Common Information based Approximate State Representations in Multi-Agent Reinforcement Learning

Hsu Kao
University of Michigan
hsukao@umich.edu

Vijay Subramanian
University of Michigan
vgsubram@umich.edu

Abstract

Due to information asymmetry, finding optimal policies for Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) is hard with the complexity growing doubly exponentially in the horizon length. The challenge increases greatly in the multi-agent reinforcement learning (MARL) setting where the transition probabilities, observation kernel, and reward function are unknown. Here, we develop a general compression framework with approximate common and private state representations, based on which decentralized policies can be constructed. We derive the optimality gap of executing dynamic programming (DP) with the approximate states in terms of the approximation error parameters and the remaining time steps. When the compression is exact (no error), the resulting DP is equivalent to the one in existing work. Our general framework generalizes a number of methods proposed in the literature. The results shed light on designing practically useful deep-MARL network structures under the "centralized learning distributed execution" scheme.

1 INTRODUCTION

Finding optimal policies for Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) is hard due to *information asymmetry*, which refers to the mismatch in the set of information each agent has in a multi-agent environment. In fact, a finite-horizon Dec-POMDP with more than one agent is NEXP-complete (Bernstein et al., 2002), implying a doubly exponential complexity growth in the horizon length. In decentralized control theory, theoretical solutions have been proposed to find the optimal control laws for Dec-POMDPs. Notably among them is the common information (CI) approach (Nayyar et al., 2013), a framework that decomposes the decision of a full policy into the decision of a "prescription policy" from the CI known by all the agents, and the "prescription" itself which is a full characterization of how the agents should act based on any realization of their own private information (PI). This approach effectively transforms the decentralized model back to a centralized one from the view of a fictitious "coordinator" who only observes the CI, and permits a coordinator level sequential decomposition using a belief state a'la POMDPs (Kumar and Varaiya, 2015).

The challenge increases greatly in the multi-agent reinforcement learning (MARL) setting where the *model* – transition probabilities, observation kernel, and reward function – is unknown. When the agents learn concurrently, information asymmetry causes another issue called the "non-stationarity issue," since the effective environment observed by each agent is time-varying as the other agents learn and update their policies. The issue can be alleviated in principle by the "centralized learning and distributed execution" scheme (Dibangoye and Buffet, 2018) as the learning is from the coordinator's viewpoint; indeed, if agents only update their policies using CI, they can perfectly track others' policies. However, there is still a big gap in applying the CI approach to the MARL setting. First, the Bayesian updates of the belief state in the CI approach require the knowledge of the model, which is not available in the MARL setting. Moreover, the linear growth of length of private histories leads to the doubly exponential growth of the space of prescriptions in time, which is explosively large even for toy-size environments and forbids any practical explorations in such space. One natural question is whether we can restrict attention to some policies (and prescriptions) that take some state variable as

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

inputs without losing much performance, where the state variables encapsulate the crucial information relevant to future decisions in a time-invariant domain, and where the representations (ways of encapsulation) can be learned without the knowledge of the model.

In this paper, we formulate good approximate common and private state representations for learning close-to-optimal policies in unknown finite-horizon Dec-POMDPs, where each agent receives its own private information plus a *common* observation. The agents also share the same commonly observed rewards; however, they may not know each others' actions. We propose conditions in Definition 3 for an approximate sufficient private state (ASPS), which compresses an agent's private information, i.e., its action observation history (AOH), and conditions in Definition 5 for an approximate sufficient common state (ASCS), which compresses the fictitious coordinator's AOH, with the actions being ASPS-based prescriptions and the observations being common observations. Critically, using Theorem 4 and Theorem 6, in Theorem 7 we derive the optimality gap in terms of the error parameters of our compressions and the remaining time steps, between the value functions of two dynamic programmings (DPs): one in Algorithm 1 for the optimal policy using the CI approach without compression, with states being the complete coordinator's AOHs and actions being the prescriptions from Nayyar et al. (2013); and the other in Algorithm 3 using our framework with states being any *valid*[1] ASCSs and the actions being ASPS-based prescriptions for valid ASPS. Our framework generalizes a number of results in the literature: first, it extends the approximate information state (AIS) framework (Subramanian and Mahajan, 2019; Subramanian et al., 2020) to the multi-agent setting; second, it extends the CI approach (Nayyar et al., 2013) and the follow-up sufficient private information (SPI) framework that compresses the private states (Tavafoghi et al., 2021), to their approximate state representation counterpart; third, it generalizes the work by Mao et al. (2020) to include *non-injective* compressions and a general approximate common state representation. Our results can provide guidance on designing Deep Learning (DL) structures to learn the (compressed) state representations and the optimal policies (using learned representations) under the centralized learning distributed execution scheme, which applies to practical offline or online MARL settings.

Related Work. The problem of state representation is well studied in the single-agent POMDP case. Stochastic control theory details the conditions an *information state* (IS) needs to satisfy so that it acts as the Markov state in an equivalent MDP so one may only consider IS-based policies without loss of generality (Mahajan and Mannan, 2016); the belief state is an example of such IS (Kumar and Varaiya, 2015). Subramanian et al. (2020) extends the idea to an *approximate information state* (AIS), where the IS conditions hold approximately; importantly, the optimality gap of running DP with any valid AIS is quantified. Based on their AIS scheme, they propose a DL framework that learns the AIS representation without knowing the model. Recent work on *Deep Bisimulation for Control* (DBC) (Zhang et al., 2021b) in the DL literature uses similar ideas: they train an encoder to predict well the instantaneous rewards and transitions, and use the encoder output to train the policies. The encoder is an encapsulation or a compression. The optimality gap established is similar to the result of the infinite horizon case in Subramanian and Mahajan (2019). There are more representation learning schemes not requiring model knowledge in the DL or RL literature, e.g. Ha and Schmidhuber (2019), with the bulk without theoretical guidance or guarantees.

In the multi-agent context, Nayyar et al. (2013) propose a belief IS for the coordinator using the CI approach, without compressing agents' private information. Tavafoghi et al. (2018) further compress private histories to *sufficient private information* (SPI) so that the corresponding spaces of the belief IS and prescriptions are time-invariant. They identify conditions such that restricting attention to SPI-based policies is without loss of optimality. However, not only do they consider a control setting where the model is required, but also only present compression of the common history to a belief state, which is a narrow class of compression schemes. Nevertheless, this work will be a starting point of our work. Mao et al. (2020) consider an information state embedding that injectively maps agents' histories to representations in a fixed domain, and quantify the effect of the embedding on the value function like Subramanian and Mahajan (2019). However, their requirement that the mapping is injective is impractical for two reasons: one, an injective mapping does not reduce the policy complexity; and two, real world applications often demand non-injective encapsulations - e.g., tiger (Kaelbling et al., 1998) where one IS is the number of right observations minus the number of left observations, which is non-injective. Moreover, they also compress the common state to a belief state, but it is unclear how this can be done in practice without model information.

Another line of work in deep-MARL literature also applies the notion of CI (also known as the common knowledge) to solving MARL problems (Schroeder de Witt et al., 2019; Foerster et al., 2019; Lerer et al., 2019; Sokota et al., 2021). They search for opti-

[1] Satisfying the approximation criteria.

mal policies for a Dec-POMDP when the model is known, while we consider designing sample efficient and lower regret learning algorithms in an offline or online MARL setting for an unknown model. Moreover, many of them involve heuristic or approximation methods without knowing the potential loss from the approximations or apply a variety of machine learning schemes without a theoretical basis or understanding.

2 PRELIMINARIES

Notation. Let $\Delta(\mathcal{X})$ denote the set of distributions on the space $\mathcal{X}$, and $\Omega(X)$ denote the space where the variable X takes values. Superscripts are used as the agent index and subscripts as the time index. The notation $X_{a:b}^{c:d}$ denotes the tuple $(X_a^c, \ldots, X_a^d, \ldots, X_b^c, \ldots, X_b^d)$. In some cases superscripts or subscripts are omitted, and if so the meaning will be clarified. Capital letters are used for random variables while lower case letters are for their realizations. For random variables X with a realization x, we use the short hand notation $\mathbb{P}(\cdot|x) \triangleq \mathbb{P}(\cdot|X=x)$ and $\mathbb{E}[\cdot|x] \triangleq \mathbb{E}[\cdot|X=x]$. If a random variable appears without realization in a place other than the operand of $\mathbb{E}$, then it means the related equation should hold for any Borel measurable subset in its domain.

2.1 Dec-POMDP Model

We consider the Dec-POMDP model with N agents, i.e., a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}_T, R, \mathcal{O}, \mathbb{P}_O, T, \mathbb{P}_I)$ where the quantities are: $\mathcal{S}$ is the state space; $\mathcal{A} = \Pi_{n=1}^N \mathcal{A}^n$ is the joint action space whose elements are joint actions $A = A^{1:N}$; $\mathbb{P}_T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel mapping a current state and a joint action to a distribution of new states; $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function mapping a current state and a joint action to a real number; $\mathcal{O} = \Pi_{n=0}^N \mathcal{O}^n$ is the joint observation space whose elements are joint observations $O = O^{0:N}$, where O^0 is commonly observed but O^n is only observed by agent n; $\mathbb{P}_O : \mathcal{S} \to \Delta(\mathcal{O})$ is the observation kernel mapping a current state to a distribution of joint observations; T is the time horizon; $\mathbb{P}_I \in \Delta(\mathcal{S})$ is the initial state distribution. Compared to the standard Dec-POMDP model (Oliehoek and Amato, 2016), we have an additional common observation (including the reward), and our observations depend only on the current state.

We assume $\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$, and T are finite and known in advance, while $\mathbb{P}_T$, R, $\mathbb{P}_O$, and $\mathbb{P}_I$ are unknown in the MARL setting. Further, agents have perfect recall. At time t, agent n observes (O_t^0, O_t^n) generated from $\mathbb{P}_O(S_t)$, then uses the policy $A_t^n = g_t^n(O_{1:t}^0, g_{1:t-1}, H_t^n)$ to select its action, where $g_s = g_s^{1:N}$ and $H_t^n = (A_{1:t-1}^n, O_{1:t}^n)$ is agent n's private history and known as its AOH. The agents receive a reward $R_t \triangleq R(S_t, A_t)$[2], and the next state S_{t+1} is generated from $\mathbb{P}_T(S_t, A_t)$. The goal is to find a policy $g = g_{1:T}$ to maximize the common cumulative reward

$$\mathbb{E}\left[\sum_{t=1}^{T} R(S_t, A_t) \middle| g\right], \tag{1}$$

where the expectation is taken over the measure generated by policy g applied to model $(\mathbb{P}_T, R, \mathbb{P}_O, \mathbb{P}_I)$.

2.2 AIS Framework

In the single-agent POMDP setting, the spaces $\mathcal{A}$ and $\mathcal{O}$ are not product spaces, and at time t the agent's policy is of the form $g_t : \Omega(H_t) \to \Omega(A_t)$, where $H_t = (A_{1:t-1}, O_{1:t})$ is the agent's AOH. Note the policy space grows exponentially in t as the length of H_t grows linearly in t. Subramanian and Mahajan (2019) give conditions of a representation encapsulating the information in H_t that is approximately sufficient for decision purposes into a time-invariant space.

Definition 1: *An (ϵ, δ)-approximate information state $\widehat{Z}_t$ is the output of a function $\widehat{Z}_t = \widehat{\vartheta}_t(H_t)$ that satisfies the following properties:*

(AIS1) *It evolves recursively $\widehat{Z}_{t+1} = \widehat{\phi}_t(\widehat{Z}_t, A_t, O_{t+1})$.*

(AIS2) *It suffices for approximate performance evaluation $|\mathbb{E}[R_t|h_t, a_t] - \mathbb{E}[R_t|\widehat{z}_t, a_t]| \leq \epsilon \ \forall \ h_t, a_t$.*

(AIS3) *It suffices for approximately predicting the observation $\mathcal{K}(\mathbb{P}(O_{t+1}|h_t, a_t), \mathbb{P}(O_{t+1}|\widehat{z}_t, a_t)) \leq \delta \ \forall \ h_t, a_t$, where $\mathcal{K}(\cdot, \cdot)$ is a distance between two distributions*[3]*.*

The value function at t obtained from Bellman equations with $\widehat{Z}$'s as states falls behind the optimal value function at the most by an expression linear in $T-t$, ϵ, and δ (Subramanian and Mahajan, 2019). When $\epsilon = \delta = 0$, the expression is 0, and the AIS $\widehat{Z}$ degenerates to an IS Z.

In Subramanian and Mahajan (2019), a DL framework is provided to find an "approximate mapping" $\widehat{\vartheta}_t(\cdot)$ for any given POMDP model. The idea is to interpret the quantities in the LHS of (AIS2) and (AIS3) as driving the learning loss in DL, and let existing DL optimization algorithms find good mappings. The resulting AIS can then be used as the state in common policy approximation methods to find a near-optimal policy.

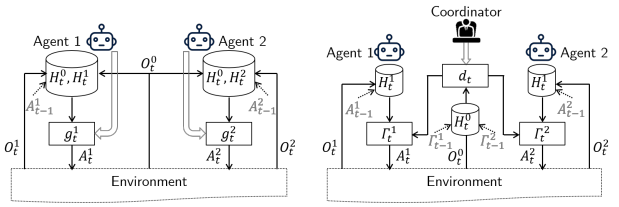
2.3 Common Information based DPs

2.3.1 DP with No Compression

In a DecPOMDP, the action decision for agent n at time t, $A_t^n = g_t^n(O_{1:t}^0, g_{1:t-1}, H_t^n)$, can be split into

[2] Or a noisy reward with mean R_t.

[3] For example, Wasserstein and total variation distances.



(a) The decision flow before decomposition. (b) The decision flow after decomposition.

Figure 1: Illustration of the CI approach.

two steps. In the first step, based on past common observations and policies $(O^0_{1:t}, g_{1:t-1})$ (using perfect recall), the agent decides g^n_t and hence $\Gamma^n_t(\cdot) \triangleq g^n_t(O^0_{1:t}, g_{1:t-1}, \cdot)$; then in the second step, it simply applies Γ^n_t to H^n_t to obtain the action $A^n_t = \Gamma^n_t(H^n_t)$. The function Γ^n_t is called the prescription (function), since it *prescribes* what the agent should do based on any possible realization of its private information.

This decomposition technique is called the CI approach (Nayyar et al., 2013). Note that the actual decision is carried out in the first step and solely upon CI (perfect recall makes policy common knowledge). One may then imagine there is a fictitious coordinator, labelled agent 0. At time t, the coordinator's policy is of the form $d_t : \Omega(H^0_t) \to \Omega(\Gamma_t)$, where $H^0_t \triangleq (O^0_{1:t}, \Gamma_{1:t-1})$ is equivalent to $(O^0_{1:t}, g_{1:t-1})$ and $\Gamma_t = \Gamma^{1:N}_t$; then it sends Γ_t to every agent, and agent n selects $A^n_t = \Gamma^n_t(H^n_t)$. It is shown that this decomposition is without loss of generality (so without loss of optimality too). The coordinator observes common observation O^0_t and chooses action Γ_t; hence, H^0_t can be seen as the coordinator's AOH and will be called the full common state (FCS), while H^n_t will be referred to as the full private state (FPS) of agent n. The transformation of the decision flows by the CI approach is illustrated in Figure 1, with the case of two agents.

From the perspective of the coordinator, the problem is now a centralized POMDP, and the goal is to find a policy $d = d_{1:T}$ that maximizes the expected cumulative reward. This permits a sequential decomposition with FCS as the state and an FPS-based prescription (meaning the prescription takes FPS as its input) as the action, which is presented in Algorithm 1.

Algorithm 1 Dynamic Programming with FCSs and FPS-based Prescriptions

$V_{T+1}(h^0_{T+1}) \triangleq 0$
for $t = T, \dots, 1$ **do**
  $Q_t(h^0_t, \gamma_t) = \mathbb{E}\big[R(S_t, \Gamma_t(H^{1:N}_t)) + V_{t+1}((H^0_t, \Gamma_t, O^0_{t+1}))|H^0_t = h^0_t, \Gamma_t = \gamma_t\big]$
  $V_t(h^0_t) = \max_{\gamma_t \in \Omega(\Gamma_t)} Q_t(h^0_t, \gamma_t)$

In practice, the coordinator is virtual and the computation of the coordinator is carried out in all agents – this is viable since the coordinator's computation only requires CI, which every agent has access to. Note the update of the state is done by direct concatenation of the incoming Γ_t and O^0_{t+1}.

2.3.2 DP with BCS

Nayyar et al. (2013) further compresses the FCS to the belief common state (BCS) $\Pi_t = \mathbb{P}(S_t, H^{1:N}_t | H^0_t)$, which is the conditional distribution on the state and the FPSs given the FCS. It is shown that restricting attention to coordinator's policy of the form $d_t : \Omega(\Pi_t) \to \Omega(\Gamma_t)$ is without loss of optimality. The DP presented thus uses this BCS as the state and an FPS-based prescription as the action – see Appendix A.2.

There are two problems with this approach when applied to the MARL setting. First, the BCS is updated via a Bayesian update using $\mathbb{P}_T$ and $\mathbb{P}_O$, which requires model knowledge. Second, the growing length of $H^{1:N}_t$ makes the spaces of Π_t and Γ_t explosively large and impossible to explore. At a conceptual level we can apply the AIS framework to the centralized POMDP of the coordinator[4]; however, the underlying decentralized information structure coupled with increasing domain of private information makes practical implementations of this scheme challenging.

2.3.3 DP with BCS and SPI

To alleviate the aforementioned dimensionality issue, Tavafoghi et al. (2018) further compresses the FPS to a representation called the sufficient private information (SPI) lying in a time-invariant domain. They identify a set of conditions for the compression so that the SPI is sufficient for decision making purposes.

Definition 2: *A sufficient private information (SPI) $Z^{1:N}_t$ is a tuple of outputs of a set of functions $Z^n_t = \vartheta^n_t(H^0_t, H^n_t)\ \forall\ n \in [N]$ satisfying the properties:*

(SPI1) *It evolves recursively, i.e., $\forall\ n \in [N]$, $Z^n_t = \phi^n_t(Z^n_{t-1}, \Gamma_{t-1}, O^0_t, A^n_{t-1}, O^n_t, g_{1:t-1})$.*

(SPI2) *It suffices for performance evaluation $\mathbb{E}[R(S_t, A_t)|h^0_t, h^n_t, a_t] = \mathbb{E}[R(S_t, A_t)|h^0_t, z^n_t, a_t]$ $\forall\ n, h^0_t, h^n_t, a_t$.*

(SPI3) *It suffices for predicting itself and the common observation $\mathbb{P}(Z^{1:N}_{t+1}, O^0_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, a_t) = \mathbb{P}(Z^{1:N}_{t+1}, O^0_{t+1}|h^0_t, z^{1:N}_t, \gamma_t, a_t)\ \forall\ h^{0:N}_t, \gamma_t, a_t$.*

(SPI4) *It suffices for predicting other agents' SPI $\mathbb{P}(Z^{-n}_t|h^0_t, h^n_t) = \mathbb{P}(Z^{-n}_t|h^0_t, z^n_t)\ \forall\ n, h^0_t, h^n_t$.*

The coordinator now considers SPI-based prescriptions $\Lambda_t = \Lambda^{1:N}_t$ where $A^n_t = \Lambda^n_t(Z^n_t)$, and the BCS is changed to $\tilde{\Pi}_t = \mathbb{P}(S_t, Z^{1:N}_t | \tilde{H}^0_t)$ where $\tilde{H}^0_t =$

[4] Strictly speaking, this requires a straightforward extension to time-varying action spaces for different time steps – see Subramanian et al. (2020) Section 5 for details.

$(O_1^0, \Lambda_1, \ldots, O_t^0)$. It is shown that restricting attention to coordinator's policy of the form $\tilde{d}_t : \Omega(\tilde{\Pi}_t) \to \Omega(\Lambda_t)$ is without loss of optimality. The resulting DP uses the BCS as the state and SPI-based prescription as the action – see Appendix A.3. Note that the compression actually leads to an action compression for the coordinator – from FPS-based prescriptions to SPI-based prescriptions – which has no loss in performance.

With $\tilde{\Pi}_t$, $Z_t^{1:N}$, and Λ_t all lying in time-invariant spaces, the complexity no longer grows with time. However, it is unclear how to find mappings satisfying Definition 2 and update the BCSs in an MARL setting. Further, the solution focuses on a decentralized setting wherein the (lossless) compression functions are consistent (common knowledge), and the performance assessments and predictions are based only on the information of any particular agent. Ensuring these properties in the RL context would require significant communication, particularly during learning the compression.

3 APPROXIMATE STATE REPRESENTATIONS

We seek to extend the idea of identifying representations sufficient for approximately optimal decision making from Section 2.2 to the multi-agent setting, and develop a general compression framework for common states and private states (hence also prescriptions) whose mappings can be learned from samples obtained by interacting with the environment alone.

In this section, we propose our general states representation framework for approximate planning and control in partially observable MARL problems. We start by compressing private histories to ASPS; for the coordinator, this induces an action compression from FPS-based prescriptions to ASPS-based prescriptions. Then based on this compression, the common history is further compressed to ASCS.

The framework we develop will be consistent with the philosophy of recent empirical MARL work wherein there is a centralized agent called the supervisor. The supervisor observes all the quantities and develops good compression of private information and common information that the coordinator can use to produce close-to-optimal prescriptions (using the compressed common information), which can be implemented by the agents using just their own compressed private information. We detail the supervisor in Section 4.1 but point out here that it has the knowledge of $H_t^{0:N}$ for all $t \in [T]$. Note that this viewpoint is consistent with the "centralized training with distributed execution" setting of the empirical MARL work.

3.1 Compressing Private States

Definition 3: *An (ϵ_p, δ_p)-approximate sufficient private state (ASPS) $\widehat{Z}_t^{1:N}$ is a tuple of outputs of a set of functions $\widehat{Z}_t^n = \widehat{\vartheta}_t^n(H_t^0, H_t^n)\ \forall\ n \in [N]$ satisfying:*

(ASPS1) *It evolves in a recursive manner, that is, $\forall\ n \in [N]$, $\widehat{Z}_t^n = \widehat{\phi}_t^n(\widehat{Z}_{t-1}^n, \Gamma_{t-1}, O_t^0, O_t^n)$.*

(ASPS2) *It suffices for approximate performance evaluation $\left|\mathbb{E}[R(S_t, A_t)|h_t^0, h_t^{1:N}, a_t] - \mathbb{E}[R(S_t, A_t)|h_t^0, \widehat{z}_t^{1:N}, a_t]\right| \le \epsilon_p/4\ \forall\ h_t^{0:N}, a_t$.*

(ASPS3) *It suffices for approximately predicting observations $\mathcal{K}\big(\mathbb{P}(O_{t+1}^{0:N}|h_t^0, h_t^{1:N}, a_t), \mathbb{P}(O_{t+1}^{0:N}|h_t^0, \widehat{z}_t^{1:N}, a_t)\big) \le \delta_p/8\ \forall\ h_t^{0:N}, a_t$.*

This definition induces the ASPS-based prescription, which is a mapping $\widehat{\lambda}_t : \Omega(\widehat{Z}_t^{1:N}) \to \Omega(A_t)$ that prescribes the action tuple for all ASPSs $A_t = \widehat{\lambda}_t(\widehat{Z}_t^{1:N}) = (\widehat{\lambda}_t^1(\widehat{Z}_t^1), \ldots, \widehat{\lambda}_t^N(\widehat{Z}_t^N))$ in a component-wise manner. One can run a DP with FCSs as states and ASPS-based prescriptions as actions – see Algorithm 2.

Algorithm 2 Dynamic Programming with FCSs and ASPS-based Prescriptions

$\widehat{V}_{T+1}(h_{T+1}^0) \triangleq 0$
for $t = T, \ldots, 1$ **do**
  $\widehat{Q}_t(h_t^0, \widehat{\lambda}_t) = \mathbb{E}\big[R_t(S_t, \widehat{\lambda}_t(\widehat{Z}_t^{1:N})) + \widehat{V}_{t+1}((H_t^0, \widehat{\Lambda}_t, O_{t+1}^0))|H_t^0 = h_t^0, \widehat{\Lambda}_t = \widehat{\lambda}_t\big]$
  $\widehat{V}_t(h_t^0) = \max_{\widehat{\lambda}_t \in \Omega(\widehat{\Lambda}_t)} \widehat{Q}_t(h_t^0, \widehat{\lambda}_t)$

The compression is characterized by functions $\widehat{\vartheta}_t^{1:N}$. These functions also relate $\Omega(\Gamma_t)$ and $\Omega(\widehat{\Lambda}_t)$ as ASPS-based prescriptions are a strict subset of FPS-based prescritions; this detail will be explained in Section 4.2. For now we note that here the conditions we set for the action compression from $\Omega(\Gamma_t)$ to $\Omega(\widehat{\Lambda}_t)$ are on the private states instead of defining an encapsulation directly on the actions (i.e., prescriptions); moreover, the compression may depend on the common state h_t^0 as well. Hence, this falls outside of the action compression scheme studied in Subramanian et al. (2020). We bound the error between the value functions obtained from Algorithm 2 and the optimal value functions obtained from Algorithm 1 in the following theorem proved in Section 4.2.

Theorem 4: *Assume the reward function R is bounded by $\bar{R}$. For any $h_t^0 \in \Omega(H_t^0)$ and $\gamma^* \in \operatorname{argmax}_\gamma Q_t(h_t^0, \gamma)$, with $\bar{t} = T - t$, there exists a $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$ such that*

$$Q_t(h_t^0, \gamma^*) - \widehat{Q}_t(h_t^0, \widehat{\lambda}) \le \frac{\bar{t}(\bar{t}+1)}{2}(\epsilon_p + T\bar{R}\delta_p) + (\bar{t}+1)\epsilon_p, \tag{2}$$

$$V_t(h_t^0) - \widehat{V}_t(h_t^0) \le \frac{\bar{t}(\bar{t}+1)}{2}(\epsilon_p + T\bar{R}\delta_p) + (\bar{t}+1)\epsilon_p. \tag{3}$$

3.2 Compressing Common States

While restricting attention to ASPS-based prescriptions, we further compress the common history to an approximate representation by applying the state compression result of Subramanian and Mahajan (2019).

Definition 5: *An (ϵ_c, δ_c)-approximate sufficient common state (ASCS) $\widehat{Z}_t^0$ is the output of a function $\widehat{Z}_t^0 = \widehat{\vartheta}_t^0(H_t^0)$ satisfying the properties:*

(ASCS1) *It evolves in a recursive manner, that is, $\widehat{Z}_t^0 = \widehat{\phi}_t^0(\widehat{Z}_{t-1}^0, \widehat{\Lambda}_{t-1}, O_t^0)$.*

(ASCS2) *It suffices for approximate performance evaluation, i.e., $\forall\ h_t^0, \widehat{\lambda}_t$, we have $\big|\mathbb{E}[R_t(S_t, A_t)|h_t^0, \widehat{\lambda}_t] - \mathbb{E}[R_t(S_t, A_t)|\hat{z}_t^0, \widehat{\lambda}_t]\big| \le \epsilon_c$.*

(ASCS3) *It suffices for approximately predicting common observation, i.e., $\forall\ h_t^0, \widehat{\lambda}_t$, we have $\mathcal{K}\big(\mathbb{P}(O_{t+1}^0|h_t^0, \widehat{\lambda}_t), \mathbb{P}(O_{t+1}^0|\hat{z}_t^0, \widehat{\lambda}_t)\big) \le \delta_c/2$.*

In our proposed representation framework, agents compress the CI and PI to ASCS $\widehat{Z}_t^0$ and ASPS $\widehat{Z}_t^{1:N}$, which can be updated recursively using the incoming CI and PI. Agents use the same policy $\breve{d}_t : \Omega(\widehat{Z}_t^0) \to \Omega(\widehat{\Lambda}_t)$ to decide the ASPS-based prescription $\widehat{\Lambda}_t$ from $\widehat{Z}_t^0$, then they apply $\widehat{\Lambda}_t$ to their own ASPS $\widehat{Z}_t^n$ to obtain the action A_t^n. Approximately optimal policies then result from the DP with ASCSs as states and ASPS-based prescriptions as actions – see Algorithm 3.

Algorithm 3 Dynamic Programming with ASCSs and ASPS-based Prescriptions

$\breve{V}_{T+1}(\hat{z}_{T+1}^0) \triangleq 0$

for $t = T, \dots, 1$ **do**

$$\breve{Q}_t(\hat{z}_t^0, \widehat{\lambda}_t) = \mathbb{E}\big[R(S_t, \widehat{\Lambda}_t(\widehat{Z}_t^{1:N})) + \breve{V}_{t+1}(\widehat{\phi}_t^0(\widehat{Z}_t^0, \widehat{\Lambda}_t, O_{t+1}^0))|\widehat{Z}_t^0 = \hat{z}_t^0, \widehat{\Lambda}_t = \widehat{\lambda}_t\big]$$

$$\breve{V}_t(\hat{z}_t^0) = \max_{\widehat{\lambda}_t \in \Omega(\widehat{\Lambda}_t)} \breve{Q}_t(\hat{z}_t^0, \widehat{\lambda}_t)$$

From Algorithm 2 to Algorithm 3, only the states are further compressed, so a gap result bounding the difference between the two DPs holds, similar to the result in Subramanian and Mahajan (2019). See Appendix C for details.

Theorem 6: *Assume the reward function R is bounded by $\bar{R}$. For any $h_t^0 \in \Omega(H_t^0)$ and $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$, with $\bar{t} = T - t$, we have*

$$\widehat{Q}_t(h_t^0, \widehat{\lambda}) - \breve{Q}_t(\widehat{\vartheta}_t^0(h_t^0), \widehat{\lambda}) \le \bar{t}(\epsilon_c + T\bar{R}\delta_c) + \epsilon_c, \quad (4)$$

$$\widehat{V}_t(h_t^0) - \breve{V}_t(\widehat{\vartheta}_t^0(h_t^0)) \le \bar{t}(\epsilon_c + T\bar{R}\delta_c) + \epsilon_c. \quad (5)$$

3.3 Main Result

Our main result bounds the optimality gap of value functions obtained from performing DP with the general common and private representations satisfying the conditions in Definition 3 and Definition 5 as in Algorithm 3, in comparison to the optimal value functions computed from Algorithm 1.

Theorem 7: *Assume the reward function R is bounded by $\bar{R}$. For any $h_t^0 \in \Omega(H_t^0)$ and $\gamma^* \in \operatorname{argmax}_\gamma Q_t(h_t^0, \gamma)$, with $\bar{t} = T - t$, there exists a $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$ such that*

$$Q_t(h_t^0, \gamma^*) - \breve{Q}_t(\widehat{\vartheta}_t^0(h_t^0), \widehat{\lambda}) \le \frac{\bar{t}(\bar{t}+1)}{2}(\epsilon_p + T\bar{R}\delta_p) + (\bar{t}+1)(\epsilon_c + \epsilon_p) + \bar{t}T\bar{R}\delta_c, \quad (6)$$

$$V_t(h_t^0) - \breve{V}_t(\widehat{\vartheta}_t^0(h_t^0)) \le \frac{\bar{t}(\bar{t}+1)}{2}(\epsilon_p + T\bar{R}\delta_p) + (\bar{t}+1)(\epsilon_c + \epsilon_p) + \bar{t}T\bar{R}\delta_c. \quad (7)$$

Proof: Combine Theorem 4 and Theorem 6. ■

We observe that the action compression induced by private state compression leads to a gap quadratic in remaining time $\bar{t} = T - t$ (Theorem 4), and common state compression causes a gap linear in remaining time. Also, note that the gap decreases to 0 as $(\epsilon_c, \delta_c, \epsilon_p, \delta_p)$ go to 0. Having developed this result, the remaining questions for learning using the sample data from the environment are: how to learn the compression mappings $\widehat{\vartheta}_{1:T}^{0:N}$ with small error; and how to learn good policies with the compressed representations. See Appendix E for a proposed scheme to answer both these questions using DL methods.

3.4 Comparisons to Existing Schemes

Nayyar et al. (2013) and Tavafoghi et al. (2018) provide lossless (performance-wise) compression. We refer to ASPS and its corresponding conditions with $\epsilon_p = \delta_p = 0$ as SPS; similarly, we refer to ASCS and its corresponding conditions with $\epsilon_c = \delta_c = 0$ as SCS. Missing proofs in this subsection are in Appendix D.

Relation to Nayyar et al. (2013). The private history is not compressed in Nayyar et al. (2013), so it is clearly a special case of SPS. The BCS proposed in Nayyar et al. (2013) is a special case of SCS as well.

Proposition 8: *The BCS $\Pi_t = \mathbb{P}(S_t, H_t^{1:N}|H_t^0)$ satisfies the conditions of an SCS in Definition 5 with $\epsilon_c = \delta_c = 0$.*

Relation to Tavafoghi et al. (2018). Our conditions of SPS and Tavafoghi et al. (2018)'s conditions of SPI both lead to performance sufficiency of the space of SPI-based (or SPS-based) prescriptions. The two sets of conditions are similar but not exactly the same. Condition (SPI1) corresponds to (SPS1); however, (SPS1) is stricter since we require *policy-independent compression*, while Tavafoghi et al.

(2018) allow policy-dependent compression. Condition (SPI3) ensures *future sufficiency* as does (SPS3). Conditions (SPI2) and (SPI4) together ensure *present sufficiency* as does (SPS3).

Proposition 9: *(SPS1) and (SPS3) imply (SPI3).*

Proposition 10: *(SPS2) and (SPI4) imply (SPI2).*

Restricting to SPS, their BCS $\tilde{\Pi}_t = \mathbb{P}(S_t, Z_t^{1:N}|H_t^0)$ is a special case of SCS as well, so a result identical to Proposition 8 holds with FPS changed to SPS.

Relation to Mao et al. (2020). Their private state embedding does not require a recursive update (ASPS1), but demands injective functions $\widehat{\vartheta}_t^{1:N}$. With this additional assumption they show linearity of the optimality gap in remaining time. For the common state, the BCS they consider is a special case of our SCS, just as the BCS of Tavafoghi et al. (2018).

A thorough comparison of the DPs proposed in this work and in the literature is summarized in Table 1.

4 OPTIMALITY GAP ANALYSIS

In this section, we outline the optimality gaps introduced in Section 3; details are in Appendix B.

4.1 Supervisor's Functions

For better exposition, we introduce another set of Q/V functions from an omniscient *supervisor*'s perspective, for the original decision problem. The supervisor can access the *union* of the information of all agents: at time t the supervisor knows $H_t^{0:N}$. In contrast, coordinator's information is the *intersection* of the information of all agents: H_t^0. The supervisor, however, only observes what is happening, lets the coordinator decide all the policies and prescriptions, and implements the coordinator's policies. Let $d_{1:T}^*$ be a coordinator's optimal policy solved using Algorithm 1, i.e. $d_t^*(h_t^0) \in \operatorname{argmax}_{\gamma_t \in \Omega(\Gamma_t)} Q_t(h_t^0, \gamma_t)$. Then the Q/V functions defined in Algorithm 1 can be rewritten as

$$Q_t(h_t^0, \gamma_t) = \mathbb{E}\left[\sum_{\tau=t}^{T} R_\tau \middle| h_t^0, \gamma_t, d_{t+1:T}^*\right], \tag{8}$$

$$V_t(h_t^0) = \mathbb{E}\left[\sum_{\tau=t}^{T} R_\tau \middle| h_t^0, d_{t:T}^*\right]. \tag{9}$$

The supervisor's Q/V functions use similar concepts, but with *supervisor's states* and *coordinator's policies*.

Definition 11: *For any $h_t^{0:N} \in \Omega(H_t^{0:N})$, $\gamma_t \in \Omega(\Gamma_t)$, define the supervisor's Q function as*

$$Q_t^S(h_t^0, h_t^{1:N}, \gamma_t) \triangleq \mathbb{E}\left[\sum_{\tau=t}^{T} R_\tau \middle| h_t^0, h_t^{1:N}, \gamma_t, d_{t+1:T}^*\right], \tag{10}$$

and the supervisor's V function as

$$V_t^S(h_t^0, h_t^{1:N}) \triangleq \mathbb{E}\left[\sum_{\tau=t}^{T} R_\tau \middle| h_t^0, h_t^{1:N}, d_{t:T}^*\right] = Q_t^S(h_t^0, h_t^{1:N}, \gamma_t^*), \tag{11}$$

where $\gamma_t^* \in \operatorname{argmax}_{\gamma_t \in \Omega(\Gamma_t)} Q_t(h_t^0, \gamma_t)$[5].

Then the coordinator's Q/V functions can be expressed as the expectation of supervisor's Q/V functions taken over the conditional distribution on FPSs given the FCS:

$$Q_t(h_t^0, \gamma_t) = \sum_{h_t^{1:N}} \mathbb{P}(h_t^{1:N}|h_t^0) Q_t^S(h_t^0, h_t^{1:N}, \gamma_t), \tag{12}$$

$$V_t(h_t^0) = \sum_{h_t^{1:N}} \mathbb{P}(h_t^{1:N}|h_t^0) V_t^S(h_t^0, h_t^{1:N}). \tag{13}$$

4.2 Proof of Theorem 4

We first determine the relationship between the space of FPS-based prescriptions $\Omega(\Gamma_t)$ and the space of ASPS-based prescriptions $\Omega(\widehat{\Lambda}_t)$. Consider a fixed h_t^0. Since the compression mappings $\widehat{Z}_t^{1:N} = \widehat{\vartheta}_t^{1:N}(H_t^0, H_t^{1:N})$ are functions, there could be multiple $h_t^{1:N}$'s that are mapped to the same $\widehat{z}_t^{1:N}$. A $\widehat{\lambda}_t \in \Omega(\widehat{\Lambda}_t)$ can thus be thought of as a special element of $\Omega(\Gamma_t)$ that prescribes the same action for all the FPSs $h_t^{1:N}$'s mapped to the same ASPS $\widehat{z}_t^{1:N}$. Hence, we can construct an injective *extension mapping* from $\Omega(\widehat{\Lambda}_t)$ to $\Omega(\Gamma_t)$ in this sense.

Definition 12: *For any $h_t^0 \in \Omega(H_t^0)$, define the extension mapping $\psi_t : \Omega(\widehat{\Lambda}_t) \times \Omega(H_t^0) \to \Omega(\Gamma_t)$ as follows: for any $h_t^{1:N}$ and $\widehat{\lambda}_t$, $\gamma_t = \psi_t(\widehat{\lambda}_t, h_t^0)$ will first compress $h_t^{1:N}$ to $\widehat{z}_t^{1:N} = \widehat{\vartheta}_t^{1:N}(h_t^0, h_t^{1:N})$ (hence ψ_t implicitly depends on $\widehat{\vartheta}_t$), then choose the action according to $\widehat{\lambda}_t(\widehat{z}_t^{1:N})$. That is,*

$$\gamma_t(h_t^{1:N}) = \psi_t(\widehat{\lambda}_t, h_t^0)(h_t^{1:N}) \triangleq \widehat{\lambda}_t(\widehat{\vartheta}_t^{1:N}(h_t^0, h_t^{1:N})).$$

Given the compression $\widehat{\vartheta}_t^{1:N}$, ψ_t is well-defined. Under this circumstance and with an abuse of notation, $\gamma_t = \psi_t(\widehat{\lambda}_t, h_t^0)$ will be written as $\gamma_{\widehat{\lambda}_t, h_t^0}$ when the considered compression is clear from the context and will be referred to as the γ_t extended from $\widehat{\lambda}_t$ under h_t^0.

The following proposition says that for any FCS one can find an ASPS-based prescription whose extension nearly achieves the same Q-value as an optimal prescription, up to a gap linear in the remaining time $\bar{t}$.

[5] The supervisor's Q function and V function are only defined when the FPS $h_t^{1:N}$ is *admissible* under h_t^0, i.e. $\mathbb{P}(h_t^{1:N}|h_t^0) > 0$. Throughout the rest of the paper, we assume that only admissible FPSs are considered.

This implies that it nearly suffices to consider the class of prescriptions extended from ASPS-based prescriptions for DP purposes.

Proposition 13: *Assume the reward function R is bounded by $\bar{R}$. For any $h_t^0 \in \Omega(H_t^0)$ and $\gamma^* \in \operatorname{argmax}_\gamma Q_t(h_t^0, \gamma)$, there exists a $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$ with*

$$\left|Q_t(h_t^0, \gamma^*) - Q_t(h_t^0, \gamma_{\widehat{\lambda}, h_t^0})\right| \leq \bar{t}(\epsilon_p + T\bar{R}\delta_p) + \epsilon_p, \quad (14)$$

which leads to

$$\left|V_t(h_t^0) - \max_{\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)} Q_t(h_t^0, \gamma_{\widehat{\lambda}, h_t^0})\right| \leq \bar{t}(\epsilon_p + T\bar{R}\delta_p) + \epsilon_p. \quad (15)$$

Before proving this critical proposition we need a few intermediate results. The first key lemma says that with the same supervisor's state, the supervisor's Q-values for two different prescriptions will be the same as long as they prescribe the same action for the given PI. In particular, the fact that the two prescriptions may prescribe different actions for other PIs becomes irrelevant after learning the given PI is realized from the supervisor's view.

Lemma 14: *For any $h_t^0 \in \Omega(H_t^0)$ and $h \in \Omega(H_t^{1:N})$, let $\gamma_1, \gamma_2 \in \Omega(\Gamma_t)$ be two prescriptions that choose the same action on h, i.e. $\gamma_1(h) = \gamma_2(h) = a$. Then*

$$Q_t^S(h_t^0, h, \gamma_1) = Q_t^S(h_t^0, h, \gamma_2). \quad (16)$$

Lemma 14 has an important implication for the structure of the optimal prescription. Let us define $Q_t^S(h_t^0, h, \gamma_1) \triangleq Q_t^S(h_t^0, h, a)$, which is a well-defined quantity from Lemma 14. Given any $h_t^0 \in \Omega(H_t^0)$ and $h_t^{1:N} \in \Omega(H_t^{1:N})$, since the choice of $\gamma_t(h_t^{1:N})$ has no bearing on other $Q_t^S(h_t^0, h, \gamma_t)$'s where $h \neq h_t^{1:N}$, we want to select an $\gamma_t(h_t^{1:N})$ that maximizes $Q_t^S(h_t^0, h_t^{1:N}, \gamma_t)$ as a result of (12). It follows that an optimal prescription $\gamma^* \in \operatorname{argmax}_{\gamma_t} Q_t(h_t^0, \gamma_t)$ is a prescription that prescribes optimal actions for all PIs:

$$\gamma^*(h) \in \operatorname*{argmax}_{a_t} Q_t^S(h_t^0, h, a_t) \ \ \forall \ h \in \Omega(H_t^{1:N}). \quad (17)$$

Lemma 14 represents a *reduction from doubly exponential to exponential* in search complexity of an optimal policy. Specifically, given any $h_t^0 \in \Omega(H_t^0)$, instead of searching γ^* in the whole space of $\Omega(\Gamma_t)$ whose size is $|\Omega(A_t)|^{|\Omega(H_t^{1:N})|}$, the structure in (17) suggests searching $\Omega(A_t)$ for each $h_t^{1:N} \in \Omega(H_t^{1:N})$, totaling a size of $|\Omega(H_t^{1:N})| \cdot |\Omega(A_t)|$[6]. This particular structure of optimal prescriptions might be helpful when designing the prescription representations for the agents in practical implementations. Further *reduction from exponential to constant* in search complexity is done by compressing public and private states to time-invariant spaces.

Next we show that the supervisor's Q-values will be nearly the same for two different FPSs that map to the same ASPS and a prescription that prescribes the same action on these two FPSs.

Lemma 15: *Assume the reward function R is bounded by $\bar{R}$. For any $h_t^0 \in \Omega(H_t^0)$, let $h_1, h_2 \in \Omega(H_t^{1:N})$ be two FPSs under h_t^0 that map to the same ASPS $\widehat{z} \in \Omega(\widehat{Z}_t^{1:N})$, i.e. $\widehat{z} = \widehat{\vartheta}_t^{1:N}(h_t^0, h_1) = \widehat{\vartheta}_t^{1:N}(h_t^0, h_2)$, and let $\gamma \in \Omega(\Gamma_t)$ be a prescription that chooses the same action on these two FPSs $\gamma(h_1) = \gamma(h_2) = a$. Then*

$$\left|Q_t^S(h_t^0, h_1, \gamma) - Q_t^S(h_t^0, h_2, \gamma)\right| \leq \bar{t}(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2. \quad (18)$$

Mixing up two FPSs with the same compressed state will incur a constant instantaneous cost resulted from (ASPS2), and a transitioning cost from (ASPS3). They will transition to two FPSs with the same compression again from (ASPS1), which suggests a continuation cost linear in $\bar{t}$ by induction.

Using the above two lemmas, we show that the supervisor's V function will differ little for two supervisor's states with the same compression of private states.

Corollary 16: *Assume the reward function R is bounded by $\bar{R}$. For any $h_t^0 \in \Omega(H_t^0)$, let $h_1, h_2 \in \Omega(H_t^{1:N})$ be two FPSs under h_t^0 that map to the same ASPS $\widehat{z} \in \Omega(\widehat{Z}_t^{1:N})$, i.e. $\widehat{z} = \widehat{\vartheta}_t(h_t^0, h_1) = \widehat{\vartheta}_t(h_t^0, h_2)$, and let $\gamma^* \in \operatorname{argmax}_\gamma Q_t(h_t^0, \gamma)$ be an optimal prescription. Then*

$$\begin{aligned} &\left|V_t^S(h_t^0, h_1) - V_t^S(h_t^0, h_2)\right| \\ &\triangleq \left|Q_t^S(h_t^0, h_1, \gamma^*) - Q_t^S(h_t^0, h_2, \gamma^*)\right| \\ &\leq \bar{t}(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2. \end{aligned} \quad (19)$$

Proof of Proposition 13: Given an optimal prescription $\gamma^* \in \operatorname{argmax}_\gamma Q_t(h_t^0, \gamma)$, we will specifically construct a $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$ that serves for the claim. For each $\widehat{z} \in \Omega(\widehat{Z}_t^{1:N})$, define

$$\mathcal{H}_{\widehat{z}} = \left\{h \in \Omega(H_t^{1:N}) : \widehat{\vartheta}_t^{1:N}(h_t^0, h) = \widehat{z}\right\} \quad (20)$$

to be the class of h's in $\Omega(H_t^{1:N})$ that are compressed to $\widehat{z}$ under the considered compression $\widehat{\vartheta}_t^{1:N}$. By the Axiom of Choice, for each $\widehat{z} \in \Omega(\widehat{Z}_t^{1:N})$ there exists a representative of the class $\mathcal{H}_{\widehat{z}}$ coming from arbitrary choice function, which we denote as $\bar{h}_{\widehat{z}}$. We then construct the $\widehat{\lambda}$ by $\widehat{\lambda}(\widehat{z}) = \gamma^*(\bar{h}_{\widehat{z}})$; the corresponding extension in $\Omega(\Gamma_t)$ will be

$$\gamma_{\widehat{\lambda}, h_t^0}(h) = \gamma^*\left(\bar{h}_{\widehat{\vartheta}_t^{1:N}(h_t^0, h)}\right) \quad \forall \ h \in \Omega(H_t^{1:N}), \quad (21)$$

[6] The size $|\Omega(H_t^{1:N})|$ is of order $\mathcal{O}(e^T)$ as the length of $H_t^{1:N}$ is linear in T; hence, $|\Omega(A_t)|^{|\Omega(H_t^{1:N})|}$ is doubly exponential in T while $|\Omega(H_t^{1:N})| \cdot |\Omega(A_t)|$ is exponential in T. The search should be performed for all $h_t^0 \in \Omega(H_t^0)$, where $|\Omega(H_t^0)|$ is also $\mathcal{O}(e^T)$; this does not change the complexities for the two cases.

that is, the prescription first compresses the input FPS and finds the representative of the corresponding compression class, then it mimics what the optimal prescription would have done with the representative. For any $h \in \Omega(H_t^{1:N})$, we have

$$\begin{aligned}
&\left|Q_t^S(h_t^0, h, \gamma^*) - Q_t^S(h_t^0, h, \gamma_{\widehat{\lambda}, h_t^0})\right| \\
\leq &\left|Q_t^S(h_t^0, h, \gamma^*) - Q_t^S(h_t^0, \bar{h}_{\widehat{\vartheta}_t^{1:N}(h_t^0, h)}, \gamma^*)\right| \\
&+ \left|Q_t^S(h_t^0, \bar{h}_{\widehat{\vartheta}_t^{1:N}(h_t^0, h)}, \gamma^*) - Q_t^S(h_t^0, \bar{h}_{\widehat{\vartheta}_t^{1:N}(h_t^0, h)}, \gamma_{\widehat{\lambda}, h_t^0})\right| \\
&+ \left|Q_t^S(h_t^0, \bar{h}_{\widehat{\vartheta}_t^{1:N}(h_t^0, h)}, \gamma_{\widehat{\lambda}, h_t^0}) - Q_t^S(h_t^0, h, \gamma_{\widehat{\lambda}, h_t^0})\right| \\
\leq &(T-t)(\epsilon_p + T\bar{R}\delta_p) + \epsilon_p,
\end{aligned}$$

as the first term is bounded by $(T-t)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2$ due to Corollary 16, the second term is 0 due to Lemma 14, and the third term is bounded by $(T-t)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2$ due to Lemma 15. If it happens to be the case that $h = \bar{h}_{\widehat{\vartheta}_t^{1:N}(h_t^0, h)}$, i.e. h is the representative, then the original term $|Q_t^S(h_t^0, h, \gamma^*) - Q_t^S(h_t^0, h, \gamma_{\widehat{\lambda}, h_t^0})|$ is 0. Taking the conditional expectation on h given h_t^0 gives the claim. ■

Proof of Theorem 4: There are three main quantities: $V_t(h_t^0)$ is the value obtained from executing optimal FPS-based prescriptions to the end, $\max_{\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)} Q_t(h_t^0, \gamma_{\widehat{\lambda}, h_t^0})$ is from executing the optimal ASPS-based prescription for step t and then optimal FPS-based prescriptions afterwards to the end, and $\widehat{V}_t(h_t^0)$ is from executing optimal ASPS-based prescriptions to the end. Proposition 13 establishes that restricting to ASPS-based prescriptions in *one step* incurs a gap (between $V_t(h_t^0)$ and $\max_{\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)} Q_t(h_t^0, \gamma_{\widehat{\lambda}, h_t^0})$) *linear* in $T-t$. Using an induction argument to accumulate this gap in *every step* from T back to t yields the gap (between $V_t(h_t^0)$ and $\widehat{V}_t(h_t^0)$) to be *quadratic* in $T-t$. See Appendix B for detailed derivations. ■

We briefly summarize the analysis framework presented in Section 4.2. In Lemma 14 and Lemma 15, we bound the differences of two supervisor's Q-functions, i.e., $\left|Q_t^S(h_t^0, h, \gamma_1) - Q_t^S(h_t^0, h, \gamma_2)\right|$ and $\left|Q_t^S(h_t^0, h_1, \gamma) - Q_t^S(h_t^0, h_2, \gamma)\right|$. The two results lead to Corollary 16, which bounds the difference of two supervisor's V-functions $\left|V_t^S(h_t^0, h_1) - V_t^S(h_t^0, h_2)\right|$. From these three intermediate results, we quantify the cost of restricting to ASPS-based prescriptions *in one step* in Proposition 13, which is linear in $T-t$. Then we induct this result through the horizon, which shows restricting to ASPS-based prescriptions all the way incurs a gap quadratic in $T-t$.

5 PRACTICAL IMPLICATIONS

The main result provides a nice theoretical support of designing practical low-regret deep-MARL algorithms, just as the way the DL schemes solves POMDP RL proposed in Subramanian et al. (2020). Algorithms using our framework will consists of two steps when learning optimal policies. In the first step, the agents use function approximation (FA) methods (e.g. DL models) to learn representations of the common and private states. Our contribution is to identify measures for "good representations" – they should satisfy the conditions of ASCS and ASPS with low error parameters $(\epsilon_c, \delta_c, \epsilon_p, \delta_p)$. The FA methods will try to predict the instantaneous reward and new observations while quantities in the LHS of the conditions are good candidates of the loss functions. In the second step, assuming the agents have learned good representations, they use policy approximation theory (Sutton and Barto, 2018) to learn good policies from the *coordinator's view*, thus alleviating the non-stationarity issue. The parameters of the policy function approximator can be updated by policy gradient theorem with a long-term reward approximation. In Subramanian et al. (2020), the two steps are performed concurrently using the concept of two time-scale algorithms (Borkar, 1997), with the first/second step being the fast/slow time-scale, as the policy approximator learns policies *based on the learned representations*. A more concrete algorithmic framework is given in Appendix E.

Note that the main result does not reflect the regret in practical implementations, as it is only the optimality gap given the approximate parameters *in one episode*. As we pointed out, in practice, one uses ML algorithms to learn representations that minimizes the approximation errors, while using policy gradient methods to learn the optimal policy (instead of solving DP directly). The true regret will then depend on both the convergence rate of the state representation learning and the policy gradient method used.

6 CONCLUSION

In this paper, we developed a general approximate state representation framework for MARL problems in a Dec-POMDP setting. We bounded the optimality gap in terms of the approximation error parameters and the number of remaining time steps. The theory provides guidance on designing deep-MARL algorithms, which has great potential in practical uses. Future directions include: exploring DL methods for applications using our framework, designing a representation for prescriptions, designing fully decentralized MARL schemes by adding communication, and extensions to general-sum games.

Acknowledgements

The authors would like to thank Aditya Mahajan, Yi Ouyang, and Chen-Yu Wei for helpful discussions. The authors would also like to thank Demosthenis Teneketzis, Ashutosh Nayyar, Hamidreza Tavafoghi, and Dengwang Tang for discussions about the CI approach and its general applicability in multi-agent systems. The authors would like to acknowledge support from NSF via grant ECCS 2038416, CNS 1955777, CCF 2008130, and support via a MIDAS grant sponsored by General Dynamics.

References

D. S. Bernstein, S. Zilberstein, and N. Immerman. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research*, 2002.

V. S. Borkar. Stochastic approximation with two time scales. *Systems and Control Letters*, 1997.

J. Dibangoye and O. Buffet. Learning to act in decentralized partially observable MDPs. In *ICML*, 2018.

N. Ferns, P. Panangaden, and D. Precup. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 2011.

J. N. Foerster, F. Song, E. Hughes, N. Burch, I. Dunning, S. Whiteson, M. Botvinick, and M. Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. In *ICML*, 2019.

D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. arXiv preprint arXiv:1809.01999, 2019.

M. Jafarnia-Jahromi, R. Jain, and A. Nayyar. Online learning for unknown partially observable MDPs. arXiv preprint arXiv:2102.12661, 2021.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998.

A. D. Kara and S. Yuksel. Near optimality of finite memory feedback policies in partially observed Markov decision processes. arXiv preprint arXiv:2010.07452, 2020.

P. R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control*. SIAM, 2015.

A. Lerer, H. Hu, J. Foerster, and N. Brown. Improving policies via search in cooperative partially observable games. arXiv preprint arXiv:1912.02318, 2019.

T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat. State representation learning for control: An overview. *Neural Networks*, 2018.

M. L. Littman, R. S. Sutton, and S. P. Singh. Predictive representations of state. In *NIPS*, 2001.

A. Mahajan and M. Mannan. Decentralized stochastic control. *Annals of Operations Research*, 2016.

W. Mao, K. Zhang, E. Miehling, and T. Başar. Information state embedding in partially observable cooperative multi-agent reinforcement learning. In *CDC*, 2020.

A. Nayyar, A. Mahajan, and D. Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 2013.

F. A. Oliehoek and C. Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

C. A. Schroeder de Witt, J. Foerster, G. Farquhar, P. Torr, W. Böehmer, and S. Whiteson. Multi-agent common knowledge reinforcement learning. In *NIPS*, 2019.

S. Sokota, E. Lockhart, F. Timbers, E. Davoodi, R. D'Orazio, N. Burch, M. Schmid, M. Bowling, and M. Lanctot. Solving common-payoff games with approximate policy iteration. In *AAAI*, 2021.

J. Subramanian and A. Mahajan. Approximate information state for partially observed systems. In *CDC*, 2019.

J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. arXiv preprint arXiv:2010.08843, 2020.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2018.

H. Tavafoghi, Y. Ouyang, and D. Teneketzis. A sufficient information approach to decentralized decision making. In *CDC*, 2018.

H. Tavafoghi, Y. Ouyang, and D. Teneketzis. A unified approach to dynamic decision problems with asymmetric information: Non-strategic agents. *IEEE Transactions on Automatic Control*, 2021.

A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello. Learning causal state representations of partially observable environments. arXiv preprint arXiv:1906.10437, 2021a.

A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine. Learning invariant representations for reinforcement learning without reconstruction. arXiv preprint arXiv:2006.10742, 2021b.

K. Zhang, Z. Yang, and T. Başar. Decentralized multi-agent reinforcement learning with networked agents: Recent advances. arXiv preprint arXiv:1912.03821, 2019.

Supplementary Material: Common Information based Approximate State Representations in Multi-Agent Reinforcement Learning

A Supplementary Details

A.1 More Related Work

Kara and Yuksel (2020) consider a special type of AIS – the N-memory, which contains the information from the last N steps. Here, the compression function is fixed but in contrast to Subramanian et al. (2020), the approximation error given each history need not be uniform. When the model is known, they provide conditions that bound the regret of N-memory policies (policies that depend on N-memory), and an algorithm that finds optimal policies within this class. The first algorithm to learn the optimal policies of POMDPs with sub-linear regret in an online setting is proposed in Jafarnia-Jahromi et al. (2021). Using a posterior sampling-based scheme, the algorithm maintains the posterior distribution on the unknown parameters of the considered POMDP, and adopts the optimal policy with respect to a set of parameters sampled from the distribution in each episode. The posterior update in the algorithm, however, heavily relies on the knowledge of the observation kernel, which is usually unknown in RL settings.

State representation for control is studied extensively in the literature (Lesort et al., 2018). Early work on predictive state representation (PSR) of POMDPs (Littman et al., 2001) only focuses on the encapsulation of the histories and does not explore its system prediction ability. The bisimulation relation clusters MDP states with similar rewards and transitions, and a bisimulation metric convexly combines the errors of the rewards and the transitions between two states (Ferns et al., 2011). The difference of the value functions of two states can be upper-bounded by the metric. The causal state representation (Zhang et al., 2021a) for POMDPs clusters the histories in the space of AOHs that will produce the same future dynamics. Using the observation history as the state, the considered POMDP can be transformed into an MDP, so that the results from the bisimulation literature can be applied.

A.2 DP with BCS

Algorithm 4 Dynamic Programming with BCSs and FPS-based Prescriptions

$\overset{\bullet}{V}_{T+1}(\pi_{T+1}) \triangleq 0$
for $t = T, \dots, 1$ **do**
  $\overset{\bullet}{Q}_t(\pi_t, \gamma_t) = \mathbb{E}\left[R(S_t, \Gamma_t(H_t^{1:N})) + \overset{\bullet}{V}_{t+1}(\eta_t(\Pi_t^0, \Gamma_t, O_{t+1}^0))|\Pi_t = \pi_t, \Gamma_t = \gamma_t\right]$
  $\overset{\bullet}{V}_t(\pi_t) = \max_{\gamma_t \in \Omega(\Gamma_t)} \overset{\bullet}{Q}_t(\pi_t, \gamma_t)$

The BCS is updated through Bayesian update with the function η_t Nayyar et al. (2013).

A.3 DP with BCS and SPI

Algorithm 5 Dynamic Programming with BCSs and SPI-based Prescriptions

$\widetilde{V}_{T+1}(\widetilde{\pi}_{T+1}) \triangleq 0$
for $t = T, \dots, 1$ **do**
  $\widetilde{Q}_t(\widetilde{\pi}_t, \lambda_t) = \mathbb{E}\left[R(S_t, \Lambda_t(Z_t^{1:N})) + \widetilde{V}_{t+1}(\widetilde{\eta}_t(\widetilde{\Pi}_t, \Lambda_t, O_{t+1}^0))|\widetilde{\Pi}_t = \widetilde{\pi}_t, \Lambda_t = \lambda_t\right]$
  $\widetilde{V}_t(\widetilde{\pi}_t) = \max_{\lambda_t \in \Omega(\Lambda_t)} \widetilde{Q}_t(\widetilde{\pi}_t, \lambda_t)$

Similar to the case of Algorithm 4, the revised version of BCS (now a distribution on the state and the SPI) is updated through Bayesian update with the function $\widetilde{\eta}_t$.

A.4 DP Comparison

Table 1: Dynamic programming comparison.

Work	Algorithm /Definition	Agent	Common State	Private State	Action	Compression Common/Private	Incurred DP Gap Common/Private
		Single	AOH	-	Action	None	0
Belief State	$\mathbb{P}(S_t\|H_t)$	Single	BS	-	Action	Lossy	0
Subramanian and Mahajan (2019)	Definition 1 with $\epsilon=\delta=0$	Single	IS	-	Action	Lossy	0
Subramanian and Mahajan (2019)	Definition 1	Single	AIS	-	Action	Lossy	Linear
CI Approach (No Compression)	Algorithm 1	Multi	FCS	FPS	FPS-pres.	None/None	0/0
Nayyar et al. (2013)	Algorithm 4	Multi	BCS	FPS	FPS-pres.	Lossy/None	0/0
Tavafoghi et al. (2018)	Algorithm 5	Multi	BCS	SPI	SPI-pres.	Lossy/Lossy	0/0
Subramanian et al. (2020)	-	Multi	ASCS	FPS	FPS-pres.	Lossy/None	Linear/0
Mao et al. (2020)	-	Multi	BCS	ASPS	ASPS-pres.	Lossy/Lossless	0/Linear
This work	Algorithm 3 with $\epsilon_c=\delta_c=0$ and $\epsilon_p=\delta_p=0$	Multi	SCS	SPS	SPS-pres.	Lossy/Lossy	0/0
This work	Algorithm 2	Multi	FCS	ASPS	ASPS-pres.	None/Lossy	0/Quadratic
This work	Algorithm 3	Multi	ASCS	ASPS	ASPS-pres.	Lossy/Lossy	Linear/Quadratic

B Omitted Analysis in Section 4.2

The following lemma shows that given the FPS, the actions the chosen prescription chooses for other FPSs does not affect the next step statistics.

Lemma 17: *Let* $h_t^0 \in \Omega(H_t^0)$, $h \in \Omega(H_t^{1:N})$, $\gamma \in \Omega(\Gamma_t)$, *and* $a = \gamma(h) \in \Omega(A_t)$. *Then*

$$\mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | H_t^0 = h_t^0, H_t^{1:N} = h, \Gamma_t = \gamma) = \mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | H_t^0 = h_t^0, H_t^{1:N} = h, A_t = a).$$

Proof: We will omit specifying the original random variables when their realizations are given in the proof.

$$\begin{aligned}
&\mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | h_t^0, h, \gamma) = \mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | h_t^0, h, \gamma, a) \\
&= \sum_{s_t} \mathbb{P}(s_t | h_t^0, h, \gamma, a) \cdot \mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | h_t^0, h, \gamma, a, s_t) \\
&= \sum_{s_t} \mathbb{P}(s_t | h_t^0, h, a) \cdot \mathbb{P}(S_{t+1} | h_t^0, h, \gamma, a, s_t) \cdot \mathbb{P}(H_{t+1}^{1:N} | h_t^0, h, \gamma, a, s_t, S_{t+1}) && (\gamma \text{ is after } s_t) \\
&= \sum_{s_t} \mathbb{P}(s_t | h_t^0, h, a) \cdot \mathbb{P}(S_{t+1} | a, s_t) \cdot \mathbb{P}(H_{t+1}^{1:N} | h_t^0, h, \gamma, a, s_t, S_{t+1}) && (\mathbb{P}_T \text{ specifies } S_{t+1} \text{ given } S_t \text{ and } A_t) \\
&= \sum_{s_t} \mathbb{P}(s_t | h_t^0, h, a) \cdot \mathbb{P}(S_{t+1} | a, s_t) \cdot \mathbb{P}(O_{t+1}^{1:N} | h_t^0, h, \gamma, a, s_t, S_{t+1}) && (H_{t+1}^{1:N} = (H_t^{1:N}, A_t, O_{t+1}^{1:N})) \\
&= \sum_{s_t} \mathbb{P}(s_t | h_t^0, h, a) \cdot \mathbb{P}(S_{t+1} | a, s_t) \cdot \mathbb{P}(O_{t+1}^{1:N} | S_{t+1}) && (\mathbb{P}_O \text{ specifies } O_{t+1}^{1:N} \text{ given } S_{t+1}) \\
&= \sum_{s_t} \mathbb{P}(s_t | h_t^0, h, a) \cdot \mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | h_t^0, h, a, s_t) \\
&= \mathbb{P}(S_{t+1}, H_{t+1}^{1:N} | h_t^0, h, a).
\end{aligned}$$

■

Proof of Lemma 14: The proof for the instantaneous part is straightforward as S_t is irrelevant to the choice

of Γ_t

$$\begin{aligned}
&\mathbb{E}\left[R_t(S_t, \Gamma_t(H_t^{1:N}))|h_t^0, h, \gamma_1\right] = \sum_{s_t} \mathbb{P}(s_t|h_t^0, h, \gamma_1) R_t(s_t, \gamma_1(h)) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h, \gamma_1(h)) R_t(s_t, a) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) R_t(s_t, a) && (\gamma_1 \text{ and } \gamma_2 \text{ are exogenously given}) \\
&= \mathbb{E}\left[R_t(S_t, \Gamma_t(H_t^{1:N}))|h_t^0, h, \gamma_2\right]. && (\text{by symmetry})
\end{aligned}$$

To show equality for the continuation part, we first define the following policy for all $\tau = t+1, \ldots, T$:

$$d'_\tau(h_\tau^0) = \begin{cases} d_\tau^*(h_t^0, \gamma_1, h_{t+1:\tau}^0) & \text{if } h_\tau^0 = (h_t^0, \gamma_2, h_{t+1:\tau}^0) \quad \forall\ h_{t+1:\tau}^0, \\ d_\tau^*(h_\tau^0) & \text{otherwise,} \end{cases}$$

where d_τ^* is an optimal policy at time step τ. Also, we have $h_{t+1:\tau}^0 = o_{t+1}^0$ when $\tau = t+1$, and $h_{t+1:\tau}^0 = (o_{t+1}^0, \gamma_{t+1}, \ldots, o_\tau^0)$ when $\tau > t+1$, so that the entire $(h_t^0, \gamma_1, h_{t+1:\tau}^0) \in \Omega(H_\tau^0)$. This policy performs the optimal policy at all times, except when γ_2 is chosen at time t, it will mimic what the optimal policy would have done if γ_1 was chosen instead; owing to perfect recall, future prescriptions can depend on past ones. Then

$$\begin{aligned}
&\mathbb{E}\left[V_{t+1}^S(H_{t+1}^0, H_{t+1}^{1:N})|h_t^0, h, \gamma_1\right] = \mathbb{E}\left[\sum_{\tau=t+1}^T R_\tau(S_\tau, A_\tau)\middle| h_t^0, h, \gamma_1, d_{t+1:T}^*\right] \\
&= \mathbb{E}\left[\sum_{\tau=t+1}^T R_\tau(S_\tau, A_\tau)\middle| h_t^0, h, \gamma_1, d'_{t+1:T}\right] \overset{(*)}{=} \mathbb{E}\left[\sum_{\tau=t+1}^T R_\tau(S_\tau, A_\tau)\middle| h_t^0, h, \gamma_2, d'_{t+1:T}\right] \\
&\leq \mathbb{E}\left[\sum_{\tau=t+1}^T R_\tau(S_\tau, A_\tau)\middle| h_t^0, h, \gamma_2, d_{t+1:T}^*\right] = \mathbb{E}\left[V_{t+1}^S(H_{t+1}^0, H_{t+1}^{1:N})|h_t^0, h, \gamma_2\right],
\end{aligned}$$

where the inequality holds as $d'_{t+1:T}$ may not be an optimal choice from the current history. By symmetry, the inequality implies that

$$\mathbb{E}\left[V_{t+1}^S(H_{t+1}^0, H_{t+1}^{1:N})|h_t^0, h, \gamma_1\right] = \mathbb{E}\left[V_{t+1}^S(H_{t+1}^0, H_{t+1}^{1:N})|h_t^0, h, \gamma_2\right].$$

The equality labeled by $(*)$ follows from the fact that under the policy $d'_{t+1:T}$, choosing γ_1 and γ_2 will generate the exact future statistics. We will show this in the following. We first prove the following claim using mathematical induction.

Claim: for all $\tau = t+1, \ldots, T$, we have

$$\mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_1, a, d'_{t+1:T}) = \mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_2, a, d'_{t+1:T}).$$

Base case: the claim holds for $\tau = t+1$.

$$\begin{aligned}
&\mathbb{P}(S_{t+1}, O_{t+1}^{0:N}, A_{t+1}|h_t^0, h, \gamma_1, a, d'_{t+1:T}) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h, \gamma_1, a, d'_{t+1:T}) \cdot \mathbb{P}(S_{t+1}, O_{t+1}^{0:N}, A_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1:T}) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}, O_{t+1}^{0:N}, A_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}) && (s_t \text{ is independent of } a \text{ given } \Gamma_t) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}) \cdot \mathbb{P}(O_{t+1}^{0:N}, A_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}, S_{t+1}) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}|s_t, a) \cdot \mathbb{P}(O_{t+1}^{0:N}, A_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}, S_{t+1}) && (\mathbb{P}_T \text{ specifies } S_{t+1} \text{ given } S_t \text{ and } A_t) \\
&= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}|s_t, a) \cdot \mathbb{P}(O_{t+1}^{0:N}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}, S_{t+1}) \cdot \mathbb{P}(A_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}, S_{t+1}, O_{t+1}^{0:N})
\end{aligned}$$

$$= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}|s_t, a) \cdot \mathbb{P}(O_{t+1}^{0:N}|S_{t+1}) \cdot \mathbb{P}(A_{t+1}|h_t^0, h, \gamma_1, a, s_t, d'_{t+1}, S_{t+1}, O_{t+1}^{0:N})$$

($\mathbb{P}_O$ specifies $O_{t+1}^{0:N}$ given S_{t+1})

$$= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}|s_t, a) \cdot \mathbb{P}(O_{t+1}^{0:N}|S_{t+1}) \cdot \mathbb{I}\left\{A_{t+1} = d'_{t+1}(h_0, \gamma_1, O_{t+1}^0)(h, a, O_{t+1}^{1:N})\right\}$$

$$= \sum_{s_t} \mathbb{P}(s_t|h_t^0, h) \cdot \mathbb{P}(S_{t+1}|s_t, a) \cdot \mathbb{P}(O_{t+1}^{0:N}|S_{t+1}) \cdot \mathbb{I}\left\{A_{t+1} = d'_{t+1}(h_0, \gamma_2, O_{t+1}^0)(h, a, O_{t+1}^{1:N})\right\} \quad \text{(definition of } d'_{t+1}\text{)}$$

$$= \mathbb{P}(S_{t+1}, O_{t+1}^{0:N}, A_{t+1}|h_t^0, h, \gamma_2, a, d'_{t+1:T}). \quad \text{(symmetric argument)}$$

Induction step: assuming the claim holds for τ, we show it holds for $\tau+1$ as well.

$$\begin{aligned}
&\mathbb{P}(S_{t+1:\tau+1}, O_{t+1:\tau+1}^{0:N}, A_{t+1:\tau+1}|h_t^0, h, \gamma_1, a, d'_{t+1:T}) \\
&= \mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_1, a, d'_{t+1:T}) \cdot \mathbb{P}(S_{\tau+1}, O_{\tau+1}^{0:N}, A_{\tau+1}|h_t^0, h, \gamma_1, a, d'_{t+1:T}, S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}) \\
&= \mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_2, a, d'_{t+1:T}) \cdot \mathbb{P}(S_{\tau+1}, O_{\tau+1}^{0:N}, A_{\tau+1}|h_t^0, h, \gamma_1, a, d'_{t+1:T}, S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}) \\
&\qquad \text{(induction hypothesis)} \\
&= \mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_2, a, d'_{t+1:T}) \cdot \mathbb{P}(S_{\tau+1}|S_\tau, A_\tau) \cdot \mathbb{P}(O_{\tau+1}^{0:N}|S_{\tau+1}) \\
&\qquad \cdot \mathbb{I}\left\{A_{\tau+1} = d'_{\tau+1}\left(h_t^0, \gamma_1, O_{t+1}^0, d'_{t+1}(h_t^0, \gamma_1, O_{t+1}^0), O_{t+2}^0, \ldots, O_{\tau+1}^0\right)(h, a, O_{t+1}^{1:N}, A_{t+1}, \ldots, O_{\tau+1}^{1:N})\right\} \\
&\overset{(\dagger)}{=} \mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_2, a, d'_{t+1:T}) \cdot \mathbb{P}(S_{\tau+1}|S_\tau, A_\tau) \cdot \mathbb{P}(O_{\tau+1}^{0:N}|S_{\tau+1}) \\
&\qquad \cdot \mathbb{I}\left\{A_{\tau+1} = d'_{\tau+1}\left(h_t^0, \gamma_2, O_{t+1}^0, d'_{t+1}(h_t^0, \gamma_2, O_{t+1}^0), O_{t+2}^0, \ldots, O_{\tau+1}^0\right)(h, a, O_{t+1}^{1:N}, A_{t+1}, \ldots, O_{\tau+1}^{1:N})\right\} \\
&= \mathbb{P}(S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}|h_t^0, h, \gamma_2, a, d'_{t+1:T}) \cdot \mathbb{P}(S_{\tau+1}, O_{\tau+1}^{0:N}, A_{\tau+1}|h_t^0, h, \gamma_2, a, d'_{t+1:T}, S_{t+1:\tau}, O_{t+1:\tau}^{0:N}, A_{t+1:\tau}) \\
&= \mathbb{P}(S_{t+1:\tau+1}, O_{t+1:\tau+1}^{0:N}, A_{t+1:\tau+1}|h_t^0, h, \gamma_2, a, d'_{t+1:T}),
\end{aligned}$$

where the equality in (†) holds due to the definition of policy d'. Note that with the CI-based approach, a generic policy d_t first maps an FCS H_t^0 to a prescription Γ_t, which in term maps an FPS $H_t^{1:N}$ to an action A_t; therefore, $d_t(H_t^0)(H_t^{1:N}) = \Gamma_t(H_t^{1:N}) = A_t$ refers to the final action A_t under the policy d_t and the supervisor's state $(H_t^0, H_t^{1:N})$. The claim implies that $\mathbb{P}(S_\tau, A_\tau|h_t^0, h, \gamma_1, d'_{t+1:T}) = \mathbb{P}(S_\tau, A_\tau|h_t^0, h, \gamma_2, d'_{t+1:T})$ for all $\tau = t+1, \ldots, T$, i.e. conditioning on $h_t^0, h, d'_{t+1:T}$, the distribution of (S_τ, A_τ) is exactly the same given γ_1 or γ_2; and (S_τ, A_τ) where $\tau = t+1, \ldots, T$ is what the expectations on both sides of $(*)$ are taken on. ■

Proof of Lemma 15: We preceed the proof by mathematical induction. The instantaneous part and the base case $t = T$ follow trivially from (ASPS2)

$$\begin{aligned}
&\left|\mathbb{E}\left[R_t(S_t, A_t)|h_t^0, h_1, \gamma\right] - \mathbb{E}\left[R_t(S_t, A_t)|h_t^0, h_2, \gamma\right]\right| \\
&\leq \left|\mathbb{E}\left[R_t(S_t, A_t)|h_t^0, h_1, \gamma\right] - \mathbb{E}\left[R_t(S_t, A_t)|h_t^0, \widehat{z}, \gamma\right]\right| + \left|\mathbb{E}\left[R_t(S_t, A_t)|h_t^0, \widehat{z}, \gamma\right] - \mathbb{E}\left[R_t(S_t, A_t)|h_t^0, h_2, \gamma\right]\right| \\
&\leq \epsilon_p/4 + \epsilon_p/4 \qquad \text{((ASPS2))} \\
&= \epsilon_p/2.
\end{aligned}$$

For the continuation part, we have

$$\begin{aligned}
&\mathbb{E}\left[V_{t+1}^S((H_t^0, \Gamma_t, O_{t+1}^0), H_{t+1}^{1:N})|h_t^0, h_1, \gamma\right] \\
&= \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_1, \gamma) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) \\
&= \sum_{o_{t+1}^{0:N}} \sum_{s_{t+1}} \mathbb{P}(o_{t+1}^{0:N}, s_{t+1}|h_t^0, h_1, \gamma) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) \\
&= \sum_{o_{t+1}^{0:N}} \sum_{s_{t+1}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_1, \gamma, s_{t+1}) \cdot \mathbb{P}(s_{t+1}|h_t^0, h_1, \gamma) \cdot V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) \\
&= \sum_{o_{t+1}^{0:N}} \sum_{s_{t+1}} \mathbb{P}(o_{t+1}^{0:N}|s_{t+1}) \cdot \mathbb{P}(s_{t+1}|h_t^0, h_1, \gamma) \cdot V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) \qquad (\mathbb{P}_O \text{ specifies } O_{t+1}^{0:N} \text{ given } S_{t+1}) \\
&= \sum_{o_{t+1}^{0:N}} \sum_{s_{t+1}} \mathbb{P}(o_{t+1}^{0:N}|s_{t+1}) \cdot \mathbb{P}(s_{t+1}|h_t^0, h_1, a) \cdot V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) \qquad \text{(Lemma 17)}
\end{aligned}$$

$$= \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_1, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})),$$

and the same equality holds for h_2. Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[V_{t+1}^S((H_t^0, \Gamma_t, O_{t+1}^0), H_{t+1}^{1:N})|h_t^0, h_1, \gamma\right] - \mathbb{E}\left[V_{t+1}^S((H_t^0, \Gamma_t, O_{t+1}^0), H_{t+1}^{1:N})|h_t^0, h_2, \gamma\right]\right| \\
&= \left|\sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_1, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) - \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_2, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_2, a, o_{t+1}^{1:N}))\right| \\
&\leq \left|\sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_1, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) - \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, \widehat{z}, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N}))\right| \\
&\quad + \left|\sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, \widehat{z}, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) - \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_2, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N}))\right| \\
&\quad + \left|\sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_2, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) - \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_2, a) V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_2, a, o_{t+1}^{1:N}))\right| \\
&:= ① + ② + ③.
\end{aligned}$$

For the first two terms, we have

$$①, ② \leq 2\|V_{t+1}\|_\infty \cdot \delta_p/8 \leq T\bar{R}\delta_p/4$$

by (ASPS3). Note that the above equation follows if $\mathcal{K}(\cdot,\cdot)$ is the total variation distance. If it is instead the Wasserstein metric, then the total variation distance will still be bounded by $\delta_p / \min_{x,y\in\Omega(O_{t+1}^{0:N}), x\neq y} \|x-y\|$; we can redefine this value as δ_p so that the total variation distance is still bounded by δ_p.

Now consider a fixed realization of $o_{t+1}^{0:N}$. We have

$$\begin{aligned}
&\widehat{\vartheta}_{t+1}^{1:N}((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) \\
&= \widehat{\phi}_{t+1}^{1:N}(\widehat{\vartheta}_t^{1:N}(h_1), h_t^0, \gamma, o_{t+1}^{0:N}) && ((\text{ASPS1})) \\
&= \widehat{\phi}_{t+1}^{1:N}(\widehat{\vartheta}_t^{1:N}(h_2), h_t^0, \gamma, o_{t+1}^{0:N}) && (\text{assumption}) \\
&= \widehat{\vartheta}_{t+1}^{1:N}((h_t^0, \gamma, o_{t+1}^0), (h_2, a, o_{t+1}^{1:N})), && ((\text{ASPS1}))
\end{aligned}$$

so that under the public FCS $(h_t^0, \gamma, o_{t+1}^0)$, the two FPSs $(h_1, a, o_{t+1}^{1:N})$ and $(h_2, a, o_{t+1}^{1:N})$ will be mapped to the same ASPS as well. Hence, by the induction hypothesis of Lemma 15 which leads to Corollary 16 at the $t+1$ step, we obtain

$$\begin{aligned}
&\left|V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) - V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_2, a, o_{t+1}^{1:N}))\right| \\
&\leq (T-t-1)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2.
\end{aligned}$$

The last term can thus be bounded by

$$\begin{aligned}
③ &\leq \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_2, a) \left|V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_1, a, o_{t+1}^{1:N})) - V_{t+1}^S((h_t^0, \gamma, o_{t+1}^0), (h_2, a, o_{t+1}^{1:N}))\right| \\
&\leq \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N}|h_t^0, h_2, a)[(T-t-1)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2] \\
&= (T-t-1)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2.
\end{aligned}$$

Combining the three terms plus the instantaneous part, it follows that

$$\begin{aligned}
\left|Q_t^S(h_t^0, h_1, \gamma) - Q_t^S(h_t^0, h_2, \gamma)\right| &\leq \epsilon_p/2 + 2 \cdot T\bar{R}\delta_p/4 + (T-t-1)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2 \\
&= (T-t)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2.
\end{aligned}$$

■

Proof of Corollary 16: Assume the optimal prescription γ^* prescribes different actions on the two FPSs $h_1, h_2 \in \Omega(H_t^{1:N})$, so that $\gamma^*(h_1) = a_1$ and $\gamma^*(h_2) = a_2$ where $a_1 \neq a_2$; otherwise, the claim directly follows by Lemma 15. Also, define $\gamma', \gamma'' \in \Omega(\Gamma_t)$ by

$$\gamma'(h) = \begin{cases} a_1 & \text{if } h = h_2, \\ \gamma^*(h) & \text{otherwise,} \end{cases} \qquad \gamma''(h) = \begin{cases} a_2 & \text{if } h = h_1, \\ \gamma^*(h) & \text{otherwise.} \end{cases}$$

Let $\upsilon_1 = Q_t^S(h_t^0, h_1, \gamma^*)$ and $\upsilon_2 = Q_t^S(h_t^0, h_2, \gamma^*)$. Denote $B_t \triangleq (T-t)(\epsilon_p + T\bar{R}\delta_p)/2 + \epsilon_p/2$ for simplicity. From (12), $Q_t(h_t, \gamma^*)$ can be expanded as

$$Q_t(h_t^0, \gamma^*) = \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)\upsilon_1 + \mathbb{P}(h_2|h_t^0)\upsilon_2.$$

Likewise, we can also expand $Q_t(h_t, \gamma')$ to

$$\begin{aligned}
& Q_t(h_t^0, \gamma') \\
&= \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma') + \mathbb{P}(h_1|h_t^0) Q_t^S(h_t^0, h_1, \gamma') + \mathbb{P}(h_2|h_t^0) Q_t^S(h_t^0, h_2, \gamma') \\
&\geq \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma') + \mathbb{P}(h_1|h_t^0) Q_t^S(h_t^0, h_1, \gamma') + \mathbb{P}(h_2|h_t^0) \left[Q_t^S(h_t^0, h_1, \gamma') - B_t \right] && \text{(Lemma 15)} \\
&= \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0) Q_t^S(h_t^0, h_1, \gamma^*) + \mathbb{P}(h_2|h_t^0) \left[Q_t^S(h_t^0, h_1, \gamma^*) - B_t \right] && \text{(Lemma 14)} \\
&= \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)\upsilon_1 + \mathbb{P}(h_2|h_t^0)(\upsilon_1 - B_t);
\end{aligned}$$

by symmetry

$$Q_t(h_t^0, \gamma'') \geq \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)(\upsilon_2 - B_t) + \mathbb{P}(h_2|h_t^0)\upsilon_2.$$

We have

$$\begin{aligned}
& \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)\upsilon_1 + \mathbb{P}(h_2|h_t^0)(\upsilon_1 - B_t) \leq Q_t(h_t^0, \gamma') \\
&\leq Q_t(h_t^0, \gamma^*) = \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)\upsilon_1 + \mathbb{P}(h_2|h_t^0)\upsilon_2
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)(\upsilon_2 - B_t) + \mathbb{P}(h_2|h_t^0)\upsilon_2 \leq Q_t(h_t^0, \gamma'') \\
&\leq Q_t(h_t^0, \gamma^*) = \sum_{h \neq h_1, h_2} \mathbb{P}(h|h_t^0) Q_t^S(h_t^0, h, \gamma^*) + \mathbb{P}(h_1|h_t^0)\upsilon_1 + \mathbb{P}(h_2|h_t^0)\upsilon_2.
\end{aligned}$$

Canceling and rearranging the terms yield

$$-B_t \leq \upsilon_1 - \upsilon_2 \leq B_t.$$

■

Proof of Theorem 4: We prove the result by induction. The base case trivially follows from Proposition 13. Note that the continuation values at $T+1$ are defined to be 0, i.e. $V_{T+1}(h_{T+1}^0) \triangleq 0$ and $\widehat{V}_{T+1}(h_{T+1}^0) \triangleq 0$ for any $h_{T+1}^0 \in \Omega(H_{T+1}^0)$. Hence, for any $h_T^0 \in \Omega(H_T^0)$ and $\gamma^* \in \Omega(\Gamma_t)$, we have

$$Q_T(h_T^0, \gamma^*) - \epsilon_p = V_T(h_T^0) - \epsilon_p \leq \max_{\widehat{\lambda} \in \Omega(\widehat{\Lambda}_T)} Q_T(h_T^0, \gamma_{\widehat{\lambda}, h_T^0}) = \max_{\widehat{\lambda} \in \Omega(\widehat{\Lambda}_T)} \widehat{Q}_T(h_T^0, \widehat{\lambda}) = \widehat{V}_T(h_T^0).$$

In the equation, $Q_T(h_T^0, \gamma_{\widehat{\lambda},h_T^0}) = \widehat{Q}_T(h_T^0, \widehat{\lambda})$ because there is no continuation value for T. Now for the induction step, we assume the induction hypothesis, i.e. the claim holds for some $t+1 \leq T$ so that we have for any $h_{t+1}^0 \in \Omega(H_{t+1}^0)$,

$$V_{t+1}(h_{t+1}^0) - \widehat{V}_{t+1}(h_{t+1}^0) \leq \frac{(T-t-1)(T-t)}{2}(\epsilon_p + T\bar{R}\delta_p) + (T-t)\epsilon_p.$$

Proposition 13 states that for any $h_t^0 \in \Omega(H_t^0)$ and optimal prescription $\gamma^* \in \operatorname{argmax}_\gamma Q_t(h_t^0, \gamma)$, there exists a $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$ such that

$$Q_t(h_t^0, \gamma^*) - Q_t(h_t^0, \gamma_{\widehat{\lambda},h_t^0}) \leq (T-t)(\epsilon_p + T\bar{R}\delta_p) + \epsilon_p.$$

Write $\mathfrak{C}_t \triangleq (T-t)(\epsilon_p + T\bar{R}\delta_p) + \epsilon_p$ for shorthand of notation. Then for this $\widehat{\lambda}$, we have

$$\begin{aligned}
&Q_t(h_t^0, \gamma^*) - \widehat{Q}_t(h_t^0, \widehat{\lambda}) = Q_t(h_t^0, \gamma^*) - Q_t(h_t^0, \gamma_{\widehat{\lambda},h_t^0}) + Q_t(h_t^0, \gamma_{\widehat{\lambda},h_t^0}) - \widehat{Q}_t(h_t^0, \widehat{\lambda}) \\
&\leq \mathfrak{C}_t + \mathbb{E}[R_t + V_{t+1}(H_{t+1}^0)|h_t^0, \gamma_{\widehat{\lambda},h_t^0}] - \mathbb{E}[R_t + \widehat{V}_{t+1}(H_{t+1}^0)|h_t^0, \gamma_{\widehat{\lambda},h_t^0}] \\
&= \mathfrak{C}_t + \sum_{h_{t+1}^0} \mathbb{P}(h_{t+1}^0|h_t^0, \gamma_{\widehat{\lambda},h_t^0}) \left[V_{t+1}(h_{t+1}^0) - \widehat{V}_{t+1}(h_{t+1}^0)\right] \\
&\leq \mathfrak{C}_t + \sum_{h_{t+1}^0} \mathbb{P}(h_{t+1}^0|h_t^0, \gamma_{\widehat{\lambda},h_t^0}) \left[\frac{(T-t-1)(T-t)}{2}(\epsilon_p + T\bar{R}\delta_p) + (T-t)\epsilon_p\right] \\
&= (T-t)(\epsilon_p + T\bar{R}\delta_p) + \epsilon_p + \frac{(T-t-1)(T-t)}{2}(\epsilon_p + T\bar{R}\delta_p) + (T-t)\epsilon_p \\
&= \frac{(T-t)(T-t+1)}{2}(\epsilon_p + T\bar{R}\delta_p) + (T-t+1)\epsilon_p.
\end{aligned}$$

■

C Omitted Analysis in Section 3.2

Proposition 18: *Assume the reward function R is bounded by $\bar{R}$. Let $h_1^0, h_2^0 \in \Omega(H_t^0)$ be two FCSs. If $\widehat{z}^0 = \widehat{\vartheta}_t^0(h_1^0) = \widehat{\vartheta}_t^0(h_2^0)$, then for any $\widehat{\lambda} \in \Omega(\widehat{\Lambda}_t)$,*

$$\left|\widehat{Q}_t(h_1^0, \widehat{\lambda}) - \widehat{Q}_t(h_2^0, \widehat{\lambda})\right| \leq 2(T-t)(\epsilon_c + T\bar{R}\delta_c) + 2\epsilon_c. \tag{22}$$

Proof: We proceed the proof again by mathematical induction. The instantaneous part as well as the base case $t = T$ trivially follow from (ASCS2)

$$\begin{aligned}
&\left|\mathbb{E}\left[R_t(S_t, A_t)|h_1^0, \widehat{\lambda}\right] - \mathbb{E}\left[R_t(S_t, A_t)|h_2^0, \widehat{\lambda}\right]\right| \\
&\leq \left|\mathbb{E}\left[R_t(S_t, A_t)|h_1^0, \widehat{\lambda}\right] - \mathbb{E}\left[R_t(S_t, A_t)|\widehat{z}^0, \widehat{\lambda}\right]\right| + \left|\mathbb{E}\left[R_t(S_t, A_t)|\widehat{z}^0, \widehat{\lambda}\right] - \mathbb{E}\left[R_t(S_t, A_t)|h_2^0, \widehat{\lambda}\right]\right| \\
&\leq \epsilon_c + \epsilon_c = 2\epsilon_c. \qquad \text{((ASCS2))}
\end{aligned}$$

For the continuation part in the induction step, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\widehat{V}_{t+1}((H_t^0, \widehat{\Lambda}_t, O_{t+1}^0))|h_1^0, \widehat{\lambda}\right] - \mathbb{E}\left[\widehat{V}_{t+1}((H_t^0, \widehat{\Lambda}_t, O_{t+1}^0))|h_2^0, \widehat{\lambda}\right]\right| \\
&= \left|\sum_{o_{t+1}^0} \mathbb{P}(o_{t+1}^0|h_1^0, \widehat{\lambda})\widehat{V}_{t+1}((h_1^0, \widehat{\lambda}, o_{t+1}^0)) - \sum_{o_{t+1}^0} \mathbb{P}(o_{t+1}^0|h_2^0, \widehat{\lambda})\widehat{V}_{t+1}((h_2^0, \widehat{\lambda}, o_{t+1}^0))\right| \\
&\leq \left|\sum_{o_{t+1}^0} \mathbb{P}(o_{t+1}^0|h_1^0, \widehat{\lambda})\widehat{V}_{t+1}((h_1^0, \widehat{\lambda}, o_{t+1}^0)) - \sum_{o_{t+1}^0} \mathbb{P}(o_{t+1}^0|\widehat{z}^0, \widehat{\lambda})\widehat{V}_{t+1}((h_1^0, \widehat{\lambda}, o_{t+1}^0))\right| \\
&\quad + \left|\sum_{o_{t+1}^0} \mathbb{P}(o_{t+1}^0|\widehat{z}^0, \widehat{\lambda})\widehat{V}_{t+1}((h_1^0, \widehat{\lambda}, o_{t+1}^0)) - \sum_{o_{t+1}^0} \mathbb{P}(o_{t+1}^0|h_2^0, \widehat{\lambda})\widehat{V}_{t+1}((h_1^0, \widehat{\lambda}, o_{t+1}^0))\right|
\end{aligned}$$

$$+\left|\sum_{o^0_{t+1}}\mathbb{P}(o^0_{t+1}|h^0_2,\widehat{\lambda})\widehat{V}_{t+1}((h^0_1,\widehat{\lambda},o^0_{t+1}))-\sum_{o^0_{t+1}}\mathbb{P}(o^0_{t+1}|h^0_2,\widehat{\lambda})\widehat{V}_{t+1}((h^0_2,\widehat{\lambda},o^0_{t+1}))\right|$$
$$:= ① + ② + ③.$$

For the first two terms, we have

$$①, ② \le 2\|\widehat{V}_{t+1}\|_\infty \cdot \delta_c/2 \le T\bar{R}\delta_c$$

by (ASCS3).

Now consider a fixed realization of o^0_{t+1}. We have

$$\begin{aligned}
&\widehat{\vartheta}^0_{t+1}((h^0_1,\widehat{\lambda},o^0_{t+1}))\\
&=\widehat{\phi}^0_{t+1}(\widehat{\vartheta}^0_t(h_1),\widehat{\lambda},o^0_{t+1}) && ((\text{ASCS1}))\\
&=\widehat{\phi}^0_{t+1}(\widehat{\vartheta}^0_t(h_2),\widehat{\lambda},o^0_{t+1}) && (\text{assumption})\\
&=\widehat{\vartheta}^0_{t+1}((h^0_2,\widehat{\lambda},o^0_{t+1})), && ((\text{ASCS1}))
\end{aligned}$$

so that the two FCSs (with ASPS-based prescription) $(h^0_1,\widehat{\lambda},o^0_{t+1})$ and $(h^0_2,\widehat{\lambda},o^0_{t+1})$ will be mapped to the same ASCS as well. Hence, by the induction hypothesis, we obtain

$$\left|\widehat{V}_{t+1}((h^0_1,\widehat{\lambda},o^0_{t+1}))-\widehat{V}_{t+1}((h^0_2,\widehat{\lambda},o^0_{t+1}))\right| \le 2(T-t-1)(\epsilon_c+T\bar{R}\delta_c)+2\epsilon_c.$$

The last term can thus be bounded by

$$\begin{aligned}
③ &\le \sum_{o^0_{t+1}}\mathbb{P}(o^0_{t+1}|h^0_2,\widehat{\lambda})\left|\widehat{V}_{t+1}((h^0_1,\widehat{\lambda},o^0_{t+1}))-\widehat{V}_{t+1}((h^0_2,\widehat{\lambda},o^0_{t+1}))\right|\\
&\le \sum_{o^0_{t+1}}\mathbb{P}(o^0_{t+1}|h^0_2,\widehat{\lambda})[2(T-t-1)(\epsilon_c+T\bar{R}\delta_c)+2\epsilon_c] = 2(T-t-1)(\epsilon_c+T\bar{R}\delta_c)+2\epsilon_c.
\end{aligned}$$

Combining the three terms plus the instantaneous part, it follows that

$$\begin{aligned}
\left|\widehat{Q}_t(h^0_1,\widehat{\lambda})-\widehat{Q}_t(h^0_2,\widehat{\lambda})\right| &\le 2\epsilon_c + 2\cdot T\bar{R}\delta_c + 2(T-t-1)(\epsilon_c+T\bar{R}\delta_c)+2\epsilon_c\\
&= 2(T-t)(\epsilon_c+T\bar{R}\delta_c)+2\epsilon_c.
\end{aligned}$$

■

Proof of Theorem 6: We proceed the proof again by mathematical induction. The base case $t=T$ trivially follows from (ASCS2). For the induction step, we have

$$\begin{aligned}
&\widehat{Q}_t(h^0_t,\widehat{\lambda})-\breve{Q}_t(\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda})\\
&=\mathbb{E}\left[R_t(S_t,A_t)|h^0_t,\widehat{\lambda}\right]-\mathbb{E}\left[R_t(S_t,A_t)|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda}\right]+\mathbb{E}\left[\widehat{V}_{t+1}(H^0_{t+1})|h^0_t,\widehat{\lambda}\right]-\mathbb{E}\left[\breve{V}_{t+1}(\widehat{\vartheta}^0_{t+1}(H^0_{t+1}))|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda}\right]\\
&\le \epsilon_c+\mathbb{E}\left[\widehat{V}_{t+1}(H^0_{t+1})|h^0_t,\widehat{\lambda}\right]-\mathbb{E}\left[\breve{V}_{t+1}(\widehat{\vartheta}^0_{t+1}(H^0_{t+1}))|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda}\right] && ((\text{ASCS2}))\\
&=\epsilon_c+\mathbb{E}\left[\widehat{V}_{t+1}(H^0_{t+1})|h^0_t,\widehat{\lambda}\right]-\mathbb{E}\left[\widehat{V}_{t+1}(H^0_{t+1})|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda}\right]\\
&\quad+\mathbb{E}\left[\widehat{V}_{t+1}(H^0_{t+1})|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda}\right]-\mathbb{E}\left[\breve{V}_{t+1}(\widehat{\vartheta}^0_{t+1}(H^0_{t+1}))|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda}\right]\\
&=\epsilon_c+\sum_{h^0_{t+1}}\left[\mathbb{P}(h^0_{t+1}|h^0_t,\widehat{\lambda})-\mathbb{P}(h^0_{t+1}|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda})\right]\widehat{V}_{t+1}(h^0_{t+1})\\
&\quad+\sum_{h^0_{t+1}}\mathbb{P}(h^0_{t+1}|\widehat{\vartheta}^0_t(h^0_t),\widehat{\lambda})\left[\widehat{V}_{t+1}(h^0_{t+1})-\breve{V}_{t+1}(\widehat{\vartheta}^0_{t+1}(h^0_{t+1}))\right]\\
&:= \epsilon_c + ① + ②.
\end{aligned}$$

The first term is bounded by (ASCS3)

$$\begin{aligned}① &= \sum_{o^0_{t+1}} \left[\mathbb{P}(o^0_{t+1}|h^0_t, \widehat{\lambda}) - \mathbb{P}(o^0_{t+1}|\widehat{\vartheta}^0_t(h^0_t), \widehat{\lambda})\right] \widehat{V}_{t+1}((h^0_t, \widehat{\lambda}, o^0_{t+1})) \\ &\leq 2\|\widehat{V}_{t+1}\|_\infty \cdot \delta_c/2 \leq T\bar{R}\delta_c,\end{aligned}$$

while the second term can be bounded by the induction hypothesis

$$\begin{aligned}② &\leq \sum_{h^0_{t+1}} \mathbb{P}(h^0_{t+1}|\widehat{\vartheta}^0_t(h^0_t), \widehat{\lambda}) \left[(T-t-1)(\epsilon_c + T\bar{R}\delta_c) + \epsilon_c\right] \\ &= (T-t-1)(\epsilon_c + T\bar{R}\delta_c) + \epsilon_c.\end{aligned}$$

Combining the terms, it follows that

$$\begin{aligned}\widehat{Q}_t(h^0_t, \widehat{\lambda}) - \breve{Q}_t(\widehat{\vartheta}^0_t(h^0_t), \widehat{\lambda}) &\leq \epsilon_c + T\bar{R}\delta_c + (T-t-1)(\epsilon_c + T\bar{R}\delta_c) + \epsilon_c \\ &= (T-t)(\epsilon_c + T\bar{R}\delta_c) + \epsilon_c.\end{aligned}$$

The V part of the claim can be obtained by considering an optimal prescription $\widehat{\lambda}^* \in \operatorname{argmax}_{\widehat{\lambda}\in\Omega(\widehat{\Lambda}_t)} \widehat{Q}_t(h^0_t, \widehat{\lambda})$ in the Q part. ■

D Omitted Analysis in Section 3.4

As mentioned in Section 3.4, when considering $\epsilon_c = \delta_c = \epsilon_p = \delta_p = 0$, we use SCS and SPS to refer to the common and private representations. Moreover, we use Z and ϑ to denote the compressed state and the compression mapping when the error parameters are 0, instead of $\widehat{Z}$ and $\widehat{\vartheta}$.

Proof of Proposition 8: For (SCS1), BCSs can be updated recursively through Bayesian updates (Nayyar et al., 2013). For (SCS2), notice that

$$\mathbb{E}[R_t(S_t, A_t)|h^0_t, \lambda_t] = \sum_{s_t, h^{1:N}_t} \mathbb{P}(s_t, h^{1:N}_t|h^0_t) R(s_t, \gamma_t(h^{1:N}_t)),$$

and the ensemble of $\mathbb{P}(s_t, h^{1:N}_t|h^0_t)$ through their spaces is exactly $\Pi_t(h^0_t) = \mathbb{P}(S_t, H^{1:N}_t|h^0_t)$. Similarly, it satisfies (SCS3) as well, since

$$\mathbb{P}(O^0_{t+1}|h^0_t, \gamma_t) = \sum_{s_t, h^{1:N}_t} \mathbb{P}(s_t, h^{1:N}_t|h^0_t) \cdot \sum_{s_{t+1}} \mathbb{P}(s_{t+1}|s_t, \gamma_t(h^{1:N}_t)) \cdot \mathbb{P}(O^0_{t+1}|s_{t+1}).$$

The quantity $\Pi_t(h^0_t) = \mathbb{P}(S_t, H^{1:N}_t|h^0_t)$ again exactly encapsulates what is needed to compute $\mathbb{P}(O^0_{t+1}|h^0_t, \gamma_t)$. ■

Proof of Proposition 9:

$$\begin{aligned}&\mathbb{P}(z^{1:N}_{t+1}, o^0_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, a_t) = \sum_{o^{0:N}_{t+1}} \sum_{s_{t+1}} \mathbb{P}(z^{1:N}_{t+1}, o^{0:N}_{t+1}, s_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, a_t) \\ &= \sum_{o^{0:N}_{t+1}} \sum_{s_{t+1}} \mathbb{P}(s_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, a_t) \cdot \mathbb{P}(o^{0:N}_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, a_t, s_{t+1}) \cdot \mathbb{P}(z^{1:N}_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, a_t, s_{t+1}, o^{0:N}_{t+1}) \\ &= \sum_{o^{0:N}_{t+1}} \sum_{s_{t+1}} \mathbb{P}(s_{t+1}|h^0_t, h^{1:N}_t, \gamma_t) \cdot \mathbb{P}(o^{0:N}_{t+1}|s_{t+1}) \cdot \mathbb{P}(z^{1:N}_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, s_{t+1}, o^{0:N}_{t+1}) \\ &\qquad\qquad \text{(redundancy of } a_t \text{ and } P_O \text{ specifies } O_{t+1} \text{ given } S_{t+1}) \\ &= \sum_{o^{0:N}_{t+1}} \sum_{s_{t+1}} \mathbb{P}(s_{t+1}|h^0_t, h^{1:N}_t, a_t) \cdot \mathbb{P}(o^{0:N}_{t+1}|s_{t+1}) \cdot \mathbb{P}(z^{1:N}_{t+1}|h^0_t, h^{1:N}_t, \gamma_t, s_{t+1}, o^{0:N}_{t+1}) \qquad \text{(Lemma 17)} \\ &= \sum_{o^{0:N}_{t+1}} \sum_{s_{t+1}} \mathbb{P}(s_{t+1}, o^{0:N}_{t+1}|h^0_t, h^{1:N}_t, a_t) \cdot \mathbb{I}\{z^{1:N}_{t+1} = \phi^{1:N}_{t+1}(\vartheta_t(h^0_t, h^{1:N}_t), \gamma_t, o^{0:N}_{t+1})\} \qquad \text{((SPS1))}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N} | h_t^0, h_t^{1:N}, a_t) \cdot \mathbb{I}\{z_{t+1}^{1:N} = \phi_{t+1}^{1:N}(z_t^{1:N}, \gamma_t, o_{t+1}^{0:N})\} \\
&= \sum_{o_{t+1}^{0:N}} \mathbb{P}(o_{t+1}^{0:N} | h_t^0, z_t^{1:N}, a_t) \cdot \mathbb{I}\{z_{t+1}^{1:N} = \phi_{t+1}^{1:N}(z_t^{1:N}, \gamma_t, o_{t+1}^{0:N})\} && \text{((SPS3))} \\
&= \mathbb{P}(z_{t+1}^{1:N}, o_{t+1}^0 | h_t^0, z_t^{1:N}, \gamma_t, a_t).
\end{aligned}
$$

Note the last equality follows as in (SPS3) it is implicitly assumed that $z_t^{1:N} = \vartheta_t(h_t^0, h_t^{1:N})$. ■

Proof of Proposition 10:

$$
\begin{aligned}
&\mathbb{E}[R(S_t, A_t) | h_t^0, h_t^n, a_t] = \sum_{s_t} R(s_t, a_t) \mathbb{P}(s_t | h_t^0, h_t^n) = \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(s_t, z_t^{-n} | h_t^0, h_t^n) \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \cdot \mathbb{P}(s_t | h_t^0, h_t^n, z_t^{-n}) \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(s_t, h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \cdot \mathbb{P}(s_t | h_t^0, h_t^n, z_t^{-n}, h_t^{-n}) \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \cdot \mathbb{P}(s_t | h_t^0, h_t^n, h_t^{-n}) && (z_t^{-n} = \vartheta_t^{-n}(h_t^0, h_t^{-n})) \\
&= \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \sum_{s_t} R(s_t, a_t) \mathbb{P}(s_t | h_t^0, h_t^n, h_t^{-n}) \\
&= \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \mathbb{E}[R(S_t, A_t) | h_t^0, h_t^{1:N}, a_t] \\
&= \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \mathbb{E}[R(S_t, A_t) | h_t^0, z_t^{1:N}, a_t] && \text{((SPS2)))} \\
&= \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \sum_{s_t} R(s_t, a_t) \mathbb{P}(s_t | h_t^0, z_t^n, z_t^{-n}) \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, h_t^n) \cdot \mathbb{P}(s_t | h_t^0, z_t^n, z_t^{-n}) \sum_{h_t^{-n}} \mathbb{P}(h_t^{-n} | h_t^0, h_t^n, z_t^{-n}) \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(z_t^{-n} | h_t^0, z_t^n) \cdot \mathbb{P}(s_t | h_t^0, z_t^n, z_t^{-n}) \cdot 1 && \text{((SPI4))} \\
&= \sum_{s_t} R(s_t, a_t) \sum_{z_t^{-n}} \mathbb{P}(s_t, z_t^{-n} | h_t^0, z_t^n) = \sum_{s_t} R(s_t, a_t) \mathbb{P}(s_t | h_t^0, z_t^n) = \mathbb{E}[R(S_t, A_t) | h_t^0, z_t^n, a_t].
\end{aligned}
$$

Note that the superscript $-n$ only contains $[N] \setminus \{n\}$ and does not contain 0. ■

E Algorithmic Framework

In this section we propose an MARL algorithmic framework using the theory developed in Section 3; the designing detail is left as future work. The framework adopts the "centralized learning distributed execution" scheme, i.e., the agents assume the omniscient supervisor's view when they learn the compressions and policies.

There are three types of functions within: the state networks $\rho^{0:N}$ modeled by recurrent neural networks (RNNs), the policy networks $\varphi^{0:N}$ modeled by deep neural networks (DNNs), and the prediction networks ψ^C and ψ^S also modeled by DNNs. The state networks $\rho^{0:N}$ serve the purpose of the compression mappings $\widehat{\phi}_t^{0:N}$ in Definition 3 and Definition 5, and their recursive evolution structures suggest an RNN modeling. The common policy network φ^0 takes $\widehat{Z}_t^0$ as input and gives the prescription $\widehat{\Lambda}_t$ as suggested by Section 3.2; the private policy network φ^n

Algorithm 6 Deep-MARL Framework

Common part: coordinator computes (done in each agent in execution phase) $\widehat{Z}_t^0 = \rho^0(O_t^0, \widehat{\Lambda}_{t-1})$, $\widehat{\Lambda}_t = \varphi^0(\widehat{Z}_t^0)$.
Private part: agent n computes $\widehat{Z}_t^n = \rho^n(O_t^n, A_{t-1}^n)$, $A_t^n = \varphi^n(\widehat{Z}_t^n, \widehat{\Lambda}_t^n)$.
if *in learning phase* **then**
    Coordinator computes $(\hat{R}_t, \hat{O}_{t+1}^0) = \psi^C(\widehat{Z}_t^0, \widehat{\Lambda}_t)$.
    $(\hat{R}_t, \hat{O}_{t+1}^0)$ is compared with ground truth (R_t, O_{t+1}^0) and loss is back-propagated to (ρ^0, ψ^C).
    Supervisor computes $(\hat{R}_t, \hat{O}_{t+1}^{0:N}) = \psi^S(\widehat{Z}_t^{0:N}, A_t^{1:N})$.
    $(\hat{R}_t, \hat{O}_{t+1}^{0:N})$ is compared with ground truth $(R_t, O_{t+1}^{0:N})$ and loss is back-propagated to $(\rho^{0:N}, \psi^S)$.
    Coordinator computes $\sum_{\tau=t-W}^{t} R_\tau$ and performs policy gradient on $\varphi^{0:N}$.

takes $\widehat{Z}_t^n$ and $\widehat{\Lambda}_t^n$ and outputs A_t^n. Here, we have to use a *variable* to represent the *prescription function*; hence, it cannot be directly applied to $\widehat{Z}_t^n$. Effective design of representing prescription function is left as future work, even though Lemma 14 provides a nice decomposition. Finally, the policy networks ψ^C and ψ^S are used to produce the predicted reward and new observations. In the learning phase, the predictions are compared with the ground truth and errors are back-propagated through the state and prediction networks. This requires full knowledge of $\widehat{Z}_t^{1:N}$ and $O_t^{1:N}$; consequently, the learning has to be centralized. A windowed (with length W) cumulative reward is summed for the computation of the loss in policy gradient methods (Sutton and Barto, 2018), which is back-propagated through the policy networks; actor-critic methods can also be employed here. In the execution phase, only state and policy networks are required, and everything can be performed in a decentralized fashion. Note that in our proof of Theorem 4 only the fact that $\widehat{Z}_t^{1:N}$ can be updated from $\widehat{Z}_{t-1}^{1:N}$ is needed.

To design a fully decentralized learning scheme, one needs conditions similar to (ASPS2) and (ASPS3) but only involving o_t^n, h_t^n and $\hat{z}_t^n$ instead of the whole $o_t^{1:N}$, $h_t^{1:N}$ and $\hat{z}_t^{1:N}$ so that individual prediction networks that does not require supervisor's view can be designed. This might demand a "consistency condition" similarly to (SPI4), and it is likely that this is only possible through agents communicating through some signal space or directly sending their parameters (Zhang et al., 2019). Further, the required expectations over the realizations of the private information conditioned on the common information could be hard to estimate. This is also left as future work.