Team Optimal Control of Coupled Subsystems with Mean-Field Sharing

Jalal Arabneydi and Aditya Mahajan

Abstract— **We investigate team optimal control of stochastic subsystems that are weakly coupled in dynamics (through the mean-field of the system) and are arbitrary coupled in the cost. The controller of each subsystem observes its local state and the mean-field of the state of all subsystems. The system has a non-classical information structure. Exploiting the symmetry of the problem, we identify an information state and use that to obtain a dynamic programming decomposition. This dynamic program determines a globally optimal strategy for all controllers. Our solution approach works for arbitrary number of controllers 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 controllers which allows us to solve problems with moderate number of controllers. We illustrate our approach by an example motivated by smart grids that consists of 100 coupled subsystems.**

I. Introduction

A. Motivation

Team optimal control of stochastic decentralized systems arises in many applications ranging from networked control systems, robotics, communication networks, transportation networks, sensor networks, and economics. There is a long and rich history of research on team theory, starting from the work of Radner [1], [2], Witsenhausen [3]–[5] and others; and continuing to various solution approaches that have been proposed in recent years. Due to space limitations, we can not provide a detailed overview of the literature; we rather refer the reader to [6], [7] for detailed overviews.

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 subsystems. These include coordination-decomposition methods [8], [9] that use iterative message passing algorithm and mean-field games that reduce the optimal control problem to a game between an individual and the mass [10], [11], and references therein.

In this paper, we introduce a solution approach that exploits symmetry to identify a low-dimensional information state. Our approach uses two steps. In the first step, we identify an equivalent centralized system using the common information approach of [12]. In the second step, we exploit the symmetry of the system to identify an information state and use that to obtain a dynamic programming decomposition.

The rest of the paper is organized as follows. We formulate the team optimal control problem in Section I-C and identify the salient features of the problem and our contributions in Sections I-D and I-E, respectively. We present the main results in Section II, and provide some generalizations in Section III. In Section IV, we present an example (motivated by smart grid applications).

B. 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). 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.

C. Problem Formulation

Consider a discrete time decentralized control system with $n \in \mathbb{N}$ homogeneous subsystems that operate for a horizon $T \in \mathbb{N}$. The state of subsystem 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 controller i, $i \in \{1, \dots, n\}$, at time t, where $\mathcal{U}$ is a finite set (that does not depend on i).

We refer to the empirical distribution of all subsystems at time t as the *mean-field* of the system and denote it by Z_t, i.e.,

$$Z_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_t^i} \tag{1}$$

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

$$Z_t(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_t^i = x) \tag{2}$$

where δ_x denotes a Dirac measure on $\mathcal{X}$ with a point mass at x. Let $|\mathcal{X}| = k$ and $\mathcal{M}_n = \{(\frac{m_1}{n}, \frac{m_2}{n}, \dots, \frac{m_k}{n}) : m_i \in \{0, \dots, n\}, \sum_{i=1}^{k} m_i = n\}$ denote the space of realizations of Z_t. Note that $\mathcal{M}_n \subset \Delta(\mathcal{X})$, the space of probability distributions on $\mathcal{X}$.

This work was supported by the Natural Sciences and Engineering Research Council of Canada through Grant NSERC-RGPIN 402753-11.

Jalal Arabneydi and Aditya Mahajan are with Department of Electrical Engineering, McGill University, 3480 University St., Montreal, QC, Canada. Email:jalal.arabneydi@mail.mcgill.ca and Email:aditya.mahajan@mcgill.ca

1) System dynamics: The subsystems are weakly coupled with each other in dynamics via the mean-field, as described below. The initial states of all subsystems are independent and distributed according to PMF (probability mass function) P_X (that does not depend on i). The state X_t^i of subsystem i evolves according to

$$X_{t+1}^i = f_t(X_t^i, U_t^i, W_t^i, Z_t), \;\; i \in \{1, \dots, n\} \tag{3}$$

where f_t is the plant function at time t and $\{W_t^i\}_{t=1}^T$ is an independent process with probability distribution P_{W_t} at time t. Note that the plant functions $\{f_t\}_{t=1}^T$and the PMFs $\{P_{W_t}\}_{t=1}^T$ do not depend on i.

The *primitive random variables* $(X_1^1, \dots, X_1^n, \{W_t^1\}_{t=1}^T, \dots, \{W_t^n\}_{t=1}^T)$ are mutually independent and defined on a common probability space.

2) Information structure: In addition to the local state of its subsystems, each controller observes the history of the mean-field. Thus, the data available at controller i, $i \in \{1, \dots, n\}$, at time t is

$$I_t^i = \{Z_{1:t}, X_t^i\}. \tag{4}$$

We refer to this information structure as *mean-field sharing*. In Section III, we consider a generalization of this information structure in which each controller observes a noisy version of the mean-field. We refer to that information structure as *partially observed mean-field sharing*.

The control action at controller i is chosen according to

$$U_t^i = g_t^i(Z_{1:t}, X_t^i). \tag{5}$$

The function g_t^i is called the *control law* of controller i at time t. In this paper, we restrict attention to identical control laws at all controllers. In particular:

***Assumption* 1** *At any time t, the control laws at all controllers 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 controller at time t as g_t.*

In view of Assumption 1, we call the collection $\mathbf{g} = (g_1, \dots, g_T)$ of control laws over time as the *control strategy* of the system.

3) Cost-structure: The subsystems are arbitrary coupled through cost. At each time step, the system incurs a cost that depends on joint state $\mathbf{X}_t = (X_t^1, \dots, X_t^n)$ and joint action $\mathbf{U}_t = (U_t^1, \dots, U_t^n)$ that is given by

$$\ell_t(\mathbf{X}_t, \mathbf{U}_t).$$

The performance of any strategy $\mathbf{g}$ is quantified by the expected total cost

$$J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=1}^T \ell_t(\mathbf{X}_t, \mathbf{U}_t)\right] \tag{6}$$

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

4) Optimization problem: We are interested in the following optimization problem.

Problem 1 *Given the information structure in* (5)*, 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 $\{P_{W_t}\}_{t=1}^T$ on the plant disturbance, identify a control strategy $\mathbf{g}^*$ to minimize the total cost $J(\mathbf{g})$ given by* (6)*.*

The above model assumes that all subsystems 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 controllers. In other applications such as multi-robot teams, all controllers may compute the mean-field in a distributed manner using methods such as consensus-based algorithms [13], [14].

We first investigate the model where the mean field is shared perfectly and develop a solution methodology for that model. In Section III, we extend the solution methodology to a more practical model in which a noisy estimate of the mean field is observed.

D. Salient Features of the Model

Our key simplifying assumption is that all control laws are identical (Assumption 1). In general, this assumption leads to a loss in performance, as is illustrated by the example below.

Example: Consider a system with n homogeneous subsystems with control horizon $T = 2$. Let state space and action space be $\mathcal{X} = \mathcal{U} = \{1, 2, \dots, n\}$ and probability distribution of initial states be uniform on $\mathcal{X}$. Suppose that the system dynamics are given by

$$X_2^i = U_1^i, \quad i \in \{1, \dots, n\}. \tag{7}$$

Let $\ell_1(\mathbf{x}_1, \mathbf{u}_1) = 0$ and $\ell_2(\mathbf{x}_2, \mathbf{u}_2) = K \cdot \mathbb{1}(z_2 \neq \{\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\})$ where K is a positive number. The asymmetric strategy $\bar{\boldsymbol{g}} = (\bar{g}_1^1, \dots, \bar{g}_1^n)$, where $\bar{g}_1^i(z_1, x_1^i) = i$, has a cost $J(\bar{\boldsymbol{g}}) = 0$. Hence, $\bar{\boldsymbol{g}}$ is optimal. On the other hand, under any symmetric strategy, $\mathbb{P}(\mathbb{1}(Z_2 \neq \{\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\}))$ 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 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 [15], [16], [17].

In the model described above, we assume that the strategies are pure (non-randomized). In general, randomized strategies are not considered in team problems because randomization does not improve performance [18, Theorem 1.6]. However, if attention is restricted to identical strategies, randomized strategies may perform better than pure strategies [15, Theorem 2.3]. In the above model, we assume that the control strategies are pure, primarily for the ease of exposition. As explained in the conclusion, our solution methodology generalizes to randomized strategies as well.

E. Contributions

In spite of the simplification provided by Assumption 1, Problem 1 is conceptually challenging because it has a non-classical information structure [4]. In general, team optimal control problems with non-classical information structure belong to NEXP complexity class [19]. Although it is possible to get a dynamic programming decomposition for problems with non-classical information structure [5], 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 [7], but even for these models the size of the information state increases exponentially with the number of controllers.

Our key contributions in this paper 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 controllers. This allows us to solve problems with moderate number of controllers. (In Section IV, we give an example with $n = 100$ controllers).
4) The solution methodology and dynamic programming decomposition extend to the scenario where all controllers observe a noisy version of the mean-field.

II. Main Results

In this section, we use the common information approach [12] to introduce an equivalent centralized problem (Problem 2) for Problem 1. Then, we find an optimal solution for the equivalent problem and translate the obtained solution back to the solution of Problem 1.

Following [12], split the information I_t^i available to controller i into two parts: the *common information* consisting of the history $Z_{1:t}$ of the mean-field process that is observed by all controllers; and the *local information* consisting of the current state X_t^i of subsystem i. Since the size of the local information does not increase with time, the model described above has a partial history sharing information structure [12]. For such systems, the structure of optimal control strategies and a dynamic programming decomposition was proposed in [12]. If we directly use these results on our model, the information state will be a posterior distribution on the global state $\mathbf{X}_t = (X_t^1, \dots, X_t^n)$ of the system. As such the complexity of the solution increases doubly exponentially with the number of controllers.

To circumvent this issue, we proceed as follows.

Step 1: We follow the common information approach proposed in [12] to convert the decentralized control problem into a centralized control problem from the point of view of a controller that observes the common information $Z_{1:t}$.

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

The details of each of these steps are presented below.

A. Step 1: An Equivalent Centralized System

Following [12], 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 Z_t and chooses a mapping $\Gamma_t : \mathcal{X} \to \mathcal{U}$ as follows

$$\Gamma_t = \psi_t(Z_{1:t}). \tag{8}$$

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 controllers. Each controller in the coordinated system is a passive agent that uses its local state X_t^i and the mapping Γ_t to generate

$$U_t^i = \Gamma_t(X_t^i), \;\; i \in \{1, \dots, n\}. \tag{9}$$

The dynamics of each subsystem and the cost function are the same as in the original problem. By a slight abuse of notation, define

$$\ell_t(\mathbf{X}_t, \Gamma_t) := \ell_t(\mathbf{X}_t, \Gamma_t(X_t^1), \dots, \Gamma_t(X_t^n)). \tag{10}$$

The performance of any coordination strategy is quantified by the total expected cost

$$\hat{J}(\boldsymbol{\psi}) = \mathbb{E}^{\boldsymbol{\psi}}[\sum_{t=1}^{T} \ell_t(\mathbf{X}_t, \Gamma_t)] \tag{11}$$

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 2 *Given the information structure in* (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 $\{P_{W_t}\}_{t=1}^T$ on the plant disturbance, identify a control strategy $\boldsymbol{\psi}^*$ to minimize the total cost $\hat{J}(\boldsymbol{\psi})$ given by* (11).

Lemma 1 ([12], Proposition 3) *Problem 1 and Problem 2 are equivalent.*

In particular, for any control strategy $\boldsymbol{g} = (g_1, \dots, g_T)$ in Problem 1, define a coordination strategy $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)$ in Problem 2 by

$$\psi_t(z_{1:t}) := g_t(z_{1:t}, \cdot), \quad \forall z_{1:t}. \tag{12}$$

Then, $J(\boldsymbol{g}) = \hat{J}(\boldsymbol{\psi})$. Similarly for any coordination strategy $\boldsymbol{\psi}$ in Problem 2, define a control strategy $\boldsymbol{g}$ in Problem 1 by

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

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

B. Step 2: Identifying an Information State and Dynamic Program

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

Lemma 2 *For any choice $\gamma_{1:t}$ of $\Gamma_{1:t}$, any realization $z_{1:t}$ of $Z_{1:t}$, and any $\mathbf{x} \in \mathcal{X}^n$,*

$$\begin{aligned}\mathbb{P}(\mathbf{X}_t{=}\mathbf{x}|Z_{1:t}{=}z_{1:t},\Gamma_{1:t}{=}\gamma_{1:t}) &= \mathbb{P}(\mathbf{X}_t{=}\mathbf{x}|Z_t{=}z_t)\\ &= \frac{1}{|H(z_t)|}\mathbb{1}(\mathbf{x} \in H(z_t))\end{aligned}$$

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

Proof outline: To prove the result, it is sufficient to show that $\mathbb{P}(\mathbf{X}_t = \mathbf{x}|Z_{1:t} = z_{1:t}, \Gamma_{1:t} = \gamma_{1:t})$ is indifferent to permutation of $\mathbf{x}$. The latter can be proved using the symmetry of the model and the control laws. ■

Using this result, we can show that

Lemma 3 *The expected per-step cost may be written as a function of Z_t and Γ_t. In particular, there exits a function $\hat{\ell}_t$ (that does not depend on strategy $\boldsymbol{\psi}$) such that*

$$\mathbb{E}[\ell_t(\mathbf{X}_t, \Gamma_t)|Z_{1:t}, \Gamma_{1:t}] =: \hat{\ell}_t(Z_t, \Gamma_t).$$

Proof outline: Consider

$$\begin{aligned}&\mathbb{E}[\ell_t(\mathbf{X}_t, \Gamma_t)|Z_{1:t} = z_{1:t}, \Gamma_{1:t} = \gamma_{1:t}]\\ &= \sum_{\mathbf{x}} \ell_t(\mathbf{x}, \gamma_t)\mathbb{P}(\mathbf{X}_t = \mathbf{x}|Z_{1:t} = z_{1:t}, \Gamma_{1:t} = \gamma_{1:t}).\end{aligned}$$

Substituting the result of Lemma 2, and simplifying gives the result. ■

Lemma 4 *For any choice $\gamma_{1:t}$ of $\Gamma_{1:t}$, any realization $z_{1:t}$ of $Z_{1:t}$, and any $z \in \mathcal{M}_n$,*

$$\mathbb{P}(Z_{t+1}{=}z|Z_{1:t}{=}z_{1:t}, \Gamma_{1:t}{=}\gamma_{1:t}){=}\mathbb{P}(Z_{t+1}{=}\ z|Z_t{=}z_t, \Gamma_t{=}\gamma_t).$$

Also, above conditional probability does not depend on strategy $\boldsymbol{\psi}$.

Proof outline: The result relies on the independence of the noise processes across subsystems and Lemma 2. ■

Based on the results in steps 1 and 2, we have that

Theorem 1 *In Problem 2, there is no loss of optimality in restricting attention to Markovian strategy i.e. $\Gamma_t = \psi_t(Z_t)$. Furthermore, an optimal strategy $\boldsymbol{\psi}^*$ is obtained by solving the following dynamic program. Define recursively value functions:*

$$V_{T+1}(z_{T+1}) := 0, \quad \forall z_{T+1} \in \mathcal{M}_n \tag{13}$$

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

$$V_t(z_t) := \min_{\gamma_t}(\hat{\ell}_t(z_t, \gamma_t) + \mathbb{E}[V_{t+1}(Z_{t+1})|Z_t = z_t, \Gamma_t = \gamma_t]) \tag{14}$$

where the minimization is over all functions $\gamma_t : \mathcal{X} \to \mathcal{U}$. Let $\psi_t^*(z_t)$ denote any argmin of the right-hand side of (14)*. Then, the coordination strategy $\boldsymbol{\psi}^* = (\psi_1^*, \dots, \psi_T^*)$ is optimal.*

Proof: Z_t is an information state for Problem 2 because:
1) As shown in Lemma 3, the per-step cost can be written as a function of Z_t and Γ_t.
2) As shown in Lemma 4, $\{Z_t\}_{t=1}^T$ is a controlled Markov process with control action Γ_t.
Thus, the result follows from standard results in Markov decision theory [20]. ■

Based on the equivalence in Lemma 1, we get

Corollary 1 *Let $\psi_t^*(z)$ be a minimizer of* (14) *at time t. Define*

$$g_t^*(z, x) := \psi_t^*(z)(x). \tag{15}$$

Then, $\mathbf{g}^*{=}(g_1^*, \dots, g_T^*)$ is an optimal strategy for Problem 1.

Remark 1 The fictitious coordinated system is described only for ease of exposition. The dynamic program of (13) and (14) uses z_t as the information state. Since z_t is observed by each controller, each controller can independently solve the dynamic program; agreeing upon a deterministic rule to break ties while using argmin ensures that all controllers compute the same optimal strategy.

Remark 2 The space $\mathcal{M}_n$ of realization of z_t is finite and has cardinality less than $(n+1)^{|\mathcal{X}|}$. Thus, the solution complexity increases polynomially with the number of controllers.

III. Generalization To Partially Observed Mean-Field Sharing

In this section, we consider a case where mean-field is not completely observable. Let $Y_t \in \mathcal{Y}$ be a noisy measurement of Z_t at time t as follows:

$$Y_t = h_t(Z_t, N_t) \tag{16}$$

where N_t is a random variable which takes value on a finite set $\mathcal{N}$. $\{N_t\}_{t=1}^T$ is an independent random process with PMF P_{N_t}, at time t, and is also mutually independent from all primitive random variables in Section I-C.1. Similar to (5), we consider the following information structure:

$$U_t^i = g_t(Y_{1:t}, X_t^i), \ \ i \in \{1, \dots, n\} \tag{17}$$

where $g_t : \mathcal{Y}^t \times \mathcal{X} \to \mathcal{U}$.

Problem 3 *Given the information structure in* (17)*, 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 $\{P_{N_t}\}_{t=1}^T$ on observation noise, and the PMFs $\{P_{W_t}\}_{t=1}^T$ on the plant disturbance, identify a control strategy $\mathbf{g}^*$ to minimize the total cost $J(\mathbf{g})$ given by* (6)*.*

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

$$\Gamma_t = \psi_t(Y_{1:t}). \tag{18}$$

The rest of the setup is same as before. Similar to Problem 2, we get

Problem 4 *Given the information structure in* (18)*, 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 $\{P_{N_t}\}_{t=1}^T$ on observation noise, and the PMFs $\{P_{W_t}\}_{t=1}^T$ on the plant disturbance, identify a control strategy $\boldsymbol{\psi}^*$ to minimize the total cost $\hat{J}(\boldsymbol{\psi})$ given by* (11).

As in Lemma 1, Problem 3 is equivalent to Problem 4. In particular, for any control strategy $\boldsymbol{g} = (g_1, \dots, g_T)$ in Problem 3, one can construct a coordination strategy $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)$ in Problem 4 that yields the same performance and vice versa.

In step 2, we show that $\Pi_t(z) := \mathbb{P}(Z_t = z | Y_{1:t}, \Gamma_{1:t-1})$ is an information state for Problem 4. In particular:

Lemma 5 *There exists a function $\tilde{\ell}_t$ (that does not depend on strategy $\boldsymbol{\psi}$) such that*

$$\mathbb{E}[\ell_t(\mathbf{X}_t, \Gamma_t) | Y_{1:t}, \Gamma_{1:t}] =: \tilde{\ell}_t(\Pi_t, \Gamma_t). \tag{19}$$

Lemma 6 *There exists a function ϕ_t (that does not depend on strategy $\boldsymbol{\psi}$) such that*

$$\Pi_{t+1} = \phi_t(\Pi_t, \Gamma_t, Y_{t+1}). \tag{20}$$

Proofs of Lemma 5 and Lemma 6 are omitted due to lack of space. Similar to Theorem 1, we have that

Theorem 2 *In Problem 4, there is no loss of optimality in restricting attention to Markovian strategy i.e. $\Gamma_t = \psi_t(\Pi_t)$. Also, optimal strategy $\boldsymbol{\psi}^*$ is obtained by solving the following dynamic program. Let $\Delta(\mathcal{M}_n)$ denote the space of probability distributions on $\mathcal{M}_n$. Define recursively value functions:*

$$V_{T+1}(\pi_{T+1}) = 0, \quad \forall \pi_{T+1} \in \Delta(\mathcal{M}_n) \tag{21}$$

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

$$V_t(\pi_t) = \min_{\gamma_t} (\tilde{\ell}_t(\pi_t, \gamma_t) + \mathbb{E}[V_{t+1}(\Pi_{t+1}) | \Pi_t = \pi_t, \Gamma_t = \gamma_t]) \tag{22}$$

where the minimization is over all functions $\gamma_t : \mathcal{X} \to \mathcal{U}$. Let $\psi_t^*(\pi_t)$ denote any argmin of the right-hand side of (22)*. Then, the coordination strategy $\boldsymbol{\psi}^* = (\psi_1^*, \dots, \psi_T^*)$ is optimal.*

Proof: Π_t is an information state for Problem 4 because:
1) As shown in Lemma 5, the expected per-step cost can be written as a function of Π_t and Γ_t.
2) As shown in Lemma 6, $\{\Pi_t\}_{t=1}^T$ is a controlled Markov process with control action Γ_t.
Thus, the result follows from standard results in Markov decision theory [20]. ■

Based on the equivalence between Problem 3 and Problem 4, we get

Corollary 2 *Let $\psi_t^*(\pi)$ be a minimizer of* (22) *at time t. Define*

$$g_t^*(\pi, x) := \psi_t^*(\pi)(x). \tag{23}$$

Then, $\boldsymbol{g}^* = (g_1^*, \dots, g_T^*)$ is an optimal strategy for Problem 3.

IV. An Example

In this section we consider an example of mean-field sharing 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.

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

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

$$\ell_t(\mathbf{X}_t, \mathbf{U}_t) = \frac{1}{n} \sum_{i=1}^n c(U_t^i) + D(Z_t \parallel \zeta)$$

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 given by (4). The objective is to choose a control strategy to minimize the infinite horizon discounted cost[1]

$$J(\boldsymbol{g}) = \mathbb{E}\left[\sum_{t=1}^\infty \beta^t \left(\frac{1}{n} \sum_{i=1}^n c(U_t^i) + D(Z_t \parallel \zeta)\right)\right] \tag{24}$$

where $\beta \in (0, 1)$ is the discounted factor.

A more elaborate variation of the above model is considered in [21] for controlling the operation of pool pumps.

Consider the above model for the following parameters

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

The optimal time-homogeneous strategy for these parameters is shown in Fig. 1. Since state space is binary, $z(1)$ is sufficient to characterise the empirical distribution $z = [z(1), z(2)]$. Hence, for ease of presentation, we plot the optimal control law and value function as a function of the first component $z(1)$ of $z = [z(1), z(2)]$.

[1] Although we have only presented the details for finite horizon setup in this paper, the results generalize naturally to infinite horizon setup under standard assumptions. See Section V-B for a brief explanation.

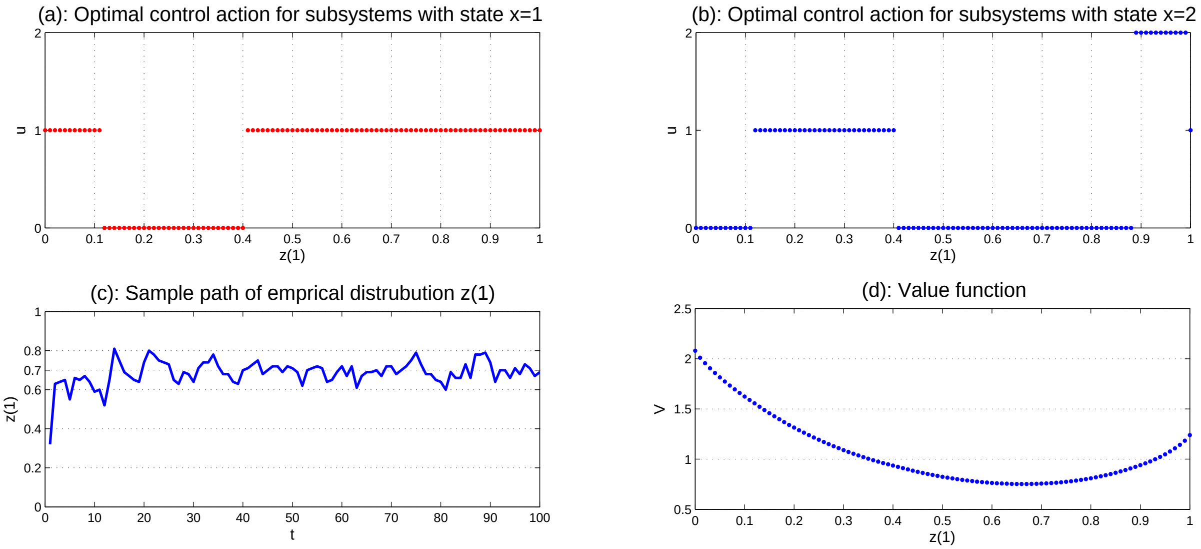

Fig. 1. Plots (a) and (b) show the optimal strategy as a function of $z(1)$. Plot (c) shows the sample path of $z(1)$ for simulation time of 100. Plot (d) depicts the value function with respect to $z(1)$.

V. Conclusion

In this paper, we considered the team optimal control of decentralized systems with mean-field sharing. We follow a two-step approach: in the first step we construct an equivalent centralized system using the common information approach of [12]; in the second step, we exploit the symmetry of the system to identify information state and dynamic programming decomposition of the problem. We generalize our result to the case of partial observation of the mean field. We illustrate our results using an example motivated by smart grids. Our results extend naturally to the following setups.

A. Randomized Strategies

As mentioned earlier, if attention is restricted to identical control laws, then randomized strategies may perform better than pure strategies [15, Theorem 2.3]. Our results extend naturally to randomized strategies by considering $\Delta(\mathcal{U})$, the space of probability distributions on $\mathcal{U}$, as the action space.

B. Infinite Horizon

The results of Lemma 3 and Lemma 4 are valid for the infinite horizon setup as well. Hence, the results of Theorem 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.

C. Multiple Types of Subsystems

We assumed that all subsystems are homogeneous. Consider a setup where subsystem i has a type $\theta^i, \theta^i \in \Theta$, and the dynamics are given by $X^i_{t+1} = f_t(\theta^i, X^i_t, U^i_t, W^i_t, Z_t)$. Our results generalize to such a setup with $Z_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{X^i_t, \theta^i}$.

References

[1] R. Radner, "Team decision problems," *JSTOR, The Annals of Mathematical Statistics*, pp. 857–881, Sept. 1962.

[2] J. Marschack and R. Radner, "Economic theory of teams," *New Haven: Yale University Press*, Feb. 1972.

[3] H. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal Of Cont. And Opt.*, vol. 6, pp. 131–147, Dec. 1968.

[4] ——, "Separation of estimation and control for discrete time systems," *Proc. of IEEE*, vol. 59, no. 11, pp. 1557–1566, Nov. 1971.

[5] H. S. Witsenhausen, "A standard form for sequential stochastic control," *Springer, Math. Sys. Theory*, vol. 7, no. 1, pp. 5–11, Mar. 1973.

[6] S. Yüksel and T. Başar, *Stochastic Networked Control Systems*. Birkhauser, 2013.

[7] A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yuksel, "Information structures in optimal decentralized control," *Proc. of Conf. on Decision and Control (CDC)*, pp. 1291–1306, Dec. 2012.

[8] J. C. Culioli and G. Cohen, "Decomposition/coordination algorithms in stochastic optimization," *SIAM Journal on Control and Optimization*, vol. 28, no. 6, pp. 1372–1403, Nov. 1990.

[9] K. Barty, P. Carpentier, and P. Girardeau, "Decomposition of large-scale stochastic optimal control problems," *RAIRO-Operations Research, Cambridge Univ Press*, vol. 44, no. 03, pp. 167–183, Jul. 2010.

[10] P. Caines, "Mean field games," *In: Samad T., Baillieul J. (Ed.) Encyclopedia of Systems and Control: SpringerReference, Springer-Verlag Berlin Heidelberg*, Oct. 2013.

[11] D. A. Gomes and J. Saude, "Mean field games models: A brief survey," *Springer, Dynamic Games and Appl.*, vol. 4, pp. 1–45, Jun. 2014.

[12] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Tran. on Automatic Control*, vol. 58, no. 7, Jul. 2013.

[13] R. Olfati-Saber, E. Franco, E. Frazzoli, and J. S. Shamma, "Belief consensus and distributed hypothesis testing in sensor networks," *Springer, Networked Embedded Sensing and Control*, vol. 331, pp. 169–182, Jul. 2006.

[14] A. N. Bishop and A. Doucet, "Distributed nonlinear consensus in the space of probability measures," *arXiv preprint arXiv:1404.0145*, 2014.

[15] F. C. Schoute, "Symmetric team problems and multi access wire communication," *Elsevier, Automatica*, vol. 14, no. 3, pp. 255–269, May 1978.

[16] Z. Shi, J. Tu, Q. Zhang, L. Liu, and J. Wei, "A survey of swarm robotics system," pp. 564–572, Jun. 2012.

[17] P. Wu and P. Antsaklis, "Symmetry in the design of large-scale complex control systems: Some initial results using dissipativity and Lyapunov stability," *Med. Conf. on Cont.*, pp. 197–202, Jun. 2010.

[18] I. I. Gihman and A. Skorohod, *Controlled stochastic processes*. Springer, 1979.

[19] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *INFORMS, Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, Nov. 2002.

[20] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2012.

[21] S. Meyn, P. Barooah, A. Bušić, Y. Chen, and J. Ehren, "Ancillary service to the grid using intelligent deferrable loads," *arXiv preprint arXiv:1402.4600*, Feb. 2014.

Appendix I
Proof of Lemma 2

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 $H(z) := \{\mathbf{x} \in \mathcal{X}^n : \frac{1}{n}\sum_{i=1}^n \delta_{x^i} = z\}$ as a set of all joint states $\mathbf{x} \in \mathcal{X}^n$ whose empirical distribution is z. Thus, at time t, we have

$$\mathbb{1}(z_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_t^i}) = \mathbb{1}(\mathbf{x}_t \in H(z_t)). \tag{25}$$

Notice that if $\mathbf{x}_t \in H(z_t)$, then one can interpret $H(z_t)$ 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

$$\begin{aligned}\mathbb{P}(\mathbf{x}_1|z_1,\gamma_1) &\overset{(a)}{=} \mathbb{P}(\mathbf{x}_1|z_1) \overset{(b)}{=} \frac{\mathbb{P}(z_1|\mathbf{x}_1)\mathbb{P}(\mathbf{x}_1)}{\mathbb{P}(z_1)} \\ &= \frac{\mathbb{1}(z_1 = \frac{1}{n}\sum_{i=1}^n \delta_{x_1^i})\mathbb{P}(\mathbf{x}_1)}{\mathbb{P}(z_1)}\end{aligned} \tag{26}$$

where (a) follows from the fact that $\gamma_1 = \psi_1(z_1)$ according to (8) and (b) follows from Bayes rule. From (25) and (26), given $\{z_1, \gamma_1\}$, we get

$$\mathbb{P}(\mathbf{x}_1|z_1,\gamma_1) = \begin{cases} 0 & \mathbf{x}_1 \notin H(z_1) \\ \alpha(z_1) & \mathbf{x}_1 \in H(z_1) \end{cases} \tag{27}$$

where $\alpha(z_1) = \frac{\mathbb{P}(\mathbf{x}_1)}{z_1}$ depends only on z_1. The reason lies in the fact that, when $\mathbf{x}_1 \in H(z_1)$, $H(z_1)$ 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 $\mathbb{P}(\mathbf{x}_1|z_1,\gamma_1)$ over $\mathbf{x}_1 \in \mathcal{X}^n$ is one, we have

$$\alpha(z_1) = \frac{1}{|H(z_1)|}. \tag{28}$$

From (27) and (28), we have

$$\mathbb{P}(\mathbf{x}_1|z_1,\gamma_1) = \mathbb{P}(\mathbf{x}_1|z_1) = \frac{\mathbb{1}(\mathbf{x}_1 \in H(z_1))}{|H(z_1)|}. \tag{29}$$

Hence, the result holds for $t=1$. Assume the result holds for step t i.e.

$$\mathbb{P}(\mathbf{x}_t|z_{1:t},\gamma_{1:t}) = \mathbb{P}(\mathbf{x}_t|z_t) = \frac{\mathbb{1}(\mathbf{x}_t \in H(z_t))}{|H(z_t)|}. \tag{30}$$

We prove that the result holds for step $t+1$ as follows.

$$\begin{aligned}&\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t+1},\gamma_{1:t+1}) \overset{(a)}{=} \mathbb{P}(\mathbf{x}_{t+1}|z_{1:t+1},\gamma_{1:t}) \\ &\overset{(b)}{=} \frac{\mathbb{P}(z_{t+1}|\mathbf{x}_{t+1})\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t},\gamma_{1:t})}{\mathbb{P}(z_{t+1}|z_{1:t},\gamma_{1:t})} \\ &= \frac{\mathbb{1}(z_{t+1} = \frac{1}{n}\sum_{i=1}^n \delta_{x_{t+1}^i})\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t},\gamma_{1:t})}{\mathbb{P}(z_{t+1}|z_{1:t},\gamma_{1:t})}\end{aligned} \tag{31}$$

where (a) follows from the fact that $\gamma_{t+1} = \psi_{t+1}(z_{1:t+1})$ according to (8) and (b) follows from Bayes rule. Similar to step $t=1$, we show that, given $\{z_{1:t+1}, \gamma_{1:t+1}\}$, above conditional probability is insensitive to permutation of $\mathbf{x}_{t+1}$. For that matter, we write the conditional probability in the numerator of (31) as follows.

$$\begin{aligned}&\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t},\gamma_{1:t}) = \sum_{\mathbf{x}_t} \mathbb{P}(\mathbf{x}_{t+1},\mathbf{x}_t|z_{1:t},\gamma_{1:t}) \\ &= \sum_{\mathbf{x}_t} \mathbb{P}(\mathbf{x}_{t+1}|\mathbf{x}_t,z_{1:t},\gamma_{1:t})\mathbb{P}(\mathbf{x}_t|z_{1:t},\gamma_{1:t}) \\ &\overset{(a)}{=} \sum_{\mathbf{x}_t,\mathbf{w}_t} \left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^i = f_t(x_t^i, \gamma_t(x_t^i), w_t^i, z_t))\right] \\ &\cdot \mathbb{P}(\mathbf{w}_t) \cdot \mathbb{P}(\mathbf{x}_t|z_{1:t},\gamma_{1:t})\end{aligned} \tag{32}$$

where (a) follows from (3) and the fact that $\mathbf{W}_t$ is independent from all data and decisions made before time t. Let $S := \sigma(1,\dots,n)$ denote an arbitrary permutation of set $\{1,\dots,n\}$ 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 $\mathbf{x}_{t+1}^S = (x_{t+1}^{S(1)},\dots,x_{t+1}^{S(n)})$ as permuted version of $\mathbf{x}_{t+1}$ with respect to S. Now, consider

$$\begin{aligned}&\mathbb{P}(\mathbf{x}_{t+1}^S|z_{1:t},\gamma_{1:t}) = \sum_{\mathbf{x}_t,\mathbf{w}_t} \left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^{S(i)} = f_t(x_t^i, \gamma_t(x_t^i), w_t^i, z_t))\right] \\ &\cdot \mathbb{P}(\mathbf{w}_t) \cdot \mathbb{P}(\mathbf{x}_t|z_{1:t},\gamma_{1:t}) \\ &\overset{(a)}{=} \sum_{\mathbf{x}_t,\mathbf{w}_t} \left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^{S(i)} = f_t(x_t^{S(i)}, \gamma_t(x_t^{S(i)}), w_t^{S(i)}, z_t))\right] \\ &\cdot \mathbb{P}(\mathbf{w}_t^S) \cdot \mathbb{P}(\mathbf{x}_t^S|z_{1:t},\gamma_{1:t}),\end{aligned} \tag{33}$$

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

$$\sum_{\mathbf{x},\mathbf{w}} D(\mathbf{x},\mathbf{w}) = \sum_{\mathbf{x}^S,\mathbf{w}^S} D(\mathbf{x}^S,\mathbf{w}^S) = \sum_{\mathbf{x},\mathbf{w}} D(\mathbf{x}^S,\mathbf{w}^S). \tag{34}$$

Now, we consider terms in (33) separately as follows.

A) Since multiplication is insensitive to permutation, the first term may be written as follows.

$$\prod_{i=1}^n \mathbb{1}(x_{t+1}^{S(i)} = f_t(x_t^{S(i)}, \gamma_t(x_t^{S(i)}), w_t^{S(i)}, z_t)) = \prod_{i=1}^n \mathbb{1}(x_{t+1}^i = f_t(x_t^i, \gamma_t(x_t^i), w_t^i, z_t))$$

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 (30), the third term may be written as follows.

$$\mathbb{P}(\mathbf{x}_t^S|z_{1:t},\gamma_{1:t}) = \frac{\mathbb{1}(\mathbf{x}_t^S \in H(z_t))}{|H(z_t)|} = \frac{\mathbb{1}(\mathbf{x}_t \in H(z_t))}{|H(z_t)|} = \mathbb{P}(\mathbf{x}_t|z_{1:t},\gamma_{1:t}).$$

Substituting (A), (B), and (C) in (33), we get

$$\begin{aligned}&\mathbb{P}(\mathbf{x}_{t+1}^S|z_{1:t},\gamma_{1:t}) = \sum_{\mathbf{x}_t,\mathbf{w}_t} \left[\prod_{i=1}^n \mathbb{1}(x_{t+1}^i = f_t(x_t^i, \gamma_t(x_t^i), w_t^i, z_t))\right] \\ &\cdot \mathbb{P}(\mathbf{w}_t) \cdot \mathbb{P}(\mathbf{x}_t|z_{1:t},\gamma_{1:t}) \overset{(b)}{=} \mathbb{P}(\mathbf{x}_{t+1}|z_{1:t},\gamma_{1:t})\end{aligned} \tag{35}$$

where (b) follows from (32).

The rest of the proof is similar to that of step $t = 1$. From (31) and (35), given $\{z_{1:t+1}, \gamma_{1:t+1}\}$, we get

$$\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t+1}, \gamma_{1:t+1}) = \begin{cases} 0 & \mathbf{x}_{t+1} \notin H(z_{t+1}) \\ \alpha(z_{t+1}) & \mathbf{x}_{t+1} \in H(z_{t+1}) \end{cases} \tag{36}$$

where $\alpha(z_{t+1}) = \frac{\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t},\gamma_{1:t})}{\mathbb{P}(z_{t+1}|z_{1:t},\gamma_{1:t})}$ depends only on z_{t+1} because, when $\mathbf{x}_{t+1} \in H(z_{t+1})$, $H(z_{t+1})$ contains nothing but permutations of $\mathbf{x}_{t+1}$ while $\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t}, \gamma_{1:t})$ is insensitive to permutation of $\mathbf{x}_{t+1}$ according to (35). Since the summation of $\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t+1}, \gamma_{1:t+1})$ over $\mathbf{x}_{t+1} \in \mathcal{X}^n$ is one, we have

$$\alpha(z_{t+1}) = \frac{1}{|H(z_{t+1})|}. \tag{37}$$

From (36) and (37), we have

$$\mathbb{P}(\mathbf{x}_{t+1}|z_{1:t+1}, \gamma_{1:t+1}) = \mathbb{P}(\mathbf{x}_{t+1}|z_{t+1}) = \frac{\mathbb{1}(\mathbf{x}_{t+1} \in H(z_{t+1}))}{|H(z_{t+1})|}.$$

■

Appendix II
Proof of Lemma 3

Consider the conditional expected cost at time t given $\{z_{1:t}, \gamma_{1:t}\}$ as follows

$$\begin{aligned} &\mathbb{E}[\ell_t(\mathbf{X}_t, \Gamma_t)|Z_{1:t} = z_{1:t}, \Gamma_{1:t} = \gamma_{1:t}] \\ &= \sum_{\mathbf{x}_t} \ell_t(\mathbf{x}_t, \gamma_t)\mathbb{P}(\mathbf{X}_t = \mathbf{x}_t|Z_{1:t} = z_{1:t}, \Gamma_{1:t} = \gamma_{1:t}) \\ &\overset{(a)}{=} \sum_{\mathbf{x}_t} \ell_t(\mathbf{x}_t, \gamma_t)\frac{\mathbb{1}(\mathbf{x}_t \in H(z_t))}{|H(z_t)|} =: \hat{\ell}_t(z_t, \gamma_t) \end{aligned}$$

where (a) follows from Lemma 2. Note that none of the above terms depend on strategy $\boldsymbol{\psi}$. ■

Appendix III
Proof of Lemma 4

For the sake of notational convenience, we denote $\mathbb{P}(A = a|B = b, C = c)$ by $\mathbb{P}(a|b, c)$ and $\mathcal{F}_t$ as the event $\{Z_{1:t} = z_{1:t}, \Gamma_{1:t} = \gamma_{1:t}\}$. Let $\alpha \in [0, 1]$ and $x \in \mathcal{X}$, then we have

$$\begin{aligned} &\mathbb{P}(Z_{t+1}(x) = \alpha|\mathcal{F}_t) = \mathbb{P}\left(\frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{1}(X^i_{t+1} = x)\right) = \alpha \,\middle|\, \mathcal{F}_t\right) \\ &\overset{(a)}{=} \sum_{\mathbf{w},\mathbf{x}} \mathbb{1}\left(\frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{1}(f_t(x^i, \gamma_t(x^i), w^i, z_t) = x)\right) = \alpha\right) \\ &\cdot \mathbb{P}(\mathbf{W}_t = \mathbf{w}, \mathbf{X}_t = \mathbf{x}|\mathcal{F}_t) \\ &\overset{(b)}{=} \sum_{\mathbf{w},\mathbf{x}} \mathbb{1}\left(\frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{1}(f_t(x^i, \gamma_t(x^i), w^i, z_t) = x)\right) = \alpha\right) \\ &\cdot \left[\prod_{i=1}^{n} \mathbb{P}(W^i_t = w^i)\right] \mathbb{P}(\mathbf{X}_t = \mathbf{x}|\mathcal{F}_t) \\ &\overset{(c)}{=} \sum_{\mathbf{w},\mathbf{x}} \mathbb{1}\left(\frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{1}(f_t(x^i, \gamma_t(x^i), w^i, z_t) = x)\right) = \alpha\right) \\ &\cdot \left[\prod_{i=1}^{n} \mathbb{P}(W^i_t = w^i)\right] \mathbb{P}(\mathbf{X}_t = \mathbf{x}|Z_t = z_t) \\ &= \mathbb{P}(Z_{t+1}(x) = \alpha|Z_t = z_t, \Gamma_t = \gamma_t) \end{aligned} \tag{38}$$

where (a) follows from (3) and the fact that $\mathbf{W}_t$ is independent from all data and decisions made before time t, and (c) follows from Lemma 2. In addition, none the terms in (38) depend on strategy $\boldsymbol{\psi}$. ■

Appendix IV
Proof of Lemma 5

Consider the conditional expectation of per-step cost

$$\begin{aligned} &\mathbb{E}[\ell_t(\mathbf{X}_t, \Gamma_t)|Y_{1:t} = y_{1:t}, \Gamma_{1:t} = \gamma_{1:t}] \\ &\overset{(a)}{=} \mathbb{E}[\hat{\ell}_t(Z_t, \Gamma_t)|Y_{1:t} = y_{1:t}, \Gamma_{1:t} = \gamma_{1:t}] \\ &\overset{(b)}{=} \mathbb{E}[\hat{\ell}_t(Z_t, \Gamma_t)|Y_{1:t} = y_{1:t}, \Gamma_{1:t} = \gamma_{1:t}, \Pi_{1:t} = \pi_{1:t}] \\ &= \sum_{z_t} \hat{\ell}_t(z_t, \gamma_t)\mathbb{P}(Z_t{=}z_t|Y_{1:t}{=}y_{1:t}, \Gamma_{1:t}{=}\gamma_{1:t}, \Pi_{1:t}{=}\pi_{1:t}) \\ &= \sum_{z_t} \hat{\ell}_t(z_t, \gamma_t)\pi_t(z_t) =: \tilde{\ell}_t(\pi_t, \gamma_t) \end{aligned}$$

where (a) follows from Lemma 3 and (b) follows from the fact that $\Pi_{1:t}$ is a function of $\{Y_{1:t}, \Gamma_{1:t}\}$. Note that none of the above terms depend on strategy $\boldsymbol{\psi}$. ■

Appendix V
Proof of Lemma 6

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

$$\begin{aligned} &\pi_{t+1}(z_{t+1}) = \mathbb{P}(z_{t+1}|y_{1:t+1}, \gamma_{1:t}) \overset{(a)}{=} \mathbb{P}(z_{t+1}|y_{1:t+1}, \gamma_{1:t}, \pi_{1:t}) \\ &= \frac{\mathbb{P}(z_{t+1}, y_{t+1}|y_{1:t}, \gamma_{1:t}, \pi_{1:t})}{\sum_{\tilde{z}_{t+1}} \mathbb{P}(\tilde{z}_{t+1}, y_{t+1}|y_{1:t}, \gamma_{1:t}, \pi_{1:t})} \\ &= \frac{\mathbb{P}(y_{t+1}|z_{t+1}, y_{1:t}, \gamma_{1:t}, \pi_{1:t})\mathbb{P}(z_{t+1}|y_{1:t}, \gamma_{1:t}, \pi_{1:t})}{\sum_{\tilde{z}_{t+1}} \mathbb{P}(y_{t+1}|\tilde{z}_{t+1}, y_{1:t}, \gamma_{1:t}, \pi_{1:t})\mathbb{P}(\tilde{z}_{t+1}|y_{1:t}, \gamma_{1:t}, \pi_{1:t})} \end{aligned} \tag{39}$$

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

$$\begin{aligned} &\mathbb{P}(y_{t+1}|\tilde{z}_{t+1}, y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \\ &\overset{(b)}{=} \sum_{n_{t+1}} \mathbb{P}_{N_{t+1}}(n_{t+1})\mathbb{1}(y_{t+1}{=}h_t(\tilde{z}_{t+1}, n_{t+1})) = \mathbb{P}(y_{t+1}|\tilde{z}_{t+1}) \end{aligned} \tag{40}$$

where (b) follows from (16). The second term can be simplified as

$$\begin{aligned} \mathbb{P}(\tilde{z}_{t+1}|y_{1:t}, \gamma_{1:t}, \pi_{1:t}) &= \sum_{\tilde{z}_t} \mathbb{P}(\tilde{z}_{t+1}, \tilde{z}_t|y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \\ &= \sum_{\tilde{z}_t} \mathbb{P}(\tilde{z}_{t+1}|\tilde{z}_t, y_{1:t}, \gamma_{1:t}, \pi_{1:t})\mathbb{P}(\tilde{z}_t|y_{1:t}, \gamma_{1:t}, \pi_{1:t}) \\ &\overset{(c)}{=} \sum_{\tilde{z}_t} \mathbb{P}(\tilde{z}_{t+1}|\tilde{z}_t, \gamma_t)\pi_t(\tilde{z}_t) =: \mathbb{P}(\tilde{z}_{t+1}|\pi_t, \gamma_t) \end{aligned} \tag{41}$$

where (c) follows from Lemma 4 and definition of Π_t. From (39), (40), and (41), we have

$$\begin{aligned} \pi_{t+1}(z_{t+1}) &= \frac{\mathbb{P}(y_{t+1}|z_{t+1})\mathbb{P}(z_{t+1}|\pi_t, \gamma_t)}{\sum_{\tilde{z}_{t+1}} \mathbb{P}(y_{t+1}|\tilde{z}_{t+1})\mathbb{P}(\tilde{z}_{t+1}|\pi_t, \gamma_t)} \\ &=: \phi_t(z_{t+1}, \pi_t, \gamma_t, y_{t+1}) \end{aligned} \tag{42}$$

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