, \mathbf{u}_t) = \ell_t(\boldsymbol{\xi}_t)$.

Proof of Lemma 2.1: Suppose $\mathcal{X} = \{s_1, \dots, s_{|\mathcal{X}|}\}$; then, the empirical distribution ξ is a vector of size $|\mathcal{X}|$ such that

$$\xi(s_m) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x^i = s_m), \quad m \in \{1, \dots, |\mathcal{X}|\}.$$

Permute the vector $\mathbf{x}$ to a vector $\mathbf{y} = (y^1, \dots, y^n) \in \prod_{i=1}^{n} \mathcal{X}$ such that for any $i \in \{1, \dots, n\}$,

$$y^i = \begin{cases} s_1, & 1 \le i \le n\xi(s_1), \\ s_2, & n\xi(s_1) < i \le n\big(\xi(s_1) + \xi(s_2)\big), \\ \vdots, & \vdots \\ s_{|\mathcal{X}|}, & n\big(\sum_{m=1}^{|\mathcal{X}|-1} \xi(s_m)\big) < i \le n. \end{cases}$$

Denote the above mapping as $\mathbf{y} = \phi(\xi)$. From $f(\mathbf{x}) = f(\sigma\mathbf{x}), \forall \sigma$, we have

$$f(\mathbf{x}) = f(\mathbf{y}) = f(\phi(\xi)) =: \bar{f}(\xi).$$

2.3 Conclusion

Due to the exchangeability property, the interaction (coupling) between the agents may be characterized (expressed) as a (lower dimensional) quantity, i.e., mean-field. Note that the notion of exchangeability is valid for any model; thus, the following interesting question arises:

Q: *What would be the mean-field of exchangeable models?*

It is shown in Proposition 2.1 that the mean-field of linear quadratic systems is the empirical average. In particular, any linear quadratic system with partially exchangeable agents—irrespective of the information structure—is equivalent to a mean-field coupled system with the same information structure. We call these systems *linear quadratic mean-field team* and in Chapter 3, we investigate the optimal control of such systems.

In addition, it is shown in Proposition 2.2 that the mean-field of controlled Markov chain systems is the empirical distribution. In particular, any controlled Markov chain system with partially exchangeable agents—irrespective of the information structure—is equivalent

to a mean-field coupled system with the same information structure. We call these systems *controlled Markov chain mean-field team* and in Chapter 4, we investigate the optimal control of such systems.

CHAPTER 3

Optimal control of linear quadratic mean-field teams

3.1 Introduction

In this chapter, we consider linear quadratic systems with partially exchangeable agents. We call these systems linear quadratic mean-field teams. In particular, we consider team optimal control of decentralized systems with linear dynamics and quadratic costs where the agents are coupled in the dynamics and cost through the mean-field (i.e., the empirical mean) of the states and actions. Two information structures are investigated: *Mean-Field Sharing Information Structure* (MFS-IS) and *Partial Mean-Field Sharing Information Structure* (PMFS-IS). In MFS-IS, all agents observe their local state and the mean-field of all sub-populations; in PMFS-IS, all agents observe their local state but the mean-field of only a subset of the sub-populations. Both information structures are non-classical and not partially nested. Nonetheless, it is shown that linear control strategies are optimal for MFS-IS and approximately optimal for PMFS-IS where the approximation error is inversely proportional to the size of the sub-populations whose mean-fields are not observed. The corresponding gains are determined by the solution of $K + 1$ Riccati equations, where K is the number of sub-populations. The dimensions of the Riccati equations do not depend on the size of the sub-populations; thus the solution complexity is independent of the number of agents. The generalizations to major-minor agents, tracking cost, weighted mean-field, and infinite horizon are provided. The results are illustrated using an example of demand response in smart grids.

Literature review

The linear quadratic systems are relevant in various research areas and have been widely investigated in the past. Herein, we review some of the models and results related to linear quadratic mean-field teams.

In [Madjidian and Mirkin, 2014], the authors consider a *homogeneous* population of *dynamically decoupled* agents which are coupled in the cost through a weighted mean-field term. Two models are investigated: (a) *hard-constraint model* where the weighted mean-field of actions must equal a pre-specified linear function of the weighted mean-field of states; and (b) *soft-constraint model* where the above hard constraint is relaxed by penalizing it in the cost. For both models, the authors show that the optimal centralized control laws are linear in the local state and the mean-field; the corresponding gains are computed by two decoupled Riccati equations. In Section 3.5.3, we generalize our results to the case when a weighted empirical mean-field is shared. In contrast to [Madjidian and Mirkin, 2014], we consider heterogeneous population and allow agents to be coupled in dynamics. Note that approximation results similar to those for partial mean-field sharing were not considered in [Madjidian and Mirkin, 2014].

Our results have similar features to those obtained for *centralized* linear quadratic mean-field control [Yong, 2013, Elliott et al., 2013]. In these models, the dynamics and the cost depend on the *statistical* mean-field of the state and action. Such a model may be viewed as a special case of our model when we restrict to a single homogeneous sub-population and consider the limit of infinite number of agents (and therefore the empirical mean and the statistical mean are the same). Our proof technique, which relies on a simple change of variables, is conceptually simpler than that of [Yong, 2013, Elliott et al., 2013].[1] It is worth highlighting that the linear quadratic mean-field control model is a centralized control problem and the results of [Yong, 2013, Elliott et al., 2013] do not apply to the multi-agent models that we consider.

Recently, an iterative bidding strategy was proposed in [Singh et al., 2015] for the optimal control multi-agent systems with decoupled dynamics that are coupled through a constraint. For LQG agents, the scheme operates as follows: at each time, a coordinator sets a price profile for all future times; agents submit a bid profile for all future times; the coordinator

[1] In [Yong, 2013], first coupled forward and backward stochastic differential equations are derived and then they are decoupled into two Riccati equations using the four step technique of [Ma and Yong, 1999]. In [Elliott et al., 2013], a matrix dynamical optimization method is used.

updates the prices and the process continues until the bids have converged. Agents choose the first value of their bid as their action and the above process is repeated at the next time step. In this scheme, agents do not need to know the system dynamics of other agents. In contrast, we assume that the system dynamics are common knowledge to all agents. However, in our model, agents only need to share the mean-field of their states (which can be computed using a consensus algorithm) rather than iteratively sharing the bid profile for all future times.

An alternative decomposition-coordination approach for optimal decentralized control of *deterministic* LQ systems was proposed in [Takahara, 1964, Cohen, 1977]. This is an iterative approach. Each iteration consists of two steps: (i) a *decomposition step* in which each agent assumes decoupled dynamics and costs and computes its local control trajectory by solving an optimal tracking problem from pre-specified linear offsets for the dynamics and a reference trajectory for the cost; (ii) a *coordination* step in which the linear offsets for the dynamics and reference trajectories for the cost are computed for all agents from the pre-specified control trajectories. It is shown that this iterative process converges to the optimal centralized solution. In contrast to such decomposition-coordination methods, our proposed solution is not iterative. The optimal gains for all agents are computed in a single step by solving Riccati equations. Furthermore, our solution methodology works for deterministic as well as stochastic systems.

A related solution approach called mean-field games (MFG) was proposed in [Huang et al., 2003, Lasry and Lions, 2006a, Lasry and Lions, 2006b, Huang et al., 2007, Li and Zhang, 2008, Caines, 2013, Gomes and Saude, 2014, Moon and Basar, 2017, Bensoussan et al., 2016] to compute approximate Nash equilibrium for large population games. The main idea is to assume an infinite large size of each sub-population and solve a set of two coupled equations: a Hamilton-Jacobi-Bellman (HJB) equation to compute the best response of a generic agent playing against a "mass trajectory" and a Fokker-Planck-Kolmogorov (FPK) equation to compute the mass trajectory from the strategy of a generic agent. It is shown that a solution to these equations exists under appropriate conditions. The resulting strategies are ε-Nash when the sub-populations are finite, where the approximation error is $\mathcal{O}(1/\sqrt{n})$, n denotes the size of smallest sub-population. For linear quadratic systems, the coupled HJB-FPK equations simplify to K Riccati equations and two coupled forward and backward ODEs. In contrast, in our solution there is an additional Riccati equation instead of the coupled forward-backward equations. The coupled equations in MFG depend on the initial

mean-field while the Riccati equation in our solution does not. It is shown in [Huang et al., 2012] that when agents have decoupled dynamics, the MFG solution is ε-team-optimal with $\varepsilon \in \mathcal{O}(1/\sqrt{n})$. We obtain a similar result for *dynamically coupled* agents with $\varepsilon \in \mathcal{O}(1/n)$.

Finally, in our opinion, the approach proposed in this chapter is easier to generalize than the approach of MFG. As a case in point, considerable technical sophistication is needed in the MFG theory to solve the so called major-minor setup [Huang, 2010] (because the coupled forward-backward ODEs become SDEs)[2]. In contrast, as we show in Section 3.5.1, the major-minor setup is simply a special case of our model. For this reason, we believe that the results presented in this chapter and the associated proof techniques may also be useful for MFG.

In many of the references cited above, the agent dynamics are assumed to be decoupled. In our model, the agent dynamics are coupled, which is significantly more challenging.[3]. When the system dynamics are coupled, the information structure is non-classical and there is no general solution methodology to obtain a team-optimal solution.

Main challenges

In general, to solve a decentralized linear quadratic system, we face two main challenges.

- **The curse of dimensionality:** For *centralized* linear quadratic systems, the computational complexity of finding the optimal solution (i.e., the size of matrices in Riccati equation) increases polynomially with the number of agents. Therefore, even if the computational complexity of a *decentralized* linear quadratic system is properly defined, in principle it must increase at least polynomially with the number of agents.[4]

- **The decentralized information:** It is conceptually difficult to establish cooperation among agents when they have discrepancy in information (perspective). In centralized linear quadratic systems, the optimal control strategy is linear in the agent's estimate of the state. The optimal gain is determined by the solution of backward Riccati equations

[2] Although [Huang, 2010] seeks the Nash strategy rather than the global optimal strategy, the technical sophistication of MFG solution would carry on to the team solution, similar to [Huang et al., 2012]. In particular, the mean-field of minor agents becomes stochastic and difficult to predict due the the presence of the major agent.

[3] Also see [Li and Zhang, 2008, Remark 13] for the difficulties in extending the proof technique used in mean-field games to systems with coupled dynamics.

[4] Note that it may not even be possible to define the computational complexity of a decentralized linear quadratic system because the space is infinite dimensional.

and the agent's estimate is updated using Kalman filtering equations. However, this is not the case for decentralized systems. As illustrated by the Witsenhausen counterexample [Witsenhausen, 1968], in decentralized systems non-linear strategies can outperform the best linear strategy. In general, when the primitive random variables are Gaussian, linear strategies are globally optimal for only partially nested information structure [Ho and Chu, 1972] and its variations [Yüksel, 2009]. Even if attention is restricted to linear strategies, the problem of finding the best linear strategies may not be convex; it is convex only for special sparsity patterns such as funnel causality [Bamieh and Voulgaris, 2005] and quadratic invariance [Rotkowitz and Lall, 2004]. Since mean-field sharing information structure (MFS-IS) is neither partially nested nor quadratic invariant, we cannot assert a priori that linear strategies are globally optimal or that the problem of finding the best linear strategies is convex. Unlike the first challenge, the second challenge exists even for a system with two agents.

Contributions and Salient features

1. The linear quadratic mean-field teams under MFS-IS are decentralized systems with non-classical information structure that are neither partially nested nor quadratic invariant; yet we show linear control laws are optimal.

2. We obtain the corresponding optimal gains by $K+1$ *decoupled* Riccati equations,: one for each sub-population and one for the mean-field term. In fact, for the decentralized implementation, each agent simply needs to solve two Riccati equations: one corresponding to its own sub-population and one to the mean-field. The dimensions of these Riccati equations do not depend on the number of agents in each sub-population. Thus, the solution complexity does not depend on the number of agents in the system (i.e., the solution is scalable for large sub-populations).

3. In general, the matrices in the dynamics and cost are allowed to depend on $\mathcal{N}$ (the number of agents) although their size only depends on K (the number of sub-populations)[5]. However, if these matrices do not depend on the number of agents, then neither do the optimal gains. Consequently, the agents *need not even be aware of the number of agents.*

[5]An example: consider a homogeneous population (i.e., $K=1$) where $x^i_{t+1} = nx^i_t + u^i_t + w^i_t \in \mathbb{R}, i \in \{1,\dots,n\}$ and n is the number of agents.

4. We show that the centralized performance is achieved by sharing only the mean-field (which can be shared using distributed algorithms such as consensus). Thus, instead of requiring that agents have the capability or the energy to communicate to a centralized controller, agents only require the capacity or the energy to communicate to their neighbours.

5. For partial mean-field sharing information structure (PMFS-IS), we propose a linear strategy that is approximately optimal where the approximation error is inversely proportional to the size of sub-populations whose mean-fields are not observed. The proposed strategy is a certainty equivalence strategy in which all agents generate an estimate of the unobserved components of the mean-field using the observed components of the mean-field. This estimate is used in the optimal strategy identified for MFS-IS. We show that the approximation error between the proposed strategy for PMFS-IS and the optimal strategy for MFS-IS is given by terms of the weighted cost of a linear system, which can be computed by a Lyapunov equation.

6. As a consequence of PMFS-IS result, when a sub-population is large, sharing its mean-field has a vanishingly small advantage. In the extreme case when all sub-populations are large, the approximately optimal solution can be implemented under a completely decentralized information structure (i.e. $(x^i_{1:t}, u^i_{1:t-1})$).

6. We show that our results generalize to several variations of the basic mean-field model including: systems where a major agent interacts with a collection of minor agents in Section 3.5.1; systems where agents have individual tracking cost in Section 3.5.2; systems where agents have individual weights in Section 3.5.3. In Section 3.6, we show our results generalize to infinite horizon setup using standard arguments.

Notation

For a set $\mathcal{N}$, $|\mathcal{N}|$ denotes its size. For a matrix A, $A^\intercal$ denotes its transpose, $\mathrm{Tr}(A)$ denotes its trace; if A is square, $A \geq 0$ (respectively $A > 0$) denotes that A is positive semi-definite (respectively positive definite). For matrices A and B of appropriate size, $A \leq B$ means $B - A \geq 0$, $\mathrm{diag}(A, B)$ denotes a block diagonal matrix with diagonal terms A and B, $\sqrt{A}$ denotes B where $A = B^\intercal B$, $A \circ B$ denotes Hadamard product, and $A \otimes B$ denotes Kronecker

product. For matrices A, B, and C with the same number of columns, $\text{rows}(A, B, C)$ denotes the matrix $[A^\intercal, B^\intercal, C^\intercal]^\intercal$. For vectors x, y, and z, $\text{vec}(x, y, z)$ denotes the vector $[x^\intercal, y^\intercal, z^\intercal]^\intercal$.

Superscripts index agents (indexed by i) or sub-populations (indexed by k). Given a set $\mathcal{N}$ of agents and states $x^i, i \in \mathcal{N}$, bold $\mathbf{x}$ denotes $\text{vec}(x^1, \dots, x^{|\mathcal{N}|})$; when all states are of the same dimension, $\langle (x^i)_{i\in\mathcal{N}} \rangle$ denotes the mean-field $\frac{1}{|\mathcal{N}|}\sum_{i=1}^{|\mathcal{N}|} x^i$ of $(x^i)_{i\in\mathcal{N}}$. For vectors and matrices, we use the short hand notation $x_{1:t}$ or $A_{1:t}$ to denote $(x_1, \dots, x_t)$ and $(A_1, \dots, A_t)$, respectively.

$\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{R}_{>0}$ denote the sets of real, non-negative real, and positive real numbers, respectively. $\mathbb{1}_{n\times m}$ denotes $n \times m$ matrix of ones, $\mathbb{I}_n$ denotes $n \times n$ identity matrix. We omit the subscripts when the dimensions are clear from the context. For a random variable x, $\mathbb{E}[x]$ and $\text{var}(x)$ denote its mean and variance, respectively.

Given horizon T and matrices $A_{1:T}$ and $Q_{1:T}$, the notation $M_{1:T} = \text{DLE}_{\text{T}}(A_{1:T}, Q_{1:T})$ means that $M_{1:T}$ is the solution of the finite horizon discrete Lyapunov equation, i.e., $M_T = Q_T$, and for $t \in \{T-1, \dots, 1\}$, $M_t = A_t{}^\intercal M_{t+1} A_t + Q_t$.

Similarly, given a horizon T and matrices $A_{1:T}$, $B_{1:T}$, $Q_{1:T}$, and $R_{1:T}$, the notation $M_{1:T} = \text{DRE}_{\text{T}}(A_{1:T}, B_{1:T}, Q_{1:T}, R_{1:T})$ means that $M_{1:T}$ is the solution of the finite horizon discrete Riccati equation, i.e., $M_T = Q_T$, and for $t \in \{T-1, \dots, 1\}$,

$$M_t = -A_t^\intercal M_{t+1} B_t \left(B_t^\intercal M_{t+1} B_t + R_t\right)^{-1} B_t^\intercal M_{t+1} A_t + A_t^\intercal M_{t+1} A_t + Q_t.$$

Given a discount factor $\beta \in (0, 1]$ and matrices A, B, Q, and R, the notation $M = \text{DALE}_\beta(A, Q)$ means that M is the solution of the discrete algebraic Lyapunov equation $M = \beta A^\intercal M A + Q$.

Similarly, the notation $M = \text{DARE}_\beta(A, B, Q, R)$ means that M is the solution of the discrete algebraic Riccati equation $M = -\beta A^\intercal M B \left(B^\intercal M B + \beta^{-1} R\right)^{-1} B^\intercal M A + \beta A^\intercal M A + Q$.

3.2 Problem formulation

System Model

Consider a population of $\mathcal{N}$ agents that are partitioned into K disjoint sub-populations $\mathcal{N}^k$, $k \in \mathcal{K} \coloneqq \{1, \dots, K\}$, such that for each $k \in \mathcal{K}$, the agents $\mathcal{N}^k$ are exchangeable. The state and action of agent $i, i \in \mathcal{N}$, at time t are denoted by x_t^i and u_t^i. Let $\mathbf{x}_t = (x_t^i)_{i\in\mathcal{N}}$ and $\mathbf{u}_t = (u_t^i)_{i\in\mathcal{N}}$ denote the state and action of the entire system. Moreover, for any sub-population $k \in \mathcal{K}$ and agent $i \in \mathcal{N}^k$, state x_t^i takes values in $\mathbb{R}^{d_x^k}$ and action u_t^i takes values

in $\mathbb{R}^{d_u^k}$.

The *mean-field* of states[6] $\bar{x}_t^k$ of sub-population k, $k \in \mathcal{K}$, is defined as the empirical mean of the states of all agents in that sub-population, i.e.,

$$\bar{x}_t^k \coloneqq \langle (x_t^i)_{i \in \mathcal{N}^k} \rangle = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} x_t^i, \quad k \in \mathcal{K}.$$

Similarly, the mean-field of the actions $\bar{u}_t^k$ of sub-population k, $k \in \mathcal{K}$, is defined as the empirical mean of the actions of all agents in that sub-population, i.e.,

$$\bar{u}_t^k \coloneqq \langle (u_t^i)_{i \in \mathcal{N}^k} \rangle = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} u_t^i, \quad k \in \mathcal{K}.$$

The mean-field of states and actions of the entire population are denoted by $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{u}}_t$ respectively, i.e.,

$$\bar{\mathbf{x}}_t = \operatorname{vec}(\bar{x}_t^1, \dots, \bar{x}_t^K), \quad \bar{\mathbf{u}}_t = \operatorname{vec}(\bar{u}_t^1, \dots, \bar{u}_t^K).$$

The dynamics of agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$ is given by

$$x_{t+1}^i = A_t^k x_t^i + B_t^k u_t^i + D_t^k \bar{\mathbf{x}}_t + E_t^k \bar{\mathbf{u}}_t + w_t^i, \tag{3.1}$$

where $w_t^i \in \mathcal{W}^k$ is the disturbance noise process. Let $\mathbf{w}_t = (w_t^i)_{i \in \mathcal{N}}$. The primitive random variables $\{\mathbf{x}_1, \{\mathbf{w}_t\}_{t=1}^T\}$ are defined on a common probability space. The per-step cost at time $t \in \{1, \dots, T-1\}$ is given by

$$c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \bar{\mathbf{x}}_t^\intercal P_t^x \bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^\intercal P_t^u \bar{\mathbf{u}}_t + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[(x_t^i)^\intercal Q_t^k x_t^i + (u_t^i)^\intercal R_t^k u_t^i \right], \tag{3.2}$$

and the per-step cost at time $t = T$ is given by

$$c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T) = \bar{\mathbf{x}}_T^\intercal P_T^x \bar{\mathbf{x}}_T + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} (x_T^i)^\intercal Q_T^k x_T^i. \tag{3.3}$$

For ease of reference, the notation is summarized in Table 3.1.

[6] In the sequel, we refer to mean-field of the states simply as mean-field.

Table 3.1 Summary of the notation used in this chapter.

Notation used for agent $i \in \mathcal{N}^k$ belonging to sub-population $k \in \mathcal{K}$	
$x_t^i \in \mathbb{R}^{d_x^k}$	State of agent i
$u_t^i \in \mathbb{R}^{d_u^k}$	Action of agent i
Notation used for sup-population $k \in \mathcal{K} = \{1, \dots, K\}$	
$\mathcal{N}^k$	Entire sub-population k
$\bar{x}_t^k = \langle (x_t^i)_{i \in \mathcal{N}^k} \rangle$	Mean-field of states at time t
$\bar{u}_t^k = \langle (u_t^i)_{i \in \mathcal{N}^k} \rangle$	Mean-field of actions at time t
Notation used for entire population	
$\mathcal{N} = \bigcup_{k \in \mathcal{K}} \mathcal{N}^k$	Entire population
$\mathbf{x}_t = (x_t^i)_{i \in \mathcal{N}}$	Joint state of entire population at time t
$\mathbf{u}_t = (u_t^i)_{i \in \mathcal{N}}$	Joint action of entire population at time t
$\bar{\mathbf{x}}_t = \mathrm{vec}(\bar{x}_t^1, \dots, \bar{x}_t^K)$	Mean-field of states of entire population at time t
$\bar{\mathbf{u}}_t = \mathrm{vec}(\bar{u}_t^1, \dots, \bar{u}_t^K)$	Mean-field of actions of entire population at time t

Information structure

We consider two information structures; in both, agents perfectly recall all data that they observe. In the first information structure, which we call *mean-field sharing* and denote by MFS-IS, every agent $i \in \mathcal{N}$ perfectly observes its local state x_t^i and the global mean-field $\bar{\mathbf{x}}_t$. Thus, the data I_t^i available to agent i at time t is given by

$$I_t^i = (x_{1:t}^i, u_{1:t-1}^i, \bar{\mathbf{x}}_{1:t}). \tag{MFS-IS}$$

In the second information structure, which we call *partial mean-field sharing* and denote by PMFS-IS, there exists a subset $\mathcal{S}$ of the sub-populations $\mathcal{K}$ such that every agent $i \in \mathcal{N}$ perfectly observes its local state x_t^i and the mean-fields of sub-populations $\mathcal{S}$, i.e., $\{\bar{x}_t^k\}_{k \in \mathcal{S}}$. We use $\mathcal{S}^c$ to denote $\mathcal{K} \backslash \mathcal{S}$. The data I_t^i available to agent i at time t is given by

$$I_t^i = (x_{1:t}^i, u_{1:t-1}^i, (\bar{x}_{1:t}^k)_{k \in \mathcal{S}}). \tag{PMFS-IS}$$

Under both information structures, agent i chooses u_t^i as follows:

$$u_t^i = g_t^i(I_t^i). \tag{3.4}$$

The function g_t^i is called the *control law of agent i* at time t. The collection $\mathbf{g}^i = (g_1^i, g_2^i, \dots, g_T^i)$ is called the *control strategy of agent i*. The collection $\mathbf{g} = (\mathbf{g}^i)_{i\in\mathcal{N}}$ is called the *control strategy of the system*. The performance of strategy $\mathbf{g}$ is given by

$$J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=1}^{T-1} c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) + c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T)\right], \tag{3.5}$$

where the expectation is with respect to the measure induced on all the system variables by the choice of strategy $\mathbf{g}$.

Remark 3.1 Note that the mean-field can be shared in a distributed manner using consensus algorithms [Xiao and Boyd, 2004].

The optimization problem

We are interested in the following optimization problem.

Problem 3.1 *In the model described above, find a strategy* $\mathbf{g}^*$ *that minimizes* (3.5)*, i.e.,*

$$J^* := J(\mathbf{g}^*) = \inf_{\mathbf{g}} J(\mathbf{g}),$$

where the infimum is taken over all strategies of form (3.4)*.*

3.3 Main results

In this section, we present the main results of this chapter.

Exact solution for MFS-IS

We impose following standard assumptions on the model:

A. 3.1 *The primitive random variables* $\{\mathbf{x}_1, \{\mathbf{w}_t\}_{t=1}^T\}$ *have zero mean and are mutually independent.*

A. 3.2 *For every* t, P_t^x, P_t^u, Q_t^k, *and* R_t^k *are symmetric matrices that satisfy*[7]

$$Q_t^k \geq 0, \ \forall k \in \mathcal{K}, \qquad \operatorname{diag}(Q_t^1, \ldots, Q_t^K) + P_t^x \geq 0, \tag{3.6}$$

$$R_t^k > 0, \ \forall k \in \mathcal{K}, \qquad \operatorname{diag}(R_t^1, \ldots, R_t^K) + P_t^u > 0. \tag{3.7}$$

Note that we do not require the initial state $\mathbf{x}_1$ and the disturbance $\mathbf{w}_t$ to be independent across agents. Nor do we require matrices P_t^x and P_t^u to be positive semi-definite as long as (3.6)–(3.7) hold.

Theorem 3.1 *Under (A.3.1), (A.3.2), and (MFS-IS), we have the following results for Problem 3.1.*

1. Structure of optimal strategy: *The optimal strategy for Problem 3.1 is unique and is linear in the local state and the mean-field of the system. In particular,*

$$u_t^i = \breve{L}_t^k (x_t^i - \bar{x}_t^k) + \bar{L}_t^k \bar{\mathbf{x}}_t, \tag{3.8}$$

   where the gains $\{\breve{L}_t^k, \bar{L}_t^k\}_{t=1}^{T-1}$ *are obtained by the solution of* $K+1$ *Riccati equations given below: one for computing each* $\breve{L}_t^k$, $k \in \mathcal{K}$, *and one for* $\bar{L}_t \coloneqq \operatorname{rows}(\bar{L}_t^1, \ldots, \bar{L}_t^K)$.

2. Riccati equations: *Let*

$$\bar{A}_t \coloneqq \operatorname{diag}(A_t^1, \ldots, A_t^K) + \operatorname{rows}(D_t^1, \ldots, D_t^K), \qquad \bar{Q}_t \coloneqq \operatorname{diag}(Q_t^1, \ldots, Q_t^K),$$
$$\bar{B}_t \coloneqq \operatorname{diag}(B_t^1, \ldots, B_t^K) + \operatorname{rows}(E_t^1, \ldots, E_t^K), \qquad \bar{R}_t \coloneqq \operatorname{diag}(R_t^1, \ldots, R^K).$$

   Then, for $t \in \{1, \ldots, T-1\}$*:*

$$\breve{L}_t^k = -\left((B_t^k)^\intercal \breve{M}_{t+1}^k B_t^k + R_t^k\right)^{-1} (B_t^k)^\intercal \breve{M}_{t+1}^k A_t^k,$$
$$\bar{L}_t = -\left(\bar{B}_t^\intercal \bar{M}_{t+1} \bar{B}_t + \bar{R}_t + P_t^u\right)^{-1} \bar{B}_t^\intercal \bar{M}_{t+1} \bar{A}_t,$$

[7] This is a sufficient condition.

where $\{\breve{M}_t^k\}_{t=1}^T$ *and* $\{\bar{M}_t\}_{t=1}^T$ *are the solutions of following Riccati equations:*

$$\breve{M}_{1:T}^k = \mathrm{DRE}_{\mathrm{T}}(A_{1:T}^k, B_{1:T}^k, Q_{1:T}^k, R_{1:T}^k), \tag{3.9}$$

$$\bar{M}_{1:T} = \mathrm{DRE}_{\mathrm{T}}(\bar{A}_{1:T}, \bar{B}_{1:T}, \bar{Q}_{1:T} + P_{1:T}^x, \bar{R}_{1:T} + P_{1:T}^u). \tag{3.10}$$

3. Optimal performance: *Let*

$$\breve{\Sigma}_t^k := \frac{1}{|\mathcal{N}^k|}\sum_{i\in\mathcal{N}^k} \mathrm{var}(w_t^i - \bar{w}_t^k),\ \bar{\Sigma}_t := \mathrm{var}(\bar{\mathbf{w}}_t),\ \breve{\Xi}^k := \frac{1}{|\mathcal{N}^k|}\sum_{i\in\mathcal{N}^k} \mathrm{var}(x_1^i - \bar{x}_1^k),\ \bar{\Xi} := \mathrm{var}(\bar{\mathbf{x}}_1).$$

Then, the optimal cost is given by

$$J^* \ = \ \sum_{k\in\mathcal{K}} \mathrm{Tr}(\breve{\Xi}^k \breve{M}_1^k) \ + \ \mathrm{Tr}(\bar{\Xi}\bar{M}_1) \ + \ \sum_{t=1}^{T-1}\Bigg[\sum_{k\in\mathcal{K}} \mathrm{Tr}(\breve{\Sigma}_t^k \breve{M}_{t+1}^k) \ + \ \mathrm{Tr}(\bar{\Sigma}_t \bar{M}_{t+1})\Bigg]. \tag{3.11}$$

The proof is presented in Section 3.4. Note that the dimensions of Riccati equations (3.9) and (3.10) do not depend on the size of the sub-populations $(|\mathcal{N}^1|, \ldots, |\mathcal{N}^K|)$. Hence, the solution complexity depends only on the number K of sub-populations and it is independent of the number of agents in each sub-population. To implement the optimal control strategies:

- all agents must compute $\bar{L}_{1:T-1}$ by solving the Riccati equation (3.10),
- agents of sub-population k must compute $\breve{L}_{1:T-1}^k$ by solving the Riccati equation (3.9)[8].

Then, an individual agent i of sub-population k, upon observing the local state x_t^i and the global mean-field $\bar{\mathbf{x}}_t$, chooses its local control action according to (3.8). Note that each agent needs to solve only two Riccati equations, although there are $K+1$ Riccati equations in Theorem 3.1.

Remark 3.2 An interesting feature of the solution is that all agents in a particular sub-population use identical control laws. This is a feature of the linear quadratic system and not

[8]The dimension of Riccati equation (3.9) depends neither on $|\mathcal{N}^k|$ nor K. The dimension of Riccati equation (3.10) does not depend on $|\mathcal{N}|$ and it depends only on K. In general, the computational complexity of solving a standard discrete-time Riccati equation is polynomial in space (i.e., dimension) and linear in horizon. For more details on the numerical solutions of standard and algebraic Riccati equations, the reader is referred to [Lancaster and Rodman, 1995, Arnold and Laub, 1984, Li et al., 2013].

of exchangeability. See [Arabneydi and Mahajan, 2014] for an example of an exchangeable system where the optimal control laws are not identical for all agents in a sub-population.

Remark 3.3 If the per-step cost has cross-terms involving $(x_t^i, \bar{\mathbf{x}}_t)$ and $(u_t^i, \bar{\mathbf{u}}_t)$, i.e.,

$$\sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[(x_t^i)^\intercal S_t^{x,k} \bar{\mathbf{x}}_t + (u_t^i)^\intercal S_t^{u,k} \bar{\mathbf{u}}_t \right],$$

then, this cost can be re-written in the form of (3.2) and (3.3), i.e., $\bar{\mathbf{x}}_t^\intercal S_t^x \bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^\intercal S_t^u \bar{\mathbf{u}}_t$, where

$$S_t^x \coloneqq \mathrm{rows}(S_t^{x,1}, \dots, S_t^{x,K}), \quad S_t^u \coloneqq \mathrm{rows}(S_t^{u,1}, \dots, S_t^{u,K}).$$

Remark 3.4 We assumed that there are no cross-terms of the form $x^\intercal S u$ in the per-step cost of (2.3) and (2.4). If such cross-terms are present, there will be cross-terms involving (x_t^i, u_t^i), $(x_t^i, \bar{u}_t)$, $(\bar{x}_t, u_t^i)$, and $(\bar{x}_t, \bar{u}_t)$ in the equivalent mean-field model presented in Proposition 2.1. These cross-terms can be treated in the standard manner as cross-terms are treated in centralized LQR.

Remark 3.5 Suppose in addition to (A.3.1), we have that $\{x_1^i, \{w_t^i\}_{t\geq 1}\}_{i\in\mathcal{N}}$ are independent and for any $k \in \mathcal{K}$, $(x_1^i)_{i\in\mathcal{N}^k}$ is i.i.d. with variance Ξ^k and $\{w_t^i\}_{i\in\mathcal{N}^k}$ is i.i.d. with variance Σ_t^k. Then, we have

$$\breve{\Sigma}_t^k = \frac{|\mathcal{N}^k|-1}{|\mathcal{N}^k|}\Sigma_t^k, \quad \bar{\Sigma}_t = \mathrm{diag}(\Sigma_t^1, \dots, \Sigma_t^K), \quad \breve{\Xi}^k = \frac{|\mathcal{N}^k|-1}{|\mathcal{N}^k|}\Xi^k, \quad \bar{\Xi} = \mathrm{diag}(\Xi^1, \dots, \Xi^K).$$

The expression of total cost (3.11) can be simplified accordingly.

Note that for some special models, the optimal strategy under (MFS-IS), given by (3.8), can be implemented under smaller information structures. Identifying these (special) models requires further study that does not lie within the scope of this chapter. However, to mention a few models, we present three examples below.

Example 1: Let $\mathcal{K} = \{1\}, D_t^1 = 0, E_t^1 = 0$, and for arbitrary $\alpha_t \in \mathbb{R}$, $P_t^x = \alpha_t Q_t^1$ and $P_t^u = \alpha_t R_t^1$. Then, $u_t^i = \breve{L}_t^1 x_t^i$. This implies there is no need to observe the mean-field $\bar{x}_t^1$ in order to compute the optimal strategy.

Example 2: Let $\mathcal{K} = \{1,2\}$, $\text{rows}(D_t^1, D_t^2) = \text{diag}(\tilde{D}_t^1, \tilde{D}_t^2)$, and $A_t^2 = -\tilde{D}_t^2$. Denote $[\bar{L}_t^{k,1} \bar{L}_t^{k,2}] := \bar{L}_t^k$. Then, $u_t^i = \breve{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^{k,1}\bar{x}_t^1, k \in \{1,2\}$. This implies agents of sub-population $\{1\}$ do not need to observe the mean-field of sub-population $\{2\}$, i.e. $\bar{x}_t^2$, in order to compute the optimal strategy.

Example 3 (No local controls): Let $B_t^k = 0$ and $R_t^k = 0, \forall k \in \mathcal{K}$. Let $\boldsymbol{\theta}_t = \text{rows}(\theta_t^1, \ldots, \theta_t^K)$, $\theta_t^k \in \mathbb{R}^{d_{\tilde{u}}} \times \mathbb{R}^{d_u}, k \in \mathcal{K}$, such that $E_t^k = \tilde{E}_t^k \boldsymbol{\theta}_t^\intercal$ for all $k \in \mathcal{K}$ and $P_t^u = \boldsymbol{\theta}_t^\intercal \tilde{P}_t^u \boldsymbol{\theta}_t$. In addition, let θ_t^{k+} denote the right inverse of θ_t^k (i.e., $\theta_t^k \theta_t^{k+} = \mathbb{I}_{\mathbb{R}^{d_{\tilde{u}}}}$), which is assumed to exist. This implies that the dynamics and cost are given as follows.

$$x_{t+1}^i = A_t^k x_t^i + D_t^k \bar{\mathbf{x}}_t + \tilde{E}_t^k \tilde{u}_t + w_t^i.$$

where $\tilde{u}_t \coloneqq \boldsymbol{\theta}_t^\intercal \bar{\mathbf{u}}_t = \sum_{k \in \mathcal{K}} \theta_t^k \bar{u}_t^k$. At time $t \in \{1, \ldots, T-1\}$, the per-step cost is given by,

$$c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \tilde{u}_t) = \bar{\mathbf{x}}_t^\intercal P_t^x \bar{\mathbf{x}}_t + \tilde{u}_t^\intercal \tilde{P}_t^u \tilde{u}_t + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} (x_t^i)^\intercal Q_t x_t^i,$$

and $t = T$,

$$c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T) = \bar{\mathbf{x}}_T^\intercal P_T^x \bar{\mathbf{x}}_T + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} (x_T^i)^\intercal Q_T x_T^i.$$

Corollary 3.1 *For the model described above, the optimal control law is given as follows.*

$$u_t^i = \theta_t^{k+} \tilde{L}_t^k \bar{x}_t^k, \quad i \in \mathcal{N}^k, k \in \mathcal{K}.$$

where $[\tilde{L}_t^1, \ldots, \tilde{L}_t^K] =: \bar{L}_t$ *is computed as in Theorem 3.1 but with* $\bar{B}_t$ *replaced by* $\tilde{B}_t = \text{rows}(\tilde{E}_t^1, \ldots, \tilde{E}_t^K)$ *and* P_t^u *replaced by* $\tilde{P}_t^u$.

The proof is presented in Section 3.4. This implies that each agent only needs to observe the mean-field of its sub-population (rather than the mean-field of entire population). Thus, this result is similar in spirit to [Asghari and Nayyar, 2016, Theorem 1].

Approximate solution for PMFS-IS

So far we have assumed that the mean-fields of all sub-populations are observed. However, in practice, it may be physically and/or economically unfeasible to collect and share the mean-fields of large sub-populations. For this reason, we consider Problem 3.1 under PMFS-IS.

Based on the results of Theorem 3.1, we propose a certainty equivalence strategy for PMFS-IS and show that the performance of this strategy is close to the optimal performance under MFS-IS. We impose the following assumptions on the model.

A. 3.3 *In addition to (A.3.1), for any $k \in \mathcal{S}$ and $k' \in \mathcal{S}^c$, initial states $(x_1^i)_{i\in\mathcal{N}^k}$ are independent of $(x_1^j)_{j\in\mathcal{N}^{k'}}$.*

A. 3.4 *The primitive random variables $\{x_1^i, \{w_t^i\}_{t=1}^T\}_{i\in\mathcal{N}}$ are independent. For any $k, k \in \mathcal{K}$, there exist finite matrices c_x^k and c_w^k such that*

$$\sup_{i\in\mathcal{N}^k} \operatorname{var}(x_1^i) \le c_x^k, \qquad \sup_{t\le T, i\in\mathcal{N}^k} \operatorname{var}(w_t^i) \le c_w^k.$$

A. 3.5 *The dynamics $\{A_t^k, B_t^k, D_t^k, E_t^k\}_{k\in\mathcal{K}}$, cost $\{Q_t^k, R_t^k,\}_{k\in\mathcal{K}}$, P_t^x and P_t^u, and covariance bounds $\{c_x^k, c_w^k\}_{k\in\mathcal{K}}$ do not depend on the sizes $(|\mathcal{N}^1|, \dots, |\mathcal{N}^K|)$ of the sub-populations.*

Since we are comparing the system performance under two information structures, we use different notation for the two. Under MFS-IS, the state and action of agent i are denoted by x_t^i and u_t^i. Assume that u_t^i is generated as per Theorem 3.1. Under PMFS-IS, the state and action of agent i are denoted by s_t^i and v_t^i. The dynamics are same as (3.1). In particular for agent i of sub-population $k \in \mathcal{K}$, $s_1^i = x_1^i$ and

$$s_{t+1}^i = A_t^k s_t^i + B_t^k v_t^i + D_t^k \bar{\mathbf{s}}_t + E_t^k \bar{\mathbf{v}}_t + w_t^i, \tag{3.12}$$

where

$$\bar{\mathbf{s}}_t = \operatorname{vec}(\bar{s}_t^1, \dots, \bar{s}_t^K), \quad \bar{s}_t^k = \langle (s_t^i)_{i\in\mathcal{N}^k} \rangle, \quad \bar{\mathbf{v}}_t = \operatorname{vec}(\bar{v}_t^1, \dots, \bar{v}_t^K), \quad \bar{v}_t^k = \langle (v_t^i)_{i\in\mathcal{N}^k} \rangle.$$

To describe the control strategy, we define a process $\{\mathbf{z}_t\}_{t=1}^T$ as follows: $\mathbf{z}_t = \operatorname{vec}(z_t^1, \dots, z_t^K)$, where for any $k \in \mathcal{K}$, $z_t^k \in \mathbb{R}^{d_x^k}$; the initial state $\mathbf{z}_1$ is given by z_1^k is $\bar{s}_1^k$ for $k \in \mathcal{S}$ and is 0 for $k \notin \mathcal{S}$.[9] The process evolves as:

$$z_{t+1}^k = \begin{cases} \bar{s}_{t+1}^k, & k \in \mathcal{S}, \\ A_t^k z_t^k + (B_t^k \bar{L}_t^k + D_t^k + E_t^k \bar{L}_t)\mathbf{z}_t, & k \in \mathcal{S}^c, \end{cases} \tag{3.13}$$

[9] If the initial states are non-zero mean, then $z_1^k = \mathbb{E}(\bar{x}_1^k)$ for $k \notin \mathcal{S}$.

where $\bar{L}_t$ is as defined in Theorem 3.1. At each time t, every agent stores $\mathbf{z}_t$; hence, all agents can compute $\mathbf{z}_{t+1}$ from (3.13) given the common information $(\{\bar{s}^k_{t+1}\}_{k\in\mathcal{S}}, \mathbf{z}_t)$.

Now, consider the following certainty equivalence strategy for PMFS-IS: for agent i of sub-population $k, k \in \mathcal{K}$,

$$v^i_t = \breve{L}^k_t(s^i_t - z^k_t) + \bar{L}^k_t \mathbf{z}_t. \tag{3.14}$$

The above strategy is similar to the optimal strategy for MFS-IS (given by (3.8) in Theorem 3.1) except that the mean-field $\{\bar{s}^k_t\}_{k\in\mathcal{K}}$ has been replaced by $\mathbf{z}_t$. For ease of exposition, let $d_x \coloneqq \sum_{k\in\mathcal{K}} d^k_x$ and matrix $H = \mathrm{rows}(H^1, \dots, H^K)$ be a binary matrix such that

$$H^k = \begin{cases} 0_{d^k_x \times d_x}, & k \in \mathcal{S}, \\ \mathbb{1}_{d^k_x \times d_x}, & k \in \mathcal{S}^c. \end{cases}$$

Let $\hat{J}$ denote the performance of strategy (3.14) and J^* denote the optimal performance under MFS-IS. Then, the difference in performance $\hat{J} - J^*$ is bounded. In particular,

Theorem 3.2 *Assume (A.3.2), (A.3.3), and (PMFS-IS). Then, the performance loss is*

$$\hat{J} - J^* = \mathrm{Tr}(\tilde{X}_1 \tilde{M}_1) + \sum_{t=1}^{T-1} \mathrm{Tr}(\tilde{W}_t \tilde{M}_{t+1}), \tag{3.15}$$

where $\tilde{X}_1 = \mathbb{1}_{2d_x \times 2d_x} \otimes [H \circ \mathrm{var}(\bar{\mathbf{x}}_1)]$, $\tilde{W}_t = \mathbb{1}_{2d_x \times 2d_x} \otimes [H \circ \mathrm{var}(\bar{\mathbf{w}}_t)]$, *and* $\tilde{M}_{1:T}$ *is the solution of following Lyapunov equation:*

$$\tilde{M}_{1:T} = \mathrm{DLE}_{\mathrm{T}}(\tilde{A}_{1:T}, \tilde{Q}_{1:T}), \tag{3.16}$$

where

$$\tilde{A} = \begin{bmatrix} \tilde{A}^1_t & -(\mathbb{1}_{d_x \times d_x} - H) \circ \tilde{A}^2_t \\ 0 & H \circ \tilde{A}^2_t \end{bmatrix}, \qquad \tilde{Q} = \begin{bmatrix} -\tilde{Q}^1_t & -0 \\ 0 & \tilde{Q}^2_t \end{bmatrix},$$

where $\tilde{A}^1_t = \bar{A}_t + \bar{B}_t \bar{L}_t$, $\tilde{A}^2_t = \bar{A}_t + \bar{B}_t \breve{L}_t$, $\tilde{Q}^1_t = P^x_t + \bar{Q}_t + \bar{L}^\intercal_t (P^u_t + \bar{R}_t)\bar{L}_t$, $\tilde{Q}^2_t = P^x_t + \bar{Q}_t + \breve{L}^\intercal_t (P^u_t + \bar{R}_t)\breve{L}_t$, *and* $\breve{L}_t = \mathrm{diag}(\breve{L}^1_t, \dots, \breve{L}^K_t)$.

The result is proved in Section 3.4. Note that when the mean-fields of all sub-populations are shared, then $\mathcal{S} = \mathcal{K}$ and, therefore, H is zero. Consequently, the approximation error given by (3.15) is zero. Hence, the result of Theorem 3.2 is consistent with Theorem 3.1.

Corollary 3.2 *When the mean-field is not shared, i.e., $\mathcal{S} = \emptyset$, the approximation error is*

$$\hat{J} - J^* = \mathrm{Tr}\left(\mathrm{var}(\bar{\mathbf{x}}_1)(\tilde{M}_1^2 - \tilde{M}_1^1)\right) + \sum_{t=1}^{T-1} \mathrm{Tr}(\mathrm{var}\left(\bar{\mathbf{w}}_t)(\tilde{M}_{t+1}^2 - \tilde{M}_{t+1}^1)\right),$$

where $\tilde{M}_{1:T}^1$ and $\tilde{M}_{1:T}^2$ are the solutions of following two decoupled Lyapunov equations:

$$\tilde{M}_{1:T}^1 = \mathrm{DLE}_{\mathrm{T}}(\tilde{A}_{1:T}^1, \tilde{Q}_{1:T}^1), \quad \tilde{M}_{1:T}^2 = \mathrm{DLE}_{\mathrm{T}}(\tilde{A}_{1:T}^2, \tilde{Q}_{1:T}^2).$$

Proof: When $\mathcal{S} = \emptyset$, H is $\mathbb{1}_{d_x \times d_x}$; thus, $\tilde{A}_t$ is block diagonal. Consequently, the Lyapunov equation (3.16) decouples into the two smaller Lyapunov equations given above. ■

Theorem 3.3 *Let $n = \min_{k \in \mathcal{S}^c}(|\mathcal{N}^k|)$. Under (A.3.2), (A.3.4), and (A.3.5),*

$$\hat{J} - J^* \in \mathcal{O}\left(\frac{T}{n}\right).$$

The result is proved in Section 3.4

Remark 3.6 As the number of agents in each sub-population $k \in \mathcal{S}^c$, becomes large, the approximation error $\hat{J} - J^*$ goes to zero; therefore, PMFS-IS is as informative as MFS-IS.

3.4 Proof of main results

Proof of the results of MFS-IS

The main idea of the proof is as follows. We construct an auxiliary system whose state, control actions, and per-step cost are equivalent to $\mathbf{x}_t$, $\mathbf{u}_t$, and $c_t(\cdot)$, respectively (modulo a change of variables that we describe later). However, this auxiliary system is centrally controlled by a single agent that has access to all the information available to the $\mathcal{N}$ decentralized agents in the original system. We show that the optimal centralized solution of this auxiliary system can be implemented in the original decentralized system, and is therefore also optimal for the decentralized system.

The auxiliary system

Define $\breve{x}_t^i = x_t^i - \bar{x}_t^k$ and $\breve{u}_t^i = u_t^i - \bar{u}_t^k$. The auxiliary system is a centralized system with state $\mathring{\mathbf{x}}_t = \mathrm{vec}((\breve{x}_t^i)_{i \in \mathcal{N}}, \bar{\mathbf{x}}_t)$ and action $\mathring{\mathbf{u}}_t = \mathrm{vec}((\breve{u}_t^i)_{i \in \mathcal{N}}, \bar{\mathbf{u}}_t)$. Note that $\mathring{\mathbf{x}}_t$ is equivalent to $\mathbf{x}_t$

and $\mathring{\mathbf{u}}_t$ is equivalent to $\mathbf{u}_t$.

The dynamics and cost of the auxiliary model are same as the model of Section 3.2. This implies that we can write,

$$\breve{x}^i_{t+1} = A^k_t \breve{x}^i_t + B^k_t \breve{u}^i_t + \breve{w}^i_t, \tag{3.17}$$

where $\breve{w}^i_t \coloneqq w^i_t - \bar{w}^k_t$ and $\bar{w}^k_t \coloneqq \langle (w^i_t)_{i \in \mathcal{N}^k} \rangle$ and

$$\bar{\mathbf{x}}_{t+1} = \bar{A}_t \bar{\mathbf{x}}_t + \bar{B}_t \bar{\mathbf{u}}_t + \bar{\mathbf{w}}_t, \tag{3.18}$$

where $\bar{\mathbf{w}}_t \coloneqq \mathrm{vec}(\bar{w}^1_t, \dots, \bar{w}^K_t)$ and $\bar{A}_t$ and $\bar{B}_t$ are defined as in Theorem 3.1. In the auxiliary system, there is a *single centralized agent* that chooses $\mathring{\mathbf{u}}_t$ based on the observations. In particular, the centralized agent observes $\mathring{\mathbf{x}}_t$ and chooses $\mathring{\mathbf{u}}_t$ according to

$$\mathring{\mathbf{u}}_t = \mathring{g}_t(\mathring{\mathbf{x}}_{1:t}, \mathring{\mathbf{u}}_{1:t-1}). \tag{3.19}$$

The performance of strategy $\mathring{\mathbf{g}} \coloneqq (\mathring{g}_1, \dots, \mathring{g}_T)$ is given by

$$\mathring{J}(\mathring{\mathbf{g}}) = \mathbb{E}^{\mathring{\mathbf{g}}} \Bigg[\sum_{t=1}^{T-1} c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) + c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T) \Bigg], \tag{3.20}$$

where the expectation is with respect to the measure induced on all system variables by the choice of strategy $\mathring{\mathbf{g}}$. We are interested in the following optimization problem.

Problem 3.2 *In the auxiliary model, find strategy $\mathring{\mathbf{g}}^*$ that minimizes (3.20), i.e.,*

$$\mathring{J}^* \coloneqq \mathring{J}(\mathring{\mathbf{g}}^*) = \inf_{\mathring{\mathbf{g}}} \mathring{J}(\mathring{\mathbf{g}}),$$

where the infimum is taken over all strategies of the form (3.19).

Let J^* and $\mathring{J}^*$ denote the optimal cost for Problem 3.1 and Problem 3.2, respectively. Since the per-step cost is the same in both cases, but Problem 3.2 is centralized, we have that $J^* \geq \mathring{J}^*$. We identify the optimal control laws for the auxiliary system and show that these laws can be implemented in, and therefore are optimal for, the original decentralized system.

A critical step in the proof is to rewrite the per-step cost $c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)$ and terminal cost $c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T)$ in terms of $\mathring{\mathbf{x}}_t$ and $\mathring{\mathbf{u}}_t$. For that matter, we need the following key result that is similar to Huygens-Steiner Theorem in mechanics [Kane and Levinson, 1985]:

Lemma 3.1 *For any* $\mathbf{x} = \mathrm{vec}(x^1, \dots, x^N)$ *and* $\bar{x} = \langle \mathbf{x} \rangle$, *let* $\breve{x}^i = x^i - \bar{x}$, $i \in \{1, \dots, N\}$. *Then, for any matrix* Q *of appropriate dimension,*

$$\frac{1}{N} \sum_{i=1}^{N} (x^i)^\intercal Q x^i = \frac{1}{N} \sum_{i=1}^{N} (\breve{x}^i)^\intercal Q \breve{x}^i + \bar{x}^\intercal Q \bar{x}.$$

Proof: The result follows from elementary algebra and the observation that $\sum_{i=1}^{N} \breve{x}^i = 0$. ■

An immediate consequence of Lemma 3.1 is the following:

Corollary 3.3 *For time* t, $t \in \{1, \dots, T\}$, *there exists function* $\mathring{c}_t$, *such that for* $t \in \{1, \dots, T-1\}$, $c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \mathring{c}_t(\mathring{\mathbf{x}}_t, \mathring{\mathbf{u}}_t)$ *and for* $t = T$, $c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T) = \mathring{c}_T(\mathring{\mathbf{x}}_T)$. *In particular, for* $t \in \{1, \dots, T-1\}$,

$$\mathring{c}_t(\mathring{\mathbf{x}}_t, \mathring{\mathbf{u}}_t) = \bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \breve{c}_t^k(\breve{x}_t^i, \breve{u}_t^i),$$

where

$$\bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \bar{\mathbf{x}}_t^\intercal (\bar{Q}_t + P_t^x) \bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^\intercal (\bar{R}_t + P_t^u) \bar{\mathbf{u}}_t,$$
$$\breve{c}_t^k(\breve{x}_t^i, \breve{u}_t^i) = \frac{1}{|\mathcal{N}^k|} \left[(\breve{x}_t^i)^\intercal Q_t^k \breve{x}_t^i + (\breve{u}_t^i)^\intercal R_t^k \breve{u}_t^i \right],$$

and for $t = T$,

$$\mathring{c}_T(\mathring{\mathbf{x}}_T) = \bar{c}_T(\bar{\mathbf{x}}_T) + \sum_{i \in \mathcal{N}^k, k \in \mathcal{K}} \breve{c}_T^k(\breve{x}_T^i),$$

where

$$\bar{c}_T(\bar{\mathbf{x}}_T) = \bar{\mathbf{x}}_T^\intercal (\bar{Q}_T + P_T^x) \bar{\mathbf{x}}_T, \quad \breve{c}_T^k(\breve{x}_T^i) = \frac{1}{|\mathcal{N}^k|} (\breve{x}_T^i)^\intercal Q_T^k \breve{x}_T^i.$$

The auxiliary system is a centralized LQR system. So, the optimal control laws are linear and the optimal gains are given by the solution of an appropriate Riccati equation. However, the dimension of the state $\mathring{\mathbf{x}}_t$, and therefore the dimension of the Riccati equation, increases with the number of agents. We present an alternative approach that involves solving $K+1$ Riccati equations that do not depend on the number of agents.

The Optimal Solution of the Auxiliary System

The auxiliary system is a stochastic linear quadratic system. From the certainty equivalence principle [Caines, 1987], we know that the optimal control law is unique and identical to the

control law in the corresponding deterministic system, whose dynamics are given as follows:

$$\breve{x}^i_{t+1} = A^k_t \breve{x}^i_t + B^k_t \breve{u}^i_t, \;\; i \in \mathcal{N}^k, k \in \mathcal{K}, \qquad \bar{\mathbf{x}}_{t+1} = \bar{A}_t \bar{\mathbf{x}}_t + \bar{B}_t \bar{\mathbf{u}}_{\mathbf{t}},$$

and whose per-step cost is $\mathring{c}_t(\mathring{\mathbf{x}}_t, \mathring{\mathbf{u}}_t)$ given by Corollary 3.3.

Note that this system consists of $(N+1)$ components: N components with state $\breve{x}^i_t$ and action $\breve{u}^i_t$, $i \in \mathcal{N}$, and one component with state $\bar{\mathbf{x}}_t$ and action $\bar{\mathbf{u}}_t$. The first N components are split into K classes of identical components—one for each sub-population. The components have decoupled dynamics and decoupled cost. Thus, the optimal control law of each class may be identified separately. In particular,

Theorem 3.4 *The optimal control strategy of the auxiliary model is unique and given by*

$$\bar{\mathbf{u}}_t = \bar{L}_t \bar{\mathbf{x}}_t \quad \textit{and for } k \in \mathcal{K}, i \in \mathcal{N}^k, \quad \breve{u}^i_t = \breve{L}^k_t \breve{x}^i_t,$$

where the gains $\{\breve{L}^k_t, \bar{L}_t\}_{t=1}^{T-1}$ *are given as in Theorem 3.1.*

To complete the proof of Theorem 3.1, note that

$$u^i_t = \breve{u}^i_t + \bar{u}^k_t = \breve{L}^k_t (x^i_t - \bar{x}^k_t) + \bar{L}^k_t \bar{\mathbf{x}}_t.$$

Thus, the control laws specified in Theorem 3.1 are the optimal *centralized* control laws, and, a fortiori, the optimal decentralized control laws.

Proof of Corollary 3.1

Under the assumptions on the model, the dynamics, given by (3.17) and (3.18), simplify to

$$\breve{x}^i_{t+1} = A^k_t \breve{x}^i_t + \breve{w}^i_t, \quad \bar{\mathbf{x}}_{t+1} = \bar{A}_t \bar{\mathbf{x}}_t + \tilde{B}_t \tilde{u}_t + \bar{\mathbf{w}}_t,$$

and $\bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)$ of Corollary 3.3 simplifies to

$$\bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \bar{\mathbf{x}}_t^\intercal (\bar{Q}_t + P^x_t) \bar{\mathbf{x}}_t + \tilde{u}_t^\intercal (\tilde{P}^u_t) \tilde{u}_t. \tag{3.21}$$

Thus, the N subsystems corresponding to $\breve{x}^i_t$ are uncontrolled and we need to identify $\tilde{u}_t$ to optimally control the dynamics of mean-field $\bar{\mathbf{x}}_t$ with per-step cost given by (3.21). Hence,

the optimal solution is given by

$$\tilde{u}_t = \bar{L}_t \bar{\mathbf{x}}_t = \sum_{k \in \mathcal{K}} \tilde{L}_t^k \bar{x}_t^k,$$

where $\bar{L}_t$ is computed as explained in Corollary 3.1. To complete the proof, note that if agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$ chooses action $u_t^i = \theta_t^{k^+} \tilde{L}_t^k \bar{x}_t^k$, then we get $\theta_t^k \bar{u}_t^k = \tilde{L}_t^k \bar{x}_t^k$; consequently, $\tilde{u}_t = \sum_{k \in \mathcal{K}} \theta_t^k \bar{u}_t^k = \sum_{k \in \mathcal{K}} \tilde{L}_t^k \bar{x}_t^k$.

Proof of the results of PMFS-IS

To simplify the presentation, for any $k \in \mathcal{K}$ and $i \in \mathcal{N}^k$, we define $\breve{x}_t^i = x_t^i - \bar{x}_t^k$, $\breve{u}_t^i = u_t^i - \bar{u}_t^k$, $\breve{s}_t^i = s_t^i - \bar{s}_t^k$ and $\breve{v}_t^i = v_t^i - \bar{v}_t^k$. Then, we have

Lemma 3.2 *For all t, $\breve{s}_t^i = \breve{x}_t^i$ and $\breve{u}_t^i = \breve{v}_t^i$.*

Proof: We prove this lemma by induction. Note that $\breve{x}_1^i = \breve{s}_1^i$ and $\breve{v}_1^i = \breve{L}_1^k \breve{s}_1^i = \breve{L}_1^k \breve{x}_1^i = \breve{u}_1^i$. This forms the basis of induction. Now assume that $\breve{s}_t^i = \breve{x}_t^i$ and $\breve{v}_t^i = \breve{u}_t^i$ and consider time $t+1$. Then,

$$\breve{s}_{t+1}^i = A_t^k \breve{s}_t^i + B_t^k \breve{v}_t^i + \breve{w}_t^i = A_t^k \breve{x}_t^i + B_t^k \breve{u}_t^i + \breve{w}_t^i = \breve{x}_{t+1}^i.$$

Moreover, $\breve{v}_{t+1}^i = \breve{L}_{t+1}^k \breve{s}_{t+1}^i = \breve{L}_{t+1}^k \breve{x}_{t+1}^i = \breve{u}_{t+1}^i$. Thus, the result is true by induction. ■

From Corollary 3.3, J^* and $\hat{J}$ may be re-written (re-arranged) as follows:

$$J^* = \mathbb{E}\big[\bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \breve{c}_t^k(\breve{x}_t^i, \breve{u}_t^i)\big], \tag{3.22}$$

$$\hat{J} = \mathbb{E}\big[\bar{c}_t(\bar{\mathbf{s}}_t, \bar{\mathbf{v}}_t) + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \breve{c}_t^k(\breve{s}_t^i, \breve{v}_t^i)\big].$$

Notice that Corollary 3.3 is just a simple re-arrangement (based on some algebraic manipulation) and it does not depend on the information structure. Now, we simplify $\hat{J} - J^*$ using Lemma 3.2.

Lemma 3.3 *The difference in performance may be written as follows:*

$$\hat{J} - J^* = \sum_{t=1}^{T} \left[\bar{c}_t(\bar{\mathbf{s}}_t, \bar{\mathbf{v}}_t) - \bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)\right]. \tag{3.23}$$

Proof: Equation (3.23) immediately follows from Lemma 3.2 and (3.22). ■

Next we simplify (3.23) in terms of the following relative errors: For any $k \in \mathcal{K}$, define

$$\zeta_t^k = \bar{x}_t^k - z_t^k \quad \text{and} \quad \xi_t^k = \bar{s}_t^k - z_t^k.$$

Let $\boldsymbol{\zeta}_t = \mathrm{vec}(\zeta_t^1, \dots, \zeta_t^K)$ and $\boldsymbol{\xi}_t = \mathrm{vec}(\xi_t^1, \dots, \xi_t^K)$. For ease of exposition, let vector $h = \mathrm{vec}(h^1, \dots, h^K)$ be binary such that $h^k = 0_{d_x^k \times 1}$ if $k \in \mathcal{S}$ and $h^k = \mathbb{1}_{d_x^k \times 1}$ if $k \in \mathcal{S}^c$.

Lemma 3.4 *Let $\tilde{A}_t$ be defined as in Theorem 3.2. Then, $\boldsymbol{\zeta}_1 = h \circ \bar{\mathbf{x}}_1$ and $\boldsymbol{\xi}_1 = h \circ \bar{\mathbf{x}}_1$ and*

$$\begin{bmatrix} \boldsymbol{\zeta}_{t+1} \\ \boldsymbol{\xi}_{t+1} \end{bmatrix} = \tilde{A}_t \begin{bmatrix} \boldsymbol{\zeta}_t \\ \boldsymbol{\xi}_t \end{bmatrix} + \begin{bmatrix} h \circ \bar{\mathbf{w}}_t \\ h \circ \bar{\mathbf{w}}_t \end{bmatrix}.$$

Proof: From (3.12) and (3.14), we get

$$\begin{aligned} \bar{s}_{t+1}^k &= A_t^k \bar{s}_t^k + B_t^k \bar{v}_t^k + D_t^k \bar{\mathbf{s}}_t + E_t^k \bar{\mathbf{v}}_t + \bar{w}_t^k, \\ \bar{v}_t^k &= \breve{L}_t^k (\bar{s}_t^k - z_t^k) + \bar{L}_t^k \mathbf{z}_t, \end{aligned} \tag{3.24}$$

where $\bar{w}_t^k \coloneqq \langle (w_t^i)_{i \in \mathcal{N}^k} \rangle$. Write (3.24) in a vectorized form,

$$\bar{\mathbf{s}}_{t+1} = \bar{A}_t \bar{\mathbf{s}}_t + \bar{B}_t \bar{\mathbf{v}}_t + \bar{\mathbf{w}}_t, \quad \bar{\mathbf{v}}_t = \breve{L}_t \boldsymbol{\xi}_t + \bar{L}_t \mathbf{z}_t,$$

where $\bar{\mathbf{w}}_t = \mathrm{vec}(\bar{w}_k^1, \dots, \bar{w}_t^K)$. From Theorem 3.1, we can write the dynamics under the optimal strategy as follows

$$\begin{aligned} \bar{x}_{t+1}^k &= A_t^k \bar{x}_t^k + (B_t^k \bar{L}_t^k + D_t^k + E_t^k \bar{L}_t) \bar{\mathbf{x}}_t + \bar{w}_t^k, \\ \bar{u}_t^k &= \bar{L}_t^k \bar{\mathbf{x}}_t, \end{aligned}$$

and in a vectorized form,

$$\bar{\mathbf{x}}_{t+1} = (\bar{A}_t + \bar{B}_t \bar{L}_t) \bar{\mathbf{x}}_t + \bar{\mathbf{w}}_t, \quad \bar{\mathbf{u}}_t = \bar{L}_t \bar{\mathbf{x}}_t.$$

Thus, the dynamics of the relative errors can be written as follows. If $k \in \mathcal{S}$,

$$\begin{aligned}\zeta_{t+1}^k &= A_t^k \zeta_t^k + (B_t^k \bar{L}_t^k + D_t^k + E_t^k \bar{L}_t)\boldsymbol{\zeta}_t - (A_t^k + B_t^k \breve{L}_t^k)\xi_t^k - (D_t^k + E_t^k \breve{L}_t)\boldsymbol{\xi}_t,\\ \xi_{t+1}^k &= 0,\end{aligned}$$

and if $k \in \mathcal{S}^c$,

$$\begin{aligned}\zeta_{t+1}^k &= A_t^k \zeta_t^k + (B_t^k \bar{L}_t^k + D_t^k + E_t^k \bar{L}_t)\boldsymbol{\zeta}_t + \bar{w}_t^k,\\ \xi_{t+1}^k &= (A_t^k + B_t^k \breve{L}_t^k)\xi_t^k + (D_t^k + E_t^k \breve{L}_t)\boldsymbol{\xi}_t + \bar{w}_t^k.\end{aligned}$$

Combining these, gives the result of the Lemma. ■

Let $\mathcal{F}_t = \{\bar{s}_{1:t}^k\}_{k\in\mathcal{S}}$ be the history of the mean-fields of sub-populations $\mathcal{S}$ that are observed.

Lemma 3.5 *For all t, $\mathbb{E}[\boldsymbol{\zeta}_t|\mathcal{F}_t] = \mathbb{E}[\boldsymbol{\xi}_t|\mathcal{F}_t] = 0$.*

Proof: If $k \in \mathcal{S}$, $\zeta_1^k = \xi_1^k = 0$ and if $k \in \mathcal{S}^c$, $\zeta_1^k = \xi_1^k = \bar{x}_1^k$, and from (A.3.3), $\mathbb{E}[\bar{x}_1^k|\mathcal{F}_1] = \mathbb{E}[\bar{x}_1^k] = 0$. Therefore, $\mathbb{E}[\boldsymbol{\zeta}_1|\mathcal{F}_1] = \mathbb{E}[\boldsymbol{\xi}_1|\mathcal{F}_1] = 0$. Thus, from Lemma 3.4 and $\mathbb{E}[\bar{\mathbf{w}}_t|\mathcal{F}_t] = 0$, we get that $\mathbb{E}[\boldsymbol{\zeta}_t|\mathcal{F}_t] = \mathbb{E}[\boldsymbol{\xi}_t|\mathcal{F}_t] = 0$. ■

Lemma 3.6 *$\mathbf{z}_t$ is measurable with respect to $\mathcal{F}_t$, therefore, $\mathbb{E}[\mathbf{z}_t|\mathcal{F}_t] = \mathbf{z}_t$.*

Proposition 3.1 *The relative loss is given*

$$\hat{J} - J^* = \mathbb{E}\left[\sum_{t=1}^{T} [\boldsymbol{\zeta}_t \quad \boldsymbol{\xi}_t]^\intercal \tilde{Q}_t [\boldsymbol{\zeta}_t \quad \boldsymbol{\xi}_t]\right].$$

Proof: Recall that $\bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \bar{\mathbf{x}}_t^\intercal(\bar{Q}_t + P_t^x)\bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^\intercal(\bar{R}_t + P_t^u)\bar{\mathbf{u}}_t$. The proof follows immediately from (3.23) and the following observation:

Lemma 3.7 *Let $\hat{Q}_t \coloneqq \bar{Q}_t + P_t^x$ and $\hat{R}_t \coloneqq \bar{R}_t + P_t^u$. Then,*

$$\mathbb{E}[\bar{\mathbf{s}}_t^\intercal \hat{Q}_t \bar{\mathbf{s}}_t - \bar{\mathbf{x}}_t^\intercal \hat{Q}_t \bar{\mathbf{x}}_t|\mathcal{F}_t] = \mathbb{E}[\boldsymbol{\xi}_t^\intercal \hat{Q}_t \boldsymbol{\xi}_t - \boldsymbol{\zeta}_t^\intercal \hat{Q}_t \boldsymbol{\zeta}_t|\mathcal{F}_t],$$

and

$$\mathbb{E}[\bar{\mathbf{v}}_t^\intercal \hat{R}_t \bar{\mathbf{v}}_t - \bar{\mathbf{u}}_t^\intercal \hat{R}_t \bar{\mathbf{u}}_t|\mathcal{F}_t] = \mathbb{E}[\boldsymbol{\xi}_t^\intercal \breve{L}_t^\intercal \hat{R}_t \breve{L}_t \boldsymbol{\xi}_t|\mathcal{F}_t] - \mathbb{E}[\boldsymbol{\zeta}_t^\intercal \bar{L}_t^\intercal \hat{R}_t \bar{L}_t \boldsymbol{\zeta}_t|\mathcal{F}_t].$$

Therefore, the proof of Proposition 3.1 is complete. ■

Proof of Lemma 3.7:

1. Substituting $\bar{\mathbf{s}}_t = \boldsymbol{\xi}_t + \mathbf{z}_t$ and $\bar{\mathbf{x}}_t = \boldsymbol{\zeta}_t + \mathbf{z}_t$, we get

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{s}}_t^\intercal \hat{Q}_t \bar{\mathbf{s}}_t - \bar{\mathbf{x}}_t^\intercal \hat{Q}_t \bar{\mathbf{x}}_t | \mathcal{F}_t] &\overset{(a)}{=} \mathbb{E}[\boldsymbol{\xi}_t^\intercal \hat{Q}_t \boldsymbol{\xi}_t - \boldsymbol{\zeta}_t^\intercal \hat{Q}_t \boldsymbol{\zeta}_t | \mathcal{F}_t] + 2\mathbb{E}[\boldsymbol{\xi}_t^\intercal \hat{Q}_t \mathbf{z}_t | \mathcal{F}_t] - 2\mathbb{E}[\boldsymbol{\zeta}_t^\intercal \hat{Q}_t \mathbf{z}_t | \mathcal{F}_t] \\ &= \mathbb{E}[\boldsymbol{\xi}_t^\intercal \hat{Q}_t \boldsymbol{\xi}_t - \boldsymbol{\zeta}_t^\intercal \hat{Q}_t \boldsymbol{\zeta}_t | \mathcal{F}_t],\end{aligned}$$

   where the last two terms in (a) are zero by Lemmas 3.5 and 3.6.

2. Substituting $\bar{\mathbf{v}}_t = \breve{L}_t \boldsymbol{\xi}_t + \bar{L}_t \mathbf{z}_t$ and $\bar{\mathbf{u}}_t = \bar{L}_t \bar{\mathbf{x}}_t = \bar{L}_t(\boldsymbol{\zeta}_t + \mathbf{z}_t)$, we get

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{v}}_t^\intercal \hat{R}_t \bar{\mathbf{v}}_t - \bar{\mathbf{u}}_t^\intercal \hat{R}_t \bar{\mathbf{u}}_t | \mathcal{F}_t] &\overset{(b)}{=} \mathbb{E}[\boldsymbol{\xi}_t^\intercal \breve{L}_t^\intercal \hat{R}_t \breve{L}_t \boldsymbol{\xi}_t - \boldsymbol{\zeta}_t^\intercal \bar{L}_t^\intercal \hat{R}_t \bar{L}_t \boldsymbol{\zeta}_t | \mathcal{F}_t] \\ &\quad + 2\mathbb{E}[\boldsymbol{\xi}_t^\intercal \breve{L}_t^\intercal \hat{R}_t \bar{L}_t \mathbf{z}_t | \mathcal{F}_t] - 2\mathbb{E}[\boldsymbol{\zeta}_t^\intercal \bar{L}_t \hat{R}_t \bar{L}_t \mathbf{z}_t | \mathcal{F}_t] = \mathbb{E}[\boldsymbol{\xi}_t^\intercal \breve{L}_t^\intercal \hat{R}_t \breve{L}_t \boldsymbol{\xi}_t - \boldsymbol{\zeta}_t^\intercal \bar{L}_t^\intercal \hat{R}_t \bar{L}_t \boldsymbol{\zeta}_t | \mathcal{F}_t],\end{aligned}$$

   where the last two terms in (b) are zero by Lemmas 3.5 and 3.6. ■

Proof of Theorem 3.2

Note that $\hat{J} - J^*$ is the expected total quadratic cost (given by Proposition 3.1) of a linear (uncontrolled) system (given by Lemma 3.4). Thus, $\hat{J} - J^*$ is given by (3.15) where $\tilde{M}_{1:T}$ is the solution of the Lyapunov equation (3.16). Note that the variance of the initial state and noises in Lemma 3.4 are given as follows:

$$\begin{aligned}\operatorname{var}(h \circ \bar{\mathbf{x}}_1, h \circ \bar{\mathbf{x}}_1) &= \mathbb{1}_{2d_x \times 2d_x} \otimes [H \circ \operatorname{var}(\bar{\mathbf{x}}_1)] =: \tilde{X}_1, \\ \operatorname{var}(h \circ \bar{\mathbf{w}}_t, h \circ \bar{\mathbf{w}}_t) &= \mathbb{1}_{2d_x \times 2d_x} \otimes [H \circ \operatorname{var}(\bar{\mathbf{w}}_t)] =: \tilde{W}_t.\end{aligned}$$

Proof of Theorem 3.3

First observe that due to (A.3.5), matrices $\tilde{A}_t$ and $\tilde{Q}_t$ do not depend on $(|\mathcal{N}^1|, \dots, |\mathcal{N}^K|)$; therefore, neither does $\tilde{M}_{1:T}$. Thus the only dependence on the size of the sub-population is due to $\tilde{X}_1$ and $\tilde{W}_t$. Under (A.3.4) and (A.3.5), for any sub-population $k \in \mathcal{K}$,

$$\begin{aligned}\operatorname{var}(\bar{x}_1^k) &= \frac{1}{|\mathcal{N}^k|^2} \sum_{i \in \mathcal{N}^k} \operatorname{var}(x_1^i) \leq \frac{c_x^k}{n}, \\ \operatorname{var}(\bar{w}_t^k) &= \frac{1}{|\mathcal{N}^k|^2} \sum_{i \in \mathcal{N}^k} \operatorname{var}(w_t^i) \leq \frac{c_w^k}{n}.\end{aligned}$$

From (A.3.4), $\mathrm{var}(\bar{\mathbf{x}}_1) = \mathrm{diag}(\mathrm{var}(\bar{x}_1^1), \dots, \mathrm{var}(\bar{x}_1^K))$ and $\mathrm{var}(\bar{\mathbf{w}}_t) = \mathrm{diag}(\mathrm{var}(\bar{w}_t^1), \dots, \mathrm{var}(\bar{w}_t^K))$. Thus

$$\tilde{X}_1 \le \frac{1}{n} \mathbb{1}_{2d_x \times 2d_x} \otimes \left[H \circ \mathrm{diag}(c_x^1, \dots, c_x^K) \right],$$
$$\tilde{W}_t \le \frac{1}{n} \mathbb{1}_{2d_x \times 2d_x} \otimes \left[H \circ \mathrm{diag}(c_w^1, \dots, c_w^K) \right].$$

Thus, $\tilde{X}_1$ and $\tilde{W}_t$ are $\mathcal{O}(\frac{1}{n})$. From (3.15), we have

$$|\hat{J} - J^*| \le \left| \mathrm{Tr}\left(\tilde{X}_1 \tilde{M}_1 \right) \right| + \sum_{t=1}^{T-1} \left| \mathrm{Tr}\left(\tilde{W}_t \ \tilde{M}_{t+1} \right) \right|,$$

where each of above absolute values is $\mathcal{O}(\frac{1}{n})$. In particular, since $\tilde{X}_1$ and $\tilde{W}_t$ are $\mathcal{O}(\frac{1}{n})$ and $\tilde{M}_{1:T}$ do not depend on n, $|\mathrm{Tr}(\tilde{X}_1 \tilde{M}_1|)$ and $|\mathrm{Tr}(\tilde{W}_t \tilde{M}_{t+1})|$ are $\mathcal{O}(\frac{1}{n})$.

3.5 Generalizations

In this section, we show that our results generalize to variations of Problem 3.1. We only present the results for MFS-IS (i.e., the analogue of Theorem 3.1). The results for PMFS (i.e., the analogue of Theorems 3.2 and 3.3) may be derived in a similar manner; in particular, by replacing the mean-field with its predicted value in the strategy for MFS-IS..

3.5.1 Major agent and a population of minor agents

Suppose there exists one sub-population, say 1, with only 1 agent, i.e., $|\mathcal{N}^1| = 1$. Then, $\bar{x}_1^1 = x_t^1$. The rest of the dynamics and cost are the same as in Section 3.2. Since the dynamics are coupled through the mean-field, the state of the agent of sub-population 1 directly influences the dynamics of all other agents and the per-step cost. For this reason, such an agent is called a *major* agent. A variation of the above model was first introduced in [Huang, 2010] and other variations have been investigated in [Caines and Kizilkale, 2014, Huang et al., 2014, Arabneydi and Mahajan, 2015]. For above model, result of Theorem 3.1 simplifies as follows.

Corollary 3.4 *For any sub-population $k \in \mathcal{K} \backslash \{1\}$ and minor agent $i \in \mathcal{N}^k$, u_t^i is given by*

(3.8). *For the major agent, the control law is given by*

$$u_t^1 = \bar{L}_t^1 \bar{\mathbf{x}}_t. \tag{3.25}$$

Note that $\breve{L}_t^1$ is not needed to compute u_t^1; so we do not need a Riccati equation to compute $\breve{M}_{1:T}^1$. To implement the optimal control strategies:

- all major and minor agents must compute $\bar{L}_{1:T-1}$ by solving the Riccati equation (3.10),
- minor agents of sub-population $k \in \mathcal{K}\backslash\{1\}$ must compute $\breve{L}_{1:T-1}^k$ by solving the Riccati equation (3.9).

Then, the major agent, upon observing the local state x_t^1 and the global mean-field $(\bar{x}_t^k)_{\in\mathcal{K}\backslash\{1\}}$ of minor agents, chooses its local control action according to (3.25). An individual minor agent i of sub-population k, upon observing the local state x_t^i, the global mean-field $(\bar{x}_t^k)_{\in\mathcal{K}\backslash\{1\}}$, and the state of major agent x_t^1, chooses its local control action according to (3.8).

3.5.2 Tracking cost function

Consider a tracking problem in which we are given a tracking signal $\{s_t^k\}_{t=1}^T$, $s_t^k \in \mathbb{R}^{d_x^k}$ for the mean-field of sub-population $k \in \mathcal{K}$ and a tracking signal $\{r_t^i\}_{t=1}^T$, $r_t^i \in \mathbb{R}^{d_x^k}$, for each agent $i \in \mathcal{N}^k$. Define $\bar{r}_t^k := \langle (r_t^i)_{i\in\mathcal{N}^k} \rangle, k \in \mathcal{K}$, $\bar{\mathbf{r}}_t := \mathrm{vec}(\bar{r}_,^1 \ldots, \bar{r}_t^K)$, and $\mathbf{s}_t = \mathrm{vec}(s_t^1, \ldots, s_t^K)$. The tracking cost is as follows. For $t \in \{1, \ldots, T-1\}$,

$$\begin{aligned} c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = (\bar{\mathbf{x}}_t - \mathbf{s}_t)^\intercal P_t^x (\bar{\mathbf{x}}_t - \mathbf{s}_t) + \bar{\mathbf{u}}_t^\intercal P_t^u \bar{\mathbf{u}}_t \\ + \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[(x_t^i - r_t^i)^\intercal Q_t^k (x_t^i - r_t^i) + (u_t^i)^\intercal R_t^k u_t^i \right], \end{aligned}$$

and for $t = T$,

$$c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T) = (\bar{\mathbf{x}}_T - \mathbf{s}_T)^\intercal P_T^x (\bar{\mathbf{x}}_T - \mathbf{s}_T) + \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} (x_T^i - r_T^i)^\intercal Q_T^k (x_T^i - r_T^i).$$

We assume that, in addition to MFS-IS specified in Section 3.2, agent i also knows signals $\{r_t^i, \bar{\mathbf{r}}_t, \mathbf{s}_t\}_{t=1}^T$. The rest of the model is the same as in Section 3.2.

Theorem 3.5 *Under (A.3.1), (A.3.2), and (MFS-IS), the optimal strategy is unique and given by*

$$u_t^i = \breve{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^k \bar{\mathbf{x}}_t + \breve{F}_t^k v_t^i + \bar{F}_t^k \bar{v}_t, \tag{3.26}$$

where the gains $\{\breve{L}_t^k, \bar{L}_t^k\}_{t=1}^{T-1}$ *are obtained by the solution of* $K+1$ *Riccati equations defined in Theorem 3.1 and the gains* $\{\breve{F}_t^k, \bar{F}_t^k\}_{t=1}^{T-1}$ *and the correction signals* $\{v_t^i, \bar{v}_t\}_{t=1}^{T}$ *are given as follows. Let* $\{\breve{M}_t^k\}_{t=1}^{T}$ *and* $\{\bar{M}_t\}_{t=1}^{T}$ *be the solutions of* $K+1$ *Riccati equations defined in Theorem 3.1. For* $t \in \{1, \dots, T-1\}$*, the gains* $\{\breve{F}_t^k, \bar{F}_t^k\}_{t=1}^{T}$ *are given by*

$$\breve{F}_t^k = \left((B_t^k)^\intercal \breve{M}_{t+1}^k B_t^k + R_t^k \right)^{-1} B_t^{k\intercal},$$

and $\mathrm{rows}(\bar{F}_t^1, \dots, \bar{F}_t^K) \coloneqq \bar{F}_t$*, where*

$$\bar{F}_t = \left(\bar{B}_t^\intercal \bar{M}_{t+1} \bar{B}_t + \bar{R}_t + P_t^u \right)^{-1} \bar{B}_t^\intercal.$$

The correction signals $\{v_t^i, \bar{v}_t\}_{t=1}^{T}$ *are given recursively as follows: for* $t = T$*,*

$$v_T^i = Q_T^k r_T^i, \quad \bar{v}_T = \bar{Q}_T \bar{\mathbf{r}}_T + P_T^x \mathbf{s}_T, \tag{3.27}$$

and for $t \in \{T-1, \dots, 1\}$*,*

$$v_t^i = (A_t^k + B_t^k \breve{L}_t^k)^\intercal v_{t+1}^i + Q_t^k r_t^i, \tag{3.28}$$

$$\bar{v}_t = (\bar{A}_t + \bar{B}_t \bar{L}_t)^\intercal \bar{v}_{t+1} + \bar{Q}_t \bar{\mathbf{r}}_t + P_t^x \mathbf{s}_t. \tag{3.29}$$

The proof is presented in Appendix A.1. To implement the optimal control strategies:

- all agents must compute $\bar{L}_{1:T-1}$ and $\bar{F}_{1:T-1}$ by solving Riccati equation (3.10) and compute the global correction signal $\bar{v}_{1:T}$ by solving backward equations (3.27) and (3.29),
- agents of sub-population k must compute $\breve{L}_{1:T-1}^k$ and $\breve{F}_{1:T-1}^k$ by solving Riccati equation (3.9),
- an individual agent i of sub-population k must compute a local correction signal $v_{1:T}^i$ by solving backward equations (3.27) and (3.28).

Then, an individual agent i of sub-population k, upon observing the local state x_t^i and the global mean-field $\bar{\mathbf{x}}_t$, chooses its local control action according to (3.26).

3.5.3 Weighted mean-field

Suppose there are weights (a^i, λ^i, b^i) associated with each agent $i \in \mathcal{N}$ such that $a^i, \lambda^i \in \mathbb{R}$ and $b^i \in \mathbb{R}_{>0}$. For each sub-population $k \in \mathcal{K}$ define the weighted mean-field of states and actions as follows.

$$\bar{x}_t^{k,\lambda} = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \lambda^i x_t^i, \qquad \bar{u}_t^{k,\lambda} = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \lambda^i u_t^i,$$
$$\bar{\mathbf{x}}_t^{\lambda} = \mathrm{vec}(\bar{x}^{1,\lambda}, \ldots, \bar{x}^{K,\lambda}), \qquad \bar{\mathbf{u}}_t^{\lambda} = \mathrm{vec}(\bar{u}^{1,\lambda}, \ldots, \bar{u}^{K,\lambda}).$$

Also, define $\bar{a}^{k,\lambda} = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \lambda^i a^i$. For sub-population $k \in \mathcal{K}$, the state of agent $i \in \mathcal{N}^k$ evolves as follows.

$$x_{t+1}^i = A_t^k x_t^i + B_t^k u_t^i + a^i (D_t^k \bar{\mathbf{x}}_t^{\lambda} + E_t^k \bar{\mathbf{u}}_t^{\lambda}) + w_t^i.$$

The per-step cost is given by

$$c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t^{\lambda}, \bar{\mathbf{u}}_t^{\lambda}) \;=\; (\bar{\mathbf{x}}_t^{\lambda})^{\intercal} P_t^x \bar{\mathbf{x}}_t^{\lambda} \;+\; (\bar{\mathbf{u}}_t^{\lambda})^{\intercal} P_t^u \bar{\mathbf{u}}_t^{\lambda} \;+\; \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{b^i}{|\mathcal{N}^k|} \left[(x_t^i)^{\intercal} Q_t^k x_t^i + (u_t^i)^{\intercal} R_t^k u_t^i \right],$$

and the terminal cost is given by

$$c_T(\mathbf{x}_T, \bar{\mathbf{x}}_T^{\lambda}) = (\bar{\mathbf{x}}_T^{\lambda})^{\intercal} P_T^x \bar{\mathbf{x}}_T^{\lambda} + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{b^i}{|\mathcal{N}^k|} \left[(x_T^i)^{\intercal} Q_T^k x_T^i \right].$$

Such models arise in applications where the interaction between two homogeneous agents is not symmetric but depends on their weights. For example, in wireless networks, the interference caused at the base-station depends on the distance of the agents from the base-station. We assume that the weights are related as follows.

A. 3.6 *For each sub-population $k \in \mathcal{K}$ and each agent $i \in \mathcal{N}^k$, $a^i b^i = \lambda^i \bar{a}^{k,\lambda}$.*

Given a sub-population $k \in \mathcal{K}$, examples of weights that satisfy (A.3.6) are: for all $i \in \mathcal{N}^k$, (i) $a^i = 0$, (ii) $a^i = 1$ and $b^i = \lambda^i$, (iii) $a^i = \lambda^i$, $b^i = 1$, and $\frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \lambda^i = 1$. To simplify the exposition, define $\mu^k := 2 - \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \frac{(\lambda^i)^2}{b^i}$.

A. 3.7 *For every t, P_t^x, P_t^u, Q_t^k, and R_t^k are symmetric matrices that satisfy*

$$Q_t^k \geq 0,\ \forall k \in \mathcal{K}, \qquad \operatorname{diag}(\mu^1 Q_t^1, \dots, \mu^K Q_t^K) + P_t^x \geq 0,$$
$$R_t^k > 0,\ \forall k \in \mathcal{K}, \qquad \operatorname{diag}(\mu^1 R_t^1, \dots, \mu^K R_t^K) + P_t^u > 0.$$

Note that if $\mu^k = 1$, (A.3.7) reduces to (A.3.2). Each agent has mean-field sharing information structure, i.e., agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$ observes the local state x_t^i and the weighted mean-field $\bar{\mathbf{x}}_t^\lambda$.

Theorem 3.6 *Under (A.3.1), (A.3.6), (A.3.7), and (MFS-IS), the optimal strategy is unique and given by*

$$u_t^i = \breve{L}_t^k \left(x_t^i - \frac{\lambda^i}{b^i} \bar{x}_t^{k,\lambda} \right) + \frac{\lambda^i}{b^i} \bar{L}_t^k \bar{\mathbf{x}}_t^\lambda, \tag{3.30}$$

where the gains $\{\breve{L}_t^k, \bar{L}_t^k\}_{t=1}^{T-1}$ are obtained by the solution of $K+1$ Riccati equations defined in Theorem 3.1 when $\bar{A}_t$, $\bar{B}_t$, $\bar{Q}_t$, and $\bar{R}_t$ are replaced by

$$\bar{A}_t \coloneqq \operatorname{diag}(A_t^1, \dots, A_t^K) + \operatorname{rows}(\bar{a}^{1,\lambda} D_t^1, \dots, \bar{a}^{K,\lambda} D_t^K), \qquad \bar{Q}_t \coloneqq \operatorname{diag}(\mu^1 Q_t^1, \dots, \mu^K Q_t^K),$$
$$\bar{B}_t \coloneqq \operatorname{diag}(B_t^1, \dots, B_t^K) + \operatorname{rows}(\bar{a}^{1,\lambda} E_t^1, \dots, \bar{a}^{K,\lambda} E_t^K), \qquad \bar{R}_t \coloneqq \operatorname{diag}(\mu^1 R_t^1, \dots, \mu^K R_t^K).$$

Proof of Theorem 3.6 is presented in Appendix A.2. To implement optimal control strategies:

- all agents must compute $\bar{L}_{1:T-1}$ by solving the Riccati equation (3.10) for new matrices $\bar{A}_t$, $\bar{B}_t$, $\bar{Q}_t$, and $\bar{R}_t$.
- agents of sub-population k must compute $\breve{L}_{1:T-1}^k$ by solving the Riccati equation (3.9).

Then, an individual agent i of sub-population k, upon observing the local state x_t^i and the global weighted mean-field $\bar{\mathbf{x}}_t^\lambda$, chooses its local control action by locally weighting the mean-fields, i.e. $\frac{\lambda^i}{b^i}$, according to (3.30).

Remark 3.7 The optimal strategy depends on the weights and, even within a sub-population, the gains of the mean-field terms are different for different agents.

Remark 3.8 If the dynamics of the agents are decoupled, i.e., $a^i = 0$ for all agents, then the results of Theorem 3.6 are similar to the model with soft constraints discussed in [Madjidian and Mirkin, 2014].

Note that if $a^i = b^i = \lambda^i = 1$ for all agents, then the weighted mean-field model reduces to the basic model described in Section 3.2 and the result of Theorem 3.6 reduces to that of Theorem 3.1.

3.6 Infinite horizon

The results presented in Section 3.3 and Section 3.5 generalize to infinite horizon setup in a natural manner. Assume that the model is time-invariant, i.e., the matrices $\{A_t^k, B_t^k, D_t^k, E_t^k, Q_t^k, R_t^k, P_t^x, P_t^u\}$ and covariances $\{\breve{\Sigma}_t^k, \bar{\Sigma}_t, \breve{\Xi}_t^k, \bar{\Xi}_t\}$ (defined in Theorem 3.1) do not depend on time; hence, we remove the subscript t. The rest of the model is as same as that in Section 3.2.

Consider the infinite horizon discounted cost and the infinite horizon long-term average setups as follows:

Problem 3.3 *Given discount factor $\beta \in (0,1)$, find a strategy $\mathbf{g}$ that minimizes the following cost:*

$$J_\beta(\mathbf{g}) = (1-\beta)\mathbb{E}^{\mathbf{g}}\left[\sum_{t=1}^{\infty} \beta^{t-1} c(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)\right],$$

where the expectation is with respect to the measure induced on all the system variables by the choice of strategy $\mathbf{g}$.

Problem 3.4 *Find a strategy $\mathbf{g}$ that minimizes the following cost:*

$$J_1(\mathbf{g}) = \lim_{T\to\infty} \mathbb{E}^{\mathbf{g}}\left[\frac{1}{T}\sum_{t=1}^{T} c(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)\right],$$

where the expectation is with respect to the measure induced on all the system variables by the choice of strategy $\mathbf{g}$.

A. 3.8 *For each sub-population $k \in \mathcal{K}$, $(\sqrt{\beta}A^k, \sqrt{\beta}B^k)$ are stabilizable and $(\sqrt{\beta}A^k, \sqrt{Q^k})$ are detectable. In addition, for $\bar{A}_t$ and $\bar{B}_t$ defined in Theorem 3.1, $(\sqrt{\beta}\bar{A}, \sqrt{\beta}\bar{B})$ are stabilizable and $(\sqrt{\beta}\bar{A}, \sqrt{\bar{Q} + P^x})$ are detectable.*

Exact solution for MFS-IS

The optimal strategy under MFS-IS is as follows.

Theorem 3.7 *Under (A.3.1), (A.3.2), (A.3.8), and (MFS-IS), the optimal strategy for Problems 3.3 and 3.4 are linear and time homogeneous and are given by*

$$u_t^i = \breve{L}^k(x_t^i - \bar{x}_t^k) + \bar{L}^k \bar{\mathbf{x}}_t, \tag{3.31}$$

where the gains $\{\breve{L}^k, \bar{L}^k\}$ *are obtained by the solution of* $K+1$ *algebraic Riccati equations given below: one for computing each* $\breve{L}^k, k \in \mathcal{K}$*, and one for* $\bar{L} \coloneqq \mathrm{rows}(\bar{L}^1, \dots, \bar{L}^K)$*. Let matrices* $\bar{A}, \bar{B}, \bar{Q}$*, and* $\bar{R}$ *be defined as in Theorem 3.1; then, given* $\beta \in (0, 1]$*,*

$$\begin{aligned}
\breve{L}^k &= -\left(B^{k\intercal}\breve{M}^k B^k + \beta^{-1}R^k\right)^{-1} B^{k\intercal}\breve{M}^k A^k, \\
\bar{L} &= -\left(\bar{B}^\intercal \bar{M}\bar{B} + \beta^{-1}(\bar{R} + P^u)\right)^{-1} \bar{B}^\intercal \bar{M}\bar{A},
\end{aligned}$$

where $\breve{M}^k$ *and* $\bar{M}$ *are the solutions of the following algebraic Riccati equations:*

$$\begin{aligned}
\breve{M}^k &= \mathrm{DARE}_\beta(A^k, B^k, Q^k, R^k), \\
\bar{M} &= \mathrm{DARE}_\beta(\bar{A}, \bar{B}, \bar{Q} + P^x, \bar{R} + P^u).
\end{aligned}$$

In addition, the optimal performance is given by

$$J_\beta^* = (1-\beta)\left[\sum_{k\in\mathcal{K}} \mathrm{Tr}\left(\breve{\Xi}^k \breve{M}^k\right) + \mathrm{Tr}(\bar{\Xi}\bar{M})\right] + \left[\sum_{k\in\mathcal{K}} \mathrm{Tr}\left(\breve{\Sigma}^k \breve{M}^k\right) + \mathrm{Tr}(\bar{\Sigma}\bar{M})\right],$$

where $\breve{\Sigma}^k$*,* $\bar{\Sigma}$*,* $\breve{\Xi}^k$*, and* $\bar{\Xi}$ *are defined as in Theorem 3.1.*

Proof: The proof follows along the same lines of the proof of Theorem 3.1. We construct an auxiliary system as in Section 3.4, which consists of $|\mathcal{N}|+1$ components with decoupled cost and dynamics coupled only through the noise. Since the costs are infinite-horizon discounted and infinite-horizon long run average, the optimal solution is given by appropriate algebraic Riccati equations.[10] ■

[10]Note that an infinite-horizon discounted problem with 4-tuple (A, B, Q, R) and discount factor β is equivalent to an undiscounted problem with 4-tuple $(\sqrt{\beta}A, \sqrt{\beta}B, Q, R)$.

Approximate solution for PMFS-IS

In this section, we propose an approximately optimal strategy for Problems 3.3 and 3.4 under PMFS-IS. Let $\breve{L} = \text{diag}(\breve{L}^1, \ldots, \breve{L}^K)$ denote a diagonal matrix with diagonal terms of $\breve{L}^k$ defined as in Theorem 3.7. We impose the following assumption.

A. 3.9 *$\sqrt{\beta}(\bar{A} + \bar{B}\breve{L})$ is Hurwitz matrix.*

Let $\hat{J}_\beta$ denote the performance of strategy (3.31) where $\bar{\mathbf{x}}_t$ is replaced by $\mathbf{z}_t$ in (3.13) and J^*_β denote the optimal performance under MFS-IS. Then, the difference in performance $\hat{J}_\beta - J^*_\beta$ is bounded. In particular, we have the following

Theorem 3.8 *Assume (A.3.2), (A.3.3), (A,3.8), (A.3.9) and (PMFS-IS). Then, for $\beta \in (0, 1]$, we have*

1. *The performance loss is given by*

$$\hat{J}_\beta - J^*_\beta = (1-\beta)\,\text{Tr}\left(\tilde{X}_1 \tilde{M}\right) + \text{Tr}\left(\tilde{W}\tilde{M}\right), \tag{3.32}$$

   where $\tilde{X}_1$ and $\tilde{W}$ are time-homogeneous and defined as in Theorem 3.2 and $\tilde{M}$ is the solution of following algebraic Lyapunov equation:

$$\tilde{M} = \text{DALE}_\beta(\tilde{A}, \tilde{Q}), \tag{3.33}$$

   where $\tilde{A}$ and $\tilde{Q}$ are defined as in Theorem 3.2 and $\breve{L} = \text{diag}(\breve{L}^1, \ldots, \breve{L}^K)$ and $\bar{L}$ are computed as in Theorem 3.7.

2. *Let $n = \min_{k \in \mathcal{S}^c}(|\mathcal{N}^k|)$. Under (A.3.4) and (A.3.5),*

$$\hat{J}_\beta - J^*_\beta \in \mathcal{O}\left(\frac{1}{n}\right).$$

Proof: The proof of part 1 follows along the same lines of the proof of Theorem 3.2. In particular, under (A.3.8) and (A.3.9), $\sqrt{\beta}\tilde{A}$ of Proposition 3.1 is Hurwtiz; hence, the performance loss may be computed by the associated algebraic Lyapunov equation given by (3.33). Note that even though $\tilde{Q}$ is not positive semi-definite, the algebraic Lyapunov equation has a solution [Hassibi et al., 1999]. The proof of part 2 follows along the same lines

of the proof of Theorem 3.3. In particular, observe that (i) $(\tilde{X}_1, \tilde{W})$ in (3.32) are $\mathcal{O}(1/n)$ due to (A.3.4), (ii) $\tilde{M}$ given by (3.33) does not depend on n due to (A.3.5) ■

Remark 3.9 Assumption (A.3.9) is always satisfied if $D_t^k = 0$ and $E_t^k = 0$ for all $k \in \mathcal{K}$. In this case,

$$\sqrt{\beta}(\bar{A} + \bar{B}\breve{L}) = \text{diag}(\sqrt{\beta}(A^1 + B^1\breve{L}^1), \ldots, \sqrt{\beta}(A^K + B^K\breve{L}^K)),$$

where each of the diagonal terms are Hurwitz by definition of $\breve{L}^k$ given in Theorem 3.7.

3.7 Numerical example: Temperature control of space heaters

To illustrate our results, we consider an example that is motivated by demand response in power systems. In demand response, the volatility in renewable generation is compensated by making small changes in the demand of a large number of loads. We model the load dynamics according to a model proposed in [Kizilkale and Malhame, 2014], but consider a different per-step cost.

Consider a homogeneous population $\mathcal{N}$ of space heaters. For space heater $i, i \in \mathcal{N}$, the state x_t^i denotes the room temperature at time t. Consider a nominal temperature x_{nom} and let u_{nom} be the control input needed to maintain the room temperature at x_{nom}. Following [Kizilkale and Malhame, 2014], we linearize dynamics around x_{nom}, i.e.,

$$x_{t+1}^i - x_{nom} = a(x_t^i - x_{nom}) + bu_t^i + w_t^i,$$

where u_t^i is control input *in addition* to u_{nom} and w_t^i is a random disturbance. We assume u_{nom} is large enough such that $(u_t^i + u_{nom})$ is positive.

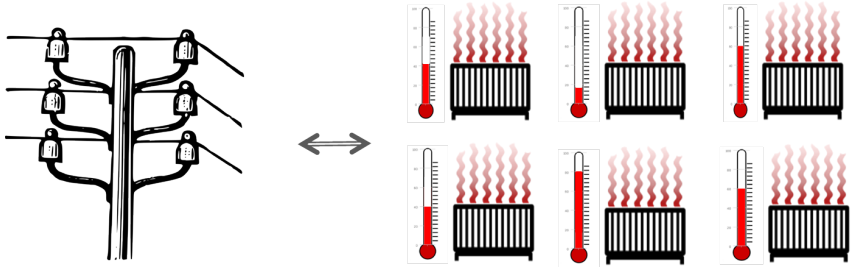

Fig. 3.1 Demand response of heaters.

Let x^i_{des} denote the desired temperature of user i. It is assumed that the mean desired temperature $\bar{x}_{des}$ is known to everyone (e.g., independent system operator (ISO) could compute it and broadcast the mean value to everyone or it could be computed in a distributed manner using a consensus algorithm). For the purpose of demand response, time is divided into epochs of length T. At the beginning of each epoch, a central authority such as an ISO generates a reference mean temperature $\bar{x}_{ref}$ and broadcasts it to all users.

During an epoch, all users collectively minimize the total expected cost $\mathbb{E}[\sum_{t=1}^{T} c_t]$, where

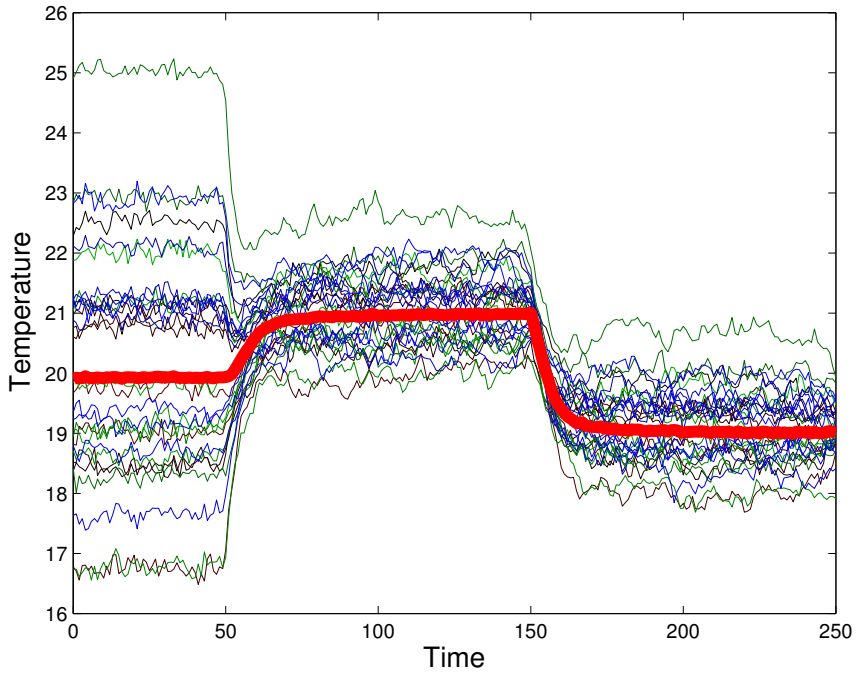

Fig. 3.2 Demand response with a population of 100 space heaters. In the initial phase, $1 \leq t \leq 50$, the system is uncontrolled. In the first epoch $50 < t \leq 150$, the system tracks a mean reference temperature of $\bar{x}_{ref} = 21$; in the second epoch $150 < t \leq 250$, the system tracks a mean reference temperature of $\bar{x}_{ref} = 19$. The thin lines show the local temperature of 30 out of the 100 space heaters. The thick red line shows the mean-temperature achieved by the optimal strategy.

the per-step cost c_t is given by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T-t}{T}\Big[q(x_t^i - x_{des}^i)^2 + r{u_t^i}^2\Big] + \frac{t}{T}p(\bar{x}_t - \bar{x}_{ref})^2.$$

The rationale for the per-step cost is that we penalize deviations from the desired temperature (which corresponds to the user's comfort level), the control effort, and deviation of the mean temperature from the reference prescribed by the ISO. The weights $(\frac{T-t}{T})$ and $\frac{t}{T}$ are so that we linearly move from preferring individual preferences to preferring global preferences.

The above problem is an optimal tracking problem and the optimal strategy is given by Theorem 3.5. As an example, we consider the following values of the parameters:

$$\begin{aligned} &n = 100, \quad a = 0.8, \quad b = 1, \quad q = 1, \quad p = 10, \quad r = 50, \\ &T = 100, \quad x_{nom} = 20, \quad w_t^i \sim \mathcal{N}(0, 0.01), \quad x_1^i \sim \mathcal{N}(20, 3), \end{aligned}$$

and consider two epochs of length $T = 100$. In the initial phase, $1 \leq t \leq 50$, space heaters are operating around their local set temperatures: In the first epoch, $50 < t \leq 150$, $\bar{x}_{ref} = 21$; in the second epoch, $150 < t \leq 250$, $\bar{x}_{ref} = 19$. The resultant trajectories of a subset of the users are shown in Fig. 3.2.

3.8 Discussion: Virtual macro and micro agents

In this section, a macro-micro framework (perspective) is adopted to provide some insight on the main results of this chapter. Let N be the number of agents with linear dynamics that wish to collaborate to minimize a common quadratic cost and K be the number of sub-populations. It is shown that $K+1$ Riccati equations must be solved (across sub-populations) to obtain the control laws of all agents; yet, only two Riccati equations must be solved for each agent: one corresponding to its own sub-population and one to the mean-field.

For each agent, consider two virtual agents: a *macro agent* and a *micro agent*. The agents of each sub-population have an identical macro agent; this means there are K macro agents in the entire population, i.e., one macro agent per each sub-population. The macro agents are in charge of macroscopic (mean-field) behaviour of the population (i.e., $\{\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t\}_{t=1}^T$). In contrast, each individual agent has a different micro agent; this means there are N micro agents. The micro agents are in charge of microscopic (individual) behaviour of agents (i.e.,

$\{x_t^i - \bar{x}_t^k, u_t^i - \bar{u}_t^k\}_{t=1}^T$). The key feature of micro agents is that their mean-field behaviour is zero. Due to this feature, the design of micro agents do not affect the design of macro agents. This implies that the design of control laws can be decoupled into macroscopic and microscopic. Note that this decoupled design holds for arbitrary size of population.

- **Macroscopic (mean-field) design:** The macro agents must collaborate to determine the mean-field behaviour of the population. The macro agents may be interpreted as leaders that control the mean-field of all sub-populations (i.e., $\{\bar{x}_{1:T}^k, \bar{u}_{1:T}^k\}_{k=1}^K$). The collaboration rule is obtained by solving one Riccati equation (3.10) across sub-populations to determine the control gains $\{\bar{L}_{1:T}^k\}_{k=1}^K$.

- **Microscopic (individual) design:** In each sub-population $k \in \{1, \dots, K\}$, the micro agents must collaborate to determine the individual behaviour of agents. For achieving that, the micro agents control the residual behaviour of agents from the mean-field of sub-population k (i.e., $\{x_t^i - \bar{x}_t^k, u_t^i - \bar{u}_t^k\}_{t=1}^T$). The collaboration rule is obtained by solving one Riccati equation (3.9) within sub-population k to determine control gains $\breve{L}_{1:T}^k$.

Agent i of sub-population k uses macroscopic and microscopic designs to decide its control action as follows:

$$u_t^i = \breve{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^k \bar{\mathbf{x}}_t.$$

Unlike the basic linear quadratic mean-field model described above, the tracking cost model in Section 3.5.2 and weighted mean-field model in Section 3.5.3 are not partially exchangeable according to exchangeability definition given in Chapter 2. However, our proof techniques work for these two cases as well.

In the case of different tracking cost in Section 3.5.2, additional terms must be determined. In macroscopic design, macro agents must compute $\bar{F}_{1:T}^k$ and correction signal $\bar{v}_{1:T}$. In microscopic design, micro agents must compute $\breve{F}_{1:T}^k$ and correction signal v_t^i. Then,

$$u_t^i = \breve{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^k \bar{\mathbf{x}}_t + \breve{F}_t^k v_t^i + \bar{F}_t^k \bar{v}_t.$$

In the case of weighted mean-field in Section 3.5.3, a weighted solution must be computed. In macroscopic design, macro agents must solve an appropriate version of Riccati equation (3.10) given in Theorem 3.6 to obtain $\{\bar{L}_{1:T}^k\}_{k=1}^K$. In microscopic design, micro agents

solve the same Riccati equation (3.9) to obtain $\breve{L}^k_{1:T}$. Agent i of sub-population k uses a weighted (localized) version of the designs as follows:

$$u^i_t = \breve{L}^k_t \left(x^i_t - \frac{\lambda^i}{b^i} \bar{x}^{k,\lambda}_t \right) + \frac{\lambda^i}{b^i} \bar{L}^k_t \bar{\mathbf{x}}^\lambda_t .$$

So far, it has been assumed that the mean-field of sub-populations is shared among agents. In the case of large sub-populations (for which collecting and sharing the mean-field may be challenging), one may decide not to share the mean-field of large sub-populations and use the approximately optimal results of PMFS-IS.

Remark 3.10 Note that the macroscopic design requires the full knowledge of the model; hence the macro agents must know the system model completely. However, the microscopic design requires only the model of sub-population k; hence, the micro agents do no need to know the complete model.

3.9 Conclusion

In this chapter, we presented the team optimal control of decentralized linear quadratic systems with two non-classical information structures. Our two main results are as follows. First, when the mean-field is observed by all agents (the MFS information structure), the linear control laws are optimal and the corresponding gains are computed by solving $K + 1$ Riccati equations, where K is the number of sub-populations. The dimensions of these Riccati equations are independent of the size of sub-populations; consequently, the solution complexity depends only on the number K of sub-populations (rather than the size of the entire population). Second, when the mean-field of a (possibly empty) subset of sub-populations is observed by all agents (the PMFS information structure), a linear control law based on certainty equivalence is approximately optimal. We generalized our results to major-minor, tracking cost, weighted mean-field, and infinite horizon. Finally, we illustrated our approach by a numerical example on demand response.

An important practical implication of these results is that they do not suffer from the curse of dimensionality. In fact, under assumption (A.3.5), the solution does not even depend on the number of agents and the optimal gains can be computed without being aware of the size of each sub-population. Consequently, the solution methodology generalizes to the setup

where the agents in a sub-population arrive and depart according to an exogenous process (e.g. number of electric vehicles plugged in for charging in smart grids).

In this chapter, the main attention is focused on the quality of the system performance, i.e., the optimal and approximately optimal performances. It is shown that to achieve the optimal centralized performance of any linear quadratic system with partially exchangeable agents, one needs to share only the mean-field. However, collecting and sharing the entire vector of the mean-field may not be feasible. Hence, it might be more practical to share a subset of the mean-field. Therefore, it is interesting to characterize the trade off between the performance loss and the communication cost. In particular,

Q: *If the agents of sub-population k do not observe the mean-field of sub-population k', how much performance will be lost?*

To answer this question, one may use the result of Theorem 3.1. In particular, since the optimal performance is computed in Theorem 3.1, one may guess (use) any arbitrary prediction for the unobserved components of the mean-field vector and compute the associated error from the optimal performance. For example, if the agents of sub-population k do not observe the mean-field of sub-population k', one may simply consider the unobserved mean-field to be zero and identify a conservative bound on the performance loss by computing the difference from the optimal performance. In the view of PMFS-IS results, for large sub-populations and independent noises, a tight bound on the performance loss is identified by predicting the deterministic evolution of the mean-field. It is shown that the performance loss goes to zero as the size of sub-populations tends to infinity (Theorems 3.2 and 3.3). However, in general, identifying a tight bound on the performance loss is still an open research question when sub-populations are not large or have correlated noises.

CHAPTER 4

Optimal control of Markov chain mean-field teams

4.1 Introduction

In this chapter, we consider controlled Markov chain systems with partially exchangeable agents. We call these systems controlled Markov chain mean-field teams. In particular, we investigate team optimal control of stochastic agents that are coupled in the dynamics and cost through the mean-field (empirical distribution) of states and actions. For the ease of exposition and to deliver the fundamental ideas, we present the main results for homogeneous population. In Section 4.5.2, we show how these results generalize to heterogeneous population. Two information structures are investigated: *Mean-Field Sharing Information Structure* (MFS-IS) and *Noisy Mean-Field Sharing Information Structure* (NMFS-IS). In MFS-IS, all agents observe their local state and the mean-field of all sub-populations; in NMFS-IS, all agents observe their local state but a noisy version of the mean-field. Both information structures are non-classical. We identify an information state and use that to obtain a dynamic programming decomposition. This dynamic program determines a globally optimal strategy for all agents. This solution approach works for arbitrary number of agents and generalizes to the setup when the mean-field is observed with noise. The size of the information state is time-invariant; thus, the results generalize to the infinite-horizon control setups as well. In addition, when the mean-field is observed without noise, the size of the corresponding information state increases polynomially (rather than exponentially) with the number of agents which allows us to solve problems with moderate number of agents. We generalize the main results to arbitrary coupled cost, heterogeneous population, and major-

minor. We illustrate our approach by an example motivated by smart grids that consists of 100 coupled agents.

Literature review

The scalability of the solution approach to large scale systems is an important consideration in team optimal control. Different approaches have been proposed to ensure that the solution complexity does not increase drastically with the number of agents. These include coordination-decomposition methods [Culioli and Cohen, 1990, Barty et al., 2010] that use iterative message passing algorithm. In [Culioli and Cohen, 1990], authors use decomposition/coordination technique where a system with N agents is decomposed to N individual independent sub-problems and one coordinator. At each iteration, the role of the coordinator is to penalize solution of each sub-problem appropriately such that after enough iteration, the resultant solution converges to a global optimum.

Another important approach is mean-field games where each agent is affected by the others through their mean-field. Such systems arise in various applications including economics [Lasry et al., 2008], robotic swarms [Shi et al., 2012], oil production [Guéant et al., 2011], wireless networks [Tembine, 2014], smart grids [Couillet et al., 2012, Kizilkale et al., 2012], emergent behaviour [Nourian et al., 2010], etc. In the mean-field games, the n-player game (n-body problem) is converted to a 2-player game between the generic player versus the mass (1-body problem). We refer the reader to [Caines, 2013, Huang et al., 2006, Guéant et al., 2011, Gomes and Saude, 2014, Nourian and Caines, 2013, Lasry and Lions, 2007] for a detailed overview of the mean-field games. Most of the analysis of mean-field games has focused on the competitive behaviour of large population systems. In this chapter, we investigate team optimal (cooperative) behaviour of such systems for arbitrary number of agents.

Contributions

The above problem is conceptually challenging because it has a non-classical information structure [Witsenhausen, 1971]. In general, team optimal control problems with non-classical information structure belong to NEXP[1] complexity class [Bernstein et al., 2002]. Although it is possible to get a dynamic programming decomposition for problems with non-classical

[1] If the horizon is larger than the number of agents, then it is NEXP-hard and if the horizon is limited to be less than the number of agents it is NEXP-complete [Bernstein et al., 2002].

information structure [Witsenhausen, 1973], the size of the corresponding information state increases with time. For some information structures, we can find information states that do not increase with time [Mahajan et al., 2012], but even for these models the size of the information state increases exponentially with the number of agents.

Our key contributions in this chapter are the following:

1. We identify a dynamic program to obtain globally optimum control strategies.
2. The size of the corresponding information state does not increase with time. Thus, our results extend naturally to infinite horizon setups.
3. The size of the corresponding information state increases polynomially with the number of agents. This allows us to solve problems with moderate number of s. (In Sections 4.7 and 4.8, we give examples with $n = 100$ agents).
4. The solution methodology and dynamic programming decomposition extend to the scenario where all agents observe a noisy version of the mean-field.
5. We generalize our results to arbitrary coupled per-step cost, heterogeneous population, major-minor, and randomized strategy.
6. We present key properties of Markov chains with exchangeable transition probability.

Notation

To distinguish between random variables and their realizations, we use upper-case letters to denote random variables (e.g. X) and lower-case letters to denote their realizations (e.g. x). For vectors x, y, and z, $\mathrm{vec}(x, y, z)$ denotes the vector $[x^\intercal, y^\intercal, z^\intercal]^\intercal$. We use the short hand notation $X_{a:b}$ for the vector $(X_a, X_{a+1}, \dots, X_b)$ and bold letters to denote vectors e.g. $\mathbf{Y} = (Y^1, \dots, Y^n)$ where n is the size of vector $\mathbf{Y}$. $\mathbb{1}(\cdot)$ is the indicator function of a set, $\mathbb{P}(\cdot)$ is the probability of an event, $\mathbb{E}[\cdot]$ is the expectation of a random variable, and $|\cdot|$ is the cardinality of a set. $\mathbb{N}$ refers to the set of natural numbers. For $x \in \mathcal{X}$, δ_x denotes a Dirac measure on $\mathcal{X}$ with a point mass at x.

4.2 Problem formulation

Consider a discrete time decentralized control system with $n \in \mathbb{N}$ homogeneous agents that operate for a horizon $T \in \mathbb{N}$. The state of agent i, $i \in \{1, \dots, n\}$, at time t, is denoted by $X_t^i \in \mathcal{X}$, where $\mathcal{X}$ is a finite set (that does not depend on i). Let $U_t^i \in \mathcal{U}$ denote the control action of agents i, $i \in \{1, \dots, n\}$, at time t, where $\mathcal{U}$ is a finite set (that does not depend on i). Denote the joint state $\mathbf{X}_t = \text{vec}(X_t^1, \dots, X_t^n)$ and the joint action $\mathbf{U}_t = \text{vec}(U_t^1, \dots, U_t^n)$. The mean-field of joint state and action of all agents is defined as the empirical distribution of the states and actions of all agents, i.e.,

$$\Xi_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_t^i, U_t^i},$$

or equivalently, for $x \in \mathcal{X}$ and $u \in \mathcal{U}$,

$$\Xi_t(x, u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_t^i = x, U_t^i = u).$$

In the sequel, we use ξ_t to denote the realization of Ξ_t at time t. We refer to the empirical distribution of states of all agents at time t as the *mean-field* of the system[2] and denote it by M_t, i.e.,

$$M_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_t^i}, \tag{4.1}$$

or equivalently, for $x \in \mathcal{X}$,

$$M_t(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_t^i = x).$$

The mean-field of states is the marginal empirical distribution of the empirical distribution of states and actions, i.e.,

$$M_t(x) = \sum_{u \in \mathcal{U}} \Xi(x, u), \quad x \in \mathcal{X}.$$

Let $|\mathcal{X}| = q$ and $\mathcal{M}_n = \{(\frac{a_1}{n}, \frac{a_2}{n}, \dots, \frac{a_q}{n}) : a_i \in \{0, \dots, n\}, \sum_{i=1}^{q} a_i = n\}$ denote the space of realizations of mean-field M_t. Note that $\mathcal{M}_n \subset \Delta(\mathcal{X})$, where $\Delta(\mathcal{X})$ denotes the space of probability distributions on $\mathcal{X}$.

[2] In the sequel, we refer to the mean-field of states simply as mean-field.

System model

The agents are coupled with each other in the dynamics and the cost via the mean-field of joint state and action, as described below. The initial states of all agents are distributed according to PMF (probability mass function) P^x (that does not depend on i). The state X_t^i of agent i evolves according to

$$X_{t+1}^i = f_t(X_t^i, U_t^i, W_t^i, \Xi_t), \quad i \in \{1, \dots, n\}, \tag{4.2}$$

where f_t is the system dynamics at time t and W_t^i is a random variable which takes value on a finite set $\mathcal{W}$. Note that the system dynamics $\{f_t\}_{t=1}^T$ do not depend on i. Let $\mathbf{W}_t = \text{vec}(W_t^i, \dots, W_t^n)$ with PMF P_t^w at time t. The primitive random variables $\{\mathbf{X}_1, \{\mathbf{W}_t\}_{t=1}^T\}$ are defined on a common probability space. At each time step, the system incurs a cost that depends on the mean-field of joint state and action that is given by

$$\ell_t(\Xi_t).$$

Remark 4.1 Note that it is shown in proposition 2.2 that any per-step cost $c_t(\mathbf{X}_t, \mathbf{U}_t)$ that is exchangeable with respect to the agents may be written as a function of Ξ_t. Herein, we present two examples:

$$c_t(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n}\sum_{i=1}^n c_t(X_t^i, U_t^i, \Xi_t) = \sum_{x\in\mathcal{X}, u\in\mathcal{U}} \Xi_t(x,u)\cdot c_t\left(x, u, \Xi_t(x,u)\right) =: \ell_t(\Xi_t),$$

and for interacting particles,

$$c_t(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n c_t(X_t^i, X_t^j, U_t^i, U_t^j) = \sum_{\substack{x\,\in\,\mathcal{X},\, u\,\in\,\mathcal{U}\\ x'\,\in\,\mathcal{X},\, u'\,\in\,\mathcal{U}}} \Xi_t(x,u)\cdot\Xi_t(x',u')\cdot c_t\left(x,u,x',u'\right) =: \ell_t(\Xi_t).$$

Information structure

We consider two information structures; in both, agents perfectly recall all data that they observe. In the first information structure, which we call *mean-field sharing* and denote by MFS-IS, every agent i perfectly observes its local state X_t^i and the global mean-field M_t.

Thus, the data I_t^i available to agent i at time t is given by

$$I_t^i = (X_t^i, U_{1:t-1}^i, M_{1:t}). \tag{MFS-IS}$$

In the second information structure, which we call *noisy mean-field sharing* and denote by NMFS-IS, every agent i perfectly observes its local state X_t^i and noisy version of the mean-field $Y_t \in \mathcal{Y}$, where $\mathcal{Y}$ is a finite set (that does not depend on i), i.e.,

$$Y_t = h_t(M_t, O_t), \tag{4.3}$$

where O_t is a random variable which takes value on a finite set $\mathcal{O}$ (that does not depend on i) with PMF P_t^o, at time t. Thus, the data I_t^i available to agent i at time t is given by

$$I_t^i = (X_t^i, U_{1:t-1}^i, Y_{1:t}). \tag{NMFS-IS}$$

Under both information structures, agent i chooses u_t^i as follows:

$$U_t^i = g_t^i(I_t^i). \tag{4.4}$$

The function g_t^i is called the *control law of agent i* at time t. The collection $\mathbf{g}^i = (g_1^i, g_2^i, \dots, g_T^i)$ is called the *control strategy of agent i*. The collection $\mathbf{g} = \{\mathbf{g}^i\}_{i=1}^n$ is called the *control strategy of the system*. The performance of strategy $\mathbf{g}$ is given by

$$J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}} \left[\sum_{t=1}^{T} \ell_t(\Xi_t) \right], \tag{4.5}$$

where the expectation is with respect to the measure induced on all the system variables by the choice of strategy $\mathbf{g}$.

Remark 4.2 The above model assumes that all agents have access to the mean-field of the system. In certain applications such as cellular communications and smart grids, a centralized authority (such as a base station in cellular communication and an independent service operator in smart grids) may measure the mean-field and transmit it to all agents. In other applications such as multi-robot teams, all agents may compute the mean-field in a distributed manner using methods such as consensus-based algorithms [Olfati-Saber et al.,

2006, Bishop and Doucet, 2014].

A. 4.1 *At any time t, the control laws at all agents are identical i.e. $g_t^i = g_t^j$ for any $i, j \in \{1, \dots, n\}$. Therefore, we drop the superscripts and denote the control law at every agent at time t as g_t.*

In general, this assumption leads to a loss in performance, as illustrated by an example below.

Example: Let $\mathcal{X} = \mathcal{U} = \{1, 2, \dots, n\}$, $T = 2$, and P^x be uniform on $\mathcal{X}$. Suppose that the system dynamics are given by

$$X_2^i = U_1^i, \quad i \in \{1, \dots, n\}.$$

Let $\ell_1(\mathbf{x}_1, \mathbf{u}_1) = 0$ and $\ell_2(\mathbf{x}_2, \mathbf{u}_2) = K \cdot \mathbb{1}(m_2 \neq \{\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\})$ where K is a positive number. The asymmetric strategy $\bar{\mathbf{g}} = (\bar{g}_1^1, \dots, \bar{g}_1^n)$, where $\bar{g}_1^i(x_1^i, m_1) = i$, has a cost $J(\bar{\mathbf{g}}) = 0$. Hence, $\bar{\mathbf{g}}$ is optimal. On the other hand, under any symmetric strategy, $\mathbb{P}(M_2 \neq \{\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\})$ is positive. Hence, a symmetric strategy is not globally optimal. By increasing K, we can make symmetric strategies perform arbitrary bad as compared to asymmetric strategies.

Although assuming identical control laws (Assumption 4.1) leads to loss in performance, it is a standard assumption in the literature on large scale systems for reasons of simplicity, fairness, and robustness. For example, similar assumption has been made in [Schoute, 1978], [Shi et al., 2012], [Wu and Antsaklis, 2010].

Remark 4.3 Note that in linear quadratic mean-field teams described in Chapter 3, A.4.1 is without loss of optimality because the optimal strategies are identical.

The optimization problem

We are interested in the following optimization problem.

Problem 4.1 *In the model described above, given the horizon T, the system dynamics $\{f_t\}_{t=1}^T$, the cost functions $\{\ell_t\}_{t=1}^T$, the PMF P^x on the initial states, the PMFs $\{P_t^w\}_{t=1}^T$ on the disturbance, and the PMFs $\{P_t^o\}_{t=1}^T$ on the observation noise, find a strategy $\mathbf{g}^*$ that minimizes (4.5), i.e.,*

$$J^* := J(\mathbf{g}^*) = \min_{\mathbf{g}} J(\mathbf{g}),$$

where the minimum is taken over all strategies of form (4.4).

4.3 Main results

In this section, we present the main results of this chapter.

Exact solution for MFS-IS

We impose the following standard assumption on the model:

A. 4.2 *The primitive random variables* $\{\mathbf{X}_1, \{\mathbf{W}_t\}_{t=1}^T\}$ *are mutually independent.*

Theorem 4.1 *Under (A.4.1), (A.4.2), and (MFS-IS), an optimal strategy for Problem 4.1 is identified by the following dynamic program. Define recursively value functions:*

$$V_{T+1}(m_{T+1}) := 0, \quad \forall m_{T+1} \in \mathcal{M}_n,$$

and for $t = T, \ldots, 1$, *and for* $m_t \in \mathcal{M}_n$,

$$V_t(m_t) := \min_{\gamma_t} \Big(\ell_t\big(\phi(m_t, \gamma_t)\big) + \mathbb{E}\big[V_{t+1}(M_{t+1}) \big| M_t = m_t, \Gamma_t = \gamma_t\big] \Big), \tag{4.6}$$

where the minimization is over all functions $\gamma_t : \mathcal{X} \to \mathcal{U}$ *and the function* ϕ *is defined as follows:*

$$\phi(m_t, \gamma_t)(x, u) = m_t(x) \mathbb{1}(u = \gamma_t(x)), \quad x \in \mathcal{X}, u \in \mathcal{U}.$$

Let $\psi_t^*(m_t)$ *denote any argmin of the right-hand side of* (4.6). *Define*

$$g_t^*(m, x) := \psi_t^*(m)(x).$$

Then, $\mathbf{g}^* = (g_1^*, \ldots, g_T^*)$ *is an optimal strategy.*

Remark 4.4 The space $\mathcal{M}_n$ of realizations m_t is a finite set and has the cardinality less than $(n+1)^{|\mathcal{X}|}$. Thus, the exploration space (i.e., the state space of dynamic program of Theorem 4.1) increases polynomially with the number of agents.

Remark 4.5 To solve the dynamic program of Theorem 4.1, one requires to compute the transition probability of the mean-field. In general, one may use Remark 4.8 to compute the transition probability. However, when the noises are independent across agents, this transition probability may be computed more efficiently by using the following property.

The probability of the summation of independent random variables is the convolution of the individual probabilities.

Remark 4.6 The dynamic program of Theorem 4.1 uses m_t as the information state. Since m_t is observed by each agent, each agent can independently solve the dynamic program; agreeing upon a deterministic rule to break ties while using arg min ensures that all agents compute the same optimal strategy.

Exact solution for NMFS-IS

We impose the following standard assumption on the model:

A. 4.3 *The observation noises $\{O_t\}_{t=1}^T$ are independent random variables and also mutually independent from $\{\mathbf{X}_1, \{\mathbf{W}_t\}_{t=1}^T\}$.*

Define the posterior probability $\Pi_t := \mathbb{P}^{\mathbf{g}_{1:t}}(M_t|Y_{1:t})$. Note that $\pi_t \in \Delta(\mathcal{M}_n)$, where $\Delta(\mathcal{M}_n)$ denotes the space of probability distributions on $\mathcal{M}_n$.

Theorem 4.2 *Under (A.4.1), (A.4.2), (A.4.3), and (NMFS-IS), an optimal strategy for Problem 4.1 is identified by following dynamic program. Define recursively value functions:*

$$V_{T+1}(\pi_{T+1}) = 0, \quad \forall \pi_{T+1} \in \Delta(\mathcal{M}_n),$$

and for $t = T, \dots, 1$, and for $\pi_t \in \Delta(\mathcal{M}_n)$,

$$V_t(\pi_t) = \min_{\gamma_t} \big(\tilde{\ell}_t(\pi_t, \gamma_t) + \mathbb{E}[V_{t+1}(\Pi_{t+1})|\Pi_t = \pi_t, \Gamma_t = \gamma_t] \big), \tag{4.7}$$

where the minimization is over all functions $\gamma_t : \mathcal{X} \to \mathcal{U}$ and

$$\tilde{\ell}_t(\pi_t, \gamma_t) = \sum_{m \in \mathcal{M}_n} \ell_t\big(\phi(m, \gamma_t)\big)\pi_t(m),$$

where ϕ is as defined in Theorem 4.1 and for any $m \in \mathcal{M}_n$,

$$\pi_{t+1}(m) = \frac{\mathbb{P}(Y_{t+1} = y_{t+1}|M_{t+1} = m)\mathbb{P}(M_{t+1} = m|\Pi_t = \pi_t, \Gamma_t = \gamma_t)}{\sum_{\tilde{m} \in \mathcal{M}_n} \mathbb{P}(Y_{t+1} = y_{t+1}|M_{t+1} = \tilde{m})\mathbb{P}(M_{t+1} = \tilde{m}|\Pi_t = \pi_t, \Gamma_t = \gamma_t)},$$

where $\mathbb{P}(M_{t+1} = m|\Pi_t = \pi_t, \Gamma_t = \gamma_t) = \sum_{\tilde{m} \in \mathcal{M}_n} \mathbb{P}(M_{t+1} = m|M_t = \tilde{m}, \Gamma_t = \gamma_t)\pi_t(\tilde{m})$. Let $\psi_t^*(\pi_t)$ denote any argmin of the right-hand side of (4.7). Define

$$g_t^*(\pi, x) := \psi_t^*(\pi)(x).$$

Then, $\mathbf{g}^*=(g_1^*, \ldots, g_T^*)$ *is an optimal strategy.*

Remark 4.7 When the primitive random variables are independent and identically distributed across agents, as the size of population n goes to infinity, the evolution of mean-field becomes deterministic due to the law of large numbers. Hence, m_t becomes predictable in large scale.

4.4 Proof of main results

Proof of the results for MFS-IS

In this section, we use the common information approach [Nayyar et al., 2013] to introduce an equivalent centralized problem (Problem 4.2) for Problem 4.1 under MFS-IS. Then, we find an optimal solution for the equivalent problem and translate the obtained solution back to the solution of Problem 4.1.

Following [Nayyar et al., 2013], split the information I_t^i available to agent i into two parts: the *common information* consisting of the history $M_{1:t}$ of the mean-field process that is observed by all agents; and the *local information* consisting of the current state X_t^i of agent i. Since the size of the local information does not increase with time, the model described above has a partial history sharing information structure [Nayyar et al., 2013]. For such systems, the structure of optimal control strategies and a dynamic programming decomposition was proposed in [Nayyar et al., 2013]. If we directly use these results on our model, the information state will be a posterior distribution on the global state $\mathbf{X}_t = (X_t^1, \ldots, X_t^n)$ of the system. As such the complexity of the solution increases doubly exponentially with the number of agents.

To circumvent this issue, we proceed as follows.

Step 1: We follow the common information approach proposed in [Nayyar et al., 2013] to convert the decentralized control problem into a centralized control problem from the point of view of a agent that observes the common information $M_{1:t}$.

Step 2: We exploit the symmetry of the problem (with respect to the agents) to show that the mean-field M_t is an information state for the centralized problem identified in Step 1. We then use this information state M_t to obtain a dynamic programming decomposition.

The details of each of these steps are presented below.

Step 1: An Equivalent Centralized System

Following [Nayyar et al., 2013], we construct a fictitious centralized *coordinated system* as follows. We refer to decision maker in the coordinated system as the *coordinator*. At time t, the coordinator observes the mean-field M_t and chooses a mapping $\Gamma_t : \mathcal{X} \to \mathcal{U}$ as follows

$$\Gamma_t = \psi_t(M_{1:t}). \tag{4.8}$$

The function ψ_t is called the *coordination rule* at time t. The collection $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)$ is called the *coordination strategy*. After the mapping Γ_t is chosen, it is communicated to all agents. Each agent in the coordinated system is a passive agent that uses its local state X_t^i and the mapping Γ_t to generate

$$U_t^i = \Gamma_t(X_t^i), \;\; i \in \{1, \dots, n\}.$$

Lemma 4.1 *There exists a function ϕ such that*

$$\Xi_t = \phi(M_t, \Gamma_t).$$

In particular, for $x \in \mathcal{X}$ and $u \in \mathcal{U}$,

$$\begin{aligned}\xi_t(x,u) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_t^i = x, u_t^i = u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_t^i = x, \gamma_t(x) = u) \\ &= \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_t^i = x)\right] \cdot \mathbb{1}\big(\gamma_t(x) = u\big) = m_t(x) \cdot \mathbb{1}\big(\gamma_t(x) = u\big).\end{aligned}$$

The dynamics of each agent and the cost function are the same as in the original problem. The performance of any coordination strategy is quantified by the total expected cost

$$\tilde{J}(\boldsymbol{\psi}) = \mathbb{E}^{\boldsymbol{\psi}}\big[\sum_{t=1}^{T} \ell_t(\phi(M_t, \Gamma_t))\big], \tag{4.9}$$

where the expectation is with respect to a joint measure induced on all system variables by the choice of $\boldsymbol{\psi}$. Consider the following optimization problem.

Problem 4.2 *Given the information structure in (4.8), the horizon T, the plant functions $\{f_t\}_{t=1}^T$, the cost functions $\{\ell_t\}_{t=1}^T$, the PMF P^x on the initial states, and the PMFs $\{P_t^w\}_{t=1}^T$ on the plant disturbance, identify a control strategy $\boldsymbol{\psi}^*$ to minimize the total cost $\hat{J}(\boldsymbol{\psi})$ given by (4.9).*

Lemma 4.2 ([Nayyar et al., 2013], Proposition 3) *Problem 4.1 and Problem 4.2 are equivalent.*

In particular, for any control strategy $\mathbf{g} = (g_1, \dots, g_T)$ in Problem 4.1, define a coordination strategy $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)$ in Problem 4.2 by

$$\psi_t(m_{1:t}) := g_t(m_{1:t}, \cdot), \quad \forall m_{1:t}.$$

Then, $J(\mathbf{g}) = \hat{J}(\boldsymbol{\psi})$. Similarly for any coordination strategy $\boldsymbol{\psi}$ in Problem 4.2, define a control strategy $\mathbf{g}$ in Problem 4.1 by

$$g_t(m_{1:t}, x_t) := \psi_t(m_{1:t})(x_t), \quad \forall m_{1:t}, \forall x_t.$$

Then, $J(\mathbf{g}) = \hat{J}(\boldsymbol{\psi})$.

Note that the fictitious coordinated system is described only for ease of exposition.

Step 2: Identifying an Information State and Dynamic Program

An important result in identifying an information state is the following:

Lemma 4.3 *Under (A.4.1), for any choice $\gamma_{1:t}$ of $\Gamma_{1:t}$, any realization $m_{1:t}$ of $M_{1:t}$, and any $m \in \mathcal{M}_n$,*

$$\mathbb{P}\big(M_{t+1} = m | M_{1:t} = m_{1:t}, \Gamma_{1:t} = \gamma_{1:t}\big) = \mathbb{P}\big(M_{t+1} = m | M_t = m_t, \Gamma_t = \gamma_t\big).$$

Proof: From (4.1) and (4.2), we have

$$M_{t+1} = \frac{1}{n} \sum_{i=1}^n \delta_{X_{t+1}^i} = \frac{1}{n} \sum_{i=1}^n \delta_{f_t(X_t^i, U_t^i, W_t^i, \Xi_t)}. \tag{4.10}$$

Note that the right hand side of (4.10) is insensitive to arbitrarily permuting agents. In particular, let S denote any arbitrary permutation of set $\{1, \dots, n\}$ and $S(i)$ denote the ith

term of vector S. Then, we have

$$M_{t+1} = \frac{1}{n}\sum_{i=1}^{n} \delta_{f_t(X_t^i, U_t^i, W_t^i, \Xi_t)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{f_t(X_t^{S(i)}, U_t^{S(i)}, W_t^{S(i)}, \Xi_t)}.$$

This implies that M_{t+1} must not depend on the order (index) of agents and it depends only on the statistical information, i.e., the empirical distribution of joint state, action, and noise. Denote

$$\zeta_t := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_t^i, U_t^i, W_t^i}.$$

Thus, M_{t+1} is a deterministic function of the empirical distribution ζ_t. For ease of exposition, let function D^1 be the deterministic function such that

$$M_{t+1} = D^1(\zeta_t). \tag{4.11}$$

In addition, we have

$$\zeta_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_t^i, U_t^i, W_t^i} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\Xi_t, W_t^i} =: D^2(\Xi_t, \mathbf{W}_t). \tag{4.12}$$

From (4.11) and (4.12), we have that

$$\begin{aligned}\mathbb{P}(M_{t+1}|\mathbf{X}_t = x_t, \mathbf{U}_t = u_t) =& \mathbb{P}\Big(D^1\big(D^2(\Xi_t, \mathbf{W}_t)\big)|\mathbf{X}_t = x_t, \mathbf{U}_t = u_t\Big) \\ \overset{(a)}{=}& \mathbb{P}(M_{t+1}|\Xi_t = \xi_t),\end{aligned} \tag{4.13}$$

where (a) follows A.4.2 that $\mathbf{W}_t$ is independent of $\{\mathbf{X}_t, \mathbf{U}_t\}$. Note that (4.13) holds for arbitrary information structure. In the case of MFS-IS, we have that

$$\begin{aligned}\mathbb{P}(M_{t+1} = m|M_{1:t} = m_{1:t}, \Gamma_{1:t} = \gamma_{1:t}) =& \mathbb{P}\Big(D^1\big(D^2(\Xi_t, \mathbf{W}_t)\big) = m\Big|M_{1:t} = m_{1:t}, \Gamma_{1:t} = \gamma_{1:t}\Big) \\ \overset{(b)}{=}& \mathbb{P}(M_{t+1} = m|M_t = m_t, \Gamma_t = \gamma_t),\end{aligned}$$

where (b) follows from Lemma 4.1. ■

Proof of Theorem 4.1

Based on the results in steps 1 and 2, we have that: M_t is an information state for Problem 4.2 because:
1) As shown in Lemma 4.1, the per-step cost can be written as a function of M_t and Γ_t.
2) As shown in Lemma 4.3, $\{M_t\}_{t=1}^T$ is a controlled Markov process with control action Γ_t.
Thus, the result follows from standard results in Markov decision theory [Bertsekas, 2012]. In particular, in Problem 4.2, there is no loss of optimality in restricting attention to Markovian strategy i.e. $\Gamma_t = \psi_t(M_t)$. Furthermore, an optimal strategy $\boldsymbol{\psi}^*$ is obtained by solving the following dynamic program. Define recursively value functions:

$$V_{T+1}(m_{T+1}) := 0, \quad \forall m_{T+1} \in \mathcal{M}_n,$$

and for $t = T, \dots, 1$, and for $m_t \in \mathcal{M}_n$,

$$V_t(m_t) := \min_{\gamma_t}(\ell_t(\phi(m_t, \gamma_t)) + \mathbb{E}[V_{t+1}(M_{t+1})|M_t = m_t, \Gamma_t = \gamma_t]),$$

where the minimization is over all functions $\gamma_t : \mathcal{X} \to \mathcal{U}$. Let $\psi_t^*(m_t)$ denote any argmin of the right-hand side of (4.6). Then, the coordination strategy $\boldsymbol{\psi}^* = (\psi_1^*, \dots, \psi_T^*)$ is optimal. Based on the equivalence in Lemma 4.2, the proof is complete.

Proof of the results for NMFS-IS

We follow the two-step approach of Section 4.4. In step 1, we construct a centralized coordinated system in which a coordinator observes $Y_{1:t}$ and chooses

$$\Gamma_t = \psi_t(Y_{1:t}). \tag{4.14}$$

The rest of the setup is same as before. Similar to Problem 4.2, we get

Problem 4.3 *Given the information structure in (4.14), the horizon T, the plant functions $\{f_t\}_{t=1}^T$, the cost functions $\{\ell_t\}_{t=1}^T$, the PMF P^x on the initial states, the PMFs $\{P_t^o\}_{t=1}^T$ on observation noise, and the PMFs $\{P_t^w\}_{t=1}^T$ on the plant disturbance, identify a control strategy $\boldsymbol{\psi}^*$ to minimize the total cost $\hat{J}(\boldsymbol{\psi})$ given by (4.9).*

As in Lemma 4.2, Problem 4.1 under NMFS-IS is equivalent to Problem 4.3. In particular, for any control strategy $\mathbf{g} = (g_1, \dots, g_T)$ in Problem 4.1, one can construct a coordination strategy $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)$ in Problem 4.3 that yields the same performance and vice versa.

In step 2, we show that $\Pi_t(m) := \mathbb{P}(M_t = m | Y_{1:t}, \Gamma_{1:t-1})$ is an information state for Problem 4.3. In particular:

Lemma 4.4 *There exists a function $\tilde{\ell}_t$ (that does not depend on strategy $\boldsymbol{\psi}$) such that*

$$\mathbb{E}[\ell_t(\Xi_t) | Y_{1:t}, \Gamma_{1:t}] =: \tilde{\ell}_t(\Pi_t, \Gamma_t).$$

Proof: Consider

$$\begin{aligned}
\mathbb{E}\big[\ell_t(\Xi_t) | Y_{1:t} = y_{1:t}, \Gamma_{1:t} = \gamma_{1:t}\big] &\overset{(a)}{=} \mathbb{E}\big[\ell_t\left(\phi(M_t, \Gamma_t)\right) | Y_{1:t} = y_{1:t}, \Gamma_{1:t} = \gamma_{1:t}\big] \\
&= \sum_{m \in \mathcal{M}_n} \ell_t\left(\phi(m_t, \gamma_t)\right) \mathbb{P}(M_t = m | Y_{1:t} = y_{1:t}, \Gamma_{1:t} = \gamma_{1:t}) \\
&= \sum_{m \in \mathcal{M}_n} \ell_t\left(\phi(m_t, \gamma_t)\right) \pi_t(m) \\
&=: \tilde{\ell}_t(\pi_t, \gamma_t),
\end{aligned}$$

where (a) follows from Lemma 4.1. Note that none of the above terms depend on strategy $\boldsymbol{\psi}$. ■

Lemma 4.5 *There exists a function ϕ_t (that does not depend on strategy $\boldsymbol{\psi}$) such that*

$$\Pi_{t+1} = \phi_t(\Pi_t, \Gamma_t, Y_{t+1}).$$

Proof: For notational convenience, denote $\mathbb{P}(A = a | B = b, C = c)$ by $\mathbb{P}(a|b, c)$. For $m_{t+1} \in \mathcal{M}_n$ and $y_{t+1} \in \mathcal{Y}$, we have

$$\begin{aligned}
\pi_{t+1}(m_{t+1}) &= \mathbb{P}(m_{t+1} | y_{1:t+1}, \gamma_{1:t}) \overset{(a)}{=} \mathbb{P}(m_{t+1} | y_{1:t+1}, \gamma_{1:t}, \pi_{1:t}) \\
&= \frac{\mathbb{P}(m_{t+1}, y_{t+1} | y_{1:t}, \gamma_{1:t}, \pi_{1:t})}{\sum_{\tilde{m}_{t+1} \in \mathcal{M}_n} \mathbb{P}(\tilde{m}_{t+1}, y_{t+1} | y_{1:t}, \gamma_{1:t}, \pi_{1:t})} \\
&= \frac{\mathbb{P}(y_{t+1} | m_{t+1}, y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \mathbb{P}(m_{t+1} | y_{1:t}, \gamma_{1:t}, \pi_{1:t})}{\sum_{\tilde{m}_{t+1} \in \mathcal{M}_n} \mathbb{P}(y_{t+1} | \tilde{m}_{t+1}, y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \mathbb{P}(\tilde{m}_{t+1} | y_{1:t}, \gamma_{1:t}, \pi_{1:t})}
\end{aligned} \tag{4.15}$$

where (a) follows from the fact that $\Pi_{1:t}$ is a function of $\{Y_{1:t}, \Gamma_{1:t}\}$. Consider the two terms of the denominator separately. The first term can be simplified as

$$\mathbb{P}(y_{t+1}|\tilde{m}_{t+1}, y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \overset{(b)}{=} \sum_{o_{t+1}} \mathbb{P}^o_{t+1}(o_{t+1}) \mathbb{1}(y_{t+1}{=}h_t(\tilde{m}_{t+1}, v_{t+1})) = \mathbb{P}(y_{t+1}|\tilde{m}_{t+1}), \tag{4.16}$$

where (b) follows from (4.3) and A.4.3. The second term can be simplified as

$$\begin{aligned} \mathbb{P}(\tilde{m}_{t+1}|y_{1:t}, \gamma_{1:t}, \pi_{1:t}) &= \sum_{\tilde{m}_t \in \mathcal{M}_n} \mathbb{P}(\tilde{m}_{t+1}, \tilde{m}_t|y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \\ &= \sum_{\tilde{m}_t} \mathbb{P}(\tilde{m}_{t+1}|\tilde{m}_t, y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \mathbb{P}(\tilde{m}_t|y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \\ &\overset{(c)}{=} \sum_{\tilde{m}_t} \mathbb{P}(\tilde{m}_{t+1}|\tilde{m}_t, \gamma_t) \pi_t(\tilde{m}_t) =: \mathbb{P}(\tilde{m}_{t+1}|\pi_t, \gamma_t), \end{aligned} \tag{4.17}$$

where (c) follows from Lemma 4.3 and definition of Π_t. From (4.15), (4.16), and (4.17), we have

$$\pi_{t+1}(m_{t+1}) = \frac{\mathbb{P}(y_{t+1}|m_{t+1}) \mathbb{P}(m_{t+1}|\pi_t, \gamma_t)}{\sum_{\tilde{m}_{t+1} \in \mathcal{M}_n} \mathbb{P}(y_{t+1}|\tilde{m}_{t+1}) \mathbb{P}(\tilde{m}_{t+1}|\pi_t, \gamma_t)} =: \phi_t(\pi_t, \gamma_t, y_{t+1})(m_{t+1}).$$

Furthermore, none of the above terms depend on strategy $\boldsymbol{\psi}$. ■

Proof of Theorem 4.2

Similar to Theorem 4.1, we have that: Π_t is an information state for Problem 4.3 because:
1) As shown in Lemma 4.4, the expected per-step cost can be written as a function of Π_t and Γ_t.
2) As shown in Lemma 4.5, $\{\Pi_t\}_{t=1}^T$ is a controlled Markov process with control action Γ_t.
Thus, the result follows from standard results in Markov decision theory [Bertsekas, 2012].

4.5 Generalizations

In this section, we show that our results generalize to variations of Problem 4.1. We only present the results for MFS-IS (i.e., the analogue of Theorem 4.1). The results for NMFS (i.e., the analogue of Theorem 4.2) may be derived in a similar manner.

4.5.1 Arbitrary coupled per-step cost

In Section 4.2, we considered the agents to be coupled in the cost only through the empirical distribution of joint state and action. In this section, we generalize the result of Theorem 4.1 to the case in which agents are arbitrary coupled in the cost, i.e. $\ell_t(X_t^1, \dots, X_t^n, U_t^1, \dots, U_t^n)$. We impose the following assumption on the model.

A. 4.4 *The initial states* $\mathbf{X}_1$ *are exchangeable random variables.*

Theorem 4.3 *Under (A.4.1), (A.4.2), (A.4.4), and (MFS-IS), an optimal strategy for Problem 4.1—when the per-step cost* ℓ_t *is arbitrary coupled—is identified by the following dynamic program. Define recursively value functions:*

$$V_{T+1}(m_{T+1}) := 0, \quad \forall m_{T+1} \in \mathcal{M}_n,$$

and for $t = T, \dots, 1$, *and for* $m_t \in \mathcal{M}_n$,

$$V_t(m_t) := \min_{\gamma_t} \Big(\hat{\ell}_t\big(m_t, \gamma_t\big) + \mathbb{E}\big[V_{t+1}(M_{t+1}) \big| M_t = m_t, \Gamma_t = \gamma_t\big] \Big),$$

where the minimization is over all functions $\gamma_t : \mathcal{X} \to \mathcal{U}$ *and the function* $\hat{\ell}_t$ *is*

$$\hat{\ell}_t(m_t, \gamma_t) := \sum_{\mathbf{x} \in \mathcal{X}^n} \ell_t\big(x^1, \dots, x^n, \gamma_t(x^1), \dots, \gamma_t(x^n)\big) \frac{1}{|H(m_t)|} \mathbb{1}(\mathbf{x} \in H(m_t)),$$

where $H(m) := \{\mathbf{x} \in \mathcal{X}^n \colon \frac{1}{n} \sum_{i=1}^n \delta_{x^i} = m\}$. *Let* $\psi_t^*(m_t)$ *denote any argmin of the right-hand side of (4.6). Define*

$$g_t^*(m, x) := \psi_t^*(m)(x).$$

Then, $\mathbf{g}^* = (g_1^*, \dots, g_T^*)$ *is an optimal strategy.*

Proof: We follow the two-step approach in Section 4.4. In particular, for the coordinated system we have

Lemma 4.6 *For any choice* $\gamma_{1:t}$ *of* $\Gamma_{1:t}$, *any realization* $m_{1:t}$ *of* $M_{1:t}$, *and any* $\mathbf{x} \in \mathcal{X}^n$,

$$\mathbb{P}(\mathbf{X}_t = \mathbf{x} | M_{1:t} = m_{1:t}, \Gamma_{1:t} = \gamma_{1:t}) = \mathbb{P}(\mathbf{X}_t = \mathbf{x} | M_t = m_t) = \frac{1}{|H(m_t)|} \mathbb{1}(\mathbf{x} \in H(m_t)),$$

where $H(m) := \{\mathbf{x} \in \mathcal{X}^n : \frac{1}{n}\sum_{i=1}^n \delta_{x^i} = m\}$. *The above probability does not depend on the control laws* $\psi_{1:t}$.

The proof is presented in Appendix B.1. Using this result, we can show that

Lemma 4.7 *The expected per-step cost may be written as a function of* M_t *and* Γ_t. *In particular,*

$$\begin{aligned}
&\mathbb{E}\big[\ell_t(X^1, \dots, X_t^n, U_t^1, \dots, U_t^n) | M_{1:t} = m_{1:t}, \Gamma_{1:t} = \gamma_{1:t}\big] \\
&\qquad = \sum_{\mathbf{x} \in \mathcal{X}^n} \ell_t\big(x^1, \dots, x^n, \gamma_t(x^1), \dots, \gamma_t(x^n)\big) \mathbb{P}(\mathbf{X}_t = \mathbf{x} | M_{1:t} = m_{1:t}, \Gamma_{1:t} = \gamma_{1:t}) \\
&\qquad = \sum_{\mathbf{x} \in \mathcal{X}^n} \ell_t\big(x^1, \dots, x^n, \gamma_t(x^1), \dots, \gamma_t(x^n)\big) \mathbb{P}(\mathbf{X}_t = \mathbf{x} | M_t = m_t) \\
&\qquad = \sum_{\mathbf{x} \in \mathcal{X}^n} \ell_t\big(x^1, \dots, x^n, \gamma_t(x^1), \dots, \gamma_t(x^n)\big) \frac{1}{|H(m_t)|} \mathbb{1}(\mathbf{x} \in H(m_t)) \\
&\qquad := \hat{\ell}_t(m_t, \gamma_t),
\end{aligned}$$

where $\hat{\ell}_t$ *that does not depend on the control laws* $\psi_{1:t}$.

Therefore, for the problem in the coordinated system we have that: M_t is an information state because:
1) As shown in Lemma 4.7, the per-step cost can be written as a function of M_t and Γ_t.
2) As shown in Lemma 4.3, $\{M_t\}_{t=1}^T$ is a controlled Markov process with control action Γ_t.
Thus, the result follows from standard results in Markov decision theory [Bertsekas, 2012].
■

Remark 4.8 One may compute the transition probability as follows,

$$\begin{aligned}
\mathbb{P}(M_{t+1} = m_{t+1} | M_t = m_t, \Gamma_t = \gamma_t) = \sum_{\mathbf{y}, \mathbf{x} \in \mathcal{X}^n} \mathbb{P}\Big(\mathbf{X}_{t+1} = \mathbf{y} | \mathbf{X}_t = \mathbf{x}, \mathbf{U}_t = \operatorname{vec}\big(\gamma_t(x^1), \dots, \gamma_t(x^n)\big)\Big) \\
\cdot \mathbb{1}\big(\mathbf{y} \in H(m_{t+1}), \mathbf{x} \in H(m_t)\big).
\end{aligned}$$

4.5.2 Heterogeneous population

So far, we assumed that agents are homogeneous. In this section, we generalize the result of Theorem 4.1 to heterogeneous agents by using a change of variable argument. Consider a

heterogeneous population with K disjoint sub-populations, where agents are homogeneous within each sub-population. We show this heterogeneous population model can be converted to homogeneous model by simply using a change of variable.

Let $\mathcal{K} = \{1, \dots, K\}$ denote the set of all sub-populations and K is the number of sub-populations. Let $\mathcal{N}^k$ denote the agents of sub-population $k \in \mathcal{K}$. Let $X_t^i \in \mathcal{X}^k$ denote the state of agents i and $u_t^i \in \mathcal{U}^k$ denote its control action. For all agents $i \in \{1, \dots, n\}$, define an augmented state $\tilde{X}_t^i := (k, X_t^i)$ that consists of the index of sub-population and the state of agent. Note that the index k is time-invariant and is always fixed for each agent. With this augmented state space, the above heterogeneous model is converted to the basic homogeneous model in Section 4.2.

In particular, define the mean-fields as follows.

$$\mathbf{\Xi}_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{k, X_t^i, U_t^i} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{\mathcal{N}^k} \delta_{k, X_t^i, U_t^i}, \qquad \mathbf{M}_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{k, X_t^i} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{\mathcal{N}^k} \delta_{k, X_t^i}.$$

The dynamics of agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$ is given by

$$X_{t+1}^i = f_t(k, X_t^i, U_t^i, W_t^i, \mathbf{\Xi}_t) =: f_t^k(X_t^i, U_t^i, W_t^i, \mathbf{\Xi}_t),$$

where the per-step cost is given by $\ell_t(\mathbf{\Xi}_t)$. The agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$ chooses action u_t^i as follows.

$$U_t^i = g_t^i(k, X_t^i, U_{1:t-1}^i, \mathbf{M}_{1:t}), \tag{MFS-IS}$$

Under identical control laws, i.e. A.4.1, we have that

$$U_t^i = g_t(k, X_t^i, U_{1:t-1}^i, \mathbf{M}_{1:t}) =: g_t^k(X_t^i, U_{1:t-1}^i, \mathbf{M}_{1:t}).$$

The above heterogeneous model is converted to homogeneous model by using augments state $\tilde{X}_t^i$. Hence, the heterogeneous version of Theorem 4.1 is as follows.

Theorem 4.4 *Under the assumption of identical control laws for each sub-population (A.4.1) (i.e., $g_t^i = g_t^j = g_t^k$, $k \in \mathcal{K}$), (A.4.2), and (MFS-IS), we have the following result. An optimal*

strategy is identified by the following dynamic program. Define recursively value functions:

$$V_{T+1}(\mathbf{m}_{T+1}) := 0, \quad \forall \mathbf{m}_{T+1} \in \mathcal{M}_n,$$

and for $t = T, \dots, 1$, *and for* $\mathbf{m}_t \in \mathcal{M}_n$,

$$V_t(\mathbf{m}_t) := \min_{\boldsymbol{\gamma}_t} \Big(\ell_t\big(\phi(\mathbf{m}_t, \boldsymbol{\gamma}_t)\big) + \mathbb{E}\big[V_{t+1}(\mathbf{M}_{t+1}) \big| \mathbf{M}_t = \mathbf{m}_t, \boldsymbol{\Gamma}_t = \boldsymbol{\gamma}_t\big] \Big), \tag{4.18}$$

where the minimization is over all functions $\boldsymbol{\gamma}_t = \mathrm{vec}(\gamma_t^1, \dots, \gamma_t^K)$ *and* $\gamma_t^k : \mathcal{X}^k \to \mathcal{U}^k$ *and the function* ϕ *is defined as follows:*

$$\phi(\mathbf{m}, \boldsymbol{\gamma})(\mathbf{x}, \mathbf{u}) = \mathbf{m}(\mathbf{x}) \prod_{k=1}^{K} \mathbb{1}\big(u^k = \gamma^k(x^k)\big), \quad \mathbf{x} \in \prod_{k=1}^{K} \mathcal{X}^k, \mathbf{u} \in \prod_{k=1}^{K} \mathcal{U}^k, x^k \in \mathcal{X}^k, u^k \in \mathcal{U}^k.$$

Let $\psi_t^*(\mathbf{m}_t)$ *denote any argmin of the right-hand side of* (4.18). *Define*

$$[g_t^{*,1}(\mathbf{m}, x^1), \dots, g_t^{*,K}(\mathbf{m}, x^k)] := \psi_t^*(\mathbf{m})(\mathbf{x}), \quad x^k \in \mathcal{X}^k.$$

Then, $\mathbf{g}^* = \{(g_1^{*,k}, \dots, g_T^{*,k})\}_{k \in \mathcal{K}}$ *is an optimal strategy.*

4.5.3 Major agent and a population of minor agents

Consider the homogeneous system in Section 4.2 where, in addition to n homogeneous agents (called minor agents), there is one major agent that directly affects the evolution of all minor agents and the cost. In mean-field games, a similar model was introduced by [Huang, 2010] and further analyzed in [Nourian and Caines, 2013].

This model is a special case of heterogeneous model described in Section 4.5.2. In particular, suppose there exist two sub-populations, i.e., $\mathcal{K} = \{1, 2\}$: major and minor. The sub-population 1 denotes the major agent and it has only 1 agent, i.e., $|\mathcal{N}^1| = 1$. Then, the mean-field of sub-population 1 is the local state of sub-population 1, i.e., X_t^1. Hence, mean-field $\mathbf{M}_t$ is a function of X_t^1 and M_t. The rest of the dynamics and cost are the same as in Section 4.2. Since the dynamics are coupled through the mean-field, the state of the agent of sub-population 1 directly influences the dynamics of all other agents and the per-step cost. For this reason, such an agent is called a *major* agent. To identify the optimal strategy for the major and minor agents, one may use Theorem 4.4. Therefore, the optimal action of

major agent $1 \in \mathcal{N}^1$ of sub-population $\{1\}$ at time t, i.e. $U_t^{1^*}$, depends on the mean-field of minor agents M_t and the state of major agent X_t^1 i.e.

$$U_t^{1^*} = g_t^1(M_t, X_t^1).$$

Moreover, the optimal action of minor agent $i \in \mathcal{N}^2$ of sub-population $\{2\}$ at time t, i.e. $U_t^{i^*}$, depends on its own local state X_t^i in addition to M_t and X_t^1 i.e.

$$U_t^{i^*} = g_t^2(M_t, X_t^1, X_t^i).$$

Remark 4.9 Note that the dynamic program of Theorem 4.4 may be simplified further in this special case. In particular, let $\Gamma_t^1 : \mathcal{X}^1 \to \mathcal{U}^1$ and $\Gamma_t^2 : \mathcal{X}^2 \to \mathcal{U}^2$. Since X_t^1 is part of the local information at the major agent as well as the common information, i.e. mean-field, the function $\Gamma_t^1 : \mathcal{X}^1 \to \mathcal{U}^1$ may be replaced by the action U_t^1 (further discussion can be found in [Arabneydi and Mahajan, 2015]).

4.5.4 Randomized strategies

In general, randomized strategies are not considered in team problems because randomization does not improve performance [Gihman and Skorohod, 1979, Theorem 1.6]. However, if attention is restricted to identical strategies, randomized strategies may perform better than pure strategies [Schoute, 1978, Theorem 2.3]. In this chapter, we assume that the control strategies are pure, primarily for the ease of exposition. In particular, our results extend naturally to randomized strategies by considering $\Delta(\mathcal{U})$, the space of probability distributions on $\mathcal{U}$, as the action space. In the model described above, we assume that the strategies are pure (non-randomized).

4.5.5 Infinite horizon

The results of Lemma 4.1 and Lemma 4.3 are valid for the infinite horizon setup as well. Hence, the result of Theorem 4.1 generalizes to infinite horizon setup and under standard assumptions, the optimal coordination strategy is time-homogeneous and is given by the solution of a fixed point equation.

4.6 Exchangeable Markov processes

In this section, we study an interesting special case in which there is no control (uncontrolled version). This means the controlled Markov chain reduces to Markov process. Consider n Markov processes that take value in $\mathcal{X}^n$. Let $\{\mathbf{X}_t\}_{t=1}^T$ denote the joint Markov process. Assume the transition probability is exchangeable, i.e., for any permutation σ,

$$\mathbb{P}(\mathbf{X}_{t+1} = \sigma\mathbf{y} | \mathbf{X}_t = \sigma\mathbf{x}), \quad x, y \in \mathcal{X}^n,$$

where $\sigma\mathbf{x}$ denotes the permutation of vector $(x^1, \ldots, x^n)$.

Theorem 4.5 *For Markov process $\{\mathbf{X}_t\}_{t=1}^T$, we have the following results.*

I. The empirical distribution $\{M_t\}_{t=1}^T$ is a Markov process, i.e.,

$$\mathbb{P}(M_{t+1} = m_{t+1} | M_{1:t} = m_{1:t}) = \mathbb{P}(M_{t+1} = m_{t+1} | M_t = m_t).$$

II. The empirical distribution M_t is the sufficient statistic to predict the empirical distribution M_{t+1}, i.e.,

$$\mathbb{P}(M_{t+1} = m_{t+1} | \mathbf{X}_{1:t} = \mathbf{x}_{1:t}) = \mathbb{P}(M_{t+1} = m_{t+1} | M_t = m_t).$$

III. If the initial states $\mathbf{X}_1$ are exchangeable, then $\{\mathbf{X}_t\}_{t=1}^T$ is called an exchangeable Markov process. For any exchangeable Markov process $\{\mathbf{X}_t\}_{t=1}^T$,

$$\mathbb{P}(\mathbf{X}_t = \mathbf{x}_t | M_{1:t} = m_{1:t}) = \mathbb{P}(\mathbf{X}_t = \mathbf{x}_t | M_t = m_t) = \frac{1}{|H(m_t)|} \mathbb{1}(\mathbf{x}_t \in H(m_t)),$$

where $H(m) := \{\mathbf{x} \in \mathcal{X}^n : \frac{1}{n} \sum_{i=1}^n \delta_{x^i} = m\}$.

Proof: Part 1 follows from Lemma 4.3, part 2 from (4.13), and part 3 from Lemma 4.6. ■

Remark 4.10 Note that results (I) and (II) do not require the initial state $\mathbf{X}_1$ to be exchangeable. In addition, all the above results hold for arbitrary number n. The property (III) implies that $\mathbf{X}_t$ is *conditionally* exchangeable.

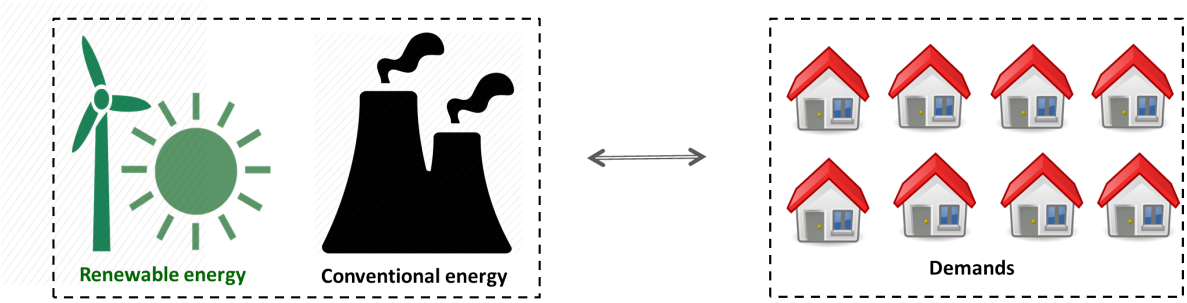

Fig. 4.1 The emergence of renewable energies causes volatility in the power grid. This volatility may be regulated by making small (local) changes at individual demands of a large group. This approach is called demand response.

4.7 Numerical example 1: Demand response

In this section, we consider an example of the basic model of Section 4.2 that is motivated by applications in smart grids. Consider a system with n-devices where $\mathcal{X} = \{1, \dots, k\}$ denotes the state space of each device and $\mathcal{U} = \{0, 1, \dots, k\}$ denotes the set of $k+1$ actions available at each device. Let $P(u)$ be the controlled transition matrix under action $u \in \mathcal{U}$, i.e.

$$[P(u)]_{xy} = \mathbb{P}(X^i_{t+1} = y \mid X^i_t = x, U^i_t = u), \quad x, y \in \mathcal{X}.$$

Action $u = 0$ is a *free action* under which each device evolves in an uncontrolled manner, i.e. $P(0) = Q$, where Q represents the *natural* dynamics of the system. Action $u \neq 0$ is a *forcing action* under which a fraction $1 - \epsilon_u$, $\epsilon_u \in [0, 1]$, of devices switch to state u, and remaining ϵ_u devices follow the natural dynamics. Thus,

$$P(u) = (1 - \epsilon_u)\mathbf{K}_u + \epsilon_u Q$$

where $\mathbf{K}_u$ is a $k \times k$ matrix where column u is all ones, and other columns are all zeros. Action $u = 0$ is free and it does not incur any cost, while action $u \neq 0$ incurs a cost $c(u)$. For notational convenience, let $c(0) = 0$.

The objective is to keep the mean-field (i.e. the empirical distribution) of the state of the devices close to a reference distribution $\zeta \in \Delta(\mathcal{X})$. The loss function is given by

$$\ell_t(\boldsymbol{\Xi}_t) = \frac{1}{n} \sum_{i=1}^{n} c(U^i_t) + D(M_t \parallel \zeta),$$

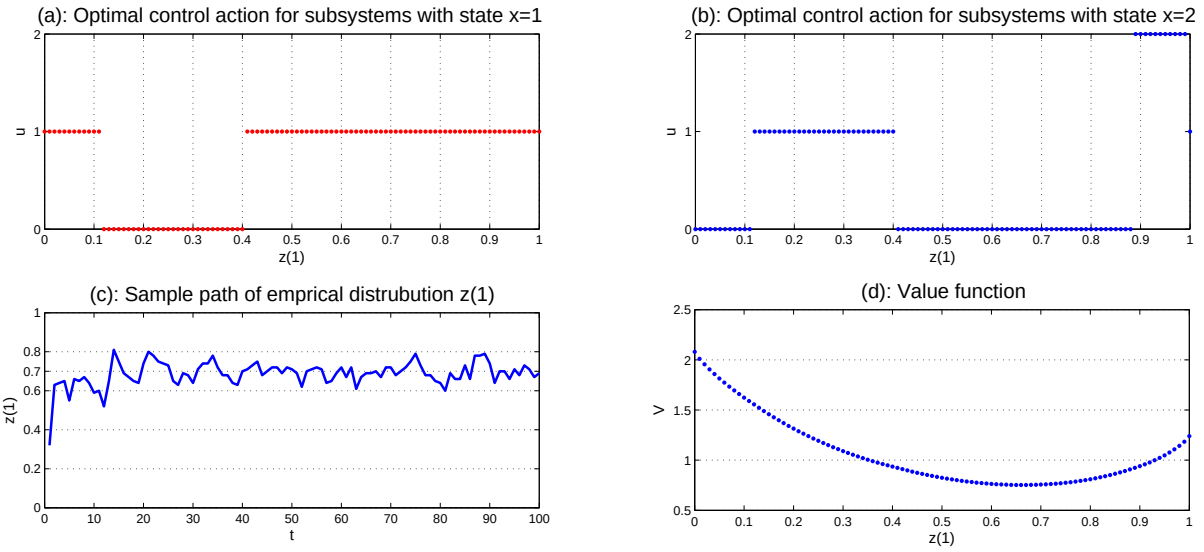

Fig. 4.2 Plots (a) and (b) show the optimal strategy as a function of $m(1)$. Plot (c) shows the sample path of $m(1)$ for simulation time of 100. Plot (d) depicts the value function with respect to $m(1)$.

where $D(p \parallel q)$ denotes the Kullback-Leibler divergence between $p, q \in \Delta(\mathcal{X})$ i.e.

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

The information structure is MFS-IS. The objective is to choose a control strategy to minimize the infinite horizon discounted cost[3]

$$J(\mathbf{g}) = \mathbb{E}\left[\sum_{t=1}^{\infty} \beta^t \left(\frac{1}{n} \sum_{i=1}^{n} c(U_t^i) + D(M_t \parallel \zeta) \right) \right],$$

where $\beta \in (0, 1)$ is the discount factor. A more elaborate variation of the above model is considered in [Meyn et al., 2014] for controlling the operation of pool pumps. Consider the

[3] Although we have only presented the details for finite horizon setup in this chapter, the results generalize naturally to infinite horizon setup under standard assumptions. See Section 4.5.5 for a brief explanation.

above model for the following parameters

$$n = 100, \quad k = 2, \quad \epsilon_1 = 0.2, \quad \epsilon_2 = 0.2, \quad c(0) = 0, \quad c(1) = 0.1, \quad c(2) = 0.2, \quad \beta = 0.9,$$
$$\zeta = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}, \quad Q = \begin{bmatrix} 0.25 & 0.75 \\ 0.375 & 0.625 \end{bmatrix}, \quad P_X = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}.$$

The optimal time-homogeneous strategy for these parameters is shown in Fig. 4.2. Since state space is binary, $m(1)$ is sufficient to characterise the empirical distribution $m = [m(1), m(2)]$. Hence, for ease of presentation, we plot the optimal control law and value function as a function of the first component $m(1)$ of $m = [m(1), m(2)]$. The reference distribution ζ is allowed to be time variant. For the above parameters, the sample path of $m(1)$ is shown in Fig. 4.3 when the reference distribution changes in time.

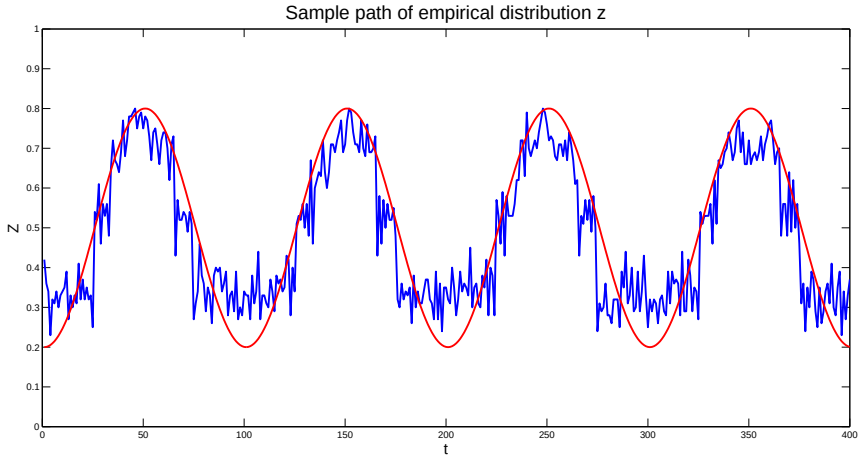

Fig. 4.3 The reference distribution of $m(1)$ is $sin(\frac{2\pi}{100}t + \frac{3\pi}{2})$ and is displayed in red. The sample path of $m(1)$ for 100 demands is displayed in blue.

4.8 Numerical example 2: Service design

We present an example of major and minor model of Section 4.5.3, where a service provider (major agent) interacts with its users (minor agents). Roughly speaking, we answer the question of how a service provider should design strategies for itself and the users such that the service is not only profitable but also customer-satisfactory.

To distinguish from minor agents, we use superscript 0 to denote the major agent. Consider a service provider (internet provider, utility provider, etc.) who serves $n \in \mathbb{N}$ users (customers). Let $X_t^i \in \mathcal{X} = \{0, 1\}$, $i \in \{1, \dots, n\}$, denote whether user i is active ($X_t^i = 1$) or passive ($X_t^i = 0$) at time t and $X_t^0 \in \mathcal{X}^0$ denote the available capacity of the system. At time t, the service provider chooses U_t^0, the capacity of the system at time $t + 1$. Thus, the dynamics of the state of the service provider is

$$X_{t+1}^0 = U_t^0,$$

where $\mathcal{X}^0 = \mathcal{U}^0$ is a finite set consisting of all feasible capacities.

The dynamics of users are represented by $P(u)$ where $P(u)$ denotes the controlled transition matrix under action $u \in \mathcal{U} = \{0, 1, 2\}$, i.e.

$$[P(u)]_{xy} = \mathbb{P}(X_{t+1}^i = y \mid X_t^i = x, U_t^i = u), \quad x, y \in \mathcal{X}.$$

The service provider can affect the evolution of user i by choosing action $u^i \in \{0, 1, 2\}$. Action $u^i = 0$ is a *free action* under which user i evolves in an uncontrolled manner, i.e. $P(0) = Q$, where Q represents the *natural* dynamics of the users. Action $u^i \neq 0$ is a *forcing action* under which the service provider forces user i to switch to state $(u^i - 1)$. However, there is $\epsilon_{u^i} \in [0, 1]$ probability that user i does *not* switch to state $(u^i - 1)$ and evolves according to its natural dynamics. Thus,

$$P(u) = (1 - \epsilon_u)\mathbf{K}_u + \epsilon_u Q,$$

where $\mathbf{K}_u$ is a $k \times k$ matrix where column $u \in \{1, 2\}$ is all ones, and other columns are all zeros. The goal is that the service provider efficiently manages the users and the capacity of the system in order to minimize costs while satisfying the users. Thus, the loss function is given by

$$\ell_t^0(M_t, X_t^0, U_t^0) + \sum_{i=1}^{n} \ell_t(U_t^i),$$

where ℓ_t^0 is as follows:

$$\ell_t^0(M_t, X_t^0, U_t^0) = S(U_t^0) + a|U_t^0 - X_t^0| - G(M_t, X_t^0),\ a \in \mathbb{R}^+,$$

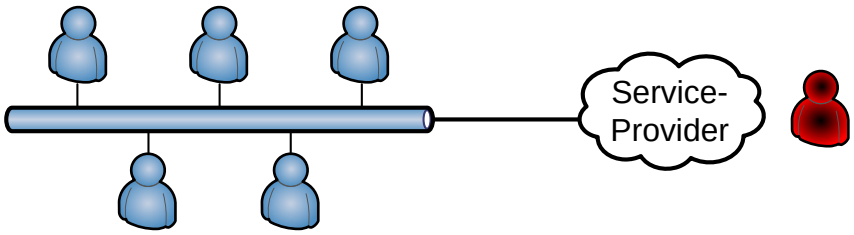

Fig. 4.4 The service provider must design strategies for itself and the users such that the service is not only profitable but also customer-satisfactory.

and $\ell_t(U_t^i) = H(U_t^i)$. In ℓ_t^0, the first term is associated with the cost of capacity and the second term refers to patching and dispatching capacity. The third term corresponds to the benefit (proportional to the number of active users) and the penalty of unavailable service (proportional to the number of active users that do not receive service) i.e.

$$G(M_t, X_t^0) = \begin{cases} bnM_t(1) & nM_t(1) \leq X_t^0 \\ bX_t^0 - c(nM_t(1) - X_t^0) & nM_t(1) > X_t^0, \end{cases}$$

where $b \in \mathbb{R}^+$ indicates the rate of benefit and $c \in \mathbb{R}^+$ determines the penalty rate of unavailable service. Note that $M_t(1) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_t^i = 1)$ is the average number of active users at time t. In addition, ℓ_t is the cost associated with forcing users. Given MFS-IS, the objective is to choose a control strategy that minimizes the infinite horizon discounted cost

$$J(\mathbf{g}) = \mathbb{E}\left[\sum_{t=1}^{\infty}\beta^t\left(\ell_t^0(M_t, X_t^0, U_t^0) + \sum_{i=1}^{n}\ell_t(U_t^i)\right)\right],$$

where $\beta \in (0,1)$ is a discount factor[4]. The optimal time-homogeneous strategies for

$$\begin{aligned}
&n = 100, \quad \mathcal{X}^0 = \{50, 100\}, \quad S(50) = 100, \quad S(100) = 300, \\
&H(0) = 0, \quad H(1) = 4, \quad H(2) = 1, \quad \epsilon_1 = 0.1, \quad \epsilon_2 = 0.1, \\
&a = 2, \quad b = 5, \quad c = 50, \quad \beta = 0.6, \qquad Q = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}
\end{aligned}$$

are presented below. Since the state space of minor agents is binary, $m(1)$ is sufficient to characterise the empirical distribution $m = [m(0), m(1)]$. Hence, for ease of presentation, we represent the optimal control law as a function of the second component $m(1)$ of $m = [m(0), m(1)]$.

$$g^*(m, x^0, x) = \begin{cases} 0 & 0 \le m(1) \le 0.53, \quad x^0 = 50, x = 0 \\ 1 & 0.53 < m(1) \le 0.76, \quad x^0 = 50, x = 0 \\ 2 & 0.76 < m(1) \le 1, \quad x^0 = 50, x = 0 \\ 0 & 0 \le m(1) \le 1, \quad x^0 = 50, x = 1 \\ 0 & 0 \le m(1) \le 0.29, \quad x^0 = 100, x = 0 \\ 2 & 0.29 < m(1) \le 1, \quad x^0 = 100, x = 0 \\ 0 & 0 \le m(1) \le 1, \quad x^0 = 100, x = 1, \end{cases}$$

and

$$g^{0*}(m, x^0) = \begin{cases} 50 & 0 \le m(1) \le 0.76, \quad x^0 = 50 \\ 100 & 0.76 < m(1) \le 1, \quad x^0 = 50 \\ 50 & 0 \le m(1) \le 0.29, \quad x^0 = 100 \\ 100 & 0.29 < m(1) \le 1, \quad x^0 = 100. \end{cases}$$

4.9 Conclusion

In this chapter, we considered the team optimal control of decentralized systems with two non-classical information structures. We followed a two-step approach: in the first step we constructed an equivalent centralized system using the common information approach

[4]Although we have only presented the details for finite horizon setup in this paper, the results generalize naturally to infinite horizon setup under standard assumption.

of [Nayyar et al., 2013]; in the second step, we exploited the symmetry of the system to identify an information state and dynamic programming decomposition of the problem. We generalized our result to the case of noisy observation of the mean-field, arbitrary coupled cost, heterogeneous population, and major-minor models. We demonstrated that our results extend naturally to randomized strategies and infinite horizon. We introduced exchangeable Markov processes (uncontrolled version) as a self interesting field of research. Finally, we illustrated our approach by two numerical examples.

In this chapter, the main attention is focused on the quality of the system performance, i.e., the optimal performances. A dynamic program is identified to find an exact optimal solution for any arbitrary finite population. It is shown the computational complexity of solving this dynamic program is polynomial with respect to the size of population (rather than exponential). This allows us to solve problems with moderate number of agents. However, in practice, finding the exact optimal solution for large population seems to be computationally very expensive. In addition, collecting and sharing the mean-field may not be physically or economically feasible in large population. For these reasons, one interesting future work is to study the approximation solutions for large population model. Thus, the following interesting question arises:

Q: *What type of approximations can be used and what will be the corresponding error?*

In general, it is difficult to answer this question completely; however, to provide some insightful answers, consider infinite population $n = \infty$ with i.i.d. noises. Due to the law of large numbers, the evolution of the mean-field becomes deterministic. One may approximate the (finite-model) transition probability by the (infinite-model) deterministic transition probability in the obtained (finite-model) dynamic program. In addition, one may approximate the mean-field by the (infinite-model) deterministic prediction (i.e., no need to share the mean-field among agents). A variation of these approximations has been widely studied in mean-field games. In particular, in mean-field games, the solution of (infinite-model) dynamic program is computed as the approximate solution for the finite model. However, in general, identifying an approximate solution with a tight approximation error bound is still an open research question for large population with correlated noises.

CHAPTER 5

Finite-state approximate solution of Partially Observable Markov Decision Process

5.1 Introduction

In this chapter, a novel approach to compute an approximate solution of Partially Observable Markov Decision Process (POMDP) is presented. First, the POMDP is converted to a countable-state Markov Decision Process (MDP), say Δ. Second, the countable-state MDP Δ is approximated by a finite-state MDP, say Δ_N, where N is the approximation index. It is shown that the error of this approximation is bounded and converges to zero as N becomes large. A key feature of this approach is the ability to use MDP solvers to find an approximate solution of POMDPs when the model is known, partly known, or not known.

Literature review

The sequential decision making emerges in many applications where an entity (agent) wishes to sequentially select the best alternative from a set of alternatives. To model the sequential decision making, Markov Decision Processes (MDPs) are widely used in the literature. The reader is referred to [Bertsekas, 2012] for standard solution approaches of MDPs (when the model in known) such as value iteration and policy iteration and to [Gosavi, 2009] for standard solution approaches of MDPs (when the model is unknown) such as TD(λ) and Q-learning. For the numerical examples, we use Q-learning algorithm. For reader's convenience, the details and the main proofs of Q-learning algorithm are presented in Appendix D.

In MDPs, it is assumed that the agent observes the state of the system *perfectly*; however

in practice, such assumption may not be realistic due to practical limitations such as measurement noises, sensor faults, noisy communication channels, or other practical issues. For that reason, Partial Observable Markov Decision Processes (POMDPs) are utilized to model the sequential decision making under uncertainty, where the agent observes the state of the system *imperfectly*. To mention a few applications using POMDPs: business [Lévesque and Maillart, 2008], machine vision [Darrell and Pentland, 1996], economy [Ray et al., 2009], medicine [Hauskrecht and Fraser, 2000], robotics [Atrash et al., 2009], social behavior [Broz et al., 2011], smart grids [Bu et al., 2011], etc.

Although POMDPs are more practical than MDPs, their solution complexity is PSpace-complete as opposed to the solution complexity of MDPs that is P-complete [Papadimitriou and Tsitsiklis, 1987]. Smallwood and Sondik [Smallwood and Sondik, 1973] introduced an algorithm to obtain an exact optimal solution of finite-horizon POMDPs in terms of $\alpha-$vectors. Later, researchers enhanced this algorithm by pruning the dominated $\alpha-$ vectors iteratively; these algorithms are known as witness algorithms [Littman, 1994b, Cassandra et al., 1997]. In general, the computational complexity of all these (exact) algorithms increases exponentially with the horizon which means they are only practical for solving problems with small horizons. Due to its high computational complexity, obtaining an exact solution for infinite-horizon POMDPs seems to be very difficult. Madani et al. [Madani et al., 1999] showed that the infinite-horizon POMDPs are undecidable.

Since the infinite-horizon POMDPs are computationally very expensive, in the past years there has been a growing interest in developing various approximate solutions for POMDPs. In the literature, there exist numerous approximation approaches, each has its own advantages and disadvantages. Herein, we mention the basic ideas behind some of the main methods. The grid-based methods [Lovejoy, 1991, Bonet, 2002] pick a (fixed) finite number of points in the belief space and approximate the value functions at these points; then, they use various interpolation techniques to construct an approximate value function over the entire belief space. One of the main advantages of grid-based methods is that their computational complexity remains fixed at each iteration and does not increases with time. However, one of the main disadvantages is that the picked points may not even be reachable. The point-based methods [Pineau et al., 2003, Shani et al., 2013] overcome the reachability drawback by restricting attention to a finite subset of reachable set. In point-based methods, a finite subset of reachable set is picked and based on the notion of $\alpha-$vectors, an approximate value function for the picked points is calculated. Then, iteratively the same procedure is

executed i.e. adding new points to the finite set and updating the value functions until a good performance is reached. In point-based methods, the (picked) points change as the value functions change unlike many grid-based methods where the picked points (i.e., the gird) remain fixed as the value functions change. The focus of point-based methods is on handling the systems with large state space. Since there is a trade-off between optimality and exponential growth in general, most of the point-based methods are case-dependent and they do not perform well for all generic cases. The above methods are structured based on approximating the value function. On the other hand, there are other methods that are based on policy search such as finite-state controllers [Hansen, 1998a, Hansen, 1998b, Kaelbling et al., 1998]. Roughly speaking, in policy-search methods, attention is restricted to a certain class (structure) of strategies. Once the class is chosen, the rest of the approach is to find the best policy in the class using different techniques including policy iteration and gradient ascent. In addition, there are many heuristic methods. For more details on POMDP solvers and related approaches, we refer reader to [Zhang, 2010, Cassandra, 1998, Poupart, 2005, Shani et al., 2013, Murphy, 2000, Lusena, 2001] (and references therein).

So far, we assumed the model of POMDP is known. However, in practice, the model may not be completely known to the agent. In this case, finding an approximate solution of POMDP is more difficult. For example, many powerful POMDP solvers as mentioned above work in the belief space while knowing the belief state itself requires the complete knowledge of the transition probabilities. Therefore, it is not clear how these solvers may be used when the model is not known. We refer reader to [Krishnamurthy, 2016] to see difficulties associated with reinforcement learning in POMDPs.

Main challenges and Contributions

In general, the computational complexity of POMDPs is PSpace-complete [Papadimitriou and Tsitsiklis, 1987] and the infinite-horizon POMDPs are undecidable [Madani et al., 1999]. These results on the computational complexity are under the assumption of complete knowledge of the model. However, when the model is not known, finding an optimal solution is more difficult. For this reason, many researchers have focused attention on approximate solutions. Nevertheless, finding a sub-optimal solution is still a hard task. Our key contributions in this chapter are the following:

1. We propose a novel finite-state approximation for POMDPs. The main difference of

our approach with the existing approaches is that our approach takes the system model into account unlike many approaches that treat all POMDPs similarly. In particular, our approach is designed based on the notion of information state (not necessarily belief state) that allows us to exploit the structure of the problems that may lead to a better approximation compared to the traditional methods.

For example, most of the POMDP solvers in the literature are designed in the belief space because belief state is a genetic information state (sufficient statistic) for POMDPs [Aoki, 1965] and the value function is *piece-wise linear* and *convex* [Smallwood and Sondik, 1973]. However, the size of belief state increases with the size of state space. In contrast, there may exist much simpler (lower-dimensional) information states than the belief state. See the two examples in Section 5.4.

2. Based on the above finite-state approximation, we propose a framework in which all the MDP reinforcement learning algorithms can be used to learn an approximate solution of POMDPs. See Example B in Section 5.5. In particular, when the model is not known, designing reinforcement learning algorithm is difficult because most of the reinforcement algorithms are designed for finite state-action MDPs and do not work for POMDPs. For example, the standard Q-learning algorithm is not implementable in POMDPs for two reasons: (a) belief space is uncountable and (b) knowing the belief state itself requires the knowledge of the model.

Notation

To distinguish the random variables from their realizations, we use capital letters for random variables (e.g. X) and small letters for respective realizations (e.g. x). We use the short-hand notation $X_{a:b}$ for the vector $(X_a, X_{a+1}, \dots, X_b)$.

5.2 Problem formulation

In this section, we present the essentials for modelling a generic POMDP problem. In this chapter, we only consider the infinite horizon, discounted POMDPs and assume discrete state, action and observation sets. The POMDP is a 7-tuple $\langle \mathcal{X}, \mathcal{U}, \mathcal{Y}, P^x, P^y, \ell, \beta \rangle$ such that

- $\mathcal{X}$ is the set of states and $\mathcal{U}$ is the set of actions. We assume the state and action spaces are finite.

- P^x is the transition probability matrix, where $P^x(x'|x,u)$ denotes the probability of transiting to state x' by performing action u at state x.
- $\mathcal{Y}$ is the set of observations. P^y denotes the observation function where $P^y(y|x,u)$ is the probability of observing y at state x when action u is executed. We assume the observation space is finite.
- $\ell(x,u)$ is a real-valued function that determines the immediate cost of action u at state x. Let ℓ_{max} and ℓ_{min} denote the maximum and the minimum of the immediate cost, respectively.

We assume that all the primitive random variables are defined on a common probability space $(\Omega, \mathcal{F}, P)$ where Ω denotes the sample space, $\mathcal{F}$ denotes the sigma-algebra of Ω, and P denotes the probability function. The performance of a policy g is measured by the expected discounted infinite-horizon cost

$$J(g) = \mathbb{E}^g \left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(X_t, U_t) \right], \tag{5.1}$$

where $\beta \in (0,1)$ is the discounting factor and X_t and U_t are the state and action at time t, respectively. The above expectation is with respect to the measure induced by the choice of policy g on the system variables. Let J^* be the minimum amount of the expected discounted infinite-horizon cost given in (5.1). The goal is to find a policy g^*_ϵ, given any $\epsilon > 0$, such that

$$J(g^*_\epsilon) - J^* \leq \epsilon,$$

where g^*_ϵ is called ϵ-optimal policy. Notice that the computational complexity of finding an ϵ-optimal policy is NP-hard [Meuleau et al., 1999, Littman, 1994a]. Prior to describing our approach, we quickly review the concept of information state that plays a key role in our approach.

Information state

In general, to obtain an optimal solution of POMDPs, one needs to track the past actions and observations. An alternative way is to find a state that is updated iteratively based on the executed actions and observed events (observations) that carries all the required

information for obtaining an optimal policy. Such state is called *information state* that is sufficient statistic. It is well-known that the posterior probability of the state of system given the history of actions and observations i.e. $\mathbb{P}(X_t|Y_{1:t}, U_{1:t-1})$, so-called belief state, is information state for POMDPs. However, given a POMDP problem, the belief state may not necessarily be the unique information state. For that reason, we denote any arbitrary information state (not necessarily belief state) by Π_t. By definition of information state, we have

1) The information state Π_t is a function of history space $\mathcal{H} := \{Y_{1:t}, U_{1:t-1}\}$.

2) There exists a function f that updates the information state Π_t based on the current action U_t and next observation Y_{t+1} such that

$$\Pi_{t+1} = f(\Pi_t, U_t, Y_{t+1}). \tag{5.2}$$

3) The current information state and action are sufficient to determine the probability of observing the next observation given history i.e.

$$\mathbb{P}(Y_{t+1}|Y_{1:t}, \Pi_{1:t}, U_{1:t}) = \mathbb{P}(Y_{t+1}|\Pi_t, U_t).$$

4) There exists a function ℓ such that

$$\mathbb{E}[\ell(X_t, U_t)|Y_{1:t} = y_{1:t}, U_{1:t} = u_{1:t}] = \hat{\ell}(\pi_t, u_t). \tag{5.3}$$

Let $\mathcal{R}$ be the reachable set of above information state i.e. $\mathcal{R}$ contains all the realizations π_t generated by $\pi_{t+1} = f(\pi_t, u, y), \forall u \in \mathcal{U}, \forall y \in \mathcal{Y}, \forall t \in \mathbb{N}$, with initial information state π_1. Then, it is well-known that there exists a stationary optimal policy $g^* : \mathcal{R} \to \mathcal{U}$[1] that satisfies the following equation

$$V(\pi_t) = \min_{u_t} \Big(\hat{\ell}(\pi_t, u_t) + \beta \mathbb{E}\big[V(f(\pi_t, u_t, Y_{t+1}))|\Pi_t = \pi_t, U_t = u_t\big] \Big), \tag{5.4}$$

where $V(\pi) : \mathcal{R} \to \mathbb{R}$ is a real-valued function, so called value function. In next section, based on the notion of information state, we present an approach that provides an ϵ-optimal

[1]By slightly abuse of notation, the domain of strategy g^* is considered to be reachable set rather than past actions and observations.

policy.

5.3 Methodology

In this section, we describe our approach by introducing a new notion that we call *Incremental Expanding Representation* (IER).

Definition 1 (Incrementally Expanding Representation (IER)) Let $\{\mathcal{S}_N\}_{N=1}^{\infty}$ be a sequence of finite sets such that $\mathcal{S}_1 \subsetneq \mathcal{S}_2 \subsetneq \ldots \subsetneq \mathcal{S}_N \subsetneq \ldots$. Let $\mathcal{S} = \lim_{N\to\infty} \mathcal{S}_N$ be the countable union of above finite sets, $B : \mathcal{S} \to \mathcal{R}$ be a sujrjective function that maps $\mathcal{S}$ to the reachable set $\mathcal{R}$, and $\tilde{f} : \mathcal{S} \times \mathcal{U} \times \mathcal{Y} \to \mathcal{S}$. The 3-tuple $\langle \{\mathcal{S}_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle$ is called an Incrementally Expanding Representation (IER), if it satisfies the following properties:

(P1) *Incremental Expansion*: For any $u \in \mathcal{U}, y \in \mathcal{Y}$, and $s \in \mathcal{S}_N$, we have that

$$\tilde{f}(s, u, y) \in \mathcal{S}_{N+1}. \tag{5.5}$$

(P2) *Consistency*: For any $(u_{1:t-1}, y_{1:t})$, let π_t be information state and s_t be state obtained by recursive application of (5.2) and (6.5) starting from $\Pi_1 = \pi_1$ and $S_1 = s_1 \in \mathcal{S}_1$, respectively. Then,

$$\pi_t = B(s_t).$$

Remark 5.1 Note that the reachable set $\mathcal{R}$ and the history space $\mathcal{H}$ (that contains all the possible sequences of actions and observations) can be represented as IERs.

One immediate result of Remark 5.1 is that there always exist at least two IERs for every POMDP. Now, we construct a countable-state MDP based on the notion of IERs.

5.3.1 Countable-state MDP Δ

Given a POMDP, choose an appropriate IER. Let the tuple $\langle \{\mathcal{S}_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle$ be the chosen IER. Then, define countable-state MDP Δ as a 5-tuple $\langle \mathcal{S}, \mathcal{U}, \tilde{f}, \tilde{\ell}, \beta \rangle$ such that

- $\mathcal{S}$ is the set of states and $\mathcal{U}$ is the set of actions.
- The initial state is s_1. If action u is executed at state s, the probability of landing in state s' is $p(s'|s, u) = \sum_{y\in\mathcal{Y}} \mathbb{1}(s' = \tilde{f}(s, u, y))\mathbb{P}(y|s, u)$, where $\mathbb{P}(y|s, u) = \mathbb{P}(y|\pi, u)$ such that $\pi = B(s)$.

- $\tilde{\ell}(s,u) := \hat{\ell}(B(s),u)$ is a real-valued function that determines the immediate cost of action u at state s.

The performance of a stationary policy $\psi : \mathcal{S} \to \mathcal{U}$ is measured by

$$\tilde{J}(\psi) = \mathbb{E}^{\psi}\left[\sum_{t=1}^{\infty} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right],$$

The above expectation is with respect to the measure induced by the choice of policy ψ on the system variables of Δ. From standard results in MDP [Bertsekas, 2012], there exists an optimal stationary ψ^*. Construct a strategy g^* such that $\psi^*(s) =: g^*(B(s)), \forall s \in \mathcal{S}$. Then, we have

Lemma 5.1 $J(g^*) = \tilde{J}(\psi^*) = J^*$.

Proof: From standard results in Markov theory [Bertsekas, 2012], there exists a stationary optimal policy ψ^* that satisfies the following dynamic program. For $s_t \in \mathcal{S}$,

$$V(s_t) = \min_{u_t}(\tilde{\ell}(s_t,u_t) + \beta\mathbb{E}[V(\tilde{f}(s_t,u_t,Y_{t+1}))|S_t = s_t, U_t = u_t]).$$

Define a strategy $g^* : \mathcal{R} \mapsto \mathcal{U}$ such that for every $s \in \mathcal{S}$

$$\psi^*(s) =: g^*(B(s))$$

Since $\mathbb{P}(y|s,u) = \mathbb{P}(y|\pi,u)$ with $\pi = B(s)$ and B is surjective, it is trivial to see that strategy g^* is an optimal solution for (5.4) with reachable set $\mathcal{R}$. ■

Remark 5.2 The cardinality of state space of Δ, i.e. $\mathcal{S}$, is allowed to be different from the cardinality of state space $\mathcal{X}$. In fact, this depends on the chosen IER. For example, given a POMDP, the cardinality of reachable set $\mathcal{R}$ depends on the cardinality of state space $\mathcal{X}$ while the cardinality of history space $\mathcal{H}$ does not. Hence, if the size of state space $\mathcal{X}$ increases, the size of state space $\mathcal{S}$ may remain fixed or increase with a different rate.

5.3.2 Finite-state MDP Δ_N

In this part, we construct a series of finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$, that approximate the countable-state MDP Δ as follows. Let Δ_N be a finite-state MDP with state space $\mathcal{S}_N$

and action space $\mathcal{U}$. The transition probability of Δ_N is constructed from the transition probability of Δ as follows. Pick any arbitrary set $\mathcal{D}_N \in \mathcal{S}_N$. For any state $s \in \mathcal{S}_N$ and action $u \in \mathcal{U}$,

$$\begin{aligned}
&\mathbb{P}_{\Delta_N}(s'|s,u) = \mathbb{P}_{\Delta}(s'|s,u), && \text{if } s' \in \mathcal{S}_N \backslash \mathcal{D}_N, \\
&\mathbb{P}_{\Delta_N}(s'|s,u) = \mathbb{P}_{\Delta}(s'|s,u) + \sum_{\tilde{s} \notin \mathcal{S}_N} \mathbb{P}_{\Delta}(\tilde{s}|s,u) q_N(s',\tilde{s},u,s), && \text{if } s' \in \mathcal{D}_N,
\end{aligned}$$

where $q_N(s',\tilde{s},u,s) \geq 0, \forall s', s \in \mathcal{S}_N, \forall u \in \mathcal{U}, \forall \tilde{s} \in \mathcal{S} \backslash \mathcal{S}_N$ and $\sum_{s' \in \mathcal{D}_N} q_N(s',\tilde{s},u,s) = 1$. In other words, we remap every transition in Δ that takes the state $s \in \mathcal{S}_N$ to $\tilde{s} \notin \mathcal{S}_N$ to a transition from $s \in \mathcal{S}_N$ to a state in $s' \in \mathcal{D}_N$. In addition, the immediate cost function of Δ_N is simply a restriction of $\tilde{\ell}$ to $\mathcal{S}_N \times \mathcal{U}$. The performance of a stationary policy $\psi_N : \mathcal{S}_N \to \mathcal{U}$ is measured by the expected discounted infinite-horizon cost

$$\tilde{J}_N(\psi_N) = \mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\infty} \beta^{t-1} \tilde{\ell}(S_t, U_t)\right].$$

5.3.3 Main results

Let $\tau_N \in \mathbb{N}$ denote a fixed horizon[2] prior to which state S_t, $t \leq \tau_N$, always remains in $\mathcal{S}_N$ under dynamics $\tilde{f}$, optimal strategy ψ^*, and any arbitrary sample path of $Y_{1:t}$. In other words, S_t can not exist $\mathcal{S}_N$ if $t \leq \tau_N$, i.e.,

$$S_t = \tilde{f}(S_{t-1}, \psi^*(S_{t-1}), Y_t) \in \mathcal{S}_N, \quad \forall t \leq \tau_N.$$

Let ψ_N^* be an optimal stationary strategy of Δ_N.

Theorem 5.1 *The difference in performance between Δ and Δ_N is bounded as follows:*

$$|\tilde{J}(\psi^*) - \tilde{J}_N(\psi_N^*)| \leq \frac{2\beta^{\tau_N}}{1-\beta}(\ell_{max} - \ell_{min}).$$

[2]Note that τ_N is not a random variable.

Proof: We have

$$\begin{aligned}\tilde{J}(\psi^*) = \min_{\psi} \tilde{J}(\psi) &= \min_{\psi} \mathbb{E}^{\psi}\left[\sum_{t=1}^{\infty} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] \\ &= \min_{\psi}\left(\mathbb{E}^{\psi}\left[\sum_{t=1}^{\tau_N} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] + \mathbb{E}^{\psi}\left[\sum_{t=\tau_N+1}^{\infty} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right]\right) \\ &\overset{(a)}{\leq} \min_{\psi_N} \mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\tau_N} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] + \min_{\psi_N, \psi_N^{-1}} \mathbb{E}^{\psi_N, \psi_N^{-1}}\left[\sum_{t=\tau_N+1}^{\infty} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] \\ &\leq \min_{\psi_N} \mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\tau_N} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] + \frac{\beta^{\tau_N}}{1-\beta}(\ell_{max} - \ell_{min}),\end{aligned}$$

where (a) follows from splitting $\psi : \mathcal{S} \to \mathcal{U}$ into two smaller strategies $\psi = (\psi_N, \psi_N^{-1})$ such that $\psi_N : \mathcal{S}_N \to \mathcal{U}$ and $\psi_N^{-1} : \mathcal{S} - \mathcal{S}_N \to \mathcal{U}$, and the fact that state S_t always stays in $\mathcal{S}_N$ for $t \leq \tau_N$.
In addition, let $\tilde{J}_N(\psi_N^*)$ be the optimal expected cost under policy ψ_N^* as follows.

$$\begin{aligned}\tilde{J}_N(\psi_N^*) &= \min_{\psi_N} \mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\infty} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] \\ &= \min_{\psi_N}\left(\mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\tau_N} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] + \mathbb{E}^{\psi_N}\left[\sum_{t=\tau_N+1}^{\infty} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right]\right) \\ &\leq \min_{\psi_N} \mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\tau_N} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right] + \frac{\beta^{\tau_N}}{1-\beta}(\ell_{max} - \ell_{min}).\end{aligned}$$

Let $a := \min_{\psi_N} \mathbb{E}^{\psi_N}\left[\sum_{t=1}^{\tau_N} \beta^{t-1}\tilde{\ell}(S_t, U_t)\right]$, then we have

$$\begin{aligned}&|\tilde{J}(\psi^*) - \tilde{J}_N(\psi_N^*)| = |J^* - \tilde{J}_N(\psi_N^*) + a - a| \leq |J^* - a| + |\tilde{J}_N(\psi_N^*) - a| \\ &\leq \frac{2\beta^{\tau_N}}{1-\beta}(\ell_{max} - \ell_{min}).\end{aligned}$$

The proof is complete. ■

The upper-bound provided in Theorem 5.1 requires the knowledge on ψ^*. However, according to (6.5), τ_N is always equal or greater than N i.e. $N \leq \tau_N$. Hence, one can obtain a more conservative error-bound (larger upper-bound) than the error-bound (upper-bound)

in Theorem 5.1 that does not require any knowledge on ψ^*.

Corollary 5.1 *The difference in performance between Δ and Δ_N is bounded as follows:*

$$|\tilde{J}(\psi^*) - \tilde{J}_N(\psi_N^*)| \leq \frac{2\beta^N}{1-\beta}(\ell_{max} - \ell_{min}).$$

From Lemma 5.1, Theorem 5.1, and Corollary 5.1, we have that

Theorem 5.2 *The approximation error is bounded as follows:*

$$|J^* - \tilde{J}_N(\psi_N^*)| \leq \epsilon_N,$$

where $\epsilon_N = \frac{2\beta^{\tau_N}}{1-\beta}(\ell_{max} - \ell_{min}) \leq \frac{2\beta^N}{1-\beta}(\ell_{max} - \ell_{min})$. The error converges to zero exponentially in N.

Proof: Let J^* be the minimum cost given by (5.1), then from Lemma 5.1 we have $J^* = \tilde{J}(\psi^*)$. The rest follows from Theorem 5.1. ■

At time t, action $u_t = \psi_N^*(s_t)$ is executed and an observation y_t is seen. Then, state s_t will transmit to next state s_{t+1} according to dynamics of MDP Δ_N. According to Lemma 5.1, the performance of policy ψ_N^* is in the ϵ-neighbourhood of the optimal performance, where $\frac{2\beta^N}{1-\beta}(\ell_{max} - \ell_{min}) \leq \epsilon$. In next section, we provide two examples where the proposed approach perform efficiently. However, notice that to ensure optimality in general, we will face the exponential growth anyway and there is no way around that.

5.4 Examples

In this section, we illustrate our approach by two examples.

5.4.1 Machine maintenance

Consider a machine with n identical independent internal components. The status of component $i \in \{1, \ldots, n\}$ is either defective or non-defective i.e. $X_t^i \in \{0, 1\}$. The probability that a component remains non-defective at each hour is p. If a component fails, it remains failed until it is either fixed or replaced. Let $\mathbf{X}_t = \sum_{i=1}^n X_t^i \in \{0, 1, \ldots, n\}$ be the number of non-defective components in the machine at hour t. The manager of the machine does not observe the status of components and he has three choices at each hour: a) continue

i.e. $u_t = C$, b) hire a serviceman to inspect and fix the machine i.e. $u_t = F$, and c) replace the machine i.e. $u_t = R$. If the manager decides to continue, it will be free of charge and the machine will evolve according to the uncontrolled dynamics. If the manager hires a serviceman to inspect the internal components and fix all the defective components, it will cost him $\gamma(n - \mathbf{X}_t) + \alpha$. The manager also has the option to buy a new machine and replace the current machine with cost c. We assume that when the machine is replaced or fixed, probability p does not change. The manager is interested to increase the number of non-defective components with the lowest cost. Hence, the immediate cost function is defined as follows.

$$\ell(\mathbf{X}_t, U_t) = C(\mathbf{X}_t, U_t) - \mathbf{X}_t,$$

where the cost of each action is as follows: $C(\mathbf{X}_t, U_t = C) = 0$, $C(\mathbf{X}_t, U_t = F) = \gamma(n - \mathbf{X}_t) + \alpha$, and $C(\mathbf{X}_t, U_t = R) = c$. At time t, the expected cost may be written (represented) as a function of posterior probability of the joint state $\mathbb{P}(X_t^1, \ldots, X_t^n | U_{1:t-1})$ or the number of non-defective components $\mathbb{P}(\mathbf{X}_t | U_{1:t-1})$. The benefit of such representations is that the value function will be piece-wise linear and convex. However, as n increases, the size of these information states increases so as the computational complexity.

Alternatively, we choose a different information state whose size does not depend on the number of components. Write (represent) the expected cost in terms of the probability of a component being non-defective i.e. M_t, where the initial state is $M_1 = 1$ (components are non-defective at the beginning). M_t evolves as follows: $M_{t+1} = 1$ if $u_t \in \{F, R\}$ and $M_{t+1} = pM_t$ if $u_t = C$. The expected cost may be written as function of M_t as follows. $\ell(M_t, U_t = C) := -\sum_{k=0}^{n} k \binom{n}{k} M_t^k (1 - M_t)^{n-k}$, $\ell(M_t, U_t = F) = \gamma n + \alpha - (1 + \gamma)\ell(M_t, U_t = C)$ and $\ell(M_t, U_t = R) = c - \ell(M_t, U_t = C)$. Note the the respective value function is not piece-wise linear and convex. Based on the information state M_t, we define IER $\langle \{\mathcal{S}_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle$ such that: A) $\mathcal{S}_N = \{1, \ldots, N\}, N \in \mathbb{N}$, B) $\forall s_t \in \mathcal{S}_N$, if $u = C$, then $s_{t+1} = \tilde{f}(s_t, u) = s_t + 1 \in \mathcal{S}_{N+1}$ and if $u \in \{F, R\}$, then $s_{t+1} = \tilde{f}(s_t, u) = 1 \in \mathcal{S}_1$, and C) the surjective mapping B is defined as $M_t = B(t), t \in \mathbb{N}$.

Construct a countable state MDP Δ with state space $\mathcal{X} = \mathbb{N}$, action space $\{C, F, R\}$, dynamics $\tilde{f}$, and cost $\tilde{\ell}(s, u) := \ell(m, u)$. Now, approximate Δ with finite-state MDP Δ_N such that the state space is $\mathcal{S}_N = \{1, \ldots, N\}$, action space $\{C, F, R\}$, if $\tilde{f}(s, u) \in \mathcal{S}_N$, then next state $s' = \tilde{f}(s, u)$; otherwise, $s' = 1$. The cost function remains as same as that of Δ. Now, we can find the optimal solution of finite-state MDP Δ_N using different approaches;

we use value iteration. We use the following numerical parameters in our simulations: $N = 100, n = 10, \gamma = 1, \alpha = 5, c = 7, p = 0.99, \beta = 0.9$. The optimal solution is to continue 13 times sequentially and then call the service man. If we change p to 0.95, the optimal solution is to continue 6 times sequentially and then replace. If we keep $p = 0.95$ and increase the number of components to $n = 20$, the optimal solution will be to continue 3 times sequentially and then replace.

5.4.2 Sensor network

In this section, we present an idealized model of a sensor network. Suppose X_t is a Markov process with transition probability P that models an environmental variable of interest (e.g., temperature, rainfall, sunlight, etc.). An estimator wants to estimate X_t, but measurements are expensive. The estimator may take a (noiseless) measurement at a cost c or not take a measurement but generate an estimate $\hat{X}_t$ of the process from previous measurements, incurring a cost $d(X_t, \hat{X}_t)$. The objective is to choose measurement strategies that minimize the total discounted cost.

This may be modelled as a POMDP with a belief state $b_t = \mathbb{P}(X_t \mid Y_{1:t})$, where Y_{t+1} denotes the measurement at time t. The belief state evolves as follows:

$$b_{t+1} = \begin{cases} b_t P, & \text{if no measurement is taken} \\ \delta_x, & \text{with probability } b_t(x) \text{if measurement is taken.} \end{cases} \tag{5.6}$$

If the initial state of the Markov process is known, then the reachable set of the belief state is

$$\mathcal{R} = \{\delta_x P^n : x \in \mathcal{X} \text{ and } n \in \mathbb{N}\}$$

Thus, we can construct a countable state MDP Δ with state space $\mathcal{S} = \mathcal{X} \times \mathbb{N}$, dynamics given by (5.6) and cost

$$\ell((x, n), u) = cu + (1 - u)\mathbb{E}[d(X_t, \hat{X}) \mid b_t = \delta_x P^n].$$

We can approximate this by a finite state MDP Δ_N such that the state space is $\mathcal{X} \times \{0, \ldots, N\}$, where the cost is as before and the dynamics are if $n < N$, then $s' = \tilde{f}((x, n), u)$ otherwise $s' = (x, 1)$ with probability $[\delta_x P^n](x)$. Now, we can find the optimal solution

of finite-state MDP Δ_N using different approaches. Note that the size of reachable set $\mathcal{R}$ increases linearly with time horizon while the size of history space increases exponentially.

5.5 Reinforcement Learning

As mentioned in Section 5.1, one of the main features of the approach introduced in this chapter is the ability to employ MDP reinforcement learning algorithms including TD(λ) and Q-learning to learn an approximate solution of POMDPs.

There are many applications where the model of the system is not completely available and the agent needs to learn the optimal solution by interacting with its environment. In general, finding an ϵ-optimal policy in such applications is much harder than when the model is fully known. Our goal is to develop Reinforcement Learning (RL) algorithms for POMDPs when the model is partially known (that we call model-based RL) or fully unknown (that we call model-free RL). We assume that the agent observes the immediate cost.

Herein, we use the finite-state approximation results presented in Section 5.3. In particular, we add one requirement to the notion of IER defined in Section 5.3 that is state space $\mathcal{X}$ and dynamics $\tilde{f}$ must not depend on the unknown parameters. Note that there always exists at least one IER that satisfies such requirement as follows.

Example A: Let $S_1 = \{\emptyset\}$, $S_2 = \{\emptyset\} \cup \{\mathcal{U} \times \mathcal{Y}\}$, and $S_{t+1} = S_t \cup \{\mathcal{U} \times \mathcal{Y}\}^t, t \in \mathbb{N}$. Let $S = \lim_{t\to\infty} S_t$ and $B : \mathcal{S} \to \mathcal{R}$ such that

$$B(\emptyset) = \pi_1, B(s_{t+1}) = f(f(...., u_{t-1}, y_t), u_t, y_{t+1}) = \pi_{t+1},$$

where $s_{t+1} = ((u_1, y_2), \ldots, (u_t, y_{t+1})) \in \mathcal{S}_{t+1}$. Define $\tilde{f}$ as follows: $\tilde{f}(s, u, y) = s \circ u \circ y$, where $\circ$ denotes concatenation. By construction, 3-tuple $\langle \{\mathcal{S}_t\}_{t=1}^{\infty}, B, \tilde{f} \rangle$ is an IER that does not depend on the model of the systems i.e. (does not depend on unknowns). Notice that there may exist more than one IER, given a POMDP, whose $\mathcal{S}$ and $\tilde{f}$ do not depend on the unknowns. In Example B, we develop a model-based RL algorithm based on a simpler IER than the generic one described above.

Once an appropriate IER is chosen, we construct the countable-state MDP Δ as described in Section 5.3.1 and approximate it by Δ_N as described in Section 5.3.2. Then, we can use various reinforcement learning algorithms to learn the optimal strategy of the finite-state MDP Δ_N.

Let $\mathcal{T}$ be a generic (model-based or model-free) RL algorithm designed for finite-state

MDPs with infinite horizon discounted cost. By a generic RL algorithm, we mean any algorithm which fits to the following framework. At each iteration $k \in \mathbb{N}$, $\mathcal{T}$ knows the state of system, selects one action, and observes an instantaneous cost and the next state. The strategy learned (generated) by $\mathcal{T}$ converges to an optimal strategy as $k \to \infty$.

Let $\mathcal{T}$ operate on MDP Δ_N such that, at iteration k, it knows the state of the system $s_k \in \mathcal{S}_N$, selects one action $u_k \in \mathcal{U}$, and observes an instantaneous cost ℓ_k (which is a realization of the incurred cost $\ell(X_k, U_k)$) . According to (5.3) and construction of $\tilde{\ell}$, we have

$$\mathbb{E}[\ell(X_k, U_k)|S_{1:k}, U_{1:k}] = \tilde{\ell}(S_k, U_k), \quad S_k \in \mathcal{S}_N.$$

Hence, the instantaneous cost ℓ_k may be interpreted as a realization of the per-step cost of Δ_N. Given dynamics $\tilde{f}$, $\mathcal{T}$ observes $y_{k+1} \in \mathcal{Y}$ and computes the next state $s_{k+1} = \tilde{f}(s_k, u_k, y_{k+1})$. If $s_{k+1} \in \mathcal{S}_{N+1} \backslash \mathcal{S}_N$, then an action (or a sequence of actions) that transmits s_k to a known state in S_N (i.e. $s_{k+1} = d^* \in D^*$) will be taken; otherwise, the system will continue from $s_{k+1} \in \mathcal{S}_N$.

Let $\psi_N^k : \mathcal{S}_N \to \mathcal{U}$ be the learned strategy associated with RL algorithm $\mathcal{T}$ operating on MDP Δ_N at iteration k. Then, $\mathcal{T}$ updates its strategy ψ_N^{k+1} based on the observed cost ℓ_k and the transmitted next state s_{k+1} by executing action u_k at state s_k. We assume $\mathcal{T}$ converges to an optimal strategy ψ_N^* as $k \to \infty$ such that

$$\lim_{k\to\infty} |\tilde{J}_N(\psi_N^k) - \tilde{J}_N(\psi_N^*)| = 0. \tag{5.7}$$

Theorem 5.3 *The approximation error associated with the learned strategy is bounded as follows:*

$$\lim_{k\to\infty} |J^* - \tilde{J}_N(\psi_N^k)| = |\tilde{J}(\psi^*) - \tilde{J}_N(\psi_N^*)| \leq \epsilon_N,$$

where $\epsilon_N = \frac{2\beta^{\tau_N}}{1-\beta}(\ell_{max} - \ell_{min}) \leq \frac{2\beta^N}{1-\beta}(\ell_{max} - \ell_{min})$.

The proof follows from Theorem 5.2 and (6.10).

Remark 5.3 Suppose Q-learning algorithm is chosen as the RL method. In this case, the size of action space $|\mathcal{U}|$ affects the number of Q functions in two places: state space $\mathcal{S}_N$ that grows exponentially with $|\mathcal{U}|$ (in worst case) and action space $|\mathcal{U}|$ that grows linearly. It is difficult to comment on how the running-time of the algorithm would be affected since the convergence time is stochastic and also depends on the model.

Algorithm 1 Finite-state reinforcement learning algorithm

1: Given $\epsilon > 0$, choose a sufficiently large $N \in \mathbb{N}$ such that $\frac{2\beta^N}{1-\beta}(\ell_{max} - \ell_{min}) \leq \epsilon$. Then, construct state space $\mathcal{S}_N$, action space $\mathcal{U}$, and dynamics $\tilde{f}$.

2: At iteration $k \in \mathbb{N}$, RL algorithm $\mathcal{T}$ picks $u_k \in \mathcal{U}$ at state $s_k \in \mathcal{S}_N$.

3: Based on the taken action, the system incurs a cost ℓ_k, evolves, and generates an observation y_{k+1} based on which the agent computes the next state as follows.

$$s_{k+1} = \tilde{f}(s_k, u_k, y_{k+1}).$$

If $s_{k+1} \notin \mathcal{S}_N$, then the agent takes an action (or a sequence of actions) that takes the state of system to a known state $s_{k+1} = d^* \in \mathcal{S}_N$; otherwise, the system proceeds from $s_{k+1} \in \mathcal{S}_N$.

4: $\mathcal{T}$ updates its strategy from ψ_N^k to ψ_N^{k+1} based on performing action u_k at state s_k and transmission to next state s_{k+1} with instantaneous cost ℓ_k.

5: $k \leftarrow k + 1$, and go to step 2 until termination.

Example B

Consider a slightly different version of machine maintenance example in Section 5.4, where the manager is interested to keep the number of non-defective components greater than a threshold $D \leq n$ for machine to work with minimum cost. Hence, if the manager decides to continue, he will be rewarded zero (no cost) unless the number of non-defective components are below D. In the latter case, the machine stops and the manager has to pay a lot of money i.e. $\ell(\mathbf{X}_t, U_t = C) = H \cdot \mathbb{1}(\mathbf{X}_t < D)$. If he calls the service man, it will cost him $\ell(\mathbf{X}_t, U_t = F) = \gamma(n - \mathbf{X}_t) - \alpha$ and if he replaces the machine, it will cost him $\ell(\mathbf{X}_t, U_t = R) = c$. The objective is to learn the optimal policy when probability p is unknown for discounted infinite horizon setup. The same IER that we use in the machine maintenance example in Section 5.4, can be used here because none of the IER components depend on p. We use the following numerical parameters in our simulations: $N = 100, n = 10, \gamma = 1, \alpha = 5, c = 7, p = 0.99(\text{unknown}), \beta = 0.9, H = 100$. If $D = 5$, the optimal solution is to continue 24 times sequentially and then replace the machine. If $D = 6$, the optimal policy is to continue 16 times sequentially and then call the service-man. If $D = 7$, the optimal policy is to continue 10 times sequentially and then call the service-man.

Remark 5.4 For on-line learning, we require a *reset property* to make sure the states and actions are visited frequently. In particular, when state s_t steps out of $\mathcal{S}_N$, this property

makes sure it can go back to $\mathcal{S}_{\mathcal{N}}$. This property has been imposed in various forms in the literature. In [Jaulmes et al., 2005], this property is motivated by the existence of an oracle that provides the current state perfectly upon request at some cost. Similar motivation is valid in sensor network systems when the sensing is cheap but communication is costly. Upon request, the exact state of the system can be delivered to the decision maker. In [Pivazyan and Shoham, 2002, Even-Dar, 2005], it is assumed there exists a reset strategy ("reset button") or an approximate reset strategy ("homing strategy").

5.6 Conclusion

In this chapter, we proposed a new approach for approximate solutions of POMDPs in the discounted infinite horizon setup. We introduced a new notion called Incrementally Expanding Representations (IERs) based on the concept of information state. Using IERs, we converted the POMDP to a countable-state MDP Δ and then approximated it by a finite-state MDP Δ_N. We showed that the error associated with this approximation is bounded and converges to zero. We illustrated our approach by two numerical examples. The proposed approach has a few salient features as follows.

- Unlike the point-based methods and other approaches that use the notion of α-vectors (i.e. rely on piece-wise linearity and convexity of value function), the proposed approach does not involve α-vectors and the value functions are updated over finite states using the conventional value iteration. This feature is similar to the grid-based method.

- Unlike the grid-based methods, the proposed approach works on a subset of reachable set, so all the states are reachable. This feature is similar to the point-based methods.

- The proposed approach can handle cases where the respective value function is not necessarily piece-wise linear and convex. In the machine maintenance example, it is shown how this feature can contribute to an efficient solution.

- The proposed approach takes the model of the system into account and exploits it by using defining an appropriate information state.

- The proposed approach extends to reinforcement learning setup.

CHAPTER 6

Decentralized reinforcement learning with partial history sharing

6.1 Introduction

In this chapter, we are interested in systems with multiple agents that wish to collaborate in order to accomplish a common task while a) the agents have different information (decentralized information) and b) the agents do not know the model of the system completely i.e., they may know the model partially or may not know it at all. The agents must learn the optimal strategies by interacting with their environment using decentralized Reinforcement Learning (RL). The presence of multiple agents with different information makes the decentralized reinforcement learning conceptually more difficult than centralized reinforcement learning. In this chapter, we develop a decentralized reinforcement learning algorithm that learns ϵ-team-optimal solution for the partial history sharing information structure, which encompasses a large class of decentralized control systems including delayed sharing, control sharing, mean-field sharing, etc. Our approach consists of two main steps. In the first step, we convert the decentralized control system to an equivalent centralized POMDP (Partially Observable Markov Decision Process) using an existing approach called common information approach. However, the resultant POMDP requires the complete knowledge of the system model. In principle, any RL algorithm of POMDP can be used in the second step. We use the finite-state approximation method developed in Chapter 5 for the second step. We illustrate the proposed approach and verify it numerically by obtaining a decentralized Q-learning algorithm for two-user Multi Access Broadcast Channel (MABC) which

is a benchmark example.

Literature review

Most of the literature decentralized decision making assumes that the system model is completely known to all decision makers; however, in practice, such knowledge may only be available partially or may not be available. Hence, it is crucial for decision makers to be able to learn the optimal solutions. In the literature, learning in centralized stochastic control is well studied and there exist many approaches such as model-predictive control, adaptive control, and reinforcement learning. This is in contrast to the learning in decentralized stochastic control; it is not immediately clear on how centralized learning approaches would work for decentralized systems. In this chapter, we propose a novel Reinforcement Learning (RL) algorithm for a class of decentralized stochastic control systems that guarantees team-optimal solution.

Existing approaches for multi-agent learning may be categorized as follows: exact methods and heuristics. The exact methods rely on the assumption that the information structure is such that all agents can consistently update the Q-function. These include approaches that rely on social convention and rules to restrict the decisions made by the agents [Spaan et al., 2002]; approaches that use communication to convey the decisions to all agents [Vlassis, 2007]; and approaches that assume that the Q-function decomposes into a sum of terms, each of which is independently updated by an agent [Kok et al., 2005]. Heuristic approaches include joint action learners heuristic [Claus and Boutilier, 1998], where each agent learns the empirical model of the system in order to estimate the control action of other agents; frequency maximum Q-value heuristic [Kapetanakis and Kudenko, 2002], where agents keep track of the frequency with which each action leads to a "good" outcome; heuristic Q-learning [Matignon et al., 2007], which assigns a rate of punishment for each agent; and distributed Q-learning [Huang et al., 2005], which uses predator-prey models to assign heuristic sub-goals to individual agents [Busoniu et al., 2006].

Main challenges and contributions

Given the complete knowledge of system model, finding team-optimal solution in decentralized control systems is conceptually challenging due to the decentralized nature of information available to the agents. The agents need to cooperate with each other to fulfill a

common objective while they have different perspectives about themselves, other agents, and the environment. This discrepancy in perspectives makes establishing cooperation among agents difficult; we refer reader to [Murphey and Pardalos, 2002] for details. Thus, finding team-optimal solution is even more challenging when agents have only partial knowledge or no knowledge of system model. Hence, it is difficult to *consistently* learn strategies in such settings.

In a more technical language, a key underlying model assumption for most of the existing RL algorithms [Sutton and Barto, 1998] is that the model is assumed to be finite state MDP; in contrary, a decentralized control system may not be MDP from each agent's eye. To overcome this technical challenge, we use common information approach for decentralized control systems with partial history sharing [Nayyar et al., 2013]. Following [Nayyar et al., 2013] the decentralized stochastic control problem is transformed into an equivalent centralized stochastic control problem. This transformation is performed from the point of view of a virtual coordinator that observes the common information among all decision makers and chooses partially evaluated functions that map local information at a decision maker to its actions. The coordinator's problem is a centralized POMDP (Partially Observable Markov Decision Process). Thus, in principle, any learning algorithm for centralized (partially observed) systems may be used for the coordinated system with an unknown model. However, as mentioned in Chapter 5, reinforcement learning in POMDP is difficult. To overcome this technical challenge, we use the method developed in Chapter 5.

Below, we mention our main contributions in this chapter.

1. We propose a novel approach to perform reinforcement learning in a large class of decentralized stochastic control systems with partial history sharing (PHS) information structure that guarantees ϵ-team-optimal solution. To the best of the author's knowledge, there is no RL approach that guarantees team-optimal solution.

2. To illustrate our approach, we illustrate how to use Q-learning on a specific example of decentralized control—two-user multi-access communication. Numerical simulations validate that the outcome of the decentralized Q-learning algorithm converges to optimal strategy.

Notation

We use upper-case letters to denote random variables (e.g. X) and lower-case letters to denote their realizations (e.g. x). We use the short-hand notation $X_{a:b}$ for the vector $(X_a, X_{a+1}, \dots, X_b)$ and bold letters to denote vectors e.g. $\mathbf{Y} = (Y^1, \dots, Y^n)$. $\mathbb{P}(\cdot)$ is the probability of an event, $\mathbb{E}[\cdot]$ is the expectation of a random variable, and $|\cdot|$ is the absolute value of a real number. $\mathbb{N}$ refers to the set of natural numbers and $\mathbb{Z}^+ = \mathbb{N} \cup \{0\}$. We use subscripts t to denote real time, N to denote approximation index, and k to denote the learning iteration.

6.2 Problem formulation

Let $X_t \in \mathcal{X}$ denote the state of a dynamical system controlled by n agents. At time t, agent i observes $Y_t^i \in \mathcal{Y}^i$ and chooses $U_t^i \in \mathcal{U}^i$. For ease of notation, we denote the joint actions and the joint observations by $\mathbf{U}_t = (U_t^1, \dots, U_t^n) \in \mathcal{U}$ and $\mathbf{Y}_t = (Y_t^1, \dots, Y_t^n) \in \mathcal{Y}$, respectively. The dynamics of the system are given by

$$X_{t+1} = f(X_t, \mathbf{U}_t, W_t^s),$$

and the observations are given by

$$\mathbf{Y}_t = h(X_t, \mathbf{U}_{t-1}, W_t^o),$$

where $\{W_t^s\}_{t=1}^\infty$ is an i.i.d. process with probability distribution function P^w, $\{W_t^o\}_{t=1}^\infty$ is an i.i.d. process with probability distribution function P^y, and X_1 is the initial state with probability distribution function P^x. The primitive random variables $\{X_1, \{W_t^s\}_{t=1}^\infty, \{W_t^o\}_{t=1}^\infty\}$ are mutually independent and defined on a common probability space.

For ease of exposition, we assume all system variables are finite valued. Let $I_t^i \subseteq \{\mathbf{Y}_{1:t}, \mathbf{U}_{1:t-1}\}$ be information available at agent i at time t. The collection $(\{I_t^i\}_{t=1}^\infty, i = 1, \dots, n)$ is called the *information structure*. In this chapter, we restrict attention to an information structure called *partial history sharing* [Nayyar et al., 2013], which will be defined later.

At time t, agent i chooses action U_t^i according to *control law* g_t^i as follows

$$U_t^i = g_t^i(I_t^i).$$

We denote $\boldsymbol{g}^i = (g^i_1, g^i_2, \ldots)$ as *strategy* of agent i and $\boldsymbol{g} = (\boldsymbol{g}^1, \ldots, \boldsymbol{g}^n)$ as joint strategy of all the agents. The performance of strategy $\mathbf{g}$ is measured by the following infinite-horizon discounted cost

$$J(\boldsymbol{g}) = \mathbb{E}^{\boldsymbol{g}} \left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(X_t, \mathbf{U}_t) \right], \tag{6.1}$$

where $\beta \in (0, 1)$ is the discount factor, ℓ is the per-step cost function, and the expectation is with respect to a joint probability distribution on $(X_{1:\infty}, \mathbf{U}_{1:\infty})$ induced by the joint probability distribution on the primitive random variables and the choice of strategy $\boldsymbol{g}$.

A strategy $\boldsymbol{g}^*$ is optimal if for any other strategy $\boldsymbol{g}$, $J(\boldsymbol{g}^*) \leq J(\boldsymbol{g})$. For $\epsilon > 0$, strategy $\boldsymbol{g}^*$ is ϵ-optimal, if for any other strategy $\boldsymbol{g}$, $J(\boldsymbol{g}^*) \leq J(\boldsymbol{g}) + \epsilon$.

Optimization problem

We will consider three different setups that differ in the assumptions about the knowledge of the model. For all the setups, we will assume that the action and the observation spaces as well as the information structure, the discount factor β, and an upper-bound on the per-step cost are common knowledge between all agents. The setups differ in the assumptions about state space $\mathcal{X}$, system dynamics and observations (f, h), probability distributions (P^x, P^w, P^y), and cost structure ℓ. These include two setups, 1) complete-knowledge of the model, and 2) incomplete-knowledge of the model which includes two sub-cases: 2a) partial-knowledge of the model and 2b) no-knowledge of the model.

In general, the complete-knowledge of the model is required to find an optimal strategy $\boldsymbol{g}^*$. However, in practice, there are many applications where such information is not completely available or is not available at all. In such applications, the agents must learn the optimal strategy by interacting with their environment. This is known as reinforcement learning (RL). If the agents have partial knowledge of the model, the setup is called *model-based* RL. If the agents have no knowledge of the model, setup is called *model-free* RL. Let $\ell_{max} \geq \max_{x,\mathbf{u}} |\ell(x, \mathbf{u})|$.. We are interested in the following problem.

Problem 6.1 *Given the information structure, action spaces $\{\mathcal{U}^i\}_{i=1}^n$, observation spaces $\{\mathcal{Y}^i\}_{i=1}^n$, discount factor β, the upper-bound ℓ_{max} on per-step cost, and any $\epsilon > 0$, develop a (model-based or model-free) reinforcement learning algorithm using which the agents learn an ϵ-optimal strategy $\boldsymbol{g}^*$.*

Preliminaries on Partial History Sharing

Herein, we present a simplified version of partial history sharing information structure, originally presented in [Nayyar et al., 2013].

Definition 2 ([Nayyar et al., 2013], Partial History Sharing (PHS)) Consider a decentralized control system with n agents. Let I_t^i denote the information available to agent i at time t. Assume $I_t^i \subseteq I_{t+1}^i$. Then, split the information at each agent into two parts: *common information* $C_t = \bigcap_{i=1}^n I_t^i$ i.e. the information shared between all agents and *local information* $M_t^i = I_t^i \backslash C_t$ that is the local information of agent i. Define $Z_t := C_{t+1} \backslash C_t$ as common observation, then $C_{t+1} = Z_{1:t}$. An information structure is called *partial history sharing* when the following conditions are satisfied:

a) The update of local information

$$M_{t+1}^i \subseteq \{M_t^i, U_t^i, Y_{t+1}^i\} \backslash Z_t, \quad i \in \{1, \dots, n\}.$$

b) For every agent i, the size of the local information M_t^i and the size of the common observation Z_t are uniformly bounded in time t.

These conditions are fairly mild and are satisfied by a large class of models. Examples include delayed sharing [Nayyar et al., 2011], periodic sharing [Ooi et al., 1997], mean-field sharing [Arabneydi and Mahajan, 2014], etc. Even for models that do not satisfy the above conditions directly, it is often possible to identify sufficient statistics that satisfy the above conditions, e.g., control sharing [Mahajan, 2013].

Remark 6.1 Note that common information between agents is allowed to be empty i.e., $C_t = \emptyset$.

Remark 6.2 Notice that finite-horizon Decentralized Partially Observable Markov Decision process (DEC-POMDP)—first introduced by [Bernstein et al., 2002]—belongs to the partial history sharing.

6.3 Methodology

In this part, we derive a RL algorithm for systems with PHS information structure. Our approach consists of two steps. In the first step, we consider the setup of the complete-

knowledge of the model and use the *common information approach* of [Nayyar et al., 2013] to convert the decentralized control problem to an equivalent centralized POMDP. In the second step, we consider the setup of incomplete-knowledge of the model and develop a finite-state RL algorithm based on the POMDP obtained in the first step.

6.3.1 Step 1: An equivalent centralized POMDP

In this section, we present common information approach of [Nayyar et al., 2013] and its main results for the setup of complete-knowledge of the model described in Section 6.2.

Let $\mathcal{M}^i$ and $\mathcal{Z}$ denote the spaces of realizations of local information of agent i and common observation, respectively. Consider a virtual *coordinator* that observes the common information C_t shared between all agents and chooses $(\Gamma_t^1, \dots, \Gamma_t^n)$, where $\Gamma_t^i : \mathcal{M}^i \to \mathcal{U}^i$ is the mapping from the local information of agent i to action of agent i at time t, according to

$$\Gamma_t^i = \psi_t^i(C_t), \quad i \in \{1, \dots, n\}.$$

We call $\boldsymbol{\psi}_t := \{\psi_t^1, \dots, \psi_t^n\}$ the *coordination law* and $\boldsymbol{\Gamma}_t = (\Gamma_t^1, \dots, \Gamma_t^n)$ the *prescription*. The agents use this prescription to choose their actions as follows:

$$U_t^i = \Gamma_t^i(M_t^i), \quad i \in \{1, \dots, n\}.$$

We denote the space of mappings Γ_t^i by $\mathcal{G}^i$ and the space of prescriptions $\boldsymbol{\Gamma}_t$ by $\mathcal{G} = \prod_{i=1}^n \mathcal{G}^i$. In the sequel, for ease of notation, we will use the following compact form for the coordinator's law,

$$\boldsymbol{\Gamma}_t = \boldsymbol{\psi}_t(C_t).$$

We call $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots\}$ as the *coordination strategy*. In the *coordinated system*, dynamics and cost function are as same as those in the original problem in Section 6.2. In particular, the infintie-horizon discounted cost in the coordinated system is as follows:

$$J(\boldsymbol{\psi}) = \mathbb{E}^{\boldsymbol{\psi}} \left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(\mathbf{X}_t, \boldsymbol{\Gamma}_t^1(M_t^1), \dots, \Gamma_t^n(M_t^n)) \right].$$

Lemma 6.1 ([Nayyar et al., 2013], Proposition 3) *The original system described in Section 6.2 with PHS information structure is equivalent to the coordinated system.*

We denote $\boldsymbol{M}_t = (M_t^1, \dots, M_t^n)$ as the joint local information. According to [Nayyar et al., 2013], $\Pi_t = \mathbb{P}(X_t, \mathbf{M}_t | Z_{1:t-1}, \mathbf{\Gamma}_{1:t-1})$ is an information state for the coordinated system. It is shown in [Nayyar et al., 2013] that:

1. There exists a function ϕ such that

$$\Pi_{t+1} = \phi(\Pi_t, \mathbf{\Gamma}_t, Z_t). \tag{6.2}$$

2. The observation Z_t only depends on (Π_t, Γ_t) i.e.

$$\mathbb{P}(Z_t | \Pi_{1:t}, \mathbf{\Gamma}_{1:t}) = \mathbb{P}(Z_t | \Pi_t, \mathbf{\Gamma}_t). \tag{6.3}$$

3. There exists a function $\hat{\ell}$ such that

$$\hat{\ell}(\pi_t, \boldsymbol{\gamma}_t) = \mathbb{E}[\ell(X_t, \mathbf{U}_t | Z_{1:t-1} = z_{1:t-1}, \mathbf{\Gamma}_{1:t} = \boldsymbol{\gamma}_{1:t})]. \tag{6.4}$$

Assume that the initial state π_1 is fixed. Let $\mathcal{R}$ denote the reachable set of above centralized POMDP that contains all the realizations of π_t generated by $\pi_{t+1} = \phi(\pi_t, \boldsymbol{\gamma}, z), \forall \boldsymbol{\gamma} \in \mathcal{G}, \forall z \in \mathcal{Z}, \forall t \in \mathbb{N}$, with initial information state π_1. Note that since all the variables are finite valued, then $\mathcal{G}$ (set of all prescriptions $\boldsymbol{\gamma}$) and $\mathcal{Z}$ (set of all observations of the coordinator) are finite sets. Hence, $\mathcal{R}$ is at most a countable set.

Theorem 6.1 ([Nayyar et al., 2013], Theorem 5) *Let $\boldsymbol{\psi}^*(\pi)$ be any argmin of the right-hand side of following dynamic program. For $\pi \in \mathcal{R}$,*

$$V(\pi) = \min_{\boldsymbol{\gamma}} (\hat{\ell}(\pi, \boldsymbol{\gamma}) + \beta \mathbb{E} V(\phi(\pi, \boldsymbol{\gamma}, Z_t)) | \Pi_t = \pi, \mathbf{\Gamma}_t = \boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma} = (\gamma^1, \dots, \gamma^n)$ and the minimization is over all functions $\gamma^i \in \mathcal{G}^i, i \in \{1, \dots, n\}$. Then, the joint stationary strategy $\boldsymbol{g}^* = (\boldsymbol{g}^{1,*}, \dots, \boldsymbol{g}^{n,*})$ is optimal such that

$$g^{i,*}(\pi, m^i) := \boldsymbol{\psi}^{i,*}(\pi)(m^i), \quad \pi \in \mathcal{R}, m^i \in \mathcal{M}^i, \forall i.$$

In the next step, we develop a finite-state RL algorithm based on the finite-state approximate solution described in Chapter 5.

6.3.2 Step 2: Finite-state reinforcement learning algorithm for POMDP

In the previous step, we identified a centralized POMDP that is equivalent to the decentralized control system with PHS information structure. However, the obtained POMDP requires the complete knowledge of the model. To circumvent this requirement, we introduce a new concept that we call *Incrementally Expanding Representation* (IER). The main feature of IER is to remove the dependency of the POMDP from the complete knowledge of the model. Based on a proper IER, in this step, we develop a finite-state RL algorithm. This step consists of three parts. In part (1), we convert the POMDP to a countable-state MDP Δ without loss of optimality. In part (2), we construct a sequence of finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$ of MDP Δ. In part (3), we use a generic RL algorithm to learn an optimal strategy of Δ_N.

Definition 3 (Incrementally Expanding Representation) Let $\{\mathcal{S}_N\}_{N=1}^{\infty}$ be a sequence of finite sets such that $\mathcal{S}_1 \subsetneq \mathcal{S}_2 \subsetneq \ldots \subsetneq \mathcal{S}_N \subsetneq \ldots$, and $\mathcal{S}_1$ is a singleton, say $\mathcal{S}_1 = \{s^*\}$. Let $\mathcal{S} = \lim_{N\to\infty} \mathcal{S}_N$ be the countable union of above finite sets, $B : \mathcal{S} \to \mathcal{R}$ be a surjective function that maps $\mathcal{S}$ to the reachable set $\mathcal{R}$, and $\tilde{f} : \mathcal{S} \times \mathcal{G} \times \mathcal{Z} \to \mathcal{S}$. The tuple $\langle \{\mathcal{S}_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle$ is called an *Incrementally Expanding Representation* (IER), if it satisfies the following properties:

(P1) *Incremental Expansion*: For any $\boldsymbol{\gamma} \in \mathcal{G}, z \in \mathcal{Z}$, and $s \in \mathcal{S}_N$, we have that

$$\tilde{f}(s, \boldsymbol{\gamma}, z) \in \mathcal{S}_{N+1}. \tag{6.5}$$

(P2) *Consistency*: For any $(\boldsymbol{\gamma}_{1:t-1}, z_{1:t-1})$, let π_t and s_t be the states obtained by recursive application of (6.2) and (6.5) starting from π_1 and s^*, respectively. Then,

$$\pi_t = B(s_t). \tag{6.6}$$

In general, every decentralized control system with PHS information structure has at least one IER. In the following example, we present a generic IER that is valid for every system with PHS information structure.

Example 1: Let $S_1 = \{\emptyset\}$, $S_2 = \{\emptyset\} \cup \{\mathcal{G} \times \mathcal{Z}\}$, and $S_{t+1} = S_t \cup \{\mathcal{G} \times \mathcal{Z}\}^t, t \in \mathbb{N}$. Let $S = \lim_{t\to\infty} S_t$ and $B : \mathcal{S} \to \mathcal{R}$ such that

$$B(\emptyset) = \pi_1, B(s_{t+1}) = \phi(\phi(...., \boldsymbol{\gamma}_{t-1}, z_{t-1}), \boldsymbol{\gamma}_t, z_t) = \pi_{t+1},$$

where $s_{t+1} = ((\boldsymbol{\gamma}_1, z_1), \ldots, (\boldsymbol{\gamma}_t, z_t)) \in \mathcal{S}_{t+1}$. Define $\tilde{f}$ as follows:

$$\tilde{f}(s, \boldsymbol{\gamma}, z) = s \circ \boldsymbol{\gamma} \circ z,$$

where $\circ$ denotes concatenation. By construction, tuple $\langle \{\mathcal{S}_t\}_{t=1}^{\infty}, B, \tilde{f} \rangle$ satisfies (P1) and (P2), and hence is an IER.

Remark 6.3 The notion of the IER is valid for any system with the partial history sharing (that contains a large class of decentralized control systems). However, in general, this notion may not be valid for a system with non-partially history sharing information structure.

Countable-state MDP Δ

Let the tuple $\langle \{\mathcal{S}_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle$ be an IER of the POMDP obtained in the first step. Then, define MDP Δ with countable state space $\mathcal{S}$, finite action space $\mathcal{G}$, and dynamics $\tilde{f}$ such that:

(F1) The initial state is singleton s^*. The state $S_t \in \mathcal{S}_k$, $k \leq t$, evolves as follows: for $\boldsymbol{\Gamma}_t \in \mathcal{G}, Z_t \in \mathcal{Z}$,

$$S_{t+1} = \tilde{f}(S_t, \boldsymbol{\Gamma}_t, Z_t), \quad S_{t+1} \in \mathcal{S}_{k+1} \tag{6.7}$$

where observation Z_t only depends on $(S_t, \boldsymbol{\Gamma}_t)$ (that is a consequence of (6.3) and consistency property in (6.6)). At time t, there is a cost depending on the current state $S_t \in \mathcal{S}$ and action $\boldsymbol{\Gamma}_t \in \mathcal{G}$ given by

$$\tilde{\ell}(S_t, \boldsymbol{\Gamma}_t) := \hat{\ell}(B(S_t), \boldsymbol{\Gamma}_t) = \hat{\ell}(\Pi_t, \boldsymbol{\Gamma}_t). \tag{6.8}$$

(F2) State space $\mathcal{S}$, action space $\mathcal{G}$, and dynamics $\tilde{f}$ do not depend on the unknowns.

The performance of a stationary strategy $\tilde{\boldsymbol{\psi}} : \mathcal{S} \to \mathcal{G}$ is quantified by

$$\tilde{J}(\tilde{\boldsymbol{\psi}}) = \mathbb{E}^{\tilde{\psi}} \left[\sum_{t=1}^{\infty} \beta^{t-1} \tilde{\ell}(S_t, \boldsymbol{\Gamma}_t) \right]. \tag{6.9}$$

There may exist more than one IER that satisfy above features. For instance, the IER of Example 1 always satisfies (F1) and (F2) (that is model-free). This IER can also be used in the model-based cases; however, in the model-based cases, due to having partial knowledge of the model, one may be able to find a simpler IER. See Section 6.5 for an example.

Lemma 6.2 *Let $\tilde{\boldsymbol{\psi}}^*$ be an optimal strategy for MDP Δ. Construct a strategy $\boldsymbol{\psi}^*$ for the coordinated system as follows:*

$$\tilde{\boldsymbol{\psi}}^*(s) =: \boldsymbol{\psi}^*(B(s)), \quad \forall s \in \mathcal{S}.$$

Then, $\tilde{J}(\tilde{\boldsymbol{\psi}}^*) = J(\boldsymbol{\psi}^*)$ and $\boldsymbol{\psi}^*$ is an optimal strategy for the coordinated system, and therefore can be used to generate an optimal strategy for the decentralized control system.

Proof follows from Lemma 5.1 and Lemma 6.1.

Finite-state incrementally expanding MDP Δ_N

In this part, we construct a series of finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$, that approximate the countable-state MDP Δ as follows. Let Δ_N be a finite-state MDP with state space $\mathcal{S}_N$ and action space $\mathcal{G}$. The transition probability of Δ_N is constructed as follows. Pick any arbitrary set $D^* \in S_N$. Remap every transition in Δ that takes the state $s \in \mathcal{S}_N$ to $s' \in \mathcal{S}_{N+1} \backslash \mathcal{S}_N$ to a transition from $s \in \mathcal{S}_N$ to any (not necessarily unique) state in D^*. In addition, the per-step cost function of Δ_N is simply a restriction of $\tilde{\ell}$ to $\mathcal{S}_N \times \mathcal{G}$.

We assume that there exists an action or a sequence of actions that if taken, the system transmits to a known state d^* in D^*. For example, suppose there is a reset action in the system. After executing the reset action, the state of the system is reset and transmitted to a known state $d^* \in D^*$.

Let $\tau_N \in \mathbb{N}$ denote the horizon prior to which state S_t, $t \leq \tau_N$, always remains in $\mathcal{S}_N$ under dynamics $\tilde{f}$, optimal strategy $\tilde{\boldsymbol{\psi}}^*$, and any arbitrary sample path of $z_{1:t-1}$. In other words, S_t can not exist $\mathcal{S}_N$ if $t \leq \tau_N$, i.e.,

$$S_t = \tilde{f}(S_{t-1}, \tilde{\boldsymbol{\psi}}^*(S_{t-1}), Z_{t-1}) \in \mathcal{S}_N, \quad \forall t \leq \tau_N.$$

Let $\tilde{\boldsymbol{\psi}}_N^*$ and $\tilde{J}_N(\tilde{\boldsymbol{\psi}}_N^*)$ be an optimal stationary strategy of Δ_N and the optimal cost (performance) of Δ_N, respectively.

Theorem 6.2 *The difference in performance between Δ and Δ_N is bounded as follows:*

$$|\tilde{J}(\tilde{\boldsymbol{\psi}}^*) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}_N^*)| \leq \frac{2\beta^{\tau_N}}{1-\beta} \ell_{max}.$$

Proof follows from the proof of Theorem 5.1.

The upper-bound provided in Theorem 6.2 requires the knowledge on $(\tilde{f},\ \tilde{\boldsymbol{\psi}}^*,\mathcal{Z})$. However, according to (6.7), τ_N is always equal or greater than N i.e. $N \leq \tau_N$. Hence, one can obtain a more conservative error-bound (larger upper-bound) than the error-bound (upper-bound) in Theorem 6.2 that does not require any knowledge on $(\tilde{f},\ \tilde{\boldsymbol{\psi}}^*,\mathcal{Z})$ as follows.

Corollary 6.1 *The difference in performance between Δ and Δ_N is bounded as follows:*

$$|\tilde{J}(\tilde{\boldsymbol{\psi}}^*) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}_N^*)| \leq \frac{2\beta^N}{1-\beta}\ell_{max}.$$

Finite-state RL algorithm

Let $\mathcal{T}$ be a generic (model-based or model-free) RL algorithm designed for finite-state MDPs with infinite horizon discounted cost. By a generic RL algorithm, we mean any algorithm which fits to the following framework. At each iteration $k \in \mathbb{N}$, $\mathcal{T}$ knows the state of system, selects one action, and observes an instantaneous cost and the next state. The strategy learned (generated) by $\mathcal{T}$ converges to an optimal strategy as $k \to \infty$.

Let $\mathcal{T}$ operate on MDP Δ_N such that, at iteration k, it knows the state of the system $s_k \in \mathcal{S}_N$, selects one action $\boldsymbol{\gamma}_k \in \mathcal{G}$, and observes an instantaneous cost ℓ_k (which is a realization of the incurred cost $\ell(X_k, \mathbf{U}_k)$ at the original decentralized system). According to (6.4) and (6.8), we have

$$\mathbb{E}[\ell(X_k, \mathbf{U}_k)|S_{1:k}, \boldsymbol{\Gamma}_{1:k}] = \tilde{\ell}(S_k, \boldsymbol{\Gamma}_k), \quad S_k \in \mathcal{S}_N.$$

Hence, the instantaneous cost ℓ_k may be interpreted as a realization of the per-step cost of Δ_N. Given dynamics $\tilde{f}$, $\mathcal{T}$ observes $z_k \in \mathcal{Z}$ and computes the next state $s_{k+1} = \tilde{f}(s_k, \boldsymbol{\gamma}_k, z_k)$. If $s_{k+1} \in \mathcal{S}_{N+1}\backslash\mathcal{S}_N$, then an action (or a sequence of actions) that transmits the state of system to a known state in S_N i.e. $s_{k+1} = d^* \in D^*$ will be taken; otherwise, the system will continue from $s_{k+1} \in \mathcal{S}_N$.

Let $\tilde{\boldsymbol{\psi}}_N^k : \mathcal{S}_N \to \mathcal{G}$ be the learned strategy associated with RL algorithm $\mathcal{T}$ operating on MDP Δ_N at iteration k. Then, $\mathcal{T}$ updates its strategy $\tilde{\boldsymbol{\psi}}_N^{k+1}$ based on the observed cost ℓ_k and the transmitted next state s_{k+1} by executing action $\boldsymbol{\gamma}_k$ at state s_k. We assume $\mathcal{T}$ converges to an optimal strategy $\tilde{\boldsymbol{\psi}}_N^*$ as $k \to \infty$ such that

$$\lim_{k\to\infty} |\tilde{J}_N(\tilde{\boldsymbol{\psi}}_N^k) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}_N^*)| = 0. \tag{6.10}$$

Now, we need to convert (translate) the strategies in Δ_N to strategies in the original decentralized control problem described in Section 6.2, where the actual learning happens. Hence, we define a strategy $\boldsymbol{g}_N^k := (g_N^{k,i}, \ldots, g_N^{k,n})$, at iteration k, as follows:

$$g_N^{k,i}(s, m^i) := \tilde{\boldsymbol{\psi}}_N^{k,i}(s)(m^i), \forall s \in \mathcal{S}_N, \forall m^i \in \mathcal{M}^i, \forall i,$$

where $\tilde{\boldsymbol{\psi}}_N^{k,i}$ denotes the ith term of $\tilde{\boldsymbol{\psi}}_N^k$.

6.3.3 Main results

Theorem 6.3 *Let J^* be the optimal performance of the original decentralized control system given in (6.1). Then, the approximation error associated with using the learned strategy of*

Algorithm 2 Finite-State RL Algorithm

1: Given $\epsilon > 0$, choose a sufficiently large $N \in \mathbb{N}$ such that $\frac{2\beta^N}{1-\beta}\ell_{max} \leq \epsilon$. Then, construct state space $\mathcal{S}_N$, action space $\mathcal{G}$, and dynamics $\tilde{f}$. Initialize $s_1 = s^*$.

2: At iteration $k \in \mathbb{N}$, RL algorithm $\mathcal{T}$ picks $\boldsymbol{\gamma}_k = (\gamma_k^1, \ldots, \gamma_k^n) \in \mathcal{G}$ at state $s_k \in \mathcal{S}_N$. Then, agent $i \in \{1, \ldots, n\}$ takes action u_k^i according to the chosen prescription γ_k^i and local information $m_k^i \in \mathcal{M}^i$ as follows:

$$u_k^i = \gamma_k^i(m_k^i), \quad \forall i.$$

3: Based on the taken actions, the system incurs a cost ℓ_k, evolves, and generates new information i.e. $(\{m_{k+1}^i\}_{i=1}^n, z_k)$. Every agent i observes z_k because it is common observation. Based on z_k, all agents consistently compute the next state

$$s_{k+1} = \tilde{f}(s_k, \boldsymbol{\gamma}_k, z_k).$$

If $s_{k+1} \notin \mathcal{S}_N$, then agents take an action (or a sequence of actions) that transmits the state of system to a state $s_{k+1} = d^* \in \mathcal{S}_N$; otherwise, the system proceeds from $s_{k+1} \in \mathcal{S}_N$. Note that during the reset process, the algorithm is paused till the system lands in a state in S_N.

4: $\mathcal{T}$ updates its strategy from $\tilde{\boldsymbol{\psi}}_N^k$ to $\tilde{\boldsymbol{\psi}}_N^{k+1}$ based on performing action $\boldsymbol{\gamma}_k$ at state s_k and transmission to next state s_{k+1} with instantaneous cost ℓ_k.

5: $k \leftarrow k + 1$, and go to step 2 until termination.

Algorithm 2 is bounded as follows:

$$\lim_{k\to\infty} |J^* - J(\boldsymbol{g}_N^k)| = |\tilde{J}(\tilde{\boldsymbol{\psi}}^*) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}_N^*)| \leq \epsilon_N,$$

where $\epsilon_N = \frac{2\beta^{\tau_N}}{1-\beta}\ell_{max} \leq \frac{2\beta^N}{1-\beta}\ell_{max}$. *Note that the error goes to zero exponentially in* N.

The proof follows from Theorem 6.1, Lemma 6.2, Theorem 6.2, and Corollary 6.1.

To illustrate the asymptotic convergence of algorithm 2 to ϵ_N-optimal solution, consider a specific learning algorithm such as Q-learning. Then, we have:

Corollary 6.2 *Suppose every pair of state* $s \in \mathcal{S}_N$ *and prescription* $\boldsymbol{\gamma} \in \mathcal{G}$ *is visited infinitely often. Then, every* Q_k *in algorithm 3 converges to* Q^* *w.p.1. Let* $\tilde{\boldsymbol{\psi}}_N(s) \in \min_{\boldsymbol{\gamma}\in\mathcal{G}} Q^*(s, \boldsymbol{\gamma})$. *Then, if every agent* $i \in \{1, \ldots, n\}$ *uses strategy* $g_N^i(s, m^i) := \tilde{\psi}_N^i(s)(m^i)$, *where* $\tilde{\psi}_N^i$ *denotes the* i*th component of* $\tilde{\boldsymbol{\psi}}_N$, *the performance of the system is guaranteed to be* ϵ_N*-optimal.*

The proof follows directly from Theorem D.4 (i.e., [Tsistsiklis, 1994, Theorem 4]) and Theorem 6.3.

6.4 Decentralized Implementation

Prior to implementation, all agents are provided with state space S_N, action space $\mathcal{G}$, and dynamics $\tilde{f}$ as described in Section 6.3.2. Note that to obtain above knowledge, every agent must only know the information structure of system, action spaces $\{\mathcal{U}^i\}_{i=1}^n$, observation spaces $\{\mathcal{Y}^i\}_{i=1}^n$, discount factor β, upper-bound ℓ_{max} on per-step cost, and $\epsilon > 0$. In addition, agents may have partial knowledge of the model of system or may not.

When the system starts operation, agents observe the instantaneous cost of the system at each time step. It is assumed that agents have access to a common shared random number generator for the purpose of exploring the system consistently. Given state space $\mathcal{S}_N$, action space $\mathcal{G}$, and dynamics $\tilde{f}$, Algorithm 2 can be executed in a distributed manner because every agent can independently run Algorithm 2; agreeing upon a deterministic rule to break ties while using argmin ensures that all agents compute the same optimal strategy. Note that no more information needs to be shared; hence, *no communication* is required. According to Remark 6.1, Algorithm 2 also works for the pure decentralized control systems, when there is no information commonly shared between agents.

Suppose that the generic RL algorithm $\mathcal{T}$ in Section 6.3.2 is Q-learning. Then, in off-line learning, every agent is allowed to have different step sizes (independently from the step sizes of other agents). However, in on-line learning, since we also need to consistently exploit the system, the step sizes should be chosen consistently, e.g., based on the number of visit to pair of state and prescription, i.e., $(s, \boldsymbol{\gamma})$.

Algorithm 3 Decentralized Q-Learning Algorithm

1: Given $\epsilon > 0$, choose a sufficiently large $N \in \mathbb{N}$ such that $\frac{2\beta^N}{1-\beta}\ell_{max} \leq \epsilon$. Then, construct state space $\mathcal{S}_N$, action space $\mathcal{G}$, and dynamics $\tilde{f}$. Initialize $s_1 = s^*$, $Q(s, \boldsymbol{\gamma}) = 0, \alpha(s, \boldsymbol{\gamma}) = 1, \forall s \in \mathcal{S}_N, \forall \boldsymbol{\gamma} \in \mathcal{G}$.

2: At iteration $k \in \mathbb{N}$, agents uniformly pick a random prescription $\boldsymbol{\gamma}_k = (\gamma_k^1, \ldots, \gamma_k^n) \in \mathcal{G}$ at state $s_k \in \mathcal{S}_N$ by means of a common shared random number generator. Then, agent $i \in \{1, \ldots, n\}$ takes action u_k^i according to the chosen prescription γ_k^i and local information $m_k^i \in \mathcal{M}^i$ as follows:

$$u_k^i = \gamma_k^i(m_k^i), \quad \forall i.$$

3: Based on the taken actions, the system incurs a cost ℓ_k, evolves, and generates new information i.e. $(\{m_{k+1}^i\}_{i=1}^n, z_k)$. Every agent i observes z_k because it is common observation. Based on z_k, all agents consistently compute the next state

$$s_{k+1} = \tilde{f}(s_k, \boldsymbol{\gamma}_k, z_k).$$

If $s_{k+1} \notin \mathcal{S}_N$, then agents take an action (or a sequence of actions) that transmits the state of the system to a state $s_{k+1} = d^* \in \mathcal{S}_N$; otherwise, the system proceeds from $s_{k+1} \in \mathcal{S}_N$. Note that during the reset process, the algorithm is paused until the system lands in a state in S_N.

4: Agents update the corresponding Q-function associated with the pair $(s_k, \boldsymbol{\gamma}_k)$ as follows:

$$Q(s_k, \boldsymbol{\gamma}_k) \leftarrow (1 - \alpha(s_k, \boldsymbol{\gamma}_k))Q(s_k, \boldsymbol{\gamma}_k) + \alpha(s_k, \boldsymbol{\gamma}_k)\left(\ell_k + \beta \min_{v \in \mathcal{G}} Q(s_{k+1}, v)\right)$$

Also, the corresponding step-size are updated:

$$\frac{1}{\alpha(s_k, \boldsymbol{\gamma}_k)} \leftarrow \frac{1}{\alpha(s_k, \boldsymbol{\gamma}_k)} + 1.$$

5: $k \leftarrow k + 1$, and go to step 2 until termination.

6.5 Numerical Example: Multi access broadcast channel

In this section, we provide an example to illustrate our approach. In this example, we consider the setup of partial knowledge of the model.

System model

Consider a two-user multi-access broadcast system. At time t, $W_t^i \in \{0,1\}$ packets arrive at each user according to independent Bernoulli processes with $\mathbb{P}(W_t^i = 1) = p^i \in (0,1)$, $i = 1,2$. Each user may store only $X_t^i \in \{0,1\}$ packets in a buffer. If a packet arrives when the user-buffer is full, the packet is dropped. Both users may transmit $U_t^i \in \{0,1\}$ packets over a shared broadcast medium. A user can transmit only if it has a packet, thus $U_t^i \leq X_t^i$. If only one user transmits at a time, the transmission is successful and the transmitted packet is removed from the queue. If both users transmit simultaneously, packets "collide" and remain in the queue. Thus, the state update for users 1 and 2 is:

$$X_{t+1}^i = \min(X_t^i - U_t^i + U_t^1 U_t^2 + W_t^i, 1), \quad i = 1,2$$

Due to the broadcast nature of the communication channel, each user observes the transmission decision of the other user i.e. information at each user at time t is $(X_t^i, \mathbf{U}_{1:t-1})$, $i \in \{1,2\}$. Each user chooses a transmission decision as

$$U_t^i = g_t^i(X_t^i, \mathbf{U}_{1:t-1}), \quad i = 1,2,$$

where only actions $U_t^i \leq X_t^i$ are feasible. Similar to Section 6.2, we denote the *control strategy* by $\boldsymbol{g} = (\boldsymbol{g}^1, \boldsymbol{g}^2)$. The per unit cost $\ell(u_t^1, u_t^2)$ is defined to reflect the quality of transmission at time t as follows:

$$\ell(\mathbf{x}_t, \mathbf{u}_t) = \begin{cases} 0 & u_t^1 = 0, u_t^2 = 0 \\ \ell^1 \leq 0 & u_t^1 = 1, u_t^2 = 0 \\ \ell^2 \leq 0 & u_t^1 = 0, u_t^2 = 1 \\ \ell^3 & u_t^1 = 1, u_t^2 = 1 \end{cases}$$

where $|\ell^j| \leq \ell_{max}, j = 1,2,3$. The performance of strategy $\boldsymbol{g}$ is measured by

$$J(\boldsymbol{g}) = \mathbb{E}^{\boldsymbol{g}}\Big[\sum_{t=1}^{\infty} \beta^{t-1} \ell(\mathbf{X}_t, \mathbf{U}_t)\Big].$$

where $\beta \in (0,1)$. The case of symmetric arrivals ($p^1 = p^2$) was considered in [Hluchyj and Gallager, 1981, Ooi and Wornell, 1996]. In recent years, the above model has been used as a benchmark for decentralized stochastic control problems [Mahajan et al., 2008, Seuken and Zilberstein, 2007, Dibangoye et al., 2009]. We are interested in the following problem.

Problem 6.2 *Given any $\epsilon > 0$, without knowing the arrival probabilities p^1 and p^2, and cost functions ℓ^1, ℓ^2, ℓ^3, develop a decentralized Q-learning algorithm for both users such that users consistently learn an ϵ-optimal strategy $\boldsymbol{g}^*$.*

Decentralized Q-learning Algorithm

In this section, we follow the proposed two-step approach to develop a finite-state RL algorithm.

An Equivalent Centralized POMDP

In this step, we follow [Mahajan, 2010] and obtain the equivalent centralized POMDP for the completely known model as described in Section 6.3.1.

The common information shared between users is $C_t = \mathbf{U}_{1:t-1}$. Define $Z_t = C_{t+1} \backslash C_t = \mathbf{U}_t$. At time t, the coordinator observes $C_t = Z_{1:t-1}$ and prescribes $\gamma_t^i \colon X_t^i \to U_t^i$ that tell each agent how to use their local information to generate the control action. For this specific model, the prescription γ^i is completely specified by $A_t^i := \gamma_t^i(1)$ (since $\gamma_t^i(0)$ is always 0). Hence,

$$U_t^i = \gamma_t^i(X_t^i) = A_t^i \cdot X_t^i \tag{6.11}$$

Therefore, we may equivalently assume that the coordinator generates actions $\mathbf{A}_t = (A_t^1, A_t^2)$. The agents are passive and generate actions (U_t^1, U_t^2) according to (6.11). Hence, at time t, the coordinator prescribes action $\mathbf{A}_t \in \{0,1\}^2$ and observes $Z_t = \mathbf{U}_t \in \{0,1\}^2$.

Following [Mahajan, 2010], define $\mathbf{\Pi}_t = (\Pi_t^1, \Pi_t^2)$, $\Pi_t^i = \mathbb{P}(X_t^i = 1 \mid \mathbf{U}_{1:t-1}, \mathbf{A}_{1:t-1})$, as information state for the coordinated system with initial state $\mathbf{\Pi}_1 = (p^1, p^2)$. It is shown in [Mahajan, 2010]:

1. The information state $\mathbf{\Pi}_t$ evolves according to

$$\mathbf{\Pi}_{t+1} = \phi(\mathbf{\Pi}_t, \mathbf{A}_t, \mathbf{U}_t)$$

where

$$\phi(\mathbf{\Pi}_t, \mathbf{A}_t, \mathbf{U}_t) = \begin{cases} (T_1\Pi_t^1, T_2\Pi_t^2) & \mathbf{A}_t = (0,0) \\ (p^1, T_2\Pi_t^2) & \mathbf{A}_t = (1,0) \\ (T_1\Pi_t^1, p^2) & \mathbf{A}_t = (0,1) \\ (1,1) & \mathbf{A}_t = (1,1), \mathbf{U}_t = (1,1) \\ (p^1, p^2) & \mathbf{A}_t = (1,1), \mathbf{U}_t \neq (1,1) \end{cases}$$

where (p^1, p^2) are arrival rates and operator T_i is given by $T_i q = (1-p^i)(1-q), \quad i = 1, 2.$

2. The expected cost function is as follows:

$$\hat{\ell}(\mathbf{\Pi}_t, \mathbf{A}_t) = \begin{cases} 0, & \mathbf{A}_t = (0,0) \\ \ell^1 \Pi_t^1, & \mathbf{A}_t = (1,0) \\ \ell^2 \Pi_t^2, & \mathbf{A}_t = (0,1) \\ \ell^1 \Pi_t^1 + \ell^2 \Pi_t^2 + (\ell^3 - \ell^1 - \ell^2)\Pi_t^1 \Pi_t^2 & \mathbf{A}_t = (1,1) \end{cases}$$

The action $(0,0)$ that corresponds to not transmitting is dominated by the actions $(1,0)$ or $(0,1)$. Therefore, with no loss of optimality, action $(0,0)$ is removed. In the sequel, we denote $\mathcal{A} := \{(0,1), (1,0), (1,1)\}$ as the action space of the coordinator.

We denote $\mathcal{R}$ as the reachable set of above centralized POMDP that contains all the realizations of $\boldsymbol{\pi}_t$ generated by $\boldsymbol{\pi}_{t+1} = \phi(\boldsymbol{\pi}_t, \boldsymbol{a}, \boldsymbol{u}), \forall \boldsymbol{a} \in \mathcal{A}, \forall \mathbf{u} \in \{0,1\}^2, \forall t \in \mathbb{N}$, with initial information state $\boldsymbol{\pi}_1 = (p^1, p^2)$. Thus, the reachable set $\mathcal{R}$ is given by

$$\mathcal{R} \coloneqq \{(1,1), (1,p^1), (p^2,1), (p^1,p^2)\} \cup \{(p^1, T_2^n p^2) \,:\, n \in \mathbb{N}\} \cup \{(T_1^n p^1, p^2) \,:\, n \in \mathbb{N}\},$$

where $T_i^n q = T_i(T_i^{n-1} q)$. According to Theorem 6.1, we have

Theorem 6.4 *Let $\boldsymbol{\psi}^*(\boldsymbol{\pi})$ be any argmin of the right-hand side of the following dynamic program. For $\boldsymbol{\pi} \in \mathcal{R}$,*

$$V(\boldsymbol{\pi}) = \min_{\boldsymbol{a}} (\hat{\ell}(\boldsymbol{\pi}, \boldsymbol{a}) + \beta \mathbb{E}[V(\phi(\boldsymbol{\pi}, \boldsymbol{a}, \mathbf{U}_t)) | \mathbf{\Pi}_t = \boldsymbol{\pi}, \boldsymbol{A}_t = \boldsymbol{a}])$$

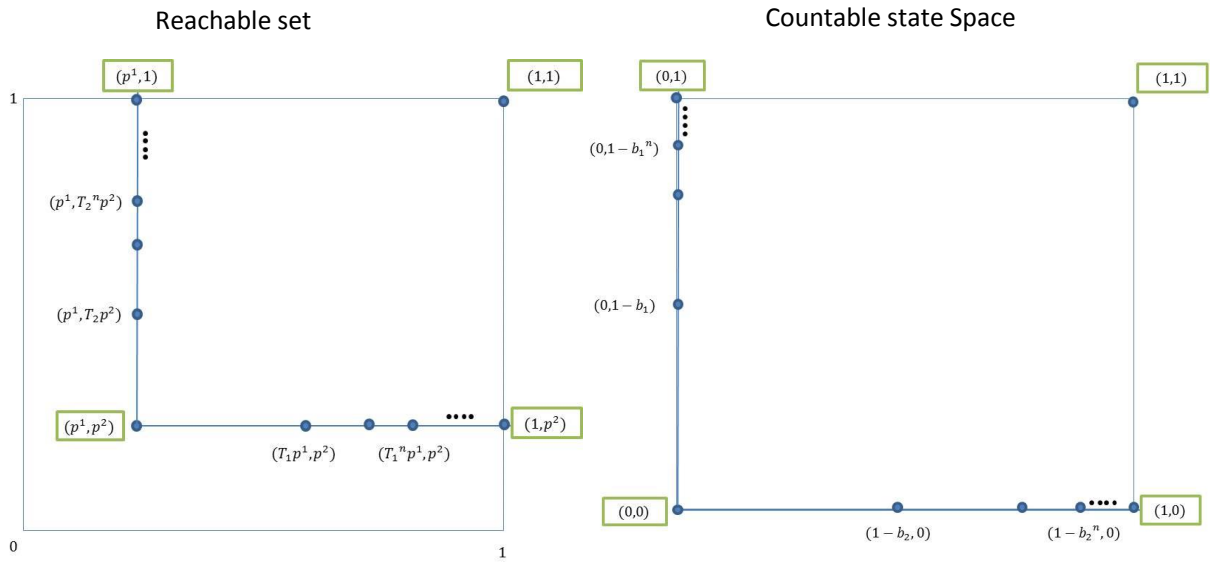

Fig. 6.1 It shows the reachable set $\mathcal{R}$ and the countable state space $\mathcal{S}$.

where $\boldsymbol{a} \in \mathcal{A}$. *The stationary strategy* $\boldsymbol{g}^* = (g^{1,*}, g^{2,*})$ *is optimal such that*

$$g^{i,*}(\boldsymbol{\pi}, x) = \boldsymbol{\psi}^{i,*}(\boldsymbol{\pi}) \cdot x, \quad \forall \boldsymbol{\pi} \in \mathcal{R}, x \in \{0,1\}, i = 1,2$$

where $\boldsymbol{\psi}^{i,*}$ *denotes ith term of* $\boldsymbol{\psi}^*$.

Q-learning algorithm for the POMDP

Let b_1, b_2 be any arbitrary number in $(0,1)$ and $B : \mathcal{R} \to \mathbb{Z}^{+2}$ be a bijective function that maps each state of $\mathcal{R}$ to a point in $\mathbb{Z}^{+2}$ as follows:

$$(0, 1-b_1^n) = B(p^1, T_2^n p^2), (1-b_2^n, 0) = B(T_1^n p^1, p^2), n \in \mathbb{N}$$

$$(0,1) = B(p^1, 1), (1,0) = B(1, p^2), (1,1) = B(1,1), (0,0) = B(p^1, p^2),$$

where $\lim_{n\to\infty} B(p^1, T_2^n p^2) = B(p^1, 1)$ and $\lim_{n\to\infty} B(T_1^n p^1, p^2) = B(1, p^2)$. Define a countable-state MDP Δ with state space $\mathcal{S}$, action space $\mathcal{A}$, dynamics $\tilde{f}$, and cost function $\tilde{\ell}$ as follows:

(F1) Let $\mathcal{S} = \{S_N\}_{N=1}^{\infty}$ be the state space, where $S_1 = \{(0,0)\}$ and

$$S_N = \{(0,0),(0,1),(1,0),(1,1),(0,1-b_1^i),(1-b_2^i,0)\}_{i=1}^{N-1}, N \geq 2.$$

The action space is $\mathcal{A} = \{(0,1),(1,0),(1,1)\}$. The initial state $S_1 = (0,0)$. The state $S_t \in \mathcal{S}_N$, $N \leq t$, evolves as follows: for $\mathbf{A}_t \in \mathcal{A}, \mathbf{U}_t \in \{0,1\}^2$,

$$S_{t+1} = \tilde{f}(S_t, \mathbf{A}_t, \mathbf{U}_t), \quad S_{t+1} \in \mathcal{S}_{N+1}.$$

For ease of exposition of dynamics $\tilde{f}$, we denote every state $S_t \in \mathcal{S}_N$ in a format of $(1 - b_2^{n_2}, 1 - b_1^{n_1})$, where n_1, n_2 take value in the set of $\{0,1,\ldots,\infty\}$. Thus,

$$\tilde{f}\left(\left(1-b_2^{(n_2)}, 1-b_1^{(n_1)}\right), \mathbf{A}_t, \mathbf{U}_t\right) = \begin{cases} (0, 1-b_1^{(n_1+1)}) & \mathbf{A}_t{=}(1,0) \\ (1-b_2^{(n_2+1)}, 0) & \mathbf{A}_t{=}(0,1) \\ (1,1) & \mathbf{A}_t{=}(1,1), \mathbf{U}_t{=}(1,1) \\ (0,0) & \mathbf{A}_t{=}(1,1), \mathbf{U}_t{\neq}(1,1), \end{cases}$$

At time t, there is a cost given by

$$\tilde{\ell}(S_t, \mathbf{A}_t) = \hat{\ell}(B^{-1}(S_t), \mathbf{A}_t).$$

It is trivial to see that the tuple $\langle \{\mathcal{S}_k\}_{k=1}^{\infty}, B^{-1}, \tilde{f} \rangle$ is an IER because of the fact that

$$\phi(\cdot, \mathbf{a}, \mathbf{u}) = B^{-1}\left(\tilde{f}\left(B(\cdot), \mathbf{a}, \mathbf{u}\right)\right), \quad \forall \mathbf{a} \in \mathcal{A}, \mathbf{u} \in \{0,1\}^2.$$

(F2) State space $\mathcal{S}$, action space $\mathcal{A}$, and dynamics $\tilde{f}$ do not depend on the unknowns i.e. $(p^1, p^2, \ell^1, \ell^2, \ell^3)$.

The performance of a stationary strategy $\tilde{\boldsymbol{\psi}} : \mathcal{S} \to \mathcal{A}$ is quantified by (6.9). According to Lemma 6.2, we can restrict attention in solving MDP Δ instead of the POMDP without loss of optimality. Let Δ_N be a finite-state MDP with state space $\mathcal{S}_N$ and action space $\mathcal{A}$. The initial state $S_1 = (0,0)$. At time t, state $S_t \in \mathcal{S}_N$ evolves as follows: for any

Algorithm 4 Decentralized Q-learning Algorithm

1: Given $\epsilon > 0$, choose a sufficiently large $N \in \mathbb{N}$ such that $\frac{2\beta^N}{1-\beta}\ell_{max} \leq \epsilon$. Then, construct state space $\mathcal{S}_N$, action space $\mathcal{A}$, and dynamics $\tilde{f}$. Let $s_1 = (0,0)$. Initialize Q-functions with zero and step-sizes α with one i.e $Q(s,\mathbf{a}) = 0, \alpha(s,\mathbf{a}) = 1, \forall s \in \mathcal{S}_N, \forall \mathbf{a} \in \mathcal{A}$.

2: At iteration $k \in \mathbb{N}$, users uniformly pick a random action $\mathbf{a}_k \in \mathcal{A}$ at state $s_k \in \mathcal{S}_N$ by means of a common shared random number generator. Then, user $i \in \{1,2\}$ takes action u_k^i according to the chosen a_k^i and local information $x_k^i \in \{0,1\}$ as follows:

$$u_k^i = a_k^i \cdot x_k^i. \quad i = 1,2.$$

3: Based on the taken actions, the system incurs a cost ℓ_k and generates $(x_{k+1}^1, x_{k+1}^2, \mathbf{u}_k = (u_k^1, u_k^2))$. Since $\mathbf{u}_k$ is observable to both users, they consistently compute the next state

$$s_{k+1} = \tilde{f}(s_k, \mathbf{a}_k, \mathbf{u}_k).$$

If $s_{k+1} \notin \mathcal{S}_N$, user 1 transmits first and then, user 2 transmits, and the state of system will be transmitted to $s_{k+1} = (1 - b_2, 0)$; otherwise, the system proceeds from $s_{k+1} \in \mathcal{S}_N$.

4: Users update the corresponding Q-function associated with the pair $(s_k, \mathbf{a}_k)$ as follows:

$$Q(s_k,\mathbf{a}_k) \leftarrow (1-\alpha(s_k,\mathbf{a}_k))Q(s_k,\mathbf{a}_k) + \alpha(s_k,\mathbf{a}_k)\left(\ell_k + \beta \min_{v\in\mathcal{A}} Q(s_{k+1},v)\right)$$

Also, the corresponding step-size are updated:

$$\frac{1}{\alpha(s_k,\mathbf{a}_k)} \leftarrow \frac{1}{\alpha(s_k,\mathbf{a}_k)} + 1.$$

5: $k \leftarrow k+1$, and go to step 2 until termination.

$\mathbf{A}_t \in \mathcal{A}, \mathbf{U}_t \in \{0,1\}^2$,

$$S_{t+1} = \begin{cases} \tilde{f}(S_t, \mathbf{A}_t, \mathbf{U}_t) & \tilde{f}(S_t, \mathbf{A}_t, \mathbf{U}_t) \in \mathcal{S}_N \\ (1-b_2, 0) & \tilde{f}(S_t, \mathbf{A}_t, \mathbf{U}_t) \in \mathcal{S}_{N+1} \backslash \mathcal{S}_N \quad \text{(reset action)} \end{cases} \tag{6.12}$$

In (6.12), whenever state s_t steps out of S_N, the users take a sequence of actions as follows: At first, user 1 transmits and user 2 does not transmit, then user 2 transmits and user 1 does not transmit. This sequence of actions takes the system to state $(1 - b_2, 0) \in S_N, N \geq 2$. Now, we use standard Q-learning algorithm as the generic RL algorithm $\mathcal{T}$ to learn the optimal strategy of Δ_N.

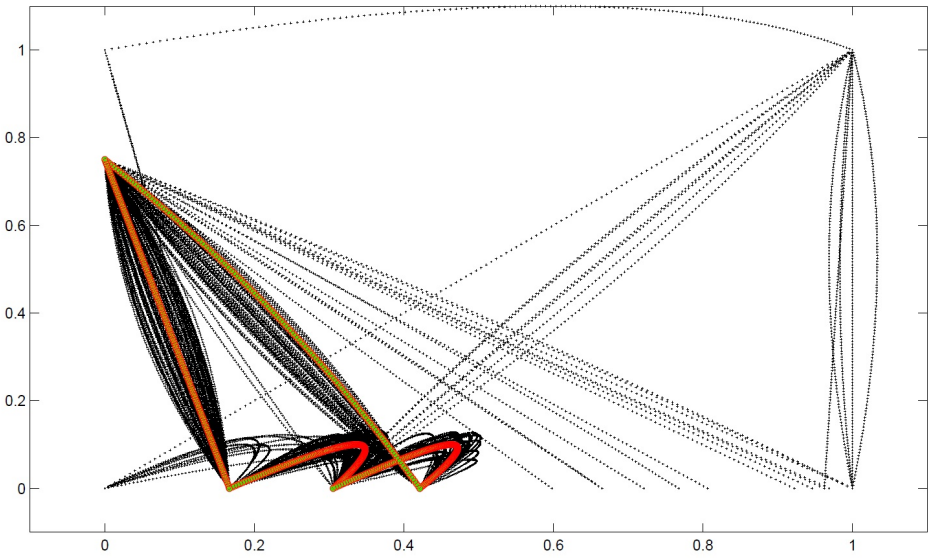

Fig. 6.2 This figure displays the learning procedure of optimal strategy in a few snapshots. It is seen that the state of the system is eventually trapped in the optimal recurrent class. The learning procedure is plotted in black and the optimal recurrent class is plotted in red. In this simulation, we use the following numerical values: $b_1 = 0.25, b_2 = 0.83, N = 20, \beta = 0.99, p^1 = 0.3, p^2 = 0.6, \ell^1 = \ell^2 = -1, \ell^3 = 0$.

According to [Tsistsiklis, 1994, Theorem 3], Q-functions in Algorithm 4 will converge to a Q^* with probability one[1]. Let Q^* be the resultant limit. Then, the optimal strategy $\tilde{\boldsymbol{\psi}}_N^*$ is as follows:

$$\tilde{\boldsymbol{\psi}}_N^* = \arg\min_{\mathbf{a}\in\mathcal{A}}(Q^*(\cdot, \mathbf{a})).$$

The strategy $\boldsymbol{g}_N^*$ is ϵ_N-optimal where

$$g_N^{i,*}(s)(x) := \tilde{\boldsymbol{\psi}}_N^{i,*}(s) \cdot x, \quad \forall s \in \mathcal{S}_N, x \in \{0,1\}, i = 1,2$$

where $\tilde{\boldsymbol{\psi}}_N^{i,*}$ denotes ith term of $\tilde{\boldsymbol{\psi}}_N^*$.

[1]In this example, every pair of (state,action) will be visited infinitely often by uniformly randomly picked actions.

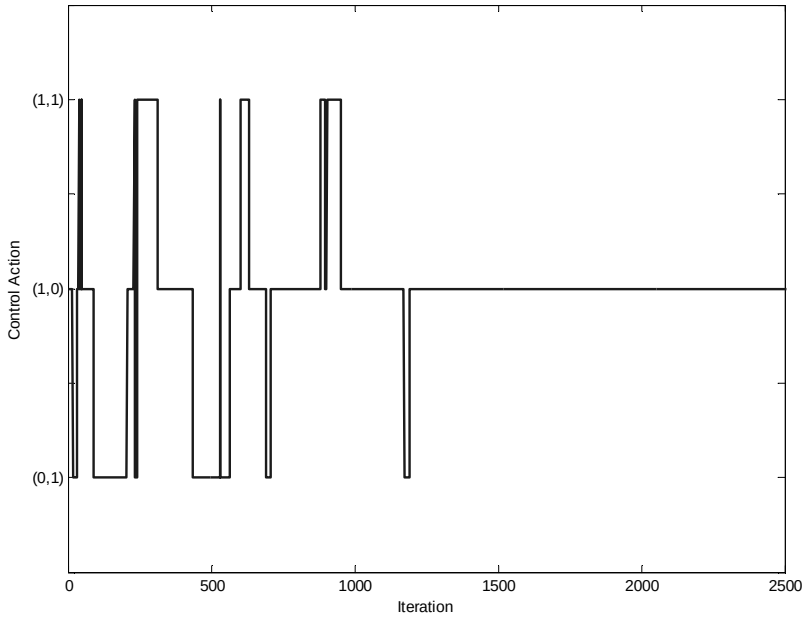

Fig. 6.3 The trajectory of the coordinator's strategy at state $(1-b_2^3, 0)$ which eventually converges to optimal action $(1, 0)$, i.e., user 1 transmits and user 2 does not.

Numerical Results

In this section, we provide a numerical simulation that shows the strategy learned by decentralized Q-learning Algorithm 4 converges to an optimal strategy when the arrival probabilities are $(p^1, p^2) = (0.3, 0.6)$ and the cost functions are $\ell^1 = \ell^2 = -1, \ell^3 = 0$.

Suppose users have no packets at the beginning. Users wait one time step to receive packets (i.e. user 1 receives a packet with $p^1 = 0.3$ probability and user 2 receives a packet with $p^2 = 0.6$ probability). At $t = 1$, action $(0, 1)$ is optimal i.e. the user 2 transmits and user 1 does not transmit. At $t \geq 2$, state s_t enters a recurrent class under the optimal strategy, and stays there forever. The recurrent class includes four states: $(0, 1-b_1^1)$, $(1-b_2^1, 0)$, $(1-b_2^2, 0)$, and $(1-b_2^3, 0)$. One immediate result is that for any $N \geq 4$, state S_t will never step out of $\mathcal{S}_N$ under the optimal strategy which implies $\tau_N = \infty$ and hence ϵ_N in Theorem 6.3 is zero (i.e. optimal strategy). For $t \geq 2$, the optimal strategy is a sequence of the following actions $(0, 1), (0, 1), (1, 0), (0, 1)$. Thus, it means that user 2 should transmit 3 times more than user 1 to minimize the number of collisions (maximize the number of successful transmission).

Figure 6.2 displays a few snapshots of state s_t governed by the strategy under the learning procedure where the learned strategy will eventually take state s_t to the optimal recurrent class.

6.6 Conclusion

In this chapter, we proposed a novel approach to develop a finite-state RL algorithm, for the partial history sharing information structure, that guarantees $\epsilon-$team-optimal solution. We presented our approach in two steps. In the first step, we used the common information approach to obtain an equivalent centralized POMDP of the decentralized control problem. However, the resultant POMDP can not be used directly because it requires the complete knowledge of the model while the agents only know the model incompletely. Thus, in the second step, we used the reinforcement learning algorithm of POMDP developed in Chapter 5. We illustrated our approach by developing a decentralized Q-learning algorithm for two-user Multi Access Broadcast Channel (MABC), a benchmark example for decentralized control systems. The numerical simulations verify that the learned strategy converges to an optimal strategy.

CHAPTER 7

Conclusion

In general, due to the limited resources, two practical dilemmas arise in teams as follows.

- **Computation (curse of dimensionality) dilemma:** When agents have access to the state of the system (i.e., centralized information), the optimization problem associated with finding the optimal control suffers from the curse of dimensionality. The dimension of exploration space increases exponentially in both time and the number of agents.
- **Communication (information) dilemma:** To deliver the state of the system to the agents, one needs to establish a network among the agents that collects and communicates the state. However, establishing such network may be physically or economically not feasible. For that reason, one may decide to deliver the state of the system incompletely (i.e. decentralized information). This makes the problem conceptually difficult because every agent may have different information (perspective) and it is difficult to establish cooperation among agents. Therefore, there is a trade-off between communicating the centralized information (the complete state) and the decentralized information (the incomplete state). The former may be expensive but conceptually easy while the latter may be cheap but conceptually difficult.

These two dilemmas are the inspiration of various fields of research such as distributed control and optimization, large scale optimization, consensus based algorithms, decentralized control, and etc.

7.1 Summary of main results

Summary of the main results of mean-field teams

In this thesis, we introduced and investigated systems, in which, the index (or the order in which agents are indexed) does not matter. Such systems emerge in many natural applications. For example, in social networks, it is not desirable that a user gets privileged or discriminated only because of its index. We called these systems mean-field teams. Below, we summarize their salient features in the view of above dilemmas.

Linear quadratic mean-field teams

- **Computational features:** In linear quadratic systems, the computational complexity of finding the optimal solution is shown to be independent of the number of agents in each sub-population. In particular, the solution is given by $K+1$ decoupled standard Riccati equations, where K is the number of sub-populations. In infinite horizon, the solution is given by $K+1$ decoupled algebraic Riccati equations. In addition, the optimal solution can be computed in a distributed manner, i.e., each agent needs to solve only two Riccati equations (rather than $K+1$): one corresponding to its own sub-population and one to the mean-field. If the matrices of the dynamics and cost do not depend on the size of population, then neither the optimal strategy. Hence, the agents do not even need to know the size of the population.

- **Communicational features:** There is no benefit of sharing any information other than the mean-field (average of local states) among agents. In particular, each agent has access to its local state (that no other agent knows) and the common mean-field (that every agent knows); hence, the solution can be implemented in a decentralized manner. In addition, the agents may use distributed algorithms such as consensus [Xiao and Boyd, 2004]—by locally communicating with their neighbors—to share the mean-field. Furthermore, in large systems, the effect of one agent on the mean-field is small; hence, the network is robust to the agent failure. Plus, the mean-field becomes predictable in asymptotically large systems; thus, it can be shared periodically with a larger sample-time than the control sample-time.

Controlled Markov chain mean-field teams

- **Computational features:** In controlled Markov chain systems, the computational complexity of finding the optimal solution is shown to be polynomial with the number of agents in each sub-population (rather than exponential). In particular, the optimal solution is given by a dynamic program. In infinite horizon, the solution is given by the fixed point theorem of the dynamic program. In addition, the dynamic program extends to the case in which the agents are arbitrary coupled in the cost. If the dynamics and cost do not depend on the size of population, then neither the optimal strategy [1]. Hence, agents do not need to know the size of the population.

- **Communicational features:** Each agent has access to its local state (that no other agent knows) and common mean-field (that every agent knows). Hence, the solution can be implemented in a decentralized manner. In addition, the agents may use distributed algorithms such as consensus [Olfati-Saber et al., 2006, Bishop and Doucet, 2014]—by locally communicating with their neighbors—to share the mean-field. Furthermore, in large systems, the effect of one agent on the mean-field is small; hence, the network is robust to the agent failure. Plus, the mean-field becomes predictable in asymptotically large systems; thus, it can be shared periodically with a larger sample-time than the control sample-time.

Remark 7.1 In the linear quadratic mean-field teams, the optimal solution under mean-field sharing information structure is the optimal solution under centralized information structure (where the agents know the complete joint state). In contrary, in the controlled Markov chain mean-field teams, the optimal solution under the mean-field sharing information structure may not be the optimal solution under the centralized information structure. However, the computational complexity of finding the optimal mean-field solution is polynomial with respect to the number of agents whereas that of the centralized solution is exponential.

Summary of the main results of decentralized reinforcement learning algorithm

In practice, the agents may not know the complete system model. We developed a decentralized reinforcement learning algorithm that guarantees $\varepsilon-$optimal performance. Hence, the agents can learn their optimal actions by interacting with their environment. As an

[1] To solve the dynamic program the agents need to know the state space of the mean-field.

intermediate step of this development, we revisited the well-known POMDP problem and proposed a novel approach that utilizes MDP solvers to learn $\varepsilon-$optimal solution when the model in known, partially known, or not known.

7.2 Future directions

Connection to game theory

It is worth highlighting that the mean-field teams are inspired by their counterpart in game theory, i.e., mean-field games [Huang et al., 2003, Lasry and Lions, 2006a, Lasry and Lions, 2006b, Huang et al., 2007, Li and Zhang, 2008, Caines, 2013, Gomes and Saude, 2014, Moon and Basar, 2017, Bensoussan et al., 2016, Salhab et al., 2015]. The key difference between mean-field teams and mean-field games is the key difference between the team theory and game theory, i.e., the solution concept: global optimal strategy versus Nash strategy. The detailed comparison between the mean-field teams and mean-field games are presented in Chapter 3 and Chapter 4. In spite of the differences, there is a close connection between the team theory and game theory. As shown in Section 3.1, the solution and associated proof techniques of the mean-field teams may be useful for the mean-field games and vice versa. As an interesting example of this connection, one could investigate the relationship between the Riccati equations of this thesis and the mixed Riccati/ordinary differential equations of [Huang et al., 2012].

Connection to Markov chains

As shown in Section 4.6, the results of Chapter 4 can be used in Markov chain systems when the transition probability is partially exchangeable. To the best of the author's knowledge, most of the results presented in Section 4.6 do not exist in the literature.

Integration of mean-field teams with distributed algorithms

As mentioned earlier, the agents may share the mean-field using distributed algorithms such as consensus by locally communicating with their neighbors. Hence, one possible future direction is to integrate the consensus algorithms with mean-field teams. In addition, the idea of micro-macro design presented in Section 3.8 may be useful for the purpose of distributed optimization.

New mean-field models

The notion of partial exchangeability is a general notion and applicable for any model. In this thesis, we only considered linear quadratic and controlled Markov chain systems. Therefore, it is possible to use this notion and define new classes of mean-field models.

Decentralized reinforcement learning for specific models

The proposed decentralized RL algorithm works under a large class of decentralized system, i.e., the partial history sharing information structure. In principle, this algorithm can be used for various specific models including mean-field teams to learn the team-optimal strategy when the complete model is not available.

Various approximation methods

In this thesis, we mostly focus on the quality of the performance, i.e., optimal or $\varepsilon-$optimal solutions. On one hand, to ensure the optimality, one must explicitly compute the terms associated with the optimal (or $\varepsilon-$optimal) solution that may require a large capacity of computation and/or communication resources. On the other hand, one can approximate these terms and obtain a sub-optimal (yet practical) solution with much less capacity. Therefore, one possible future direction is to study different approximation approaches. In Section 3.9 and Section 4.9, we mention a few of such approximations.

Different applications

As mentioned in Chapter 1, the team theory emerges in many various applications including modern power systems, social networks, robotics, network controlled systems, sensor networks, and economics. Hence, one possible future direction is to use the results of this thesis in different applications.

Appendices

APPENDIX A

Linear quadratic mean field team

A.1 Proof of Theorem 3.5

As in the proof of Theorem 3.1 described in Section 3.4, define $\breve{x}_t^i = x_t^i - \bar{x}_t^k$, $\breve{u}_t^i = u_t^i - \bar{u}_t^k$, $\mathring{\mathbf{x}}_t = \mathrm{vec}((\breve{x}_t^i)_{i\in\mathcal{N}}, \bar{\mathbf{x}}_t)$, and $\mathring{\mathbf{u}}_t = \mathrm{vec}((\breve{u}_t^i)_{i\in\mathcal{N}}, \bar{\mathbf{u}}_t)$. We identify a cost function $\mathring{c}_t(\mathring{\mathbf{x}}_t, \mathring{\mathbf{u}}_t)$ as in Corollary 3.3.

Lemma A.1 *For time t, $t \in \{1, \ldots, T\}$, there exists function $\mathring{c}_t$, such that for $t \in \{1, \ldots, T-1\}$, $c_t(\mathbf{x}_t, \mathbf{u}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \mathring{c}_t(\mathring{\mathbf{x}}_t, \mathring{\mathbf{u}}_t)$ and for $t = T$, $c_T(\mathbf{x}_T) = \mathring{c}_T(\mathring{\mathbf{x}}_T)$. In particular, for $t \in \{1, \ldots, T-1\}$,*

$$\mathring{c}_t(\mathring{\mathbf{x}}_t, \mathring{\mathbf{u}}_t) = \bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) + \sum_{i\in\mathcal{N}^k, k\in\mathcal{K}} \breve{c}_t^k(\breve{x}_t^i, \breve{u}_t^i) - \sum_{k\in\mathcal{K}} (\bar{r}_t^k)^{\mathsf{T}} Q_t^k \bar{r}_t^k,$$

and for $t = T$,

$$\mathring{c}_T(\mathring{\mathbf{x}}_T) = \bar{c}_T(\bar{\mathbf{x}}_T) + \sum_{i\in\mathcal{N}^k, k\in\mathcal{K}} \breve{c}_T^k(\breve{x}_T^i) - \sum_{k\in\mathcal{K}} (\bar{r}_T^k)^{\mathsf{T}} Q_T^k \bar{r}_T^k.$$

To describe $\bar{c}_t(\cdot)$, *define* $\mathbf{y}_t \coloneqq \left[\begin{smallmatrix} \bar{\mathbf{x}}_t - \bar{\mathbf{r}}_t \\ \bar{\mathbf{x}}_t - \mathbf{s}_t \end{smallmatrix}\right]$. *Then,*

$$\bar{c}_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = \mathbf{y}_t^\intercal \begin{bmatrix} \bar{Q}_t & 0 \\ 0 & P_t^x \end{bmatrix} \mathbf{y}_t + \bar{\mathbf{u}}_t^\intercal (\bar{R}_t + P_t^u) \bar{\mathbf{u}}_t,$$

$$\bar{c}_T(\bar{\mathbf{x}}_T) = \mathbf{y}_T^\intercal \begin{bmatrix} \bar{Q}_T & 0 \\ 0 & P_T^x \end{bmatrix} \mathbf{y}_T.$$

Moreover,

$$\breve{c}_t^k(\breve{x}_t^i, \breve{u}_t^i) = \frac{1}{|\mathcal{N}^k|} \left[(\breve{x}_t^i - r_t^i)^\intercal Q_t^k (\breve{x}_t^i - r_t^i) + (\breve{u}_t^i)^\intercal R_t^k \breve{u}_t^i \right],$$
$$\breve{c}_T^k(\breve{x}_T^i) = \frac{1}{|\mathcal{N}^k|} \left[(\breve{x}_T^i - r_T^i)^\intercal Q_T^k (\breve{x}_T^i - r_T^i) \right].$$

Note that the per-step cost is decomposed into terms that depend only on $(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)$ and terms that depend only on $(\breve{x}_t^i, \breve{u}_t^i)$ (and terms that do not depend on the control strategy). The rest of the proof follows along the same lines of the proof of Theorem 3.1. In particular, we consider a deterministic dynamical system and split it into $K+1$ classes. The agents in class $k, k \in \mathcal{K}$, are solving a tracking problem whose solution is given by

$$\breve{u}_t^i = \breve{L}_t^k \breve{x}_t^i + \breve{F}_t^k v_t^i.$$

The mean-field component is also solving a tracking problem whose solution is given by

$$\bar{\mathbf{u}}_t = \bar{L}_t \bar{\mathbf{x}}_t + \bar{F}_t \bar{v}_t.$$

The result of the Theorem follows from combining the above equations. Therefore, from standard results in LQR tracking problem, the optimal control law of agent $i \in \mathcal{N}^k$ of sub-population $k \in \mathcal{K}$ is given by

$$u_t^i = \breve{u}_t^i + \bar{u}_t^k = \left[\breve{L}_t^k (x_t^i - \bar{x}_t^k) + \breve{F}_t^k v_t^i \right] + \left[\bar{L}_t^k \bar{\mathbf{x}}_t + \bar{F}_t^k \bar{v}_t \right],$$

where gains $\{\breve{L}_t^k, \bar{L}_t^k, \breve{F}_t^k, \bar{F}_t^k\}_{t=1}^{T-1}$ are identical for all agents of sub-population k, $\bar{v}_t$ is identical for all agents of all sub-populations, and v_t^i may be different for each agent.

A.2 Proof of Theorem 3.6

The proof follows the same lines as the proof of Theorem 3.1 with the following differences. The mean-field is defined as $\bar{x}_t^{k,\lambda} = \frac{1}{|\mathcal{N}^k|}\sum_{i\in\mathcal{N}^k}\lambda^i x_t^i$ (similar interpretations hold for $\bar{u}_t^{k,\lambda}$ and $\bar{w}_t^{k,\lambda}$) and the breve variables are defined as $\breve{x}_t^i = x_t^i - \frac{\lambda^i}{b^i}\bar{x}_t^{k,\lambda}$ (similar interpretations hold for $\breve{u}_t^i$ and $\breve{w}_t^i$). Note that due to (A.3.6), the dynamics of $\breve{x}_t^i$ and $\bar{\mathbf{x}}_t^\lambda$ are still given by (3.17) and (3.18), respectively, where $\bar{A}_t$ and $\bar{B}_t$ are defined as in Theorem 3.6. The equivalent of Lemma 3.1 is the following:

Lemma A.2 *Let $(\lambda^1,\dots,\lambda^N)\in\mathbb{R}^N$ and $(b^1,\dots,b^N)\in\mathbb{R}^N_{>0}$. In addition, for any $\mathbf{x} = \mathrm{vec}(x^1,\dots,x^N)$ and $\bar{x}^\lambda = \langle(\lambda^i x^i)_{i=1}^N\rangle$, let $\breve{x}^i = x^i - \frac{\lambda^i}{b^i}\bar{x}^\lambda$, $i\in\{1,\dots,N\}$. Then, for any matrix Q of appropriate dimension,*

$$\frac{1}{N}\sum_{i=1}^N b^i (x^i)^\intercal Q x^i = \frac{1}{N}\sum_{i=1}^N b^i(\breve{x}^i)^\intercal Q\breve{x}^i + (\bar{x}^\lambda)^\intercal \mu Q\bar{x}^\lambda,$$

where $\mu \coloneqq 2 - \frac{1}{N}\sum_{i=1}^N \frac{(\lambda^i)^2}{b^i}$.

Consequently, the equivalent of Corollary 3.3 is the following

Corollary A.1 *For time t, $t\in\{1,\dots,T\}$, there exists function $\mathring{c}_t$, such that for $t\in\{1,\dots,T-1\}$, $c_t(\mathbf{x}_t,\mathbf{u}_t,\bar{\mathbf{x}}_t^\lambda,\bar{\mathbf{u}}_t^\lambda) = \mathring{c}_t(\mathring{\mathbf{x}}_t,\mathring{\mathbf{u}}_t)$ and for $t=T$, $c_T(\mathbf{x}_T,\bar{\mathbf{x}}_T^\lambda) = \mathring{c}_T(\mathring{\mathbf{x}}_T)$. In particular, for $t\in\{1,\dots,T-1\}$,*

$$\mathring{c}_t(\mathring{\mathbf{x}}_t,\mathring{\mathbf{u}}_t) = \bar{c}_t(\bar{\mathbf{x}}_t^\lambda,\bar{\mathbf{u}}_t^\lambda) + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}^k}\breve{c}_t^i(\breve{x}_t^i,\breve{u}_t^i),$$

where

$$\begin{aligned}\bar{c}_t(\bar{\mathbf{x}}_t^\lambda,\bar{\mathbf{u}}_t^\lambda) &= (\bar{\mathbf{x}}_t^\lambda)^\intercal(\bar{Q}_t + P_t^x)\bar{\mathbf{x}}_t^\lambda + (\bar{\mathbf{u}}_t^\lambda)^\intercal(\bar{R}_t + P_t^u)\bar{\mathbf{u}}_t^\lambda,\\ \breve{c}_t^i(\breve{x}_t^i,\breve{u}_t^i) &= \frac{b^i}{|\mathcal{N}^k|}\left[(\breve{x}_t^i)^\intercal Q_t^k\breve{x}_t^i + (\breve{u}_t^i)^\intercal R_t^k\breve{u}_t^i\right],\end{aligned}$$

and for $t=T$,

$$\mathring{c}_T(\mathring{\mathbf{x}}_T) = \bar{c}_T(\bar{\mathbf{x}}_T^\lambda) + \sum_{i\in\mathcal{N}^k,k\in\mathcal{K}}\breve{c}_T^i(\breve{x}_T^i),$$

where

$$\bar{c}_T(\bar{\mathbf{x}}_T^\lambda) = (\bar{\mathbf{x}}_T^\lambda)^\mathsf{T}(\bar{Q}_T + P_T^x)\bar{\mathbf{x}}_T^\lambda,$$
$$\breve{c}_T^i(\breve{x}_T^i) = \frac{b^i}{|\mathcal{N}^k|}\left[(\breve{x}_T^i)^\mathsf{T} Q_T^k \breve{x}_T^i\right],$$

where $\bar{Q}_t$ and $\bar{R}_t$ are defined as in Theorem 3.6.

The rest of the proof is the same as in Section 3.4. We can show that the optimal control strategy of auxiliary model is given by

$$\bar{\mathbf{u}}_t^\lambda = \bar{L}_t \bar{\mathbf{x}}_t^\lambda \quad \text{and for } k \in \mathcal{K}, i \in \mathcal{N}^k, \quad \breve{u}_t^i = \breve{L}_t^k \breve{x}_t^i,$$

where the gains $\{\breve{L}_t^k, \bar{L}_t\}_{t=1}^{T-1}$ are given as in Theorem 3.6. To complete the proof of Theorem 3.6, note that

$$u_t^i = \breve{u}_t^i + \frac{\lambda^i}{b^i}\bar{u}_t^{k,\lambda} = \breve{L}_t^k\left(x_t^i - \frac{\lambda^i}{b^i}\bar{x}_t^{k,\lambda}\right) + \frac{\lambda^i}{b^i}\bar{L}_t^k \bar{\mathbf{x}}_t^\lambda.$$

Thus, the control laws specified in Theorem 3.6 are the optimal *centralized* control laws, and, a fortiori, the optimal decentralized control laws.

APPENDIX B

Controlled Markov chain mean field team

B.1 Proof of Lemma 4.6

We use induction to prove the result. For notational convenience, we denote $\mathbb{P}(A = a|B = b, C = c)$ by $\mathbb{P}(a|b, c)$. Define $H(m) := \{\mathbf{x} \in \mathcal{X}^n : \frac{1}{n}\sum_{i=1}^n \delta_{x^i} = m\}$ as a set of all joint states $\mathbf{x} \in \mathcal{X}^n$ whose empirical distribution is m. Thus, at time t, we have

$$\mathbb{1}(m_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_t^i}) = \mathbb{1}(\mathbf{x}_t \in H(m_t)). \tag{B.1}$$

Notice that if $\mathbf{x}_t \in H(m_t)$, then one can interpret $H(m_t)$ as a collection of all permutations of $\mathbf{x}_t$ (such interpretation is critical for our proof).

In the first step, $t = 1$, we have

$$\mathbb{P}(\mathbf{x}_1|m_1, \gamma_1) \overset{(a)}{=} \mathbb{P}(\mathbf{x}_1|m_1) \overset{(b)}{=} \frac{\mathbb{P}(m_1|\mathbf{x}_1)\mathbb{P}(\mathbf{x}_1)}{\mathbb{P}(m_1)} = \frac{\mathbb{1}(m_1 = \frac{1}{n}\sum_{i=1}^n \delta_{x_1^i})\mathbb{P}(\mathbf{x}_1)}{\mathbb{P}(m_1)} \tag{B.2}$$

where (a) follows from the fact that $\gamma_1 = \psi_1(m_1)$ according to (4.8) and (b) follows from Bayes rule. From (B.1) and (B.2), given $\{m_1, \gamma_1\}$, we get

$$\mathbb{P}(\mathbf{x}_1|m_1, \gamma_1) = \begin{cases} 0 & \mathbf{x}_1 \notin H(m_1) \\ \alpha(m_1) & \mathbf{x}_1 \in H(m_1) \end{cases} \tag{B.3}$$

where $\alpha(m_1) = \frac{\mathbb{P}(\mathbf{x}_1)}{\mathbb{P}(m_1)}$ depends only on m_1. The reason lies in the fact that, when $\mathbf{x}_1 \in H(m_1)$, $H(m_1)$ contains nothing but permutations of $\mathbf{x}_1$ while joint probability distribution of initial states $\mathbb{P}(\mathbf{x}_1) = \prod_{i=1}^{n} P_X(x_1^i)$ is insensitive to permutation of $\mathbf{x}_1$.

Since the summation of $\mathbb{P}(\mathbf{x}_1|m_1, \gamma_1)$ over $\mathbf{x}_1 \in \mathcal{X}^n$ is one, we have

$$\alpha(m_1) = \frac{1}{|H(m_1)|}. \tag{B.4}$$

From (B.3) and (B.4), we have

$$\mathbb{P}(\mathbf{x}_1|m_1, \gamma_1) = \mathbb{P}(\mathbf{x}_1|m_1) = \frac{\mathbb{1}(\mathbf{x}_1 \in H(m_1))}{|H(m_1)|}.$$

Hence, the result holds for $t = 1$. Assume the result holds for step t i.e.

$$\mathbb{P}(\mathbf{x}_t|m_{1:t}, \gamma_{1:t}) = \mathbb{P}(\mathbf{x}_t|m_t) = \frac{\mathbb{1}(\mathbf{x}_t \in H(m_t))}{|H(m_t)|}. \tag{B.5}$$

We prove that the result holds for step $t+1$ as follows.

$$\begin{aligned}\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t+1}, \gamma_{1:t+1}) &\overset{(a)}{=} \mathbb{P}(\mathbf{x}_{t+1}|m_{1:t+1}, \gamma_{1:t}) \overset{(b)}{=} \frac{\mathbb{P}(m_{t+1}|\mathbf{x}_{t+1})\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t}, \gamma_{1:t})}{\mathbb{P}(m_{t+1}|m_{1:t}, \gamma_{1:t})} \\ &= \frac{\mathbb{1}(m_{t+1} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{t+1}^i})\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t}, \gamma_{1:t})}{\mathbb{P}(m_{t+1}|m_{1:t}, \gamma_{1:t})},\end{aligned} \tag{B.6}$$

where (a) follows from the fact that $\gamma_{t+1} = \psi_{t+1}(m_{1:t+1})$ according to (4.8) and (b) follows from Bayes rule. Similar to step $t = 1$, we show that, given $\{m_{1:t+1}, \gamma_{1:t+1}\}$, above conditional probability is insensitive to permutation of $\mathbf{x}_{t+1}$. For that matter, we write the conditional probability in the numerator of (B.6) as follows.

$$\begin{aligned}\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t}, \gamma_{1:t}) &= \sum_{\mathbf{x}_t} \mathbb{P}(\mathbf{x}_{t+1}, \mathbf{x}_t|m_{1:t}, \gamma_{1:t}) = \sum_{\mathbf{x}_t} \mathbb{P}(\mathbf{x}_{t+1}|\mathbf{x}_t, m_{1:t}, \gamma_{1:t})\mathbb{P}(\mathbf{x}_t|m_{1:t}, \gamma_{1:t}) \\ &\overset{(a)}{=} \sum_{\mathbf{x}_t, \mathbf{w}_t} \left[\prod_{i=1}^{n} \mathbb{1}(x_{t+1}^i = f_t(x_t^i, \gamma_t(x_t^i), w_t^i, m_t))\right] \cdot \mathbb{P}(\mathbf{w}_t) \cdot \mathbb{P}(\mathbf{x}_t|m_{1:t}, \gamma_{1:t}),\end{aligned} \tag{B.7}$$

where (a) follows from (4.2) and the fact that $\mathbf{W}_t$ is independent from all data and decisions made before time t. Let $S := \sigma(1, \ldots, n)$ denote an arbitrary permutation of set $\{1, \ldots, n\}$ and $S(i)$ denote the ith term of vector S. We use superscript S to denote the permuted

version of variables. For example, we denote $\mathbf{x}_{t+1}^S = (x_{t+1}^{S(1)}, \ldots, x_{t+1}^{S(n)})$ as permuted version of $\mathbf{x}_{t+1}$ with respect to S. Now, consider

$$\begin{aligned}\mathbb{P}(\mathbf{x}_{t+1}^S|m_{1:t},\gamma_{1:t}) =& \sum_{\mathbf{x}_t,\mathbf{w}_t}\left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^{S(i)} = f_t(x_t^i,\gamma_t(x_t^i),w_t^i,m_t))\right]\cdot\mathbb{P}(\mathbf{w}_t)\cdot\mathbb{P}(\mathbf{x}_t|m_{1:t},\gamma_{1:t})\\ \overset{(a)}{=}& \sum_{\mathbf{x}_t,\mathbf{w}_t}\left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^{S(i)} = f_t(x_t^{S(i)},\gamma_t(x_t^{S(i)}),w_t^{S(i)},m_t))\right]\cdot\mathbb{P}(\mathbf{w}_t^S)\cdot\mathbb{P}(\mathbf{x}_t^S|m_{1:t},\gamma_{1:t}),\end{aligned} \tag{B.8}$$

where (a) follows from the fact that summation is insensitive to permutation. In particular, if $D(\mathbf{x},\mathbf{w})$ is any arbitrary function of $(\mathbf{x},\mathbf{w})$, then we have

$$\sum_{\mathbf{x},\mathbf{w}} D(\mathbf{x},\mathbf{w}) = \sum_{\mathbf{x}^S,\mathbf{w}^S} D(\mathbf{x}^S,\mathbf{w}^S) = \sum_{\mathbf{x},\mathbf{w}} D(\mathbf{x}^S,\mathbf{w}^S).$$

Now, we consider terms in (B.8) separately as follows.

A) Since multiplication is insensitive to permutation, the first term may be written as follows.

$$\prod_{i=1}^n \mathbb{1}(x_{t+1}^{S(i)}=f_t(x_t^{S(i)},\gamma_t(x_t^{S(i)}),w_t^{S(i)},m_t)) = \prod_{i=1}^n \mathbb{1}(x_{t+1}^{i}=f_t(x_t^{i},\gamma_t(x_t^{i}),w_t^{i},m_t))$$

B) The second term may be written as follows.

$$\mathbb{P}(\mathbf{w}_t^S) = \prod_{i=1}^n P_{W_t}(w_t^{S(i)}) = \prod_{i=1}^n P_{W_t}(w_t^i) = \mathbb{P}(\mathbf{w}_t).$$

C) According to (B.5), the third term may be written as follows.

$$\mathbb{P}(\mathbf{x}_t^S|m_{1:t},\gamma_{1:t}) = \frac{\mathbb{1}(\mathbf{x}_t^S\in H(m_t))}{|H(m_t)|} = \frac{\mathbb{1}(\mathbf{x}_t\in H(m_t))}{|H(m_t)|} = \mathbb{P}(\mathbf{x}_t|m_{1:t},\gamma_{1:t}).$$

Substituting (A), (B), and (C) in (B.8), we get

$$\mathbb{P}(\mathbf{x}_{t+1}^S|m_{1:t},\gamma_{1:t}) = \sum_{\mathbf{x}_t,\mathbf{w}_t}\left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^{i}=f_t(x_t^{i},\gamma_t(x_t^{i}),w_t^{i},m_t))\right]\cdot\mathbb{P}(\mathbf{w}_t)\cdot\mathbb{P}(\mathbf{x}_t|m_{1:t},\gamma_{1:t}) \overset{(b)}{=} \mathbb{P}(\mathbf{x}_{t+1}|m_{1:t},\gamma_{1:t}), \tag{B.9}$$

where (b) follows from (B.7). The rest of the proof is similar to that of step $t = 1$. From

(B.6) and (B.9), given $\{m_{1:t+1}, \gamma_{1:t+1}\}$, we get

$$\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t+1}, \gamma_{1:t+1}) = \begin{cases} 0 & \mathbf{x}_{t+1} \notin H(m_{t+1}) \\ \alpha(m_{t+1}) & \mathbf{x}_{t+1} \in H(m_{t+1}) \end{cases} \tag{B.10}$$

where $\alpha(m_{t+1}) = \frac{\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t},\gamma_{1:t})}{\mathbb{P}(m_{t+1}|m_{1:t},\gamma_{1:t})}$ depends only on m_{t+1} because, when $\mathbf{x}_{t+1} \in H(m_{t+1})$, $H(m_{t+1})$ contains nothing but permutations of $\mathbf{x}_{t+1}$ while $\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t}, \gamma_{1:t})$ is insensitive to permutation of $\mathbf{x}_{t+1}$ according to (B.9).

Since the summation of $\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t+1}, \gamma_{1:t+1})$ over $\mathbf{x}_{t+1} \in \mathcal{X}^n$ is one, we have

$$\alpha(m_{t+1}) = \frac{1}{|H(m_{t+1})|}. \tag{B.11}$$

From (B.10) and (B.11), we have

$$\mathbb{P}(\mathbf{x}_{t+1}|m_{1:t+1}, \gamma_{1:t+1}) = \mathbb{P}(\mathbf{x}_{t+1}|m_{t+1}) = \frac{\mathbb{1}(\mathbf{x}_{t+1} \in H(m_{t+1}))}{|H(m_{t+1})|}.$$

APPENDIX C

Alternate proof of common information approach

Problem formulation

Consider a decentralized dynamic stochastic system with $n \in \mathbb{N}$ agents where each agent has only access to a partial information of system (incomplete information) and the objective of all agents is to minimize a common cost function. We assume all system variables are finite. Let $\mathcal{X}$ be a finite state space and $X_t \in \mathcal{X}$ be the state of system at time $t \in \mathbb{N}$. Let $u_t^i \in \mathcal{U}_t^i$ be control action of agent $i \in \{1, \dots, n\}$ at time t and $\mathcal{U}_t^i$ be a finite set of all admissible control actions of agent i at time t. The initial state X_1 is a random variable. The system evolves as follows:

$$X_{t+1} = f_t(X_t, \mathbf{U}_t, W_t), \tag{C.1}$$

where $\mathbf{U}_t$ denotes vector $(U_t^1, \dots, U_t^n)$ and $\{W_t\}_{t=1}^{\infty}$ is an independent random process that is independent of the initial state X_1. At time t, agent i observes partial and noisy version of state of system i.e.

$$Y_t^i = h_t^i(X_t, V_t^i), \;\; i = 1, \dots, n, \tag{C.2}$$

where $\{V_t^i\}_{t=1}^{\infty}$ is an independent random processes. Let $\mathbf{V_t} = \text{vec}(V_t^1, \dots, V_t^n)$ and $\mathbf{Y_t} = \text{vec}(Y_t^1, \dots, Y_t^n)$. Let $\{X_1, W_t, \mathbf{V}_t\}_{t=1}^{\infty}$ be mutually independent. Denote I_t^i as the set of all the information available to agent i up to time t, i.e.,

$$I_t^i \subseteq \{\mathbf{Y}_{1:t}, \mathbf{U}_{1:t-1}\}, \;\; i = 1, \dots, n,$$

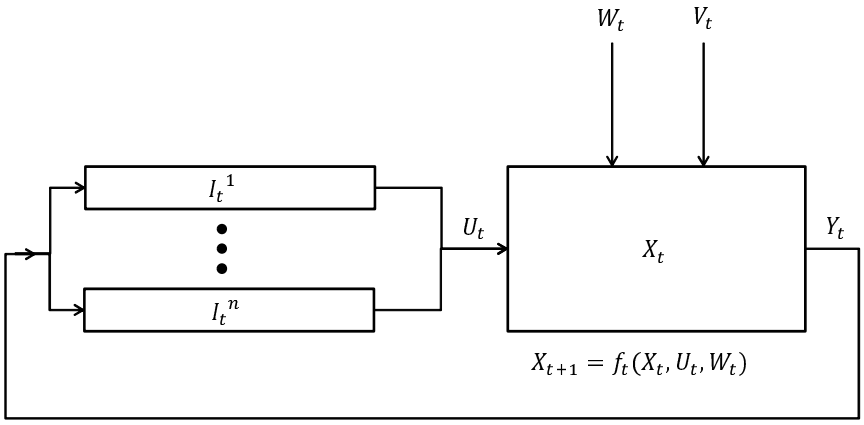

Fig. C.1 The control scheme of Problem 1.

Let g_t^i be control law of agent i at time t, so $g_t^i : I_t^i \mapsto \mathcal{U}_t^i$ i.e.

$$U_t^i = g_t^i(I_t^i). \tag{C.3}$$

Let $\ell_t(X_t, \mathbf{U}_t)$ be the incurred cost when action $\mathbf{U}_t$ is performed in state X_t at time $t \in \mathbb{N}$. We are interested in the following optimization problem.

Problem 1: Given the dynamics $\{f_t\}_{t=1}^T$, observation functions $\{\{h_t^i\}_{t=1}^T\}_{i=1}^n$, probability distributions of primitive random variables, cost functions $\{\ell_t\}_{t=1}^T$, and the horizon T, find a policy $g := (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_T)$, where $\mathbf{g}_t = (g_t^1, \ldots, g_t^n)$, such that

$$J(g, x_1) = \mathbb{E}^g[\sum_{t=1}^{T} \ell_t(X_t, \mathbf{U}_t) | X_1 = x_1],$$

is minimized.

Block diagram below shows the control structure of Problem 1. The main difficulty in such systems is the fact that every agent has different set of information which makes the control block in Fig. C.1 a decentralized controller with n different inputs.

Common information approach

In this section, we use an alternate proof to solve Problem 1 and obtain the results in [Nayyar et al., 2013]. At the first step, we split control block in Fig. C.1 into two blocks as shown in

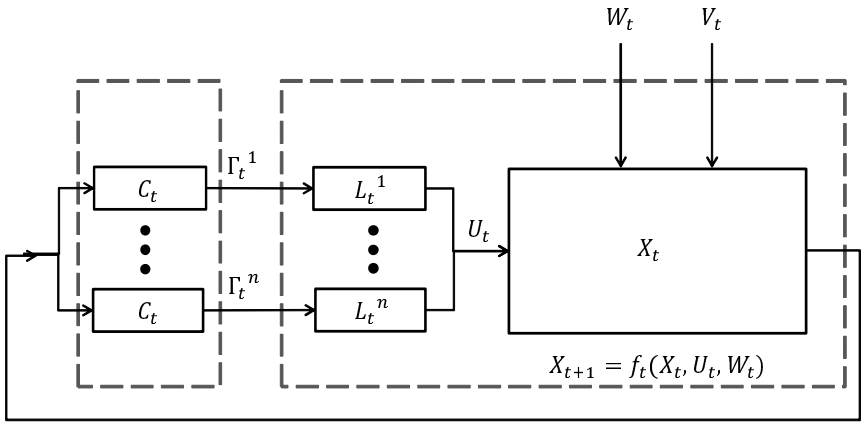

Fig. C.2 Splitting the control scheme of Fig. C.1 into two parts: coordinator and coordinated system (extended plant).

Fig. C.2:

1. The first block contains the part of information shared between all agents (*common information*) at time t i.e. $C_t := \cap_i I_t^i$. This block is called *coordinator* (centralized agent).

2. The second block contains the remaining part of information of all agents i.e. $\mathbf{L}_t := \{L_t^i, i = 1, \dots, n\}$ where $L_t^i := I_t^i \backslash C_t$ is the local information of agent i at time t. This block is added to the plant and the resultant block is called *coordinated system* (extended plant).

Notice that Fig. C.1 and Fig. C.2 only differ in labels. The advantage of this labeling is that coordinator is a centralized agent because its only input is C_t. The price of this simplicity (transformation of decentralized problem into centralized problem) is that the extended plant (coordinated system) is more complicated than the original plant. In particular, the input of extended plant is a function rather than a variable and the size of extended plant is, in general, larger than the size of original plant. Based on this split, policy g will also be split correspondingly. The share of coordinator of splitting policy g is defined as coordinator's policy ψ. Let $\boldsymbol{\psi}_t$ denote coordinator's control law at time t where $\boldsymbol{\psi}_t := (\psi_t^1, \dots, \psi_t^n)$ and ψ_t^i is the share of coordinator from g_t^i. Hence, ψ_t^i can be obtained from (C.3) and the fact that $I_t^i = \{C_t, L_t^i\}$ as follows:

$$\psi_t^i(C_t)(\cdot) := g_t^i(C_t, \cdot), \;\; i = 1, \dots, n$$

which implies $\boldsymbol{\psi}_t := \mathbf{g}_t(C_t, \cdot)$. Let $\boldsymbol{\Gamma}_t := (\Gamma_t^1, \dots, \Gamma_t^n)$ be the remaining part of the split of policy g defined as follows:

$$\Gamma_t^i(L_t^i) := g_t^i(C_t, L_t^i) = \psi_t^i(C_t)(L_t^i), \;\; i = 1, \dots, n$$

where $\Gamma_t^i : L_t^i \mapsto \mathcal{U}_t^i$, $i = 1, \dots, n$ i.e.

$$U_t^i = \Gamma_t^i(L_t^i), \;\; i = 1, \dots, n. \tag{C.4}$$

A. C.1 *We assume $C_t \subseteq C_{t+1}$.*

Notice that we have only labeled problem 1 differently; hence, finding coordinator's policy $\psi := (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_T)$ is as equivalent as finding policy g. Let $Z_{t+1} \subseteq \{\mathbf{L}_t, \mathbf{Y}_t, \mathbf{U}_t\}$ be coordinator's observation of the extended plant at time $t+1$. For now we assume the coordinator stores the history of common information by time t, i.e., $Z_{1:t}$; later, we see $Z_{1:t}$ can be compressed into an updateable information state Π_t that only requires Z_{t+1} (rather than the history $Z_{1:t+1}$) to update to the next information state Π_{t+1}. At time $t = 1, \dots, T$, the coordinator observes $(Z_{1:t}, \boldsymbol{\Gamma}_{1:t-1})$ and chooses function $\boldsymbol{\Gamma}_t$.

Now, we impose the following assumption to make the coordinator's problem a POMDP. In particular, we want the coordinated system (extended plant) evolves in Markovian manner so that we can use the standard result of POMDPs.

A. C.2 (Key Assumption) $\mathbb{P}(\mathbf{L}_{t+1} | X_{1:t}, \mathbf{L}_{1:t}, \mathbf{U}_{1:t}) = \mathbb{P}(\mathbf{L}_{t+1} | X_t, \mathbf{L}_t, \mathbf{U}_t)$. *This transition probability does not depend on policy $\psi_{1:t}$.*

From Fig. C.2, it is seen that the extended plant at time t is identified by $S_t := (\mathbf{L}_t, X_t)$. From (C.1), the original plant evolves in MDP manner and from (C.2) and A.(C.2), the update of local information would also evolve in MDP manner. In particular,

Lemma C.1 (Lemma 1 of [Nayyar et al., 2013]) *Extended plant evolves in MDP manner i.e.*

1. $\mathbb{P}(S_{t+1} | S_{1:t}, \boldsymbol{\Gamma}_{1:t}) = \mathbb{P}(S_{t+1} | S_t, \boldsymbol{\Gamma}_t)$.
2. $P(Z_{t+1} | S_{1:t}, \boldsymbol{\Gamma}_{1:t}) = P(Z_{t+1} | S_t, \boldsymbol{\Gamma}_t)$.

3. $\mathbb{E}[\ell_t(X_t, \mathbf{U}_t)|S_{1:t}, \mathbf{\Gamma}_{1:t}] = \mathbb{E}[\ell_t(X_t, \mathbf{U}_t)|S_t, \mathbf{\Gamma}_t] =: \hat{\ell}_t(S_t, \mathbf{\Gamma}_t)$.

Proof: Property (1) follows from (C.1) and (C.4) and assumption A.(C.2). Property (2) follows from $Z_{t+1} \subseteq \{\mathbf{L}_t, \mathbf{Y}_t, \mathbf{U}_t\}$, (C.2), and (C.4). Property (3) follows (C.4). ■

Therefore, from Lemma C.1, we can conclude that Fig. C.2 is a standard POMDP problem where $\{S_t\}_{t=1}^T$ is a controlled Markov process under $\mathbf{\Gamma}_t$. The coordinator is a centralized controller with partial observation Z_{t+1}. Thus, from standard results in POMDP, the belief state is an information state, i.e.,

$$\Pi_t(s_t) := \mathbb{P}(S_t = s_t | Z_{1:t}, \mathbf{\Gamma}_{1:t-1}).$$

Hence, we can develop a dynamic program to compute Γ_t as follows: for $t = T$,

$$V_T(\pi_T) = \min_{\boldsymbol{\gamma}_T}(\ell'_T(\pi_T, \boldsymbol{\gamma}_T)),$$

and for $t = T-1, \dots, 1$,

$$V_{t+1}(\pi_t) = \min_{\boldsymbol{\gamma}_t}(\ell'_t(\pi_t, \boldsymbol{\gamma}_t) + \mathbb{E}[V_t(\pi_{t+1})|\pi_t, \boldsymbol{\gamma}_t]),$$

where $\ell'_t(\pi, \boldsymbol{\gamma}) := \sum_s \hat{\ell}_t(s, \boldsymbol{\gamma})\pi(s)$ and $\boldsymbol{\gamma}_t = \mathrm{vec}(\gamma_t^1, \dots, \gamma_t^n)$ and $\gamma_t^i : l_t^i \mapsto u_t^i$.

Remark C.1 Note that we assumed the coordinator stores $Z_{1:t}$. Alternatively, the coordinator could store the information state Π_t rather than the history $Z_{1:t}$. In this case, the strategy given by above dynamic program is implementable by all agents as long as (Π_{t-1}, Z_t) belongs to common information C_t.

To ensure the belief state is valid in infinite horizon, we impose the following assumption:

A. C.3 (Definition 2 of [Mahajan and Mannan, 2016]) *The size of local information L_t is uniformly bounded i.e there exists a k such that $|L_t^i| \leq k, \forall i = 1, \dots, n$.*

According to Definition 2 of [Mahajan and Mannan, 2016], we have that

Definition C.1 (Partial History Sharing) Any information structure that satisfies above setting is called partial history sharing.

Therefore, the partial history sharing encompasses a large class of decentralized information structures because Assumptions A.(C.2) and A.(C.3) are mild conditions.

Notice that the common information approach compresses the common information C_t into the belief state π_t which is updatable; consequently, there is no need to keep the history of the common information. In addition, the minimization of each step of above dynamic program is a pure decentralized problem; hence, its computational complexity increases doubly exponential in general with the number of agents. This implies the common information approach is not useful to solve the pure decentralized control systems (when the common information that is the centralized information is empty) because the decentralized information are intact in this approach and relabelled as the partially evaluated functions.

APPENDIX D

Preliminaries on Q-Learning (discounted and time average)

Problem Statement

Consider a time-homogeneous dynamical system where $\mathcal{U}$ is a set of all admissible actions, $\mathcal{X}$ is state space, and $x_t \in \mathcal{X}$ is a state of system at time $t \in \mathbb{N} \cup \{0\}$. Let $u_t \in \mathcal{U}$ be control action at time t which is chosen according to a control law g, i.e. $\forall x_t \in \mathcal{X}, \;\; u_t = g_t(x_t)$, and a cost c_t is incurred by executing action u_t in state x_t. Model of such system is given by:

$$\begin{cases} X_{t+1} = f(X_t, U_t, W_t^x), \\ C_t = \ell(X_t, U_t, W_t^c). \end{cases} \tag{D.1}$$

Let $\{W_t^x\}_{t=1}^{\infty}$ be i.i.d stochastic process with time-invariant probability mass function P_w^x, and $\{W_t^c\}_{t=1}^{\infty}$ be i.i.d stochastic process with time-invariant probability mass function P_w^c. Let $x_0 \in \mathcal{X}$ be initial state, and $\{X_0, W_t^x, W_t^c\}$, $t = 1, 2, \ldots$, be independent.

Given a $\beta \in (0, 1)$, performance of a policy $g = (g_0, g_1, g_2, \ldots, g_t, \ldots)$, where $g_t : \mathcal{X} \to \mathcal{U}$, with initial state x_0 is defined as follows:

$$J_\beta(g)(x_0) = \mathbb{E}^g\{\sum_{t=0}^{\infty} \beta^t \ell(X_t, g_t(X_t), W_t^c) | X_0 = x_0\}.$$

Objective: We wish to find a policy g^* that minimizes the above-mentioned expected total cost, i.e. $J_\beta(g^*)(x_0) = \min_g J_\beta(g)(x_0)$.

From dynamic programming [1], we know there exist optimal policies that are time-

invariant, i.e. $\forall t \;\; g_t = g$, and such policies can be obtained by following dynamic programming equation:

$$V^*(x) = \min_{u \in \mathcal{U}} \{\mathbb{E}[\ell(x, u, W^c) + \beta V^*(f(x, u, W^x))]\}. \tag{D.2}$$

A time-invariant policy $g = (g^*, g^*, \ldots)$ is optimal where g^* belongs to the following set:

$$g^*(x) \in \arg\min_{u \in \mathcal{U}} \{\mathbb{E}[\ell(x, u, W^c) + \beta V^*(f(x, u, W^x))]\}. \tag{D.3}$$

Let us define:

$$Q^*(x, u) := \mathbb{E}[\ell(x, u, W^c) + \beta V^*(f(x, u, W^x))]. \tag{D.4}$$

Then, (D.2), (D.3), and (D.4) yield:

$$V^*(x) = \min_{u \in \mathcal{U}} Q^*(x, u), \tag{D.5}$$

$$g^*(x) \in \arg\min_{u \in \mathcal{U}} Q^*(x, u). \tag{D.6}$$

One difficulty to find optimal policy using (D.4) is that system model (f, ℓ, P_w^x, and P_w^c) must be known. To overcome this difficulty, Q-learning approach computes $Q^*(x, u)$ based on a reformulation of (D.4) without requiring above-mentioned information of system.

In brief, Q-learning is a stochastic iterative algorithm that at each step assigns every pair of state and action a Q function, $Q_n(x, u)$, and updates it iteratively. Under a few reasonable assumptions, every $Q_n(x, u)$, associated with state x and action u, converges to a $Q^*(x, u)$ with probability one. In other words,

$$\forall x \in \mathcal{X}, \quad \forall u \in \mathcal{U}(x), \quad \lim_{n \to \infty} Q_n(x, u) = Q^*(x, u) \quad w.p.1,$$

where n is the index of successive steps in updating process.

Q-Learning

Given the problem in D, we are interested in computing optimal policy when no model of system, (f, ℓ, P_w^x, and P_w^c), is available, but we have access to a system simulator. A simulator is a black box which receives u_n^i and x_n^i as inputs and gives out x_{n+1}^o and c_{n+1}^o as outputs at iteration n, i.e. $(X_{n+1}^o, C_{n+1}^o) = S(x_n^i, u_n^i, W_n^s)$, where x_{n+1}^o is the next state reached by

performing action u_n^i in state x_n^i, c_{n+1}^o is incurred cost at iteration n, and $\{W_n^s\}_{n=1}^{\infty}$ is i.i.d stochastic process with time invariant probability mass function. Note that independency of $\{W_n^s\}_{n=1}^{\infty}$ implies each invocation of the simulator is independent. The simulator should be consistent with the model in (D.1) where the consistency is defined as follows:

Definition D.1 A simulator (S, W^s) is consistent with a dynamical model (f, l, P_w^x, P_w^c) if:

$$P^s(X^o = \widetilde{x}, C^o = \widetilde{c} | X^i = x, U^i = u) = P(f(x, u, W^x) = \widetilde{x}, \ell(x, u, W^c) = \widetilde{c}).$$

In addition, we assume state space $\mathcal{X}$ and action space $\mathcal{U}$ are finite.
The rest of this chapter is organized as follows: Section D.1 describes Q-learning algorithm in five steps and section D.2 briefly states the relationship between Q-learning and stochastic approximation theory. Section D.3 introduces stochastic approximation theory, and its associated model and assumptions. Section **??** shows that Q-learning is in the form of stochastic approximation and provides a sketch of proof of Q-learning algorithm based on the results of stochastic approximation theory.

D.1 Q-Learning algorithm

Q-learning consists of a "learning function" γ_n, and the following algorithm to update Q-functions. For each pair of state and action, (x, u), iteratively define a sequence of functions Q_n , $n \in \mathbb{N}$, where $Q_n : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. Let (x_n^i, u_n^i) be inputs and (x_{n+1}^o, c_{n+1}^o) be outputs of the simulator, and $\lambda_n^i \in [0, 1]$ be the rate of learning according to which Q-function associated with (x_n^i, u_n^i) is updated at iteration n. Let $\mathcal{F}_n$ denote $\sigma(x_{1:n}^i, u_{1:n}^i, x_{1:n}^o, c_{1:n}^o, \lambda_{1:n-1})$ as the history of the algorithm up to iteration n, and let $\mathcal{F}_n$ and Q_n be defined on a common probability space $(\Omega, \mathcal{F}, P)$. Then, Q-learning algorithm is defined through the following basic steps.
Step 1: Arbitrarily initialize Q_0, e.g. $\forall x \in \mathcal{X}, \forall u \in \mathcal{U} \to \quad Q_0(x, u) = 0$.
Step 2: At iteration n, $n \in \mathbb{N}$, "Learner" chooses a pair of state and action (x_n^i, u_n^i) and updates the corresponding Q-function with a learning rate of λ_n^i. To do this, learning function γ_n is used to compute triple $(x_n^i, u_n^i, \lambda_n^i)$ as follows:

$$(x_n^i, u_n^i, \lambda_n^i) = \gamma_n(x_{1:n-1}^i, u_{1:n-1}^i, x_{1:n-1}^o, c_{1:n-1}^o, \lambda_{1:n-1}).$$

Step 3: Run the simulator with (x_n^i, u_n^i) as inputs to get (x_{n+1}^o, c_{n+1}^o).
Step 4: Update Q-functions as follows:

$$Q_{n+1}(x,u) = \begin{cases} Q_n(x,u) + \lambda_n^i[c_{n+1}^o + \beta \min\limits_{v \in \mathcal{U}(x_{n+1}^o)} Q_n(x_{n+1}^o, v) - Q_n(x,u))] & \text{if } (x,u) = (x_n^i, u_n^i) \\ Q_n(x,u) & \text{if } (x,u) \neq (x_n^i, u_n^i). \end{cases} \tag{D.7}$$

Step 5: Go back to step 2 until termination.
Note that if we define:

$$\alpha_n(x,u) := \begin{cases} \lambda_n^i & \text{if } (x,u) = (x_n^i, u_n^i) \\ 0 & \text{Otherwise} \end{cases} \tag{D.8}$$

Then, we can simply convert (D.7) into the following unified form:

$$Q_{n+1}(x,u) = Q_n(x,u) + \alpha_n(x,u)[c_{n+1}^o + \beta \min_{v \in \mathcal{U}(x_{n+1}^o)} Q_n(x_{n+1}^o, v) - Q_n(x,u))], \forall x \in \mathcal{X}, u \in \mathcal{U}.$$

Theorem D.1 *Under the following conditions:*

(C1) $\sum_{n=0}^{\infty} \alpha_n(x,u) = \infty$, $\sum_{n=0}^{\infty} \alpha_n^2(x,u) \leq K$, *$K$ is constant, w.p.1.*

(C2) The learning rule γ_n *is such that each* (x^i, u^i) *occurs infinitely often.*

$\{Q_n\}$ *converges to the* Q^* *corresponding to the model in* (D.1) *w.p.1. Furthermore, there exist special cases where Q-learning algorithm converges when* $\beta = 1$. *To see these cases, check Theorem D.4.*

One trivial example of γ_n is to sequentially cycle through all possible values of (x^i, u^i), and let $\{\lambda_n^i\}_{n=1}^{\infty}$ be a deterministic sequence satisfying (C1). According to (D.8), every sequence $\{\alpha_n(x,u)\}_{n=1}^{\infty}$ consists of some zeros and a subsequence of $\{\lambda_n^i\}_{n=1}^{\infty}$, so if $\{\lambda_n^i\}_{n=1}^{\infty}$ satisfies (C1), every $\{\alpha_n(x,u)\}_{n=1}^{\infty}$ will satisfy (C1). Another example of γ_n is to choose (x^i, u^i) at iteration n such that $\#(x^i, u^i, n)$ is the smallest, where $\#(x^i, u^i, n)$ is the number of times action u^i is performed in state x^i up to time n, and then pick $\lambda_n^i = \frac{1}{\#(x^i, u^i, n)}$.

One advantage of Q-learning is rather than trying to learn the model of system, it aims directly to estimate value function and optimal policy. On the other hand, there are other approaches which estimate the model first and then obtain optimal policy by solving (D.2). As far as computational expenses are concerned, Q-learning can be very more efficient than those approaches when number of states is considerably greater than number of actions.

D.2 Relationship between Q-learning and Stochastic Approximation Theory

Stochastic approximation algorithms often have a structure defined as follows :

$$x^i := x^i + \alpha^i(F^i(x) - x^i + w^i),$$

where $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$, $F^1, \ldots, F^d$ are mappings from $\mathbb{R}^d \to \mathbb{R}$, w^i is a random noise and α^i is a small, usually decreasing, step size. The Q-learning algorithm is exactly of the above-mentioned form, with the mapping $F = (F^1, \ldots, F^d)$ being closely related to dynamic programming operator associated with a Markov decision process.

Due to such a close relationship between stochastic approximation theory and Q-learning, Section D.3 describes model and assumptions of stochastic approximations, and it is proven that under mild conditions, associated stochastic process x converges to fixed point of F.

Section D.4 shows that Q-learning can be reformulated and converted into a standard model of stochastic approximation model, and it satisfies all the required conditions of stochastic approximation theory to converge to fixed point of the corresponding mapping F. In Q-learning, the fixed point of the corresponding mapping F is nothing but $Q^*(x, u)$, through which value function and optimal policy can be easily computed by (D.5) and (D.6).

D.3 Asynchronous Stochastic Approximation Theory

The algorithm consists of noisy updates of vector $x \in \mathbb{R}^d$, for the purpose of finding the fixed point of a function $F : \mathbb{R}^d \to \mathbb{R}^d$. Let $F^1, \ldots, F^d : \mathbb{R}^d \to \mathbb{R}$ be the corresponding component mappings i.e. $F(x) = (F^1(x), \ldots, F^d(x)) \ \ \forall x \in \mathbb{R}^d$. Let x_n be the value of the vector x at iteration n and let x_n^i denote its ith element. Let T^i be an infinite subset of non-negative integers indicating the set of times at which an update of x^i is executed. We assume that:

$$x_{n+1}^i = x_n^i, \quad n \notin T^i. \tag{D.9}$$

With respect to the times that x^i is updated, we postulate an update equation of the form:

$$x_{n+1}^i = x_n^i + \alpha_n^i(F^i(\hat{x}_n^i) - x_n^i + w_n^i), \quad n \in T^i. \tag{D.10}$$

Here, α_n^i is a stepsize parameter belonging to $[0,1]$, w_n^i is a noise term, and $\hat{x}_n^i$ is a vector of possibly outdated components of x. In particular, we assume that:

$$\hat{x}_n^i = (x^1_{\tau_n^{1,i}}, x^2_{\tau_n^{2,i}}, \ldots, x^d_{\tau_n^{1,d}}), \quad n \in T^i,$$

where each $\tau_n^{j,i}$ is an integer satisfying $0 \leq \tau_n^{j,i} \leq n$. If no information is outdated, we achieve $\tau_n^{j,i} = n$ and $\hat{x}_n^i = x_n$ for all n. In order to bring (D.9) and (D.10) into a unified form, it is convenient to assume that α_n^i, w_n^i, and $\tau_n^{j,i}$ are defined for every i, j, and n, but that $\alpha_n^i = 0$ for $n \notin T^i$.

We will now continue with our assumptions. All variables introduced so far $(x_n, \tau_n^{j,i}, \alpha_n^i, w_n^i)$ are viewed as random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and assumptions deal primarily with dependencies between these random variables. Our assumptions also involve an increasing sequence $\{\mathcal{F}_n\}_{n=0}^{\infty}$ of sub-fields of $\mathcal{F}$. Intuitively, $\mathcal{F}_n$ is meant to represent the history of the algorithm up to n, and including the point at which the step-sizes α_n^i for nth iteration are selected, but just before the noise term w_n^i is generated.

For any vector $v = (v_1, \ldots, v_d)$ we define a norm $\|.\|_v$ on $\mathbb{R}^d$ by letting: $\|x\|_v = \max_i \frac{|x_i|}{|v_i|}$, $v_i \neq 0$ and $x \in \mathbb{R}^d$. If all components of v are equal to 1, $\|.\|_v$ is the same as $\|.\|_\infty$.

A. D.1 *For any i and j, $\lim_{n\to\infty} \tau_n^{j,i} = \infty$, with probability 1.*

This assumption guarantees that even though information can be outdated, any old information is eventually discarded.

A. D.2 *Assume the following:*

(a) x_0 is $\mathcal{F}(0)$-measurable.

(b) For every i and n, w_n^i is $\mathcal{F}_{n+1}$-measurable.

(c) For every i, j, and n, α_n^i, $\{n \in T^i\}$, and $\tau_n^{j,i}$ are $\mathcal{F}_n$-measurable.

(d) For every i and n, we have $\mathbb{E}[w_n^i | \mathcal{F}_n] = 0$.

(e) There exist (deterministic) constants A and B such that:

$$\mathbb{E}[(w_n^i)^2 | \mathcal{F}_n] \leq A + B \max_j \max_{\tau \leq n} |x_\tau^j|^2 \quad \forall i, n.$$

Assumption A.(D.2) allows for the possibility of deciding whether to update a particular component x_n^i at time n, based on the past history of the process. In this case, the step-size α_n^i becomes a random variable. However, part (c), of the assumption requires that the choice of the component to be updated must be made without anticipatory knowledge of the noise variables w^i that have not been realized yet.

A. D.3 *Assume:*

(a) For every i, $\sum_{n=0}^{\infty} \alpha_n^i = \infty$ *w.p.1.*

(b) There exists (deterministic) constant K such that for every i:

$$\sum_{n=0}^{\infty} (\alpha_n^i)^2 \leq K, \quad w.p.1.$$

A. D.4 *Assume:*

(a) The mapping F is monotone; that is, if $x \leq y$, then $F(x) \leq F(y)$.

(b) The mapping F is continuous.

(c) The mapping F has a unique fixed point x^*.

(d) If $e \in \mathbb{R}^n$ is the vector with all components equal to 1, and r is a positive scalar, then

$$F(x) - re \leq F(x - re) \leq F(x + re) \leq F(x) + re.$$

A. D.5 *There exists a vector $x^* \in \mathbb{R}^d$, a positive vector v, and a scalar $\rho \in [0, 1)$, such that:*

$$\|F(x) - x^*\|_v \leq \rho \|x - x^*\|, \ \ \forall x \in \mathbb{R}^d.$$

Lemma D.1 *If $\exists\, x^*, v$ satisfying assumption A.(D.5), then x^* is a unique fixed point of F. To see that x^* is a fixed point, let $x = x^*$:*

$$x = x^* \longrightarrow \|F(x^*) - x^*\|_v \leq \rho \|x^* - x^*\| = 0 \Longrightarrow F(x^*) = x^*.$$

To see that x^* is the unique fixed point, let $y^* \in \mathbb{R}^d$ be a different fixed point of F, then from assumption 5 we have:

$$F(y^*) = y^* \longrightarrow \|F(y^*) - x^*\|_v = \|y^* - x^*\|_v \leq \rho\|y^* - x^*\|_v \Longrightarrow x^* = y^*.$$

A. D.6 *Suppose there exists a cost-free absorbing state, say state s, i.e. $\mathbb{P}(s|s,u) = 1, \ell(s,u) = 0 \forall u \in \mathcal{U}$. We say that a stationary policy is proper if the probability of being at the absorbing state converges to s as time converges to infinity; otherwise, we say that the policy is improper.*

a) There exists at least one proper stationary policy.

b) Every improper stationary policy yields infinite expected cost for at least one initial state.

Theorem D.2 *[Tsistsiklis, 1994, Theorem 2] Let assumptions A.(D.1), A.(D.2), A.(D.3), and A.(D.4) hold. Furthermore, suppose sequence $\{x_n\}_{n=1}^{\infty}$ generated by stochastic approximation theory is bounded with probability 1. Then, $\{x_n\}_{n=1}^{\infty}$ converges to x^* with probability 1 where x^* is the unique point of F.*

Theorem D.3 *[Tsistsiklis, 1994, Theorem 3] Let assumptions A.(D.1), A.(D.2), A.(D.3), and A.(D.5) hold. Then, the sequence $\{x_n\}_{n=1}^{\infty}$ generated by stochastic approximation theory converges to x^* with probability 1 where x^* is the unique point of F.*

D.4 Main theorems including sketch of the proofs

Theorem D.4 *[Tsistsiklis, 1994, Theorem 4] Consider the Q-learning algorithm and let*

$$Q^*(x,u) = \mathbb{E}[\ell(x,u,W^c)] + \beta \sum_{y \in \mathcal{X}} \mathbb{P}(y|x,u)V^*(y).$$

Then, $\{Q_n(x,u)\}_{n=1}^{\infty}$ converges to $Q^*(x,u)$ with probability 1, for every x and u, in each of the following cases:

(a) $\beta < 1$. This case represents discounted problems.

(b) $\beta = 1$, $Q_0(s,u) = 0 \, \forall u \in \mathcal{U}$, and all policies are proper where "s" denotes the cost-free absorbing state. This case represents time average problems.

Proof: We first show that Q-learning algorithm is in the form of stochastic approximation i.e. (D.9) and (D.10). Hence, we can use the results of stochastic approximation theory.

Let $d := |\mathcal{X}| \times |\mathcal{U}|$ be total number of pairs of (action,state), and $Q_n \in \mathbb{R}^d$ with components of $Q_n(x,u) \in \mathbb{R}$. Let F be a mapping from $\mathbb{R}^d \to \mathbb{R}^d$ with components $F^{(x,u)}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ defined by:

$$F^{(x,u)}(Q_n) := \mathbb{E}[\ell(x,u,W^c)] + \beta \mathbb{E}[\min_{v \in \mathcal{U}} Q_n(f(x,u,W^x),v)], \tag{D.11}$$

where $f(x,u,W^x)$ is the next state reached by executing action u in state x and $\ell(x,u,W^c)$ is the incurred cost of such transition. We also define:

$$w_n(x_n^i,u_n^i) := C_{n+1}^o - \mathbb{E}[\ell(x,u,W^c)] + \beta \min_{v \in \mathcal{U}} Q_n(X_{n+1}^o,v) - \beta \mathbb{E}[\min_{v \in \mathcal{U}} Q_n(f(x,u,W^x),v)]. \tag{D.12}$$

Now, we re-write (D.7) in terms of (D.11) and (D.12).

$$\begin{cases} Q_{n+1}(x,u) = Q_n(x,u) + \lambda_n^i [F^{(x,u)}(Q_n) - Q_n(x,u) + w_n(x,u)] & \text{if } (x,u) = (x_n^i,u_n^i) \\ Q_{n+1}(x,u) = Q_n(x,u) & \text{if } (x,u) \neq (x_n^i,u_n^i). \end{cases} \tag{D.13}$$

Notice that (D.7) and (D.13) are identical. It is trivial to see that Q-learning algorithm (D.13) is precisely in the form of stochastic approximation algorithm i.e. (D.9) and (D.10) where $T^{(x,u)} = \{n : (x,u) = (x_n^i,u_n^i)\}$.

Now, it can be shown that F in (D.11) and w in (D.12) always satisfy assumption A.(D.2). Also, if $\beta < 1$, then F is a contraction which implies F satisfies assumption A.(D.5). Furthermore, if assumption A.(D.6) holds, then F satisfies assumption A.(D.4). To see more details of these proofs, we refer reader to [Tsistsiklis, 1994] and [Bertsekas and Tsitsiklis, 1991]. Notice that assumptions A.(D.1) and A.(D.3) are exogenous assumptions and independent of function F. Assumption A.(D.3) is up to the designer and corresponds to condition (C1) in Theorem D.1. Assumption A.(D.1) depends on the system being fully explored and it corresponds to condition (C2) in Theorem D.1. The proof of part (b) is based on the fact that when all the policies are proper, dynamic programming operator is a contraction. We refer reader to [Bertsekas and Tsitsiklis, 1991] for more details. So, the proof of part (b) follows from Theorem D.3. ■

Theorem D.5 *[Bertsekas, 1998, Proposition 1:] Consider a uni-chain model with* ***time-average*** *cost criterion. If there exists a special state s such that s is recurrent under all the*

policies, without loss of optimality, the problem can be converted to a problem with a cost-free absorbing state where all the policies are proper. Consequently, Q-learning algorithm can be developed for the converted problem and the proof of its convergence follows from part (b).

(c) $\beta = 1$, $Q_0(s,u) = 0 \; \forall u \in \mathcal{U}$, *assumption A.(D.6) holds, and* $\{Q_n(x,u)\}_{n=1}^{\infty}$ *is guaranteed to be bounded with probability* 1.

The actual restrictive assumption is assumption A.(D.6) because according to the following theorem, assumption A.(D.6) implies the boundedness of Q-functions whenever the stage costs are non-negative.

Theorem D.6 *[Tsistsiklis, 1994, Lemma 9] Suppose that* $\beta = 1$, *assumption A.D.6 holds, and* $Q_0(s,u) = 0 \; \forall u \in \mathcal{U}$. *Furthermore, suppose that all one-stage costs* $\ell(x,u,w^c)$ *are non-negative with probability* 1, *and that all initial Q-functions are non-negative i.e.* $Q_0(x,u) \geq 0$. *Then, the sequence* $\{Q_n(x,u)\}_{n=1}^{\infty}$ *generated by Q-learning algorithm in D.1 is bounded with probability* 1. *Hence, if assumption A.D.6 holds and all stage costs are non-negative,then Q functions are bounded w.p.1. Plus, if assumption A.D.6 holds, F satisfies A.D.4. Thus, the proof of part (c) follows from Theorem D.2.*

It is worth mentioning that Q-learning can be used in finite horizon time problems, too. Let $T \in \mathbb{N}$ be the finite horizon. We define a cost-free absorbing state which absorbs all the probabilities after time T i.e. $t = T + 1$ and stays there forever. Since the transition probabilities and cost functions should not be time varying, we need to define an extended state as a collection of (state,time). Consequently, Q-functions will be defined in terms of (state, time, action). Now, we can run Q-learning algorithm for the extended system (where transition probabilities and cost functions are modified appropriately) and the convergence of Q-learning algorithm is guaranteed by part (b) in Theorem D.4.

References

[Aoki, 1965] Aoki, M. (1965). Optimal control of partially observable Markovian systems. *Journal of The Franklin Institute*, 280(5):367–386.

[Arabneydi and Mahajan, 2014] Arabneydi, J. and Mahajan, A. (2014). Team optimal control of coupled subsystems with mean-field sharing. *IEEE Conference on Decision and Control*, pages 1669–1674.

[Arabneydi and Mahajan, 2015] Arabneydi, J. and Mahajan, A. (2015). Team optimal control of coupled major-minor subsystems with mean-field sharing. *Indian Control Conference (ICC)*, pages 95–100.

[Arnold and Laub, 1984] Arnold, W. F. and Laub, A. J. (1984). Generalized eigenproblem algorithms and software for algebraic riccati equations. *Proceedings of the IEEE*, 72(12):1746–1754.

[Asghari and Nayyar, 2016] Asghari, S. M. and Nayyar, A. (2016). Decentralized control problems with substitutable actions. *arXiv preprint arXiv:1601.02250*.

[Atrash et al., 2009] Atrash, A., Kaplow, R., Villemure, J., West, R., Yamani, H., and Pineau, J. (2009). Development and validation of a robust speech interface for improved human-robot interaction. *International Journal of Social Robotics*, 1(4):345–356.

[Bamieh and Voulgaris, 2005] Bamieh, B. and Voulgaris, P. G. (2005). A convex characterization of distributed control problems in spatially invariant systems with communication constraints. *Elsevier, System and Control Letters*, 54(6):575–583.

[Barty et al., 2010] Barty, K., Carpentier, P., and Girardeau, P. (2010). Decomposition of large-scale stochastic optimal control problems. *RAIRO-Operations Research, Cambridge Univ Press*, 44(03):167–183.

[Bensoussan et al., 2016] Bensoussan, A., Sung, K. C. J., Yam, S. C. P., and Yung, S. P. (2016). Linear-quadratic mean field games. *Journal of Optimization Theory and Applications*, 169(2):496–529.

[Bernstein et al., 2002] Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *INFORMS, Mathematics of operations research*, 27(4):819–840.

[Bertsekas, 1998] Bertsekas, D. P. (1998). A value iteration method for the average cost dynamic programming problem. *SIAM Journal on control optimization*, 36(2):742–759.

[Bertsekas, 2012] Bertsekas, D. P. (2012). *Dynamic programming and optimal control*. Athena Scientific.

[Bertsekas and Tsitsiklis, 1991] Bertsekas, D. P. and Tsitsiklis, J. N. (1991). An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595.

[Bishop and Doucet, 2014] Bishop, A. N. and Doucet, A. (2014). Distributed nonlinear consensus in the space of probability measures. *arXiv preprint arXiv:1404.0145*.

[Bonet, 2002] Bonet, B. (2002). An epsilon-optimal grid-based algorithm for partially observable Markov decision processes. *Proc. 19th International Conference on Machine Learning*, pages 51–58.

[Broz et al., 2011] Broz, F., Nourbakhsh, I., and Simmons, R. (2011). Designing POMDP models of socially situated tasks. *2011IEEE RO-MAN*, pages 39–46.

[Bu et al., 2011] Bu, S., Yu, F. R., Liu, P. X., and Zhang, P. (2011). Distributed scheduling in smart grid communications with dynamic power demands and intermittent renewable energy resources. *2011 IEEE International Conference on Communications Workshops (ICC)*, pages 1–5.

[Busoniu et al., 2006] Busoniu, L., Babuska, R., and Schutter, B. D. (2006). Multi-agent reinforcement learning: A survey. *9th International Control, Automation, Robotics and Vision (ICARCV '06)*, 1:108–113.

[Caines, 2013] Caines, P. (2013). Mean field games. *In: Samad T., Baillieul J. (Ed.) Encyclopedia of Systems and Control: SpringerReference, Springer-Verlag Berlin Heidelberg*.

[Caines, 1987] Caines, P. E. (1987). *Linear stochastic systems*. John Wiley and Sons, Inc.

[Caines and Kizilkale, 2014] Caines, P. E. and Kizilkale, A. C. (2014). Mean field estimation for partially observed LQG systems with major and minor agents. *19th IFAC World Congress*, 47(3):8705–8709.

[Cassandra et al., 1997] Cassandra, A., Littman, M. L., and Zhang, N. L. (1997). Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. *Thirteenth conference on Uncertainty in artificial intelligence*, pages 54–61.

[Cassandra, 1998] Cassandra, A. R. (1998). A survey of POMDP applications. *Working Notes of AAAI 1998 Fall Symposium on Planning with Partially Observable Markov Decision Processes*, pages 17–24.

[Claus and Boutilier, 1998] Claus, C. and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. *15th National conference on AI and 10th conference on Innovative Applications of AI (AAAI/IAAI-98)*, pages 746–752.

[Cohen, 1977] Cohen, G. (1977). On an algorithm of decentralized optimal control. *Journal Math. Anal. and Appl.*, 59(2):242–259.

[Couillet et al., 2012] Couillet, R., Perlaza, S. M., Tembine, H., and Debbah, M. (2012). Electrical vehicles in the smart grid: a mean field game analysis. *IEEE Journal on Selected Areas in Communications*, 30(6):1086–1096.

[Culioli and Cohen, 1990] Culioli, J. C. and Cohen, G. (1990). Decomposition/coordination algorithms in stochastic optimization. *SIAM Journal on Control and Optimization*, 28(6):1372–1403.

[Darrell and Pentland, 1996] Darrell, T. and Pentland, A. (1996). Active gesture recognition using partially observable Markov decision processes. *13th International Conference onPattern Recognition*, 3:984–988.

[Dibangoye et al., 2009] Dibangoye, J. S., Mouaddib, A.-I., and Chai-draa, B. (2009). Point-based incremental pruning heuristic for solving finite-horizon dec-POMDPs. *8th International Conference on Autonomous Agents and Multiagent Systems*, 1:569–576.

[Elliott et al., 2013] Elliott, R., Li, X., and Ni, Y.-H. (2013). Discrete time mean-field stochastic linear-quadratic optimal control problems. *Automatica*, 49(11):3222–3233.

[Even-Dar, 2005] Even-Dar, E. (2005). Algorithms for reinforcement learning. *Ph.D. thesis, TAV university.*

[Gihman and Skorohod, 1979] Gihman, I. I. and Skorohod, A. (1979). *Controlled stochastic processes.* Springer.

[Gomes and Saude, 2014] Gomes, D. A. and Saude, J. (2014). Mean field games models: A brief survey. *Springer, Dynamic Games and Appl.*, 4:1–45.

[Gosavi, 2009] Gosavi, A. (2009). Reinforcement learning: A tutorial survey and recent advances. *INFORMS Journal on Computing*, 21(2):178–192.

[Guéant et al., 2011] Guéant, O., Lasry, J.-M., and Lions, P.-L. (2011). Mean field games and applications. *Springer, Paris-Princeton Lectures on Mathematical Finance 2010*, pages 205–266.

[Hansen, 1998a] Hansen, E. A. (1998a). Finite-memory control of partially observable systems. *proceedings of the conference uncertainty in artificial intelligence*, pages 211–219.

[Hansen, 1998b] Hansen, E. A. (1998b). Finite-memory control of partially observable systems. *PhD dissertation, University of Massachusetts Amherst*.

[Hassibi et al., 1999] Hassibi, B., Sayed, A. H., and Kailath, T. (1999). *Indefinite-Quadratic Estimation and Control: A Unified Approach to $\mathcal{H}_2$ and $\mathcal{H}_\infty$ Theories*. SIAM.

[Hauskrecht and Fraser, 2000] Hauskrecht, M. and Fraser, H. (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine*, 18(3):221–244.

[Hluchyj and Gallager, 1981] Hluchyj, M. G. and Gallager, R. G. (1981). Multiacces of a slotted channel by finitely many users. *National Telecommunication Conference*, pages 421–427.

[Ho and Chu, 1972] Ho, Y. C. and Chu, K. h. (1972). Team decision theory and information structures in optimal control problems–part I. *IEEE Transactions on Automatic Control*, 17(1):15–22.

[Huang et al., 2014] Huang, J., Wang, S., and Wu, Z. (2014). Mean field linear-quadratic-Gaussian LQG Games: Major and minor players. *arXiv preprint arXiv:1403.3999*.

[Huang et al., 2005] Huang, J., Yang, B., and Liu, D. Y. (2005). A distributed q-learning algorithm for multi-agent team coordination. *IEEE 40th Int. conf. on Machine Learning and Cybernetics*, 1:108–113.

[Huang, 2010] Huang, M. (2010). Large-population LQG games involving a major player: the Nash certainty equivalence principle. *SIAM Journal on Control and Optimization*, 48(5):3318–3353.

[Huang et al., 2003] Huang, M., Caines, P. E., and Malhamé, R. P. (2003). Individual and mass behaviour in large population stochastic wireless power control problems: centralized and Nash equilibrium solutions. *IEEE Conference on Decision and Control*, pages 98–103.

[Huang et al., 2007] Huang, M., Caines, P. E., and Malhamé, R. P. (2007). Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria. *IEEE Transactions on Automatic and Control*, 52(9):1560–1571.

[Huang et al., 2012] Huang, M., Caines, P. E., and Malhamé, R. P. (2012). Social optima in mean field LQG control: centralized and decentralized strategies. *IEEE Transactions on Automatic and Control*, 57(7):1736–1751.

[Huang et al., 2006] Huang, M., Malhamé, R. P., Caines, P. E., et al. (2006). Large population stochastic dynamic games: closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems, International Press of Boston*, 6(3):221–252.

[Jaulmes et al., 2005] Jaulmes, R., Pineau, J., and Precup, D. (2005). Active learning in partially observable Markov decision processes. *European Conference on Machine Learning*, pages 601–608.

[Kaelbling et al., 1998] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Elsevier, Artificial Intelligence*, 101(1):99–134.

[Kane and Levinson, 1985] Kane, T. R. and Levinson, D. A. (1985). *Dynamics, theory and applications*. McGraw Hill.

[Kapetanakis and Kudenko, 2002] Kapetanakis, S. and Kudenko, D. (2002). Reinforcement learning of coordination in cooperative multi-agent systems. *18th National Conference on AI and 14th conference on Innovative Applications of AI (AAAI/IAAI-02)*, pages 326–331.

[Kizilkale and Malhame, 2014] Kizilkale, A. C. and Malhame, R. P. (2014). Collective target tracking mean field control for Markovian jump-driven models of electric water heating loads. *19th IFAC World Congress*, 47(3):1867–1872.

[Kizilkale et al., 2012] Kizilkale, A. C., Mannor, S., and Caines, P. E. (2012). Large scale real-time bidding in the smart grid: A mean field framework. *IEEE 51st Annual Conference on Decision and Control*, pages 3680–3687.

[Kok et al., 2005] Kok, J. R., Spaan, M. T., and Vlassis, N. (2005). Non-communicative multirobot coordination in dynamic environment. *Robotics and Autonomous Systems*, 50(2-3):99–114.

[Krishnamurthy, 2016] Krishnamurthy, V. (2016). *Partially Observed Markov Decision Processes*. Cambridge University Press.

[Lancaster and Rodman, 1995] Lancaster, P. and Rodman, L. (1995). *Algebraic riccati equations*. Clarendon press.

[Lasry and Lions, 2006a] Lasry, J. M. and Lions, P. L. (2006a). Jeux á champ moyen. I – le cas stationnaire. *C. R. Acad. Sci. Paris, Ser. I*, 343:619–625.

[Lasry and Lions, 2006b] Lasry, J. M. and Lions, P. L. (2006b). Jeux á champ moyen. II – horizon fini et contrôle optimal. *C. R. Acad. Sci. Paris, Ser. I*, 343:679–684.

[Lasry and Lions, 2007] Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Springer, Japanese Journal of Mathematics*, 2(1):229–260.

[Lasry et al., 2008] Lasry, J.-M., Lions, P.-L., Guéant, O., et al. (2008). Application of mean field games to growth theory. *(hal-00348376)*.

[Lessard and Lall, 2015] Lessard, L. and Lall, S. (2015). Optimal control of two-player systems with output feedback. *IEEE Transactions on Automatic Control*, 60(8):2129–2144.

[Lévesque and Maillart, 2008] Lévesque, M. and Maillart, L. M. (2008). Business opportunity assessment with costly, imperfect information. *IEEE Transactions on Engineering Management*, 55(2):279–291.

[Li et al., 2013] Li, T., wah Chu, E. K., Lin, W.-W., and Weng, P. C.-Y. (2013). Solving large-scale continuous-time algebraic riccati equations by doubling. *Journal of Computational and Applied Mathematics*, 237(1):373–383.

[Li and Zhang, 2008] Li, T. and Zhang, J.-F. (2008). Asymptotically optimal decentralized control for large population stochastic multiagent systems. *IEEE Transactions on Automatic and Control*, 53(7):1643–1660.

[Lipsa and Martins, 2011a] Lipsa, G. M. and Martins, N. C. (2011a). Optimal memoryless control in gaussian noise: A simple counterexample. *Automatica*, 47(3):552–558.

[Lipsa and Martins, 2011b] Lipsa, G. M. and Martins, N. C. (2011b). Remote state estimation with communication costs for first-order LTI systems. *IEEE Transactions on Automatic and Control*, 56(9):2013–2025.

[Littman, 1994a] Littman, M. L. (1994a). Memoryless policies: Theoretical limitations and practical results. *Third International Conference on Simulation of Adaptive Behavior*, 3:238–245.

[Littman, 1994b] Littman, M. L. (1994b). The witness algorithm: Solving partially observable Markov decision processes. *Brown University, Providence, RI*.

[Lovejoy, 1991] Lovejoy, W. S. (1991). Computationally feasible bounds for partially observed Markov decision processes. *Operations research*, 39(1):162–175.

[Lusena, 2001] Lusena, C. (2001). Finite memory policies for partially observable Markov decision processes. *PhD dissertation, University of Kentucky*.

[Ma and Yong, 1999] Ma, J. and Yong, J. (1999). *Forward-backward stochastic differential equations and their applications*, volume 1702. Springer, Lecture Notes in Mathematics.

[Madani et al., 1999] Madani, O., Hanks, S., and Condon, A. (1999). On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. *Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, pages 541–548.

[Madjidian and Mirkin, 2014] Madjidian, D. and Mirkin, L. (2014). Distributed control with low-rank coordination. *IEEE Transactions on Control of Network Systems*, 1(1):53—63.

[Mahajan, 2010] Mahajan, A. (2010). Optimal transmission policies for two-user multiple access broadcast using dynamic team theory. *IEEE 48th Annual Allerton Conference on Communication, Control, and Computing*, pages 806–813.

[Mahajan, 2013] Mahajan, A. (2013). Optimal decentralized control of coupled subsystems with control sharing. *IEEE Transactions on Automatic Control*, 58(9):2377–2382.

[Mahajan and Mannan, 2016] Mahajan, A. and Mannan, M. (2016). Decentralized stochastic control. *Springer, Annals of Operations Research*, 241:109—-126.

[Mahajan et al., 2012] Mahajan, A., Martins, N. C., Rotkowitz, M. C., and Yuksel, S. (2012). Information structures in optimal decentralized control. *Proc. of Conf. on Decision and Control (CDC)*, pages 1291–1306.

[Mahajan et al., 2008] Mahajan, A., Nayyar, A., and Teneketzis, D. (2008). Identifying tractable decentralized control problems on the basis of information structure. *46th Annual Allerton Conf. Communication, Control, and Computing*, pages 1440–1449.

[Marschack and Radner, 1972] Marschack, J. and Radner, R. (1972). Economic theory of teams. *New Haven: Yale University Press*.

[Matignon et al., 2007] Matignon, L., Laurent, G. J., and Fort-Piat, N. L. (2007). Hysteretic q-learning : an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. *IEEE/RSJ International conference on Intelligent Robots and Systems*, pages 326–331.

[Meuleau et al., 1999] Meuleau, N., Kim, K.-E., Kaelbling, L. P., and Cassandra, A. R. (1999). Solving POMDPs by searching the space of finite policies. *Fifteenth conference on Uncertainty in artificial intelligence*, pages 417–426.

[Meyn et al., 2014] Meyn, S., Barooah, P., Bušić, A., Chen, Y., and Ehren, J. (2014). Ancillary service to the grid using intelligent deferrable loads. *arXiv preprint arXiv:1402.4600*.

[Moon and Basar, 2017] Moon, J. and Basar, T. (2017). Linear quadratic risk-sensitive and robust mean field games. *To appear in IEEE Transactions on Automatic Control*, 62(6).

[Murphey and Pardalos, 2002] Murphey, R. and Pardalos, P. M. (2002). Cooperative control and optimization. *Springer Science and Business Media, ISBN 978-0-306-47536-8*, 66.

[Murphy, 2000] Murphy, K. P. (2000). A survey of POMDP solution techniques. *Technical report, U.C. Berkeley.*

[Nayyar et al., 2011] Nayyar, A., Mahajan, A., and Teneketzis, D. (2011). Optimal control strategies in delayed sharing information structures. *IEEE Transactions on Automatic Control*, 56(7):1606–1620.

[Nayyar et al., 2013] Nayyar, A., Mahajan, A., and Teneketzis, D. (2013). Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658.

[Nourian and Caines, 2013] Nourian, M. and Caines, P. E. (2013). ϵ-nash mean field game theory for nonlinear stochastic dynamical systems with major and minor agents. *SIAM Journal on Control and Optimization*, 51(4):3302–3331.

[Nourian et al., 2010] Nourian, M., Caines, P. E., Malhamé, R. P., and Huang, M. (2010). Leader-follower Cucker-Smale type flocking synthesized via mean field stochastic control theory. *Springer, Brain, Body and Machine*, pages 283–298.

[Olfati-Saber et al., 2006] Olfati-Saber, R., Franco, E., Frazzoli, E., and Shamma, J. S. (2006). Belief consensus and distributed hypothesis testing in sensor networks. *Springer, Networked Embedded Sensing and Control*, 331:169–182.

[Ooi et al., 1997] Ooi, J. M., Verbout, S. M., Ludwig, J. T., Wornell, G. W., et al. (1997). A separation theorem for periodic sharing information patterns in decentralized control. *IEEE transactions on Automatic Control*, 42(11):1546–1550.

[Ooi and Wornell, 1996] Ooi, J. M. and Wornell, G. W. (1996). Decentralized control of a multiple access broadcast channel: performance bounds. *35th Conference on Decision and Control*, pages 293–298.

[Ouyang and Teneketzis, 2015] Ouyang, Y. and Teneketzis, D. (2015). Signaling for decentralized routing in a queueing network. *Annals of Operations Research*, pages 1–39.

[Papadimitriou and Tsitsiklis, 1987] Papadimitriou, C. H. and Tsitsiklis, J. N. (1987). The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450.

[Pineau et al., 2003] Pineau, J., Gordon, G., Thrun, S., et al. (2003). Point-based value iteration: An anytime algorithm for POMDPs. *International Joint Conference on Artificial Intelligence*, 3:1025–1032.

[Pivazyan and Shoham, 2002] Pivazyan, K. and Shoham, Y. (2002). Polynomial-time reinforcement learning of near-optimal policies. *AAAI/IAAI*, pages 205–210.

[Poupart, 2005] Poupart, P. (2005). *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*. Ph.D. thesis, University of Toronto, Toronto, Ont., Canada, Canada.

[Radner, 1962] Radner, R. (1962). Team decision problems. *The Annals of Mathematical Statistics*, pages 857–881.

[Ray et al., 2009] Ray, D., King-Casas, B., Montague, P. R., and Dayan, P. (2009). Bayesian model of behaviour in economic games. *Advances in neural information processing systems*, pages 1345–1352.

[Rotkowitz and Lall, 2004] Rotkowitz, M. and Lall, S. (2004). On computation of optimal controllers subject to quadratically invariant sparsity constraints. *IEEE American Control Conference*, 6:5659–5664.

[Salhab et al., 2015] Salhab, R., Malham, R. P., and Le Ny, J. (2015). A dynamic game model of collective choice in multi-agent systems. *54th IEEE Conference on Decision and Control (CDC)*, pages 4444–4449.

[Schoute, 1978] Schoute, F. C. (1978). Symmetric team problems and multi access wire communication. *Elsevier, Automatica*, 14(3):255–269.

[Seuken and Zilberstein, 2007] Seuken, S. and Zilberstein, S. (2007). Memory-bounded dynamic programming for dec-POMDPs. *20th international joint conference on Artificial intelligence*, pages 2009–2015.

[Shani et al., 2013] Shani, G., Pineau, J., and Kaplow, R. (2013). A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51.

[Shi et al., 2012] Shi, Z., Tu, J., Zhang, Q., Liu, L., and Wei, J. (2012). A survey of swarm robotics system. *Springer, Advances in Swarm Intelligence*, pages 564–572.

[Singh et al., 2015] Singh, R., Kumar, P. R., and Xie, L. (2015). The ISO problem: Decentralized stochastic control via bidding schemes. *arXiv:1510.00983*.

[Smallwood and Sondik, 1973] Smallwood, R. D. and Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088.

[Spaan et al., 2002] Spaan, M. T. J., Vlassis, N., and Groen, F. C. A. (2002). High level coordination of agents based on multiagent Markov decision processes with roles. *Workshop on Cooperative Robotics, IEEE/RSJ Int. conf. on Intelligent Robots and Systems*, pages 66–73.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

[Takahara, 1964] Takahara, Y. (1964). Multi-level approach to dynamic optimization. *Technical report, DTIC Document*.

[Tembine, 2014] Tembine, H. (2014). Energy-constrained mean field games in wireless networks. *Strategic Behavior and the Environment*, 4(2):187–211.

[Tsistsiklis, 1994] Tsistsiklis, J. N. (1994). Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16:185–202.

[Vlassis, 2007] Vlassis, N. (2007). *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*. Morgan and Claypool Publishers, 1st edition.

[Whittle and Rudge, 1974] Whittle, P. and Rudge, J. (1974). The optimal linear solution of a symmetric team control problem. *Journal of Applied Probability*, pages 377–381.

[Witsenhausen, 1968] Witsenhausen, H. (1968). A counterexample in stochastic optimum control. *SIAM Journal Of Control And Optimization*, 6:131–147.

[Witsenhausen, 1971] Witsenhausen, H. (1971). Separation of estimation and control for discrete time systems. *Proc. of IEEE*, 59(11):1557–1566.

[Witsenhausen, 1973] Witsenhausen, H. S. (1973). A standard form for sequential stochastic control. *Springer, Math. Sys. Theory*, 7(1):5–11.

[Wu and Antsaklis, 2010] Wu, P. and Antsaklis, P. (2010). Symmetry in the design of large-scale complex control systems: Some initial results using dissipativity and Lyapunov stability. *18th Mediterranean Conference on Conference on Control*, pages 197–202.

[Xiao and Boyd, 2004] Xiao, L. and Boyd, S. (2004). Fast linear iterations for distributed averaging. *Elsevier Systems and Control Letters*, 53(1):65–78.

[Yong, 2013] Yong, J. (2013). Linear-quadratic optimal control problems for mean-field stochastic differential equations. *SIAM on Control and Optimization*, 51(4):2809–2838.

[Yüksel, 2009] Yüksel, S. (2009). Stochastic nestedness and the belief sharing information pattern. *IEEE Transactions on Automatic Control*, 54(12):2773–2786.

[Yüksel and Başar, 2013] Yüksel, S. and Başar, T. (2013). *Stochastic Networked Control Systems*. Birkhauser.

[Yuksel and Tatikonda, 2009] Yuksel, S. and Tatikonda, S. (2009). A counterexample in distributed optimal sensing and control. *IEEE Transactions on Automatic Control*, 54(4):841–844.

[Zhang, 2010] Zhang, H. (2010). Partially observable Markov decision processes: A geometric technique and analysis. *Operations Research*, 58(1):214–228.