Les Cahiers du GERAD ISSN: 0711–2440

Decentralized stochastic control

A. Mahajan
M. Mannan

G–2014–87

November 2014

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Decentralized stochastic control

Aditya Mahajan

Mehnaz Mannan

GERAD & Electrical and Computer Engineering, McGill University, Montréal (Québec) Canada, H3A 0E9

aditya.mahajan@mcgill.ca
mehnaz.mannan@mail.mcgill.ca

November 2014

Les Cahiers du GERAD

G–2014–87

Abstract: Decentralized stochastic control refers to the multi-stage optimization of a dynamical system by multiple controllers that have access to different information. Decentralization of information gives rise to new conceptual challenges that require new solution approaches. In this expository paper, we use the notion of an *information-state* to explain the two commonly used solution approaches to decentralized control: the person-by-person approach and the common-information approach.

Key Words: Decentralized stochastic control, dynamic programming, team theory, information structures.

Acknowledgments: This work was supported by Fonds de recherche du Québec–Nature et technologies (FRQ-NT) Establishment of New Researcher Grant 166065 and by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant 402753-11. The authors are grateful to A. Nayyar, D. Teneketzis, and S. Yüksel for useful discussions.

1 Introduction

Centralized stochastic control refers to the multi-stage optimization of a dynamical system by a single controller. Stochastic control, and the associated principle of dynamic programming, have roots in statistical sequential analysis [2] and have been used in various application domains including operations research [23], economics [29], engineering [6], computer science [26], and mathematics [5]. The fundamental assumption of centralized stochastic control is that the decisions at each stage are made by a single controller that has *perfect recall*, that is, a controller that remembers its past observations and decisions. This fundamental assumption is violated in many modern applications where decisions are made by multiple controllers. The multi-stage optimization of such systems is called *decentralized stochastic control* or *dynamic team theory*.

Decentralized stochastic control started with seminal work of Marschak and Radner [17, 25] on static systems that arise in organizations and of Witsenhausen [34–36] on dynamic systems that arise in systems and control. We refer the reader to [4, 12] for a discussion of the history of decentralized stochastic control and to [14, 21, 39] for survey of recent results.

Decentralized stochastic control is fundamentally different from, and significantly more challenging than, centralized stochastic control. Dynamic programming, which is the primary solution concept of centralized stochastic control, does not directly work in decentralized stochastic control. New ways of thinking need to be developed to address information decentralization. The focus of this expository paper is to highlight the conceptual challenges of decentralized control and explain the intuition behind the solution approaches. No new results are presented in this paper; rather we present new insights and connections between existing results. Since the focus is on conceptual understanding, we do not present proofs and ignore the technical details, in particular, measurability concerns, in our description.

We use the following notation. Random variables are denoted by upper case letters; their realizations by the corresponding lower case letters; and their space of realizations by the corresponding calligraphic letters. For integers $a \le b$, $X_{a:b}$ is a short hand for the set $\{X_a, X_{a+1}, \dots, X_b\}$. When $a > b$, $X_{a:b}$ refers to the empty set. In general, subscripts are used as time index while superscripts are used to index controllers. $\mathbb{P}(\cdot)$ denotes the probability of an event and $\mathbb{E}[\cdot]$ denotes the expectation of a random variable. For a collection of functions $\mathbf{g}$, the notations $\mathbb{P}^{\mathbf{g}}(\cdot)$ and $\mathbb{E}^{\mathbf{g}}[\cdot]$ indicate that the probability measure and the expectation depend on the choice of the functions $\mathbf{g}$. $\mathbb{Z}_{>0}$ denotes the set of positive integers and $\mathbb{R}$ denotes the set of real numbers.

2 Decentralized stochastic control: Models and problem formulation

2.1 State, observation, and control processes

Consider a dynamical system with n controllers. Let $\{X_t\}_{t=0}^{\infty}$, $X_t \in \mathcal{X}$, denote the state process of the system. Controller i, $i \in \{1, \dots, n\}$, observes the process $\{Y_t^i\}_{t=0}^{\infty}$, $Y_t^i \in \mathcal{Y}^i$, and generates a control process $\{U_t^i\}_{t=0}^{\infty}$, $U_t^i \in \mathcal{U}^i$. The system yields a reward $\{R_t\}_{t=0}^{\infty}$. These processes are related as follows:

1. Let $\mathbf{U}_t \coloneqq \{U_t^1, \dots, U_t^n\}$ denote the control action of all controllers at time t. Then, the reward at time t depends only on the current state X_t, the future state X_{t+1}, and the current control actions $\mathbf{U}_t$. Furthermore, the state process $\{X_t\}_{t=0}^{\infty}$ is a controlled Markov process given $\{\mathbf{U}_t\}_{t=0}^{\infty}$, i.e., for any $\mathcal{A} \subseteq \mathcal{X}$ and $\mathcal{B} \subseteq \mathbb{R}$, and any realization $x_{1:t}$ of $X_{1:t}$ and $\mathbf{u}_{1:t}$ of $\mathbf{U}_{1:t}$, we have that
$$\mathbb{P}\big(X_{t+1} \in \mathcal{A}, R_t \in \mathcal{B} \mid X_{1:t} = x_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t}\big) = \mathbb{P}\big(X_{t+1} \in \mathcal{A}, R_t \in \mathcal{B} \mid X_t = x_t, \mathbf{U}_t = \mathbf{u}_t\big). \qquad (1)$$

2. The observations $\mathbf{Y}_t \coloneqq \{Y_t^1, \dots, Y_t^n\}$ depend only on current state X_t and previous control actions $\mathbf{U}_{t-1}$, i.e., for any $\mathcal{A}^i \subseteq \mathcal{Y}^i$ and any realization $x_{1:t}$ of $X_{1:t}$ and $\mathbf{u}_{1:t-1}$ of $\mathbf{U}_{1:t-1}$, we have that
$$\mathbb{P}\Big(\mathbf{Y}_t \in \prod_{i=1}^{n} \mathcal{A}^i \;\Big|\; X_{1:t} = x_{1:t}, \mathbf{U}_{1:t-1} = \mathbf{u}_{1:t-1}\Big) = \mathbb{P}\Big(\mathbf{Y}_t \in \prod_{i=1}^{n} \mathcal{A}^i \;\Big|\; X_t = x_t, \mathbf{U}_{t-1} = \mathbf{u}_{t-1}\Big). \qquad (2)$$

2.2 Information structure

At time t, controller i, $i \in \{1, \dots, n\}$, has access to information I^i_t which is a superset of the history $\{Y^i_{1:t}, U^i_{1:t-1}\}$ of the observations and control actions at controller i and a subset of the history $\{\mathbf{Y}_{1:t}, \mathbf{U}_{1:t-1}\}$ of the observations and control actions at all controllers, i.e.,

$$\{Y^i_{1:t}, U^i_{1:t-1}\} \subseteq I^i_t \subseteq \{\mathbf{Y}_{1:t}, \mathbf{U}_{1:t-1}\}.$$

The collection $(I^i_t, i \in \{1, \dots, n\}, t = 0, 1, \dots)$, which is called the *information structure* of the system, captures *who knows what about the system and when.* A decentralized system is characterized by its information structure.

2.3 Control strategies and problem formulation

Based on the information I^i_t available to it, controller i chooses action U^i_t using a *control law* $g^i_t \colon I^i_t \mapsto U^i_t$. The collection of control laws $\mathbf{g}^i \coloneqq (g^i_0, g^i_1, \dots)$ is called a *control strategy of controller* i. The collection $\mathbf{g} \coloneqq (\mathbf{g}^1, \dots, \mathbf{g}^n)$ is called the *control strategy of the system.*

The optimization objective is to pick a control strategy $\mathbf{g}$ to maximize the expected discounted reward

$$\Lambda(\mathbf{g}) \coloneqq \mathbb{E}^{\mathbf{g}}\Big[\sum_{t=0}^{\infty} \beta^t R_t\Big] \tag{3}$$

for a given discount factor $\beta \in (0, 1)$.

2.4 Relationship to other models

The decentralized control problem formulated above is closely related to dynamic games; in particular to dynamic cooperative games. The key difference between the two models is that in decentralized control all controllers have a common objective while in game theory each player has an individual objective. To highlight this fact, decentralized control problems are also referred to as *dynamic teams.*

In cooperative game theory, the concepts of bargaining and contracts are used to study when coalitions are formed and how members of the coalition split the value. In decentralized stochastic control, splitting of the value between the members is not modeled. In this regard, decentralized control is simpler than cooperative games.

In dynamic game theory, the concepts of sequential rationality and consistency of beliefs are used to refine Nash equilibria. In decentralized control, all controllers have the same objective so many of the conceptual difficulties of non-cooperative game theory do not arise.

Although decentralized control is conceptually simpler than the corresponding game theoretic setup, the optimization problem formulated above is non-trivial and the corresponding setup of dynamic cooperative games with incomplete information is an open area of research.

2.5 An example

To illustrate these concepts, let's consider a stylized example of a communication system in which two devices transmit over a multiple access channel.

Packet arrival at the devices. Packets arrive at device i, $i \in \{1, 2\}$, according to Bernoulli processes $\{W^i_t\}^{\infty}_{t=0}$ with success probability p^i. Device i may store $N^i_t \in \{0, 1\}$ packets in a buffer. If a packet arrives when the buffer is full, the packet is dropped.

Channel model. At time t, the channel-state $S_t \in \{0, 1\}$ may be idle ($S_t = 0$) or busy ($S_t = 1$). The channel-state process $\{S_t\}^{\infty}_{t=0}$ is a Markov process with known initial distribution and transition matrix $\mathbf{P} = \left[\begin{smallmatrix} \alpha_0 & 1-\alpha_0 \\ 1-\alpha_1 & \alpha_1 \end{smallmatrix}\right]$. The channel-state process is independent of the packet-arrival process at the device.

System dynamics. At time t, device i, $i \in \{1, 2\}$, may transmit $U^i_t \in \{0, 1\}$ packets, $U^i_t \le N^i_t$. If only one device transmits and the channel is idle, the transmission is successful and the transmitted packet is removed from the buffer. Otherwise the transmission is unsuccessful. The state of each buffer evolves as

$$N^i_{t+1} = \min\{N^i_t - U^i_t(1 - U^j_t)(1 - S_t) + W^i_t, 1\}, \quad \forall i \in \{1, 2\}, \quad j = 3 - i. \tag{4}$$

Each transmission costs c and a successful transmission yields a reward r. Thus, the total reward *for both devices* is

$$R_t = -(U^1_t + U^2_t)c + (U^1_t \oplus U^2_t)(1 - S_t)r$$

where $\oplus$ denotes the XOR operation.

Observation model. Controller i, $i \in \{1, 2\}$, perfectly observes the number N^i_t of packets in the buffer. In addition, *both* controllers observe the one-step delayed control actions (U^1_{t-1}, U^2_{t-1}) of each other and the channel state if *either of devices transmit*. Let H_t denote this additional observation. Then $H_t = S_{t-1}$ if $U^1_{t-1} + U^2_{t-1} > 0$, otherwise $H_t = \mathfrak{E}$ (which denotes no channel-state observation).

Information structure and objective. The information I^i_t available at device i, $i \in \{0, 1\}$, is given by $I^i_t = \{N^i_{1:t}, H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}\}$. Based on the information available to it, device i chooses control action U^i_t using a control law $g^i_t \colon I^i_t \mapsto U^i_t$. The collection of control laws $(\mathbf{g}^1, \mathbf{g}^2)$, where $\mathbf{g}^i := (g^i_0, g^i_1, \dots)$, is called a *control strategy*. The objective is to pick a control strategy $(\mathbf{g}^1, \mathbf{g}^2)$ to maximize the expected discounted reward

$$\Lambda(\mathbf{g}^1, \mathbf{g}^2) := \mathbb{E}^{(\mathbf{g}^1, \mathbf{g}^2)}\Big[\sum_{t=0}^{\infty} \beta^t R_t\Big].$$

We make the following assumption in the paper.

(A) The arrival process at the two controllers is independent.

2.6 Conceptual difficulties in finding an optimal solution

There are two conceptual difficulties in the optimal design of decentralized stochastic control:

1. The optimal control problem is a functional optimization problem where we have to choose an infinite sequence of control laws $\mathbf{g}$ to maximize the expected total reward.
2. In general, the domain I^i_t of control laws g^i_t increases with time. Therefore, it is not immediately clear if we can solve the above optimization problem; even if it is solved, it is not immediately clear if we can implement the optimal solution.

Similar conceptual difficulties arise in centralized stochastic control where they are resolved by identifying an appropriate *information-state* process (see Definition 1 below) and solving a corresponding dynamic program. It is not possible to directly apply such an approach to decentralized stochastic control problems.

In order to better understand the difficulties in extending the solution techniques of centralized stochastic control to decentralized stochastic control, we revisit the main results of centralized stochastic control in the next section.

3 Overview of centralized stochastic control

A centralized stochastic control system is a special case of a decentralized stochastic control system in which there is only one controller ($n = 1$), and the controller has perfect recall ($I^1_t \subseteq I^1_{t+1}$), i.e., the controller remembers everything that it has seen and done in the past. For ease of notation, we drop the superscript i and denote the observation, information, control action, and control law of the controller by Y_t, I_t, U_t, and g_t, respectively. Using this notation, the information available to the controller at time t is given by $I_t = \{Y_{1:t}, U_{1:t-1}\}$. The controller uses a control law $g_t \colon I_t \mapsto U_t$ to choose a control action U_t. The collection $\mathbf{g} = (g_0, g_1, \dots)$ of control laws is called a *control strategy*.

The optimization objective is to pick a control strategy $\mathbf{g}$ to maximize the expected discounted reward

$$\Lambda(\mathbf{g}) \coloneqq \mathbb{E}^{\mathbf{g}}\Big[\sum_{t=0}^{\infty} \beta^t R_t\Big] \tag{5}$$

for a given discount factor $\beta \in (0,1)$.

In the centralized stochastic control literature, the above model is sometimes referred to a partially observable Markov decision process (POMDP). The solution to a POMDP is obtained in two steps [6].

1. Consider a simpler model in which the controller perfectly observes the state of the system, i.e., $Y_t = X_t$. Such a model is called a Markov decision process (MDP). Show that there is no loss of optimality in restricting attention to *Markov strategies*, i.e., control laws of the form $g_t \colon X_t \mapsto U_t$. Obtain an optimal control strategy of this form by solving an appropriate dynamic program.
2. Define a *belief state* of a POMDP as the posterior distribution of X_t given the information at the controller, i.e., $B_t(\cdot) = \mathbb{P}(X_t = \cdot \mid I_t)$. Show that the belief state is a MDP, and use the results for MDP.

A slightly more general approach is identify an *information-state* process of the system and present the solution in terms of the information state. We present this approach below.

Definition 1 *A process $\{Z_t\}_{t=0}^{\infty}$, $Z_t \in \mathcal{Z}_t$, is called an* information-state *process if it satisfies the following properties:*

1. *Z_t is a function of the information I_t available at time t, i.e., there exist a series of functions $\{f_t\}_{t=0}^{\infty}$ such that*
$$Z_t = f_t(I_t). \tag{6}$$
2. *The process Z_t is a controlled Markov process controlled by $\{U_t\}_{i=0}^{\infty}$, that is for any $\mathcal{A} \subseteq \mathcal{Z}_{t+1}$ and any realization i_t of I_t and any choice u_t of U_t, we have that*
$$\mathbb{P}(Z_{t+1} \in \mathcal{A} \mid I_t = i_t, U_t = u_t) = \mathbb{P}(Z_{t+1} \in \mathcal{A} \mid Z_t = f_t(i_t), U_t = u_t). \tag{7}$$
3. *Z_t absorbs the effect all the available information on the current rewards, i.e., for any $\mathcal{B} \subseteq \mathbb{R}$, and any realization i_t of I_t and any choice u_t of U_t, we have that*
$$\mathbb{P}(R_t \in \mathcal{B} \mid I_t = i_t, U_t = u_t) = \mathbb{P}(R_t \in \mathcal{B} \mid Z_t = f_t(i_t), U_t = u_t). \tag{8}$$

In general, a system may have more than one information-state process. The following theorems hold for any information-state process.

Theorem 1 (Structure of optimal control laws) *Let $\{Z_t\}_{t=0}^{\infty}$, $Z_t \in \mathcal{Z}_t$, be an information-state process. Then,*

1. *The information state absorbs the effect of available information on expected future rewards, i.e., for any realization i_t of the information state I_t, any choice u_t of U_t and any choice of future strategy $\mathbf{g}_{(t)} = (g_{t+1}, g_{t+2}, \dots)$, we have that*
$$\mathbb{E}^{\mathbf{g}_{(t)}}\Big[\sum_{\tau=t}^{\infty} \beta^\tau R_\tau \;\Big|\; I_t = i_t, U_t = u_t\Big] = \mathbb{E}^{\mathbf{g}_{(t)}}\Big[\sum_{\tau=t}^{\infty} \beta^\tau R_\tau \;\Big|\; Z_t = f_t(i_t), U_t = u_t\Big]. \tag{9}$$
2. *Therefore, Z_t is a sufficient statitistic for performance evaluation and there is no loss of optimality in restricting attention to control laws of the form $g_t \colon Z_t \mapsto U_t$.*

Theorem 2 (Dynamic programming decomposition) *Assume that the probability distributions in the right-hand side of (1), (2), (7) and (8) are time homogeneous. Let $\{Z_t\}_{t=0}^{\infty}$, be an information-state process such that the space of realization of Z_t is time-invariant, i.e., $Z_t \in \mathcal{Z}$.*

1. *For any choice of future strategy $\mathbf{g}_{(t)} = (g_{t+1}, g_{t+2}, \dots)$, where g_τ, $\tau > t$, is of the form $g_\tau \colon Z_\tau \mapsto U_\tau$ and for any realization z_t of Z_t and any choice u_t of U_t, we have that*

$$\mathbb{E}^{\mathbf{g}_{(t)}}\Bigg[\mathbb{E}^{\mathbf{g}_{(t+1)}}\Big[\sum_{\tau=t+1}^{\infty} \beta^\tau R_\tau \;\Big|\; Z_{t+1}, U_{t+1} = g_{t+1}(Z_{t+1})\Big]\Bigg| Z_t = z_t, U_t = u_t\Bigg]$$
$$= \mathbb{E}^{\mathbf{g}_{(t)}}\Big[\sum_{\tau=t+1}^{\infty} \beta^\tau R_\tau \Big| Z_t = z_t, U_t = u_t\Big] \quad (10)$$

2. *There exists a time-invariant optimal strategy $\mathbf{g}^* = (g^*, g^*, \dots)$ that is given by*

$$g^*(z) = \arg\sup_{u \in \mathcal{U}} Q(z,u), \quad \forall z \in \mathcal{Z} \quad (11a)$$

where Q is the fixed point solution of the following dynamic program[1]

$$Q(z,u) = \mathbb{E}[R_t + \beta V(Z_{t+1}) \mid Z_t = z, U_t = u], \quad \forall z \in \mathcal{Z},\; u \in \mathcal{U}; \quad (11b)$$
$$V(z) = \sup_{u \in \mathcal{U}} Q(z,u), \quad \forall z \in \mathcal{Z}. \quad (11c)$$

The dynamic program can be solved using different methods such as value-iteration, policy-iteration, or linear-programming. See [24] for details.

The information-state based solution approach presented above is equivalent to the standard description of centralized stochastic control. In particular, the current state X_t and the belief state $\mathbb{P}(X_t = \cdot \mid I_t)$ are, respectively, the information states in MDP and POMDP.

An important property of the information state is that the conditional future reward, which is given by (9), does not depend on the past and current control strategy $(g_0, g_1, \dots, g_t)$. This *strategy independence* of future cost is critical to obtain a recurrence relation for the conditional future cost (10) that does not depend on the current control law g_t. Based on this recurrence, we can convert the functional optimization problem of finding the best control law g_t into a set of parametric optimization problem of finding the best control action U_t for each realization of the information state Z_t. This resolves the first conceptual difficulty described in Section 2.6.

In addition, when the information-state process as well as the probability distributions in the right hand side of (7) and (8) are time-homogeneous, time-invariant strategies perform as well as time-varying strategies. Restricting attention to time-invariant strategies resolves the second conceptual difficulty described in Section 2.6.

3.1 An example

To illustrate the concepts described above, consider an example of a device transmitting over a communication channel. This may be considered as a special case of the example of Section 2.5 in which one of the devices never transmits.

Packet arrival at the device. The packet arrival model is the same as that of Section 2.5. Since there is only one device, we omit the superscripts in W_t, N_t, and p.

Channel model. The channel model is exactly same as that of Section 2.5.

System dynamics. At time t, the device transmits $U_t \in \{0,1\}$ packets, $U_t \le N_t$. If the device transmits when the channel is idle, the transmission is successful and the transmitted packet is removed from the buffer. Otherwise, the transmission is unsuccessful. Thus, the state of the buffer evolves as

$$N_{t+1} = \min\{N_t - U_t(1 - S_t) + W_t, 1\}$$

and the total reward is given by

$$R_t = U_t[-c + r(1 - S_t)].$$

[1] In general, a dynamic program may not have an unique solution, or any solution at all. In this paper, we ignore the issue of existence of such a solution and refer the reader to [11] for details.

Observation model. The controller perfectly observes the number N_t of packets in the buffer. In addition, it observes a channel-state *only if it transmits.* Let H_t denote this additional observation. Then $H_t = S_{t-1}$ if $U_{t-1} = 1$, otherwise $H_t = \mathfrak{E}$ (which denotes no observation).

Information structure. The information I_t available at the device is given by $I_t = \{N_{1:t}, U_{1:t-1}, H_{1:t}\}$. The device chooses U_t using a control law $g_t \colon I_t \mapsto U_t$. The objective is to pick a control strategy $\mathbf{g} = (g_0, g_1, \dots)$ to maximize the expected discounted reward.

The model described above is a centralized stochastic control system with state $X_t = (N_t, S_t)$, observation $Y_t = (N_t, H_t)$, reward R_t, and control U_t; one may verify that these processes satisfy (1) and (2) (with $n = 1$).

Let $\xi_t \in [0, 1]$ denote the posterior probability that the channel is busy, i.e.,

$$\xi_t \coloneqq \mathbb{P}(S_t = 1 \mid H_{1:t}).$$

One may verify that $Z_t = (N_t, \xi_t)$ is an information state that satisfies (7) and (8). So, there is no loss of optimality in using control laws of the form $g_t : (N_t, \xi_t) \mapsto U_t$. The information state takes value in the uncountable space $\{0, 1\} \times [0, 1]$. Since ξ_t is a posterior distribution, we can use the computational techniques of POMDPs [28, 40] to numerically solve the corresponding dynamic program.

However, a simpler dynamic programming decomposition is possible by characterizing the reachable set of ξ_t, which is given by

$$\mathcal{Q} \coloneqq \{q_{0,m} \mid m \in \mathbb{Z}_{>0}\} \cup \{q_{1,m} \mid m \in \mathbb{Z}_{>0}\} \tag{12a}$$

where

$$q_{s,m} \coloneqq \mathbb{P}(S_m = 1 \mid S_0 = s), \quad \forall s \in \{0, 1\},\ m \in \mathbb{Z}_{>0}. \tag{12b}$$

Therefore, $\{(N_t, \xi_t)\}_{t=0}^{\infty}$, $(N_t, \xi_t) \in \{0, 1\} \times \mathcal{Q}$, is an alternative information-state process. In this alternative characterization, the information state is denumerable and we may use finite-state approximations to solve the corresponding dynamic program [8–10, 27, 33].

The dynamic program for this alternative characterization is given below. Let $\overline{p} = 1-p$ and $\overline{q}_{s,m} = 1-q_{s,m}$. Then for $s \in \{0, 1\}$ and $m \in \mathbb{Z}_{>0}$, we have that[2]

$$V(0, q_{s,m}) = \beta\bigl[\overline{p}V(0, q_{s,m+1}) + pV(1, q_{s,m+1})\bigr] \tag{13a}$$

$$V(1, q_{s,m}) = \max\bigl\{\beta V(1, q_{s,m+1}), \overline{q}_{s,m}r - c + \beta W(q_{s,m})\bigr\} \tag{13b}$$

where

$$W(q_{s,m}) = \overline{p}\,\overline{q}_{s,m}V(0, q_{0,1}) + p\,\overline{q}_{s,m}V(1, q_{0,1}) + q_{s,m}V(1, q_{1,1}).$$

The first alternative in the right hand side of (13b) corresponds to choosing $u = 0$ while the second corresponds to choosing $u = 1$. The resulting optimal strategy for $\beta = 0.9$, $\alpha_0 = \alpha_1 = 0.75$, $r = 1$, $p = 0.3$, and $c = 0.4$ is given by

$$g^*(1, q_{s,m}) = \begin{cases} 0, & \text{if } s = 1 \text{ and } m \le 2 \\ 1, & \text{otherwise.} \end{cases}$$

As is illustrated by the above example, a general solution methodology for centralized stochastic control is as follows:

1. Identify an information-state process for the given system.
2. Obtain a dynamic program corresponding to the information-state process.
3. Either obtain an exact analytic solution of the dynamic program (which is only possible for simple stylized models), or obtain an approximate numerical solution of the dynamic program (as was done in the example above), or prove qualitative properties of the optimal solution (e.g., in the above example, for appropriate values of c, r, and $\mathbf{P}$, the set $T(s, n) = \{m \in \mathbb{Z}_{>0} \mid g^*(n, q_{s,m}) = 1\}$ is convex).

In the rest of this paper, we explore whether a similar solution approach is possible for decentralized stochastic control problems.

[2] Note that $\{q_{s,m} \mid s \in \{0, 1\}$ and $m \in \mathbb{Z}_{>0}\}$ is equivalent to the reachable set $\mathcal{Q}$ of ξ_t.

4 Conceptual difficulties in dynamic programming for decentralized stochastic control

Recall the two conceptual difficulties that arise in decentralized stochastic control and were described in Section 2.6. Similar difficulties arise in centralized stochastic control, where they are resolved by identifying an appropriate information-state process. It is natural to ask if a similar simplification is possible in decentralized stochastic control. In particular:

1. Is it possible to identify an information state Z^i_t, $Z^i_t \in \mathcal{Z}^i_t$, such that there is no loss of optimality in restricting attention to controllers of the form $g^i_t \colon Z^i_t \mapsto U^i_t$?
2. If the probability distributions in the right hand side of (1) and (2) are time-homogeneous, is it possible to identify a time-homogeneous information-state process and a corresponding dynamic programming that determines a time-invariant optimal control strategies for all controllers?

The second question is significantly more important, and considerably harder, than the first. There are two approaches to find a dynamic programming decomposition. The first approach is to find a set of coupled dynamic programs, where each dynamic program is associated with a controller and determines the "optimal" control strategy at that controller. The second approach is to find a dynamic program that simultaneously determines the optimal control strategy at all controllers.

It is not obvious how to identify such dynamic programs. Let's conduct a thought experiment in which we assume that such dynamic programs have been identified and let's try to identify the implications. The description below is qualitative; the mathematical justification is presented later in the paper.

Consider the first approach. Suppose we are able to find a set of coupled dynamic programs, where the dynamic program for controller i, which we refer to as $\mathcal{D}^i$, determines the "optimal" strategy $\mathbf{g}^i$ for controller i. We use the term optimal in quotes because we cannot isolate an optimization problem for controller i until we specify the control strategy $\mathbf{g}^{-i}$ for all other controllers. Therefore, dynamic program $\mathcal{D}^i$ determines the *best response strategy* $\mathbf{g}^i$ for a particular choice of control strategies $\mathbf{g}^{-i}$ for other controllers. With a slight abuse of notation, we can write this as

$$\mathbf{g}^i = \mathcal{D}^i(\mathbf{g}^{-i}).$$

Any fixed-point $\mathbf{g}^* = (\mathbf{g}^{*,1}, \dots, \mathbf{g}^{*,n})$ of these coupled dynamic programs has the property that every controller i, $i \in \{1, \dots, n\}$, is playing its best response strategy to the strategies of other controllers. Such a strategy is called a *person-by-person optimal* strategy (which is related to the notion of local optimum in optimization theory and the notion of Nash equilibrium in game theory). In general, a person-by-person optimal strategy need not be globally optimal; in fact, a person-by-person strategy may perform arbitrarily bad as compared to the globally optimal strategy. So, unless we impose further restrictions on the model, a set of coupled dynamic programs cannot determine a globally optimal strategy.

Now, consider the second approach. Suppose we are able to find a dynamic program similar to (11a)–(11c) that determines the optimal control strategies for all controllers. All controllers must be able to use this dynamic program to find their control strategy. Therefore, the information-state process $\{Z_t\}_{t=0}^{\infty}$ of such a dynamic program must have the following property: *Z_t is a function of the information I^i_t available to every controller i, $i \in \{1, \dots, n\}$.* In other words, the information state must be measurable with respect to the *common knowledge* (in the sense of Aumann [3]) between the controllers.

If we follow the methodology of centralized stochastic control and restrict attention to control laws of the form $g^i_t \colon Z_t \mapsto U^i_t$, then each controller would be ignoring its *local information* (i.e., the information not commonly known to all controllers). Hence, if the dynamic program similar to (11a)–(11c) determines a globally optimal strategy, then the step corresponding to (11c) cannot be a parametric optimization problem that finds an optimal U^i_t for each Z_t.

Now let's try to characterize the nature of the optimization problem corresponding to (11c). Let L^i_t denote the *local information* at each controller so that Z_t and L^i_t are sufficient to determine I^i_t. Then, for a particular realization z of the information-state, the step corresponding to (11c) of the dynamic program must determine functions $(\gamma^1, \dots, \gamma^n)$ such that: (i) computing $(\gamma^1, \dots, \gamma^n)$ for each realization of the information state is equivalent to choosing $(g^1, \dots, g^n)$. (ii) γ^i gives instructions to controller i on how to use its local information L^i_t to determine the control action U^i_t, i.e., γ^i maps L^i_t to U^i_t. Thus, the step corresponding to (11c) is a functional optimization problems.

The above discussion shows that dynamic programming for decentralized stochastic control will be different from that for centralized stochastic control. Either we must be content with a person-by-person optimal strategy; or, if we pursue global optimality, then we must be willing to solve functional optimization problems in the step corresponding to (11c) in an appropriate dynamic program.

In the literature, the first approach is called the *person-by-person approach* and the second approach is called the *common-information approach*. We describe both these approaches in the next section.

5 The person-by-person approach

The person-by-person approach is motivated by the computational approaches for finding Nash equilibrium in game theory. It was proposed by Marschak and Radner [17, 25] in the context of static systems with multiple controllers and has been subsequently used in dynamic systems as well. This approach is used to identify structural results as well as identify coupled dynamic programs to find person-by-person optimal (or equilibrium) strategies.

5.1 Structure of optimal control strategies

To find the structural results, proceed as follows. Pick a controller that has perfect recall, say i; *arbitrarily* fix the control strategies $\mathbf{g}^{-i}$ of all controllers except controller i and consider the sub-problem of finding the *best response* strategy $\mathbf{g}^i$ at controller i. Since controller i has perfect recall, this sub-problem is centralized. Suppose that we identify an information-state process $\{\tilde{I}^i_t\}_{t=0}^\infty$ for this sub-problem. Then, there is no loss of (best-response) optimality in restricting attention to control laws of the form $\tilde{g}^i_t \colon \tilde{I}^i_t \to U^i_t$ at controller i.

The choice of control strategies $\mathbf{g}^{-i}$ was completely arbitrary. Hence, if the structure of $\tilde{g}^i_t$ does not depend on the choice of (the arbitrarily chosen) control strategies $\mathbf{g}^{-i}$ of other controllers, then there is no loss of (global) optimality in restricting attention to control laws of the form $\tilde{g}^i_t$ at controller i.

Repeat this procedure at all controllers that have perfect recall. Let $\{\tilde{I}^i_t\}_{t=0}^\infty$ be the information-state processes identified at controller i, $i \in \{1, \dots, n\}$. Then there is no loss of global optimality in restricting attention to the information structure $(\tilde{I}^i_t, i \in \{1, \dots, n\}, t = 0, 1, \dots)$.

5.2 An example

To illustrate this approach, consider the example of the decentralized control system of Section 2.5. Arbitrarily fix the control strategy $\mathbf{g}^j$ of controller j, $j \in \{1, 2\}$. The next step is to identify an information-state process for the centralized sub-problem of finding the best response strategy $\mathbf{g}^i$ of controller i, $i = 3 - j$.

Assumption (A) (which states that the packet-arrival processes at the two devices are independent) implies that

$$\mathbb{P}(N^1_{1:t}, N^2_{1:t} \mid H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}) = \mathbb{P}(N^1_{1:t} \mid H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1})\mathbb{P}(N^2_{1:t} \mid H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}) \tag{14}$$

Using the above conditional independence, we can show that for any choice of control strategy $\mathbf{g}^j$, $\tilde{I}^i_t = \{N^i_t, H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}\}$ is an information state for controller i. By Theorem 1, we get that there is no loss of optimality (for best response strategy) in restricting attention to control laws of the form $\tilde{g}^i_t \colon \tilde{I}^i_t \mapsto U^i_t$. Since the structure of the optimal best response strategy does not depend on the choice of $\mathbf{g}^j$, there is no loss of global optimality in restricting attention to control laws of the form $\tilde{g}^i_t$. Equivalently, there is no loss of optimality in assuming that the system has a simplified information structure $(\tilde{I}^i_t, i \in \{1, 2\}, t = 0, 1, \dots)$.

5.3 Coupled dynamic program for person-by-person optimal solution

As discussed in Section 4, we can *in principle* identify coupled dynamic programs that determine a person-by-person optimal solution. Such coupled dynamic programs have been used to find person-by-person optimal strategies in sequential detection problems [30, 31]. In this section, we highlight two salient features of this approach.

Suppose as a first step, we use the person-by-person approach to find the structure of optimal controllers $\tilde{g}^i_t \colon \tilde{I}^i_t \mapsto U^i_t$. Pick a controller, say i. Arbitrarily fix the control strategies $\tilde{\mathbf{g}}^{-i}$ of all controllers other than i and consider the sub-problem of finding the best response strategy $\tilde{\mathbf{g}}^i$. In general, the information-state process $\{\tilde{I}^i_t\}_{t=0}^{\infty}$ may not be time-homogeneous (as is the case in the above example where $\tilde{I}^i_t = \{N^i_t, H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}\}$. A fortiori, we cannot show that restricting attention to time-invariant strategies is without loss of optimality.

Suppose that the information-state process $\{\tilde{I}^i_t\}_{t=0}^{\infty}$ of every controller i, $i \in \{1, \dots, n\}$, is time homogeneous. Even then, when we *arbitrarily* fix the control strategies $\tilde{\mathbf{g}}^{-i}$ of all other controllers, the dynamical model seen by controller i is not time homogeneous. For the dynamic model from the point of view of controller i to be time-homogeneous, we must *a priori* restrict attention to time-invariant strategies at each controller.

Thus, a time-invariant person-by-person optimal strategy obtained by the coupled dynamic programs described in Section 4 need not be globally optimal for two reasons. First, there might be other time-invariant person-by-person strategies that achieve a higher expected discounted reward. Second, there might be other time-*varying* strategies that achieve higher expected discounted reward.

6 The common-information approach

The common-information approach was proposed by Nayyar, Mahajan, and Teneketzis [15, 18–20] and provides a dynamic programming decomposition (that determines optimal control strategies for all controllers) for a subclass of decentralized control systems. Variation of this approach had been used for specific information structures including delayed state sharing [1], partially nested systems with common past [7], teams with sequential partitions [37], periodic sharing information structure [22], and belief sharing information structure [38].

This approach formalizes the intuition presented in Section 4: to obtain a dynamic program that determines optimal control strategies for all controllers, the information-state process must be measurable at all controllers and, at each step of the dynamic program, we must solve a functional optimization problem that determines instructions to map local information to control action for each realization of the information state.

To formally describe this intuition, split the information available at each controller into two parts: the *common information*

$$C_t = \bigcap_{\tau \ge t} \bigcap_{i=1}^{n} I^i_\tau$$

and the *local information*

$$L^i_t = I^i_t \setminus C_t, \quad \forall i \in \{1, \dots, n\}.$$

By construction, the common and local information determine the total information, i.e., $I^i_t = C_t \cup L^i_t$ and the common information is nested, i.e., $C_t \subseteq C_{t+1}$.

The common information approach applies to decentralized control systems that have a *partial history sharing* information structure [19, 20].

Definition 2 *An information structure is called* partial history sharing *when the following conditions are satisfied:*

1. *For any set of realizations $\mathcal{A}$ of L^i_{t+1} and any realization c_t of C_t, ℓ^i_t of L^i_t, u^i_t of U^i_t and y^i_{t+1} of Y^i_{t+1}, we have*

$$\mathbb{P}(L^i_{t+1} \in \mathcal{A} \mid C_t = c_t, L^i_t = \ell^i_t, U^i_t = u^i_t, Y^i_{t+1} = y^i_{t+1}) = \mathbb{P}(L^i_{t+1} \in \mathcal{A} \mid L^i_t = \ell^i_t, U^i_t = u^i_t, Y^i_{t+1} = y^i_{t+1})$$

2. *The size of the local information is uniformly bounded*[3]*, i.e., there exists a k such that for all t and all $i \in \{1, \dots, n\}$, $|\mathcal{L}^i_t| \le k$, where $\mathcal{L}^i_t$ denotes the space of realizations of L^i_t.*

An example of partial history sharing is the celebrated k-step delayed sharing [35] information structure. Let $J^i_t = \{Y^i_{1:t}, U^i_{1:t-1}\}$ denote the *self information* of controller i. In k-step delay sharing, each controller has access to the k-step delayed self information of all other controllers, i.e.,

$$I^i_t = J^i_t \cup \Big(\bigcup_{\substack{j=1 \\ j \ne i}}^{n} J^j_{t-k} \Big), \quad \forall i \in \{1, \dots, n\}.$$

Another example is k-step periodic sharing, where all controllers periodically share their self information after every k steps, i.e.,

$$I^i_t = J^i_t \cup \Big(\bigcup_{\substack{j=1 \\ j \ne i}}^{n} J^j_{\lfloor t/k \rfloor k} \Big), \quad \forall i \in \{1, \dots, n\}.$$

The example described in Section 2.5 does not have partial history sharing information structure. However, if we follow the person-by-person approach of Section 5.2 and restrict attention to the information structure $(\tilde{I}^i_t, i \in \{1,2\}, t = 0, 1, \dots)$ where $\tilde{I}^i_t = \{N^i_t, H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}\}$, then the model has partial history sharing information structure.

To identify a dynamic program that determines optimal control strategies for all controllers, the common-information approach exploits the fact that planning is centralized, i.e., the control strategies for all controllers are chosen before the system starts running and, therefore, optimal strategies can be searched in a centralized manner.

The construction of an appropriate dynamic program relies on partial evaluation of a function defined below.

Definition 3 *For any function $f\colon (x, y) \mapsto z$ and a value x_0 of x, the* partial evaluation *of f and $x = x_0$ is a function $g\colon y \mapsto z$ such that for all values of y,*

$$g(y) = f(x_0, y).$$

For example, if $f(x, y) = x^2 + xy + y^2$, then the partial evaluation of f at $x = 2$ is $g(y) = y^2 + 2y + 4$.

The common-information approach proceeds as follows [19, 20]:

1. Construct an equivalent centralized coordinated system.
   The first step of the common-information approach is to construct an equivalent centralized stochastic control system which we call the *coordinated system*. The controller of this system, called the *coordinator*, observes the common information C_t and chooses the partially evaluated control laws g^i_t, $i \in \{1, \dots, n\}$, at C_t. Denote these partial evaluations by Γ^i_t and call them *prescriptions*. These prescriptions tell the controllers how to map their local information information into control actions; in particular $U^i_t = \Gamma^i_t(L^i_t)$. The decision rule $\psi_t\colon C_t \mapsto (\Gamma^1_t, \dots, \Gamma^n_t)$ that chooses the prescriptions is called a *coordination law* and the choice of $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots)$ is called a *coordination strategy*.
   Note that the prescription Γ^i_t is a partial evaluation of the control law g^i_t at the common information C_t. Hence, for any coordination strategy $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots)$, we can construct an equivalent control strategy $\mathbf{g}^{i,*} = (g^{i,*}_1, g^{i,*}_2, \dots)$, $i \in \{1, \dots, n\}$ by choosing

$$g^{i,*}_t(c_t, \ell^i) = \psi^{i,*}_t(c_t)(\ell^i),$$

[3] This condition is needed to ensure that the information-state is time-homogeneous and, as such, may be ignored for finite horizon models [20].

where $\psi_t^{i,*}$ denotes the i-th component of ψ_t^*. The coordination strategy $\boldsymbol{\psi}$ is equivalent to the control strategy $\mathbf{g}^*$ in the following sense. For any realization of the primitive random variables of the system, the reward process in the original system under $\mathbf{g}^*$ has the same realization as the reward process in coordinated system under ψ. Therefore, the problem of finding the optimal decentralized control strategy in the original system is equivalent to that of finding the optimal coordination strategy in the coordinated system.

The coordinated system has only one controller, the coordinator, which has perfect recall; the controllers of the original system are passive agents that simply use the prescriptions given by the coordinator. Hence, the coordinated system is a centralized stochastic control system with the state process $\{(X_t, L_t^1, \dots, L_t^n)\}_{t=0}^{\infty}$, the observation process $\{C_t\}_{t=0}^{\infty}$, the reward process $\{R_t\}_{t=0}^{\infty}$, and the control process $\{(\Gamma_t^1, \dots, \Gamma_t^n)\}_{t=0}^{\infty}$.

2. Identify an information state of the coordinated system

The coordinated system is a centralized system in which the control process is a sequence of functions. Let $\{Z_t\}_{t=0}^{\infty}$, $Z_t \in \mathcal{Z}_t$, be any information-state process for the coordinated system.[4] Then, by Theorem 1, there is no loss of optimality in restricting attention to coordination laws of the form

$$\psi_t \colon Z_t \mapsto (\Gamma_t^1, \dots, \Gamma_t^n).$$

Suppose the probability distributions on the right hand side of (1) and (2) are time-homogeneous, the evolution of Z_t is time-homogeneous, and the state space $\mathcal{Z}_t$ of the realizations of Z_t is time-invariant, i.e., $\mathcal{Z}_t = \mathcal{Z}$. Then, by Theorem 2, there exists a time-invariant coordination strategy $\boldsymbol{\psi}^* = (\psi^*, \psi^*, \dots)$ where ψ^* is given by

$$\psi^*(z) = \arg \sup_{(\gamma^1, \dots, \gamma^n)} Q(z, (\gamma^1, \dots, \gamma^n)), \quad \forall z \in \mathcal{Z} \tag{15a}$$

where Q is the unique fixed point of the following set of equations: $\forall z \in \mathcal{Z}$ and $\forall \boldsymbol{\gamma} = (\gamma^1, \dots, \gamma^n)$

$$Q(z, \boldsymbol{\gamma}) = \mathbb{E}[R_t + \beta V(Z_{t+1}) \mid Z_t = z, \Gamma_t^1 = \gamma^1, \dots, \Gamma_t^n = \gamma^n], \tag{15b}$$

$$V(z) = \sup_{\boldsymbol{\gamma}} Q(z, \boldsymbol{\gamma}). \tag{15c}$$

As explained in the previous step, the optimal time-invariant control strategies $\mathbf{g}^{i,*} = (g^{i,*}, g^{i,*}, \dots)$, $i \in \{1, \dots, n\}$, for the original decentralized system are given by

$$g^{i,*}(z, \ell^i) = \psi^{i,*}(z)(\ell^i)$$

where $\psi^{i,*}$ denotes the i-th component of ψ^*.

Note that step (15c) of the above dynamic program is a functional optimization problem. In contrast, step (11c) of the dynamic program for centralized stochastic control was a parametric optimization problem.

Remark 1 The coordinated system and the coordinator described above are fictitious and used only as a tool to explain the approach. The computations carried out at the coordinator are based on the information known to all controllers. Hence, each controller can carry out the computations attributed to the coordinator. As a consequence, it is possible to describe the above approach without considering a coordinator, but in our opinion thinking in terms of a fictitious coordinator makes it easier to understand the approach.

6.1 An example

To illustrate this approach, consider the decentralized control example of Section 2.5. Start with the simplified information structure $\tilde{I}_t^i = \{N_t^i, H_{1:t}, U_{1:t-1}^1, U_{1:t-1}^2\}$ obtained using the person-by-person approach. The common information is given by

$$C_t = \bigcap_{\tau \ge t} (\tilde{I}_\tau^1 \cap \tilde{I}_\tau^2) = \{H_{1:t}, U_{1:t-1}^1, U_{1:t-1}^2\}$$

[4] For example, the process $\{\pi_t\}_{t=0}^{\infty}$, where π_t is the conditional probability measure on $(X_t, L_t^1, \dots, L_t^n)$ conditioned on C_t, is always an information-state process.

and the local information is given by

$$L^i_t = \tilde{I}^i_t \setminus C_t = \{N^i_t\}, \quad \forall i \in \{1,2\}.$$

Thus, in the coordinated system, the coordinator observes C_t and uses the coordination law $\psi_t \colon C_t \mapsto (\gamma^1_t, \gamma^2_t)$, where γ^i_t maps the local information N^i_t to U^i_t. Note that γ^i_t is completely specified by $D^i_t = \gamma^i_t(1)$ because the constraint $U^i_t \le N^i_t$ implies that $\gamma^i_t(0) = 0$. Therefore, we may assume that the coordinator uses a coordination law $\psi_t \colon C_t \mapsto (D^1_t, D^2_t)$, $D^i_t \in \{0,1\}$, $i \in \{1,2\}$ and each device then chooses a control action according to $U^i_t = N^i_t D^i_t$. The system dynamics and the reward process are same as in the original decentralized system.

Since the coordinator has perfect recall, the problem of finding the best coordination strategy is a centralized stochastic control problem. To simplify this centralized stochastic control problem, we need to identify an information state as described in Definition 1.

Let $\zeta^i_t \in [0,1]$ denote the posterior probability that device i, $i \in \{1,2\}$ has a packet in its buffer given the channel feedback, i.e.,

$$\zeta^i_t = \mathbb{P}(N^i_t = 1 \mid H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}), \quad \forall i \in \{1,2\}.$$

Moreover, as in the centralized case, let $\xi_t \in [0,1]$ denote the posterior probability that the channel is busy given the channel feedback, i.e.,

$$\xi_t = \mathbb{P}(S_t = 1 \mid H_{1:t}, U^1_{1:t-1}, U^2_{1:t-1}) = \mathbb{P}(S_t = 1 \mid H_{1:t}).$$

One may verify that $(\zeta^1_t, \zeta^2_t, \xi_t)$ is an information state that satisfies (7) and (8). So, there is no loss of optimality in using coordination laws of the form $\gamma \colon (\zeta^1_t, \zeta^2_t, \xi_t) \mapsto (D^1_t, D^2_t)$. This information state takes values in the uncountable space $[0,1]^3$. Since each component ζ^1_t, ζ^2_t, and ξ_t of the information state is a posterior distribution, we can use the computational techniques of POMDPs [28,40] to numerically solve the corresponding dynamic program.

However, a simpler dynamic programming decomposition is possible by characterizing the reachable set of the information state. The reachable set of ζ^i_t is given by

$$\mathcal{R}^i \coloneqq \{z^i_k \mid k \in \mathbb{Z}_{>0}\} \cup \{1\} \tag{16a}$$

where

$$z^i_k \coloneqq \mathbb{P}(N^i_k = 1 \mid N^i_0 = 0, D^i_{0:k-1} = (0,\dots,0)), \quad \forall s \in \{0,1\},\ k \in \mathbb{Z}_{>0} \tag{16b}$$

and the reachable set of ξ_t is given by $\mathcal{Q}$ defined in (12). For ease of notation, define $z^i_\infty = 1$.

Therefore, $\{(\zeta^1_t, \zeta^2_t, \xi_t)\}^\infty_{t=0}$, $(\zeta^1_t, \zeta^2_t, \xi_t) \in \mathcal{R}^1 \times \mathcal{R}^2 \times \mathcal{Q}$, is an alternative information-state process. In this alternative characterization, the information state is denumerable and we may use finite-state approximations to solve the corresponding dynamic program [8–10,27,33].

The dynamic program for this alternative characterization is given below. Let $\bar{q}_{s,m} = 1 - q_{s,m}$ and $\bar{z}^i_k = 1 - z^i_k$. Then for $s \in \{0,1\}$ and $k, \ell \in \mathbb{Z}_{>0} \cup \{\infty\}$ and $m \in \mathbb{Z}_{>0}$, we have that

$$V(z^1_k, z^2_\ell, q_{s,m}) = \max\big\{Q_{00}(z^1_k, z^2_\ell, q_{s,m}), Q_{10}(z^1_k, z^2_\ell, q_{s,m}), Q_{01}(z^1_k, z^2_\ell, q_{s,m}), Q_{11}(z^1_k, z^2_\ell, q_{s,m})\big\} \tag{17a}$$

where $Q_{d^1d^2}(z^1_k, z^2_\ell, q_{s,m})$ corresponds to choosing the prescription (d^1, d^2) and is given by

$$Q_{00}(z^1_k, z^2_\ell, q_{s,m}) = \beta V(z^1_{k+1}, z^2_{\ell+1}, q_{s,m+1}); \tag{17b}$$

$$\begin{aligned} Q_{10}(z^1_k, z^2_\ell, q_{s,m}) &= z^1_k\, \bar{q}_{s,m}\, r - z^1_k\, c + \beta\Big[\bar{z}^1_k V(z^1_1, z^2_{\ell+1}, q_{s,m+1}) \\ &\qquad + z^1_k\, \bar{q}_{s,m} V(z^1_1, z^2_{\ell+1}, q_{0,1}) + z^1_k\, q_{s,m} V(z^1_\infty, z^2_{\ell+1}, q_{1,1})\Big]; \end{aligned} \tag{17c}$$

$$\begin{aligned} Q_{01}(z^1_k, z^2_\ell, q_{s,m}) &= z^2_\ell\, \bar{q}_{s,m}\, r - z^2_\ell\, c + \beta\Big[\bar{z}^2_\ell V(z^1_{k+1}, z^2_1, q_{s,m+1}) \\ &\qquad + z^2_\ell\, \bar{q}_{s,m} V(z^1_{k+1}, z^2_1, q_{0,1}) + z^2_\ell\, q_{s,m} V(z^1_{k+1}, z^2_\infty, q_{1,1})\Big]; \end{aligned} \tag{17d}$$

$$\begin{aligned} Q_{11}(z_k^1, z_\ell^2, q_{s,m}) = [z_k^1\,\overline{z}_\ell^2 + \overline{z}_k^1\,z_\ell^2]\,\overline{q}_{s,m}\,r - [z_k^1 + z_\ell^2]\,c + \beta\Big[&\overline{z}_k^1\,\overline{z}_\ell^2 V(z_1^1, z_1^2, q_{s,m+1}) \\ &+ [z_k^1\,\overline{z}_\ell^2 + \overline{z}_k^1\,z_\ell^2]\,\overline{q}_{s,m} V(z_1^1, z_1^2, q_{0,1}) + z_k^1 z_\ell^2 \overline{q}_{s,m} V(z_\infty^1, z_\infty^2, q_{0,1}) \\ &+ z_k^1 \overline{z}_\ell^2 q_{s,m} V(z_\infty^1, z_1^2, q_{1,1}) + \overline{z}_k^1 z_\ell^2 q_{s,m} V(z_1^1, z_\infty^2, q_{1,1}) \\ &+ z_k^1 z_k^2 q_{s,m} V(z_\infty^1, z_\infty^2, q_{1,1})\Big]. \end{aligned} \tag{17e}$$

The optimal strategies obtained by solving (17) for $\beta = 0.9$, $\alpha_0 = \alpha_1 = 0.75$, $r = 1$, $p_1 = p_2 = 0.3$, and $c = 0.4$ is given by

$$g^*(z_k^1, z_\ell^2, q_{s,m}) = \begin{cases} (0,0), & \text{if } s = 1 \text{ and } m \le 2 \\ d(z_k^1, z_\ell^2), & \text{otherwise,} \end{cases}$$

where

$$d(z_k^1, z_\ell^2) = \begin{cases} (1,0), & \text{if } k > \ell \\ (0,1), & \text{if } k < \ell \\ (1,0) \text{ or } (0,1), & \text{if } k = \ell. \end{cases}$$

Remark 2 As we argued in Section 4, if a single dynamic program determines the optimal control strategies at all controllers, then the step (15c) must be a functional optimization problem. Consequently, the dynamic program for decentralized stochastic control is significantly more difficult to solve than dynamic programs for centralized stochastic control. When the observation and control processes are finite valued (as in the above example), the space of functions from L_t^i to U_t^i are finite and step (15c) can be solved by exhaustively searching over all alternatives.

Remark 3 As in centralized stochastic control, the information-state in decentralized control is sensitive to the modeling assumptions. For example, in the above example, if we remove assumption (A) (which states that the packet-arrival processes at the two devices are independent), then the conditional independence in (14) is not valid; therefore, we cannot use the person-by-person approach to show that $\{N_t^i, U_{1:t-1}^1, U_{1:t-1}^2, H_{1:t}\}_{t=0}^\infty$ is an information state for controller i. In the absence of this result, the information structure is not partial history sharing. So, we cannot identify a dynamic program for the infinite horizon problem.

7 Conclusion

Decentralized stochastic control gives rise to new conceptual challenges as compared to centralized stochastic control. There are two solution methodologies to overcome these challenges: (i) the person-by-person approach and (ii) the common-information approach. The person-by-person approach provides the structure of globally optimal control strategies and coupled dynamic programs that determine person-by-person optimal control strategies. The common-information approach provides the structure of globally optimal control strategies as well as a dynamic program that determines globally optimal control strategies. A functional optimization problem needs to be solved to solve the dynamic program.

In practice, both the person-by-person approach and the common information approach need to be used in tandem to solve a decentralized stochastic control problem. For example, in the example of Section 2.5 we first used the person-by-person approach to simplify the information structure of the system and then used the common-information approach to find a dynamic programming decomposition. Neither approach could give a complete solution on its own. A similar tandem approach has been used for simplifying specific information structures [13], real-time communication [32], networked control systems [16].

Therefore, a general solution methodology for decentralized stochastic control is as follows.

1. Use the person-by-person approach to simplify the information structure of the system.
2. Use the common-information approach on the simplified information structure to identify an information-state process for the system.

3. Obtain a dynamic program corresponding to the information-state process.
4. Either obtain an exact analytic solution of the dynamic program (as in the centralized case, this is possible only for very simple models), or obtain an approximate numerical solution of the dynamic program (as was done in the example above), or prove qualitative properties of optimal solution.

This approach is similar to the general solution approach of centralized stochastic control, although the last step is significantly more difficult.

The above methodology applies only to systems with partial-history sharing and to systems that reduce to partial-history sharing by a person-by-person approach. Identifying solution techniques for other subclasses of decentralized stochastic control remains an active area of research.

References

[1] Aicardi, M., Davoli, F., Minciardi, R.: Decentralized optimal control of Markov chains with a common past information set. IEEE Transactions on Automatic Control 32(11), 1028–1031 (1987)

[2] Arrow, K.J., Blackwell, D., Girshick, M.A.: Bayes and minimax solutions of sequential decision problems. Econometrica 17(3/4), 213–244 (1949)

[3] Aumann, R.J.: Agreeing to disagree. Annals of Statistics (4), 1236–1239 (1976)

[4] Başar, T., Bansal, R.: The theory of teams: A selective annotated bibliography. In: T. Başar, P. Bernhard (eds.) Differential Games and Applications, *Lecture Notes in Control and Information Sciences*, 119, 186–201. Springer (1989)

[5] Bellman, R.: Dynamic Programming. Princeton University Press (1957)

[6] Bertsekas, D.P.: Dynamic Programming and Optimal Control, vol. 1. Athena Scientific, Belmont, MA (1995)

[7] Casalino, G., Davoli, F., Minciardi, R., Puliafito, P., Zoppoli, R.: Partially nested information structures with a common past. IEEE Transactions on Automatic Control 29(9), 846–850 (1984)

[8] Cavazos-Cadena, R.: Finite-state approximations for denumerable state discounted markov decision processes. Applied Mathematics and Optimization 14(1), 1–26 (1986)

[9] Flåm, S.D.: Finite State Approximations for Countable State Infinite Horizon Discounted Markov Decision Processes. Modeling, Identification and Control 8(2), 117–123 (1987)

[10] Hernández-Lerma, O.: Finite-state approximations for denumerable multidimensional state discounted markov decision processes. Journal of Mathematical Analysis and Applications 113(2), 382–389 (1986)

[11] Hernández-Lerma, O., Lasserre, J.: Discrete-Time Markov Control Processes. Springer-Verlag (1996)

[12] Ho, Y.C.: Team decision theory and information structures. Proceedings of the IEEE 68(6), 644–654 (1980)

[13] Mahajan, A.: Optimal decentralized control of coupled subsystems with control sharing. IEEE Transactions on Automatic Control (to appear) (2013)

[14] Mahajan, A., Martins, N., Rotkowitz, M., Yüksel, S.: Information structures in optimal decentralized control. In: Proc. 51st IEEE Conf. Decision and Control, 1291–1306. Maui, Hawaii (2012)

[15] Mahajan, A., Nayyar, A., Teneketzis, D.: Identifying tractable decentralized control problems on the basis of information structure. In: Proc. 46th Annual Allerton Conf. Communication, Control, and Computing, 1440–1449. Monticello, IL (2008)

[16] Mahajan, A., Teneketzis, D.: Optimal performance of networked control systems with non-classical information structures. SIAM Journal of Control and Optimization 48(3), 1377–1404 (2009)

[17] Marschak, J., Radner, R.: Economic Theory of Teams. Yale University Press, New Haven (1972)

[18] Nayyar, A.: Sequential decision making in decentralized systems. Ph.D. thesis, University of Michigan, Ann Arbor, MI (2011)

[19] Nayyar, A., Mahajan, A., Teneketzis, D.: The common-information approach to decentralized stochastic control. In: B. Bernhardsson, G. Como, A. Rantzer (eds.) Information and Control in Networks. Springer Verlag (2013)

[20] Nayyar, A., Mahajan, A., Teneketzis, D.: Decentralized stochastic control with partial history sharing: A common information approach. IEEE Transactions on Automatic Control 58(7), 1644–1658 (2013)

[21] Oliehoek, F.A., Spaan, M.T.J., Amato, C., Whiteson, S.: Incremental clustering and expansion for faster optimal planning in decentralized POMDPs. Journal of Artificial Intelligence Research 46, 449–509 (2013)

[22] Ooi, J.M., Verbout, S.M., Ludwig, J.T., Wornell, G.W.: A separation theorem for periodic sharing information patterns in decentralized control. IEEE Transactions on Automatic Control 42(11), 1546–1550 (1997)

[23] Powell, W.B.: Approximate Dynamic Programming: Solving the curses of dimensionality, vol. 703. John Wiley & Sons (2007)

[24] Puterman, M.: Markov decision processes: Discrete Stochastic Dynamic Programming. John Wiley and Sons (1994)

[25] Radner, R.: Team decision problems. Annals of Mathmatical Statistics 33, 857–881 (1962)

[26] Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Prentice Hall (1995)

[27] Sennott, L. I.: Stochastic dynamic programming and the control of queueing systems. Wiley, New York, NY (1999)

[28] Shani, G., Pineau, J., Kaplow, R.: A survey of point-based POMDP solvers. Autonomous Agents and Multi-Agent Systems 27(1), 1–51 (2013)

[29] Stokey, N.L., Lucas Robert E, J.: Recursive methods in economic dynamics. Harvard University Press (1989)

[30] Teneketzis, D., Ho, Y.: The decentralized Wald problem. Information and Computation (formerly Information and Control) 73(1), 23–44 (1987)

[31] Teneketzis, D., Varaiya, P.: The decentralized quickest detection problem. IEEE Transactions on Automatic Control AC-29(7), 641–644 (1984)

[32] Walrand, J.C., Varaiya, P.: Optimal causal coding-decoding problems. IEEE Transactions on Information Theory 29(6), 814–820 (1983)

[33] White, D.: Finite state approximations for denumerable state infinite horizon discounted Markov processes. Journal of Mathematical Analysis and Applications 74(1), 292–295 (1980)

[34] Witsenhausen, H.S.: On information structures, feedback and causality. SIAM Journal of Control 9(2), 149–160 (1971)

[35] Witsenhausen, H.S.: Separation of estimation and control for discrete time systems. Proceedings of the IEEE 59(11), 1557–1566 (1971)

[36] Witsenhausen, H.S.: A standard form for sequential stochastic control. Mathematical Systems Theory 7(1), 5–11 (1973)

[37] Yoshikawa, T.: Decomposition of dynamic team decision problems. IEEE Transactions on Automatic Control 23(4), 627–632 (1978)

[38] Yüksel, S.: Stochastic nestedness and the belief sharing information pattern. IEEE Transactions on Automatic Control, 2773–2786 (2009)

[39] Yüksel, S., Başar, T.: Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints. Birkhäuser, Boston, MA (2013)

[40] Zhang, W.: Algorithms for partially observed Markov decision processes. Ph.D. thesis, Hong Kong University of Science and Technology (2001)