Solving Hierarchical Information-Sharing Dec-POMDPs: An Extensive-Form Game Approach

Johan Peralez[1] **Aurélien Delage**[1] **Olivier Buffet**[2] **Jilles S. Dibangoye**[3]

Abstract

A recent theory shows that a multi-player decentralized partially observable Markov decision process can be transformed into an equivalent single-player game, enabling the application of Bellman's principle of optimality to solve the single-player game by breaking it down into single-stage subgames. However, this approach entangles the decision variables of all players at each single-stage subgame, resulting in backups with a double-exponential complexity. This paper demonstrates how to disentangle these decision variables while maintaining optimality under hierarchical information sharing, a prominent management style in our society. To achieve this, we apply the principle of optimality to solve any single-stage subgame by breaking it down further into smaller subgames, enabling us to make single-player decisions at a time. Our approach reveals that extensive-form games always exist with solutions to a single-stage subgame, significantly reducing time complexity. Our experimental results show that the algorithms leveraging these findings can scale up to much larger multi-player games without compromising optimality.

The multi-player decentralized partially observable Markov decision process (Dec-POMDP) is a general game-theoretic setting for decision-making by a team of collaborative players (Amato et al., 2013). In this multi-player game, players must coordinate while they can neither see the actual state of the world nor explicitly share what they see or do with each other due to communication costs, latency, or noise. This so-called silent coordination dilemma provides a partial explanation of the worst-case complexity results—*i.e.,* infinite-horizon cases are undecidable, finite-horizon ones are NEXP-hard, and finding ϵ-approximations remains hard (Bernstein et al., 2002; Rabinovich et al., 2003). Methods for Dec-POMDPs are split between local and global, each with strengths and weaknesses.

Local methods trade global optima, or ϵ-approximations for weaker solution concepts, *e.g.,* local optima, Nash equilibria, or any arbitrary feasible solution. While they share core ideas with global methods, their primary focus is on solving relaxations of the original multi-player game, *e.g.,* independent planners reason in isolation, policy gradient targets first-order solutions of non-convex functions (Tan, 1998; Peshkin et al., 2001; Bono et al., 2018). Of particular attention, local methods using deep neural networks can apply effectively to virtually any non-critical application, *e.g.,* online services, logistics, or board games (Lowe et al., 2017; Foerster et al., 2018; Rashid et al., 2018).

On the other hand, in many critical and high-stakes applications, *e.g.,* search and rescue, security, and healthcare, global methods can find solutions with the required theoretical guarantees, but scalability remains a significant issue. These algorithms recast the original multi-player game into an equivalent single-player one, which overcomes the silent coordination dilemma and allows the principle of optimality to apply. Intuitively, this principle decomposes the single-player game into single-stage subgames and solves them recursively. Yet, doing so, in return, virtually entangles decision variables of all players at each single-stage subgame, resulting in double-exponential complexity. Because they update decision variables of all players in sync at every single-stage subgame, even a single update can be prohibitively expensive (Szer & Charpillet, 2005; MacDermed & Isbell, 2013; Nayyar et al., 2013; Oliehoek, 2013). To somewhat mitigate this burden, branch-and-bound search algorithms and mixed-integer linear programs were introduced, but the limitation remains (Oliehoek et al., 2010; Dibangoye et al., 2009; 2013; 2016). In many cases, however, real-world environments contain significant structure that can be exploited (Amato et al., 2013).

Indeed, several forms of structure have been investigated in

[1] Université de Lyon, INSA Lyon and Inria, CITI, F-69000 Lyon
[2] Université de Lorraine, CNRS, INRIA, LORIA, F-54000 Nancy
[3] Bernoulli Institute, University of Groningen, Nijenborgh 4, NL-9747 AG Groningen, Netherlands. Correspondence to: Jilles S. Dibangoye <j.s.dibangoye@rug.nl>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the past—*e.g.,* dynamics independence (Becker et al., 2004; Dibangoye et al., 2012), weak-separability (Nair et al., 2005; Dibangoye et al., 2014), and delayed information-sharing (Nayyar et al., 2010). Algorithms that use such structures can optimally solve structured multi-player games much faster than generic ones. This paper exploits HIS structure, a dominant management style in our society for corporations, governments, criminal enterprises, armies, and religions. This management style involves each player being aware of what its subordinate knows, and this knowledge is passed down the chain of command. In other words, player n at the top of the hierarchy knows all that player $n-1$ knows; player $n-2$ knows all that player $n-3$ knows, and so forth. Moreover, HIS is equivalent to one-sidedness when only two players are involved, which was previously recognized as a tractable structure for two-person partially observable stochastic games (Horák et al., 2017; Horák & Bošanskỳ, 2019; Hadfield-Menell et al., 2016; Malik et al., 2018; Xie et al., 2020). Still, little is known about how HIS affects existing theory and algorithms.

The main contribution of this paper is the proof that under the HIS assumption, perfect-information extensive-form games always exist with solutions to single-stage subgames, resulting in a significant reduction in time complexity. When expressed as extensive-form games, one can optimize all decision variables in isolation while preserving optimality, resulting in an exponential drop in time complexity, hence generalizing to multiple players a similar property to that available under one-sidedness (Xie et al., 2020). To show this result, we apply the principle of optimality to solve any single-stage subgame by breaking it down further into smaller subgames, enabling us to make one-player decisions at a time. In the resulting perfect-information extensive-form game, we exhibit concise representations of states and actions along with Bellman's optimality equations to solve the game. Finally, we present a point-based value-iteration algorithm for solving the original multi-player game leveraging HIS properties. Experiments show that algorithms exploiting these findings scale up to much larger multi-player games without compromising optimality.

1. Background

This section presents state-of-the-art multi- and single-player formulations for Dec-POMDPs under HIS.

Notations. For integers $t_1 \leqslant t_2$, $\kappa_{t_1:t_2}$ is a shorthand for $(\kappa_{t_1}, \kappa_{t_1+1}, \ldots, \kappa_{t_2})$. Let $\kappa_{t_1:t_2}$ be a complete vector, shorthands $\kappa_{t_1:}$ and $\kappa_{:t_2}$ denote suffix and prefix, respectively. For two variables a and b, we denote by δ_a^b the Kronecker delta, which is 1 if a equals b, and 0 otherwise.

1.1. Multi-Player Formulation

An n-player Dec-POMDP is given by tuple $M \doteq \langle n, X, U, Z, p, r, s_0, \gamma, \ell \rangle$, where X is a finite set of hidden states; U^i is the finite actions set for player i, where $U = U^1 \times \cdots \times U^n$ specifies the set of joint actions $u = (u^1, \ldots, u^n)$; Z^i is the finite observation set for player i, where $Z = Z^1 \times \cdots \times Z^n$ specifies the set of joint observations $z = (z^1, \ldots, z^n)$; function $p\colon X \times U \to \triangle(X \times Z)$ describes a transition function with conditional probability distribution $p(y, z|x, u)$ defining the probability of transitioning from state x to y after taking joint action u and seeing z; function $r\colon X \times U \to \mathbb{R}$ is a reward model with $r(x, u)$ being the immediate reward received after taking joint action u from state x; s_0 is the initial state distribution, γ is the discount factor, and ℓ is the number of stages.

In the remainder, we consider M under the HIS assumption. That is, every player $0 < i \leqslant n$ has instantaneous and cost-free access to its subordinate's action $u_{\tau-1}^{i-1}$ and observation z_τ^{i-1} at every stage τ. Consequently, there exists a function ζ^i that maps z_τ^i to $\zeta^i(z_\tau^i) = (u_{\tau-1}^{i-1}, z_\tau^{i-1})$. Player 1 is at the bottom of the hierarchy, *i.e.,* the player whose actions and observations are public to all other players, and player n is at the top of the hierarchy, *i.e.,* the player that sees all actions and observations. These characteristics are embodied in many real-world applications, including autonomous vehicle platooning, assembly line optimization, or railway traffic control. Consider an autonomous vehicle platooning that relies on a leading vehicle followed by a group of autonomous vehicles; see Figure 1. Autonomous vehicles involved in platooning can exchange information between vehicles using Vehicle-to-Everything (V2X) communications in an HIS fashion (Wang et al., 2015) . That is the total data transit from each autonomous vehicle i to its following autonomous vehicle $i+1$. The objective of platoon control is to determine the control input of the following autonomous vehicles so that all the vehicles move at the same speed while maintaining the desired distances between each pair of preceding and following vehicles. Platooning constitutes an efficient technique for increasing road capacity, reducing fuel consumption, and enhancing driving safety and comfort.

1.2. Limitations of Multi-Player Formulations

While HIS is a dominant management style in our society, little is known about how HIS affects existing theories and algorithms. In general, solving M aims at finding optimal joint policy $a_{0:} \doteq (a_0, \ldots, a_{\ell-1})$, *i.e.,* an n-tuple of sequences of private decision rules $a_{0:}^i \doteq (a_0^i, \ldots, a_{\ell-1}^i)$, one per player. For each player i, private decision rule $a_\tau^i : o_\tau^i \mapsto u_\tau^i$ depends on τ-step histories $o_\tau^i \doteq (u_{0:\tau-1}^i, z_{1:\tau}^i)$, with 0-step private history being $o_0^i \doteq \varnothing$. A joint policy is optimal if it maximizes the expected cumulative reward starting at

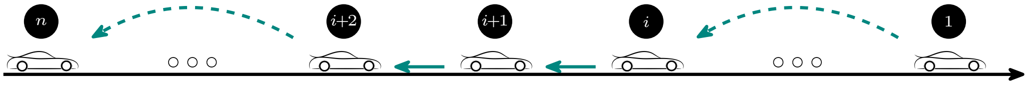

Figure 1. V2X information transmitted to vehicles in the platooning.

initial state distribution s_0 onward and given by $\upsilon_0^{a_{0:}}(s_0) \doteq \mathbb{E}_{(x_0,o_0)\sim\Pr\{\cdot|s_0,a_{0:}\}}\{\alpha_0^{a_{0:}}(x_0,o_0)\}$ where $\alpha_\tau^{a_{\tau:}}(x_\tau,o_\tau) \doteq \mathbb{E}_{(x_{\tau:\ell-1},u_{\tau:\ell-1})\sim\Pr\{\cdot|x_\tau,o_\tau,a_{\tau:}\}}\{\sum_{t=\tau}^{\ell-1}\gamma^{\tau-t}\cdot r(x_t,u_t)\}$ for any game stage τ. Unfortunately, optimally solving M in its multi-player formulation is non-trivial because of the silent coordination dilemma (Rabinovich et al., 2003). Indeed, no statistics on what the players see and do are sufficient to solve M optimally. Private histories o_τ^i are not geared to perform policy evaluation, let alone policy ordering. In addition, joint histories $o_\tau \doteq (o_\tau^1,\ldots,o_\tau^n)$ cannot ensure policy disentanglement, *i.e.,* an individual policy per player. To better understand this, notice that multi-player coordination is based on common ground, *i.e.,* knowledge, beliefs, and assumptions shared among players about the environment at each stage, making it possible to perform policy ordering and disentanglement. Paradoxically, M aims to coordinate agents without common ground, thus explaining the silent coordination dilemma. The motivation for a single-player reformulation is twofold: first, to provide the central planner with the common ground at the offline planning phase, which achieves policy ordering and disentanglement, then to ease the transfer of theories and algorithms from single- to multi-player formulations.

1.3. Single-Player Reformulation

The single-player reformulation describes M from the perspective of an offline central planner (Szer & Charpillet, 2005; Nayyar et al., 2013; Oliehoek, 2013; Dibangoye et al., 2013; 2016). This planner reasons for all players in sync, prescribing a joint decision rule and receiving rewards and public observations. Game M lies in some underlying state and players have experienced a joint history at each plan-time stage. Unfortunately, the central planner can see neither the state nor the joint history. Yet, it can still prescribe to players what joint decision rule to follow based only upon the joint policy it has prescribed to players so far. Upon executing the prescribed joint decision rule, the central planner receives the expected immediate reward and the next public observation, *i.e.,* the information of player 1. This process follows at the next plan-time stage, but the game has another underlying state, and players are experiencing another joint history. This process repeats until the number of stages is exhausted. The summary of the history of prescribed joint decision rules and received public observations, *i.e.,* occupancy state, describes a Markov decision process. Occupancy states proved to be common ground for coordinating players under the silent coordination dilemma, *i.e.,* occupancy states are sufficient statistics for optimal decision-making in Dec-POMDPs (Dibangoye et al., 2013; 2016).

Markov decision process $M' \doteq \langle S, A, \boldsymbol{T}, \boldsymbol{R}, s_0, \gamma, \ell\rangle$ *w.r.t.* M consists of the occupancy-state space S, where occupancy states are conditional probability distribution over hidden states and joint histories; the action space A prescribing joint decision rules; the transition probability $\boldsymbol{T}\colon S \times A \to \triangle(S)$, where $\boldsymbol{T}(s_\tau,a_\tau,s_{\tau+1}) \doteq \sum_{o,z}\delta_{\rho(s_\tau,a_\tau,z^1)}^{s_{\tau+1}}\sum_{x,y}s_\tau(x,o)\cdot p(y,z|x,a_\tau(o))$, where the next occupancy state $s_{\tau+1} \doteq \rho(s_\tau,a_\tau,z_{\tau+1}^1)$ follows from taking joint decision rule a_τ in occupancy state s_τ and then receiving public observation $z_{\tau+1}^1$, *i.e.,* for any arbitrary hidden state y and joint history (o,u,z), we have $s_{\tau+1}(y,(o,u,z))\propto\sum_x s_\tau(x,o)\cdot\delta_{z_{\tau+1}^1}^{z^1}\cdot\delta_{a_\tau(o)}^{u}\cdot p(y,z|x,u)$; and finally, $\boldsymbol{R}\colon S\times A\to\mathbb{R}$ is the expected immediate reward function, *i.e.,* $\boldsymbol{R}(s_\tau,a_\tau)\doteq\sum_{x,o}s_\tau(x,o)\cdot r(x,a_\tau(o))$. Recasting the original multi-player game into an equivalent single-player one allows the principle of optimality to solve the single-player game by breaking it down into single-stage subgames and solving them recursively. Consequently, optimally solving M aims at finding solutions $V_\tau^*(s_\tau)$ of single-stage subgame for every occupancy state s_τ, *i.e.,* $V_\tau^*(s_\tau) = \max_{a_\tau} Q_\tau^*(s_\tau,a_\tau)$, where $Q_\tau^*(s_\tau,a_\tau) \doteq \boldsymbol{R}(s_\tau,a_\tau) + \gamma\sum_{s_{\tau+1}}\boldsymbol{T}(s_\tau,a_\tau,s_{\tau+1})\cdot V_{\tau+1}^*(s_{\tau+1})$ with boundary condition $V_\ell^*(\cdot) \doteq 0$. Each occupancy state s_τ has its corresponding single-stage subgame $G_{s_\tau} \doteq \langle n, A, Q_\tau^*(s_\tau,\cdot)\rangle$, whose solution is $V_\tau^*(s_\tau)$. Optimally solving single-stage subgame G_{s_τ} is significantly more efficient by leveraging the piecewise-linearity and convexity property of action-value functions Q_τ^*.

Lemma 1.1. *For every game stage τ, the optimal value function $Q_\tau^*\colon S\times A\to\mathbb{R}$ is piecewise-linear and convex over occupancy states and joint decision rules. Alternatively, there exists a finite collection $\mathcal{Q}_\tau \subseteq \{\beta_\tau^{a_{\tau+1:}}|a_{\tau+1:}\in A_{\tau+1:}\}$ of action-value functions $\beta_\tau^{a_{\tau+1:}}$ under joint policy $a_{\tau+1:}$, such that: for occupancy state s_τ and joint decision rule a_τ,*

$$Q_\tau^*(s_\tau,a_\tau) = \max\nolimits_{\beta_\tau\in\mathcal{Q}_\tau}\ \mathbb{E}_{(x,o,u)\sim\Pr\{\cdot|s_\tau,a_\tau\}}\{\beta_\tau(x,o,u)\}$$

$$\beta_\tau^{a_{\tau+1:}}(x,o,u) = r(x,u)+\gamma\mathbb{E}_{(y,z)\sim p(\cdot|x,u)}\{\alpha_{t+1}^{a_{\tau+1:}}(y,(o,z))\}$$

with boundary condition $\alpha_\ell^{\cdot}(\cdot) = \beta_\ell^{\cdot}(\cdot) \doteq 0$.

Lemma 1.1 allows us to optimally solve a single-stage subgame G_{s_τ} by taking the best among solutions of single-stage subgames $G_{s_\tau}^{\beta_\tau} \doteq \langle n, A, Q_{\beta_\tau}(s_\tau,\cdot)\rangle$ induced by action-value function $\beta_\tau\in\mathcal{Q}_\tau$ under a fixed joint policy, where

$Q_{\beta_\tau}(s_\tau, \cdot)\colon a_\tau \mapsto \mathbb{E}_{(x,o,u)\sim \Pr\{\cdot|s_\tau,a_\tau\}}\{\beta_\tau(x,o,u)\}$. In particular, the linearity of $Q_{\beta_\tau}(s_\tau, \cdot)$ over joint decision rules will play a crucial role in disentangling decision variables.

1.4. Limitations of Single-Player Reformulations

The single-player reformulation applies under HIS, but the curse of dimensionality restricts its scalability in the face of games with many players. To better understand this, notice that the complexity of optimally solving a single-player reformulation depends on two operators: the point-based backup operator, which optimally solves single-stage subgame $G^{\beta_\tau}_{s_\tau}$, and the estimation operator, which updates all decision variables involved in the common ground, *i.e.,* occupancy states. In either case, the single-player reformulation is not geared to exploit HIS. State-of-the-art approaches to solving $G^{\beta_\tau}_{s_\tau}$ perform either brute-force or implicit enumeration and evaluation of double-exponentially many joint decision rules (Oliehoek et al., 2010; Dibangoye et al., 2009; 2013; 2016). This provides an intuitive explanation for the negative complexity results: optimally solving $G^{\beta_\tau}_{s_\tau}$ is NP-hard, and finding ϵ-approximations remains hard (Tsitsiklis, 1984). The estimation operator also suffers from the curse of dimensionality. Indeed, the number of decision variables of all players in the common ground under the silent coordination dilemma grows exponentially with time and team size. In this paper, we investigate the following question.

> *How can we improve the representations of common ground and Bellman optimality equations to scale-up point-based backup and estimation operators to optimally solving $G^{\beta_\tau}_{s_\tau}$ and eventually M' (resp. M) under HIS?*

2. Hierarchical Information Sharing

This section explores the ramifications of HIS assumption in achieving an optimal solution for a single-stage subgame, specifically, $G^{\beta_\tau}_{s_\tau}$.

2.1. From Single-Stage to Extensive-Form Games

While the principle of optimality allows us to break down the single-player reformulation M' into smaller subgames $G^{\beta_\tau}_{s_\tau}$ per stage, an alternative approach is to segment single-stage subgames $G^{\beta_\tau}_{s_\tau}$ per player further. That allows the centralized planner to act sequentially for each player, starting from player 1 up to player n. In addition, instead of choosing a decision rule for each player based on the current occupancy state and decision rules selected thus far, the planner can independently branch over each history that HIS makes available to the current player, without compromising optimality. A formal description of this process follows.

Starting from player 1 at the bottom of the hierarchy, *cf.* Figure 2, the planner chooses action u^1_τ according to its total available information $\varsigma^1_\tau = (s_\tau, o^1_\tau)$. It then moves to player 2, the next player in the reversed order of the hierarchy, but now it randomly lands on total available information $\varsigma^2_\tau = (\varsigma^1_\tau, u^1_\tau, o^2_\tau)$ and chooses action u^2_τ. The process continues until the planner reaches player n at the top of the hierarchy, where it randomly lands on total available information $\varsigma^n_\tau = (\varsigma^{n-1}_\tau, u^{n-1}_\tau, o^n_\tau)$ and chooses action u^n_τ and receives expected rewards $R(\varsigma^n_\tau, u^n_\tau) \doteq \mathbb{E}_{x\sim \Pr\{\cdot|\varsigma^n_\tau,u^n_\tau\}}\{\beta_\tau(x,o,u)\}$ upon taking action u^n_τ in information state ς^n_τ. Upon acting sequentially for i players, the total information available to the planner denoted $\varsigma^{i+1}_\tau \doteq (\varsigma^i_\tau, u^i_\tau, o^{i+1}_\tau)$, is the current occupancy state s_τ of the single-stage subgame $G^{\beta_\tau}_{s_\tau}$ along with the sequence of actions that the planner selected and private histories that the planner received according to probability $T(\varsigma^{i+1}_\tau|\varsigma^i_\tau) \doteq \Pr\{o^{i+1}_\tau|s_\tau, o^1_\tau, \ldots, o^i_\tau\} \cdot \delta^{\varsigma^{i+1}_\tau}_{\varsigma^i_\tau, u^i_\tau, o^{i+1}_\tau}$.

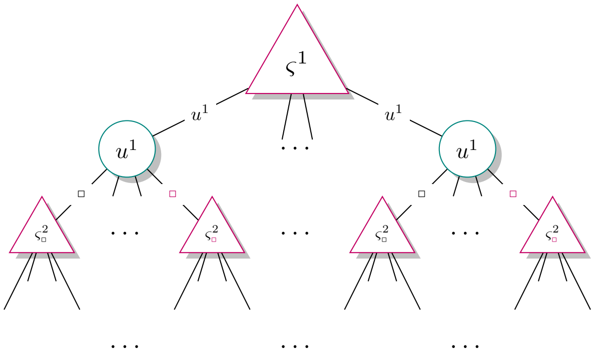

Figure 2. The search space for a single-stage subgame from a centralized planner acting sequentially one player at a time, illustrated as an AND/OR tree. OR nodes (triangle) represent alternative ways to solve $\bar{G}^{\beta_\tau}_{s_\tau}$. AND nodes (circle) represent subproblem alternatives to be solved. **Best viewed in color.**

The total available information of the sequential-move central planner when solving a single-stage subgame describes a common-payoff perfect-information extensive-form game (Shoham & Leyton-Brown, 2008).

Definition 2.1. The common-payoff perfect-information extensive-form game[1] *w.r.t.* $G^{\beta_\tau}_{s_\tau}$ is a tuple $\bar{G}^{\beta_\tau}_{s_\tau} \doteq \langle n, \Sigma, \Psi, T, R\rangle$ where: n is the number of players; Σ is the set of nodes that occupancy state s_τ induces; $\Psi\colon \Sigma \to 2^{\cup^n_{i=1} U^i}$ is a function that specifies the allowed actions from each node $\varsigma \in \Sigma$; transition function $T\colon \Sigma \times (\cup^n_{i=1} U^i) \times \Sigma \to [0,1]$ specifies the probability of a successor node; reward function $R\colon \Sigma \times (\cup^n_{i=1} U^i) \to \mathbb{R}$ specifies the common payoff received upon taking an action in a node.

[1] This definition differs from the formal definition of an extensive form game (EFG). To recover the standard EFG formalism, note that: i) agents act in sequence, ii) rewards are zero everywhere except at the leaves of the game tree, iii) stochastic transitions correspond to the presence of a chance player between two agents.

2.2. Optimally Solving $G^{\beta_\tau}_{s_\tau}$ As $\bar{G}^{\beta_\tau}_{s_\tau}$

Optimally solving a common-payoff perfect-information extensive-form game aims at finding the action-value functions $\beta^{1:n,*}_\tau$ mapping nodes and actions to optimal values. Unlike the original single-stage subgame $G^{\beta_\tau}_{s_\tau}$, the perfect information extensive form game $\bar{G}^{\beta_\tau}_{s_\tau}$ makes the HIS structure explicit. Every time the planner acts on behalf of a player, that player is perfectly informed about all the histories that have previously occurred—*i.e.,* all histories of its subordinates. Hence, the total information nodes include the actions the planner selected for the subordinates of the current player, along with the histories of its subordinates. Nonetheless, both games yield the same solution.

Theorem 2.2. *Any optimal solution for $\bar{G}^{\beta_\tau}_{s_\tau}$ is also an optimal solution for $G^{\beta_\tau}_{s_\tau}$. Besides, the optimal action-value functions $\beta^{1:n,*}_\tau$ of $\bar{G}^{\beta_\tau}_{s_\tau}$ is the solution of the Bellman's optimality equations: at any i, ς^i_τ, and u^i_τ,*

$$\beta^{i,*}_\tau(\varsigma^i_\tau, u^i_\tau) = \mathbb{E}_{\varsigma^{i+1}_\tau \sim T(\cdot|\varsigma^i_\tau, u^i_\tau)}\{\max_{u^{i+1}_\tau} \beta^{i+1,*}_\tau(\varsigma^{i+1}_\tau, u^{i+1}_\tau)\},$$

with boundary condition $\beta^{n,*}_\tau\colon (\varsigma^n_\tau, u^n_\tau) \mapsto R(\varsigma^n_\tau, u^n_\tau)$. Also, greedy decision rule $a^{i,*}_\tau$ for any player i at o^i_τ is:

$$a^{i,*}_\tau(o^i_\tau) \in \mathrm{argmax}_{u^i_\tau}\ \beta^{i,*}_\tau(\varsigma^i_\tau, u^i_\tau),$$

where $\varsigma^i_\tau \doteq \langle s_\tau, o^{1:i}_\tau, a^{1:i-1,*}_\tau(o^{1:i-1}_\tau)\rangle$.

Proof. The proof proceeds in two steps. First, it shows that the original game $G^{\beta_\tau}_{s_\tau}$ can alternatively be solved via a sequential-move central planner, which breaks $G^{\beta_\tau}_{s_\tau}$ down into smaller subgames $\langle G^{\beta_\tau}_{s_\tau,\varnothing}, G^{\beta_\tau}_{s_\tau,a^1_\tau}, \ldots, G^{\beta_\tau}_{s_\tau,a^{1:n-1}_\tau}\rangle$, one subgame per player. To this end, recall the goal of optimally solving $G^{\beta_\tau}_{s_\tau}$, *i.e.,* finding a joint decision rule which yields the highest performance index, $V_{\beta_\tau}(s_\tau) \doteq \max_{a_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau)$. The expansion of joint decision rule a_τ as a n-tuple of private decision rules $(a^1_\tau, a^2_\tau, \ldots, a^n_\tau)$ allows to rewite the objectif of $G^{\beta_\tau}_{s_\tau}$ as follows, $V_{\beta_\tau}(s_\tau) = \max_{a^1_\tau} \max_{a^2_\tau} \ldots \max_{a^n_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau)$. Let $Q^i_{\beta_\tau}(s_\tau, \cdot)\colon a^{1:i}_\tau \mapsto \max_{a^{i+1:n}_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau)$ be a sequential action-value function. Then, it follows that

$$\begin{aligned} V_{\beta_\tau}(s_\tau) &= \max_{a^1_\tau} \max_{a^2_\tau} \ldots \max_{a^n_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau), \\ &= \max_{a^1_\tau}\ \left[\max_{a^2_\tau} \ldots \max_{a^n_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau)\right], \\ &= \max_{a^1_\tau}\ Q^1_{\beta_\tau}(s_\tau, a^1_\tau). \end{aligned}$$

Interestingly, for every player $i \in \{1, 2, \ldots, n-1\}$, the action-value functions $Q^i_{\beta_\tau}(s_\tau, a^{1:i}_\tau)$ satisfy the following recursion

$$\begin{aligned} Q^i_{\beta_\tau}(s_\tau, a^{1:i}_\tau) &= \max_{a^{i+1}_\tau} \max_{a^{i+2}_\tau} \ldots \max_{a^n_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau), \\ &= \max_{a^{i+1}_\tau}\ \left[\max_{a^{i+2}_\tau} \ldots \max_{a^n_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau)\right], \\ &= \max_{a^{i+1}_\tau}\ Q^{i+1}_{\beta_\tau}(s_\tau, a^{1:i+1}_\tau), \end{aligned}$$

with boundary condition $Q^n_{\beta_\tau}(s_\tau, a_\tau) \doteq Q_{\beta_\tau}(s_\tau, a_\tau)$. For any arbitrary player $i \in \{2, 3, \ldots, n\}$, define game $G^{\beta_\tau}_{s_\tau,a^{1:i-1}_\tau} \doteq \langle i, A^i, Q^i_{\beta_\tau}(s_\tau, a^{1:i-1}_\tau, \cdot)\rangle$ to be the subgame upon the sequential-move central planner selected decision rules $a^{1:i-1}_\tau$ starting in game $G^{\beta_\tau}_{s_\tau}$, with boundary condition $G^{\beta_\tau}_{s_\tau,\varnothing} \doteq \langle 1, A^1, Q^1_{\beta_\tau}(s_\tau, \cdot)\rangle$. Consequently, optimally solving the original game $G^{\beta_\tau}_{s_\tau}$ can be performed by optimally solving smaller subgames $\langle G^{\beta_\tau}_{s_\tau,\varnothing}, G^{\beta_\tau}_{s_\tau,a^1_\tau}, \ldots, G^{\beta_\tau}_{s_\tau,a^{1:n-1}_\tau}\rangle$, one subgame per player, recursvively.

Next, we shall prove that the best decision rule in any arbitrary sequential-move subgame $G^{\beta_\tau}_{s_\tau,a^{1:i-1}_\tau}$ depends on the current occupancy state s_τ along with previously selected decision rules $a^{1:i-1}_\tau$, only through the corresponding nodes $\varsigma^i_\tau \doteq (s_\tau, u^{1:i-1}_\tau, o^{1:i}_\tau)$ of the perfect information extensive form game $\bar{G}^{\beta_\tau}_{s_\tau}$. In other words, instead of selecting actions for all private histories of player i in sync, one can choose the best action for each private history independently without compromising optimality. The proof of this statement proceeds by induction from player n to player 1. At player n, the greedy decision rule $\hat{a}^n_\tau$ satisfies the following:

$$\begin{aligned} \hat{a}^n_\tau &\in \mathrm{argmax}_{a^n_\tau}\ Q^n_{\beta_\tau}(s_\tau, a_\tau), \\ &\in \mathrm{argmax}_{a^n_\tau}\ Q_{\beta_\tau}(s_\tau, a_\tau), \\ &\in \mathrm{argmax}_{a^n_\tau}\ \mathbb{E}_{(x,o,u)\sim \Pr\{\cdot|s_\tau, a_\tau\}}\{\beta_\tau(x, o, u)\}. \end{aligned}$$

Expanding over private histories of player n, we have that

$$\hat{a}^n_\tau(o^n_\tau) \in \mathrm{argmax}_{u^n_\tau}\ \mathbb{E}_{(x,o,u)\sim \Pr\{\cdot|s_\tau, o^n_\tau, a_\tau\}}\{\beta_\tau(x, o, u)\}.$$

Leveraging information available to player n as provided by the HIS assumption, we know that the knowledge of private history o^n_τ implies the knowledge of histories of all other players $o^{1:n-1}_\tau$, hence the joint history o_τ, *i.e.,*

$$\hat{a}^n_\tau(o^n_\tau) \in \mathrm{argmax}_{u^n_\tau}\ \mathbb{E}_{x\sim \Pr\{\cdot|s_\tau, o_\tau, a_\tau\}}\{\beta_\tau(x, o, u)\}.$$

In addition, the knowledge of $o^{1:n-1}_\tau$ together with the decision rules $a^{1:n-1}_\tau$ the sequential-move central planner selected previously, makes it possible to access node $\varsigma^n_\tau \doteq \langle s_\tau, o^{1:n}_\tau, a^{1:n-1}_\tau(o^{1:n-1}_\tau)\rangle$ such that:

$$\hat{a}^n_\tau(o^n_\tau) \in \mathrm{argmax}_{u^n_\tau}\ \beta^n_\tau(\varsigma^n_\tau, u^n_\tau),$$

where $\beta^n_\tau\colon (\varsigma^n_\tau, u^n_\tau) \mapsto \mathbb{E}_{x\sim \Pr\{\cdot|\varsigma^n_\tau, u^n_\tau\}}\{\beta_\tau(x, o, u)\}$, which proves the statement holds at player n. Define function

$\alpha^n_\tau \colon \varsigma^n_\tau \mapsto \max_{u^n_\tau} \beta^n_\tau(\varsigma^n_\tau, u^n_\tau)$ at player n. Notice that the value of the sequential-move subgame $G^{\beta_\tau}_{s_\tau, a^{1:n-1}_\tau}$ can be rewritten as follows:

$$\begin{aligned} Q^{n-1}_{\beta_\tau}(s_\tau, a^{1:n-1}_\tau) &= \max_{a^n_\tau} Q^n_{\beta_\tau}(s_\tau, a_\tau) \\ &= \mathbb{E}_{\varsigma^n_\tau \sim \Pr\{\cdot | s_\tau, a^{1:n-1}_\tau\}}\{\max_{u^n_\tau} \beta^n_\tau(\varsigma^n_\tau, u^n_\tau)\} \\ &= \mathbb{E}_{\varsigma^n_\tau \sim \Pr\{\cdot | s_\tau, a^{1:n-1}_\tau\}}\{\alpha^n_\tau(\varsigma^n_\tau)\}. \end{aligned}$$

Suppose the statement holds for any player $i > 1$, with greedy decision rule $\hat{a}^i_\tau(o^i_\tau) \in \operatorname{argmax}_{u^i_\tau} \; \beta^i_\tau(\varsigma^i_\tau, u^i_\tau)$. Define function $\alpha^i_\tau \colon \varsigma^i_\tau \mapsto \max_{u^i_\tau} \beta^i_\tau(\varsigma^i_\tau, u^i_\tau)$ at player i. Also, the value of the sequential-move subgame $G^{\beta_\tau}_{s_\tau, a^{1:i-1}_\tau}$ can be rewritten by expanding over the sequential-move nodes $\varsigma^i_\tau \doteq \langle s_\tau, o^{1:i}_\tau, a^{1:i-1}_\tau(o^{1:i-1}_\tau)\rangle$, *i.e.*,

$$Q^{i-1}_{\beta_\tau}(s_\tau, a^{1:i-1}_\tau) = \mathbb{E}_{\varsigma^i_\tau \sim \Pr\{\cdot | s_\tau, a^{1:i-1}_\tau\}}\{\alpha^i_\tau(\varsigma^i_\tau)\}.$$

We are now ready to prove the statement also holds at player $i-1$. From the sequential-move central planner's viewpoint, decision rule $\hat{a}^{i-1}_\tau$ satisfies the following expression:

$$\begin{aligned} \hat{a}^{i-1}_\tau &\in \operatorname{argmax}_{a^{i-1}_\tau} \; Q^{i-1}_{\beta_\tau}(s_\tau, a^{1:i-1}_\tau), \\ &\in \operatorname{argmax}_{a^{i-1}_\tau} \mathbb{E}_{\varsigma^i_\tau \sim \Pr\{\cdot | s_\tau, a^{1:i-1}_\tau\}}\{\alpha^i_\tau(\varsigma^i_\tau)\}. \end{aligned}$$

Similarly to player n, the knowledge of $o^{1:i-1}_\tau$ together with the decision rules $a^{1:i-2}_\tau$ the sequential-move central planner selected previously, makes it possible to access node $\varsigma^{i-1}_\tau \doteq \langle s_\tau, o^{1:i-1}_\tau, a^{1:i-2}_\tau(o^{1:i-2}_\tau)\rangle$ such that:

$$\hat{a}^{i-1}_\tau(o^{i-1}_\tau) \in \operatorname{argmax}_{u^{i-1}_\tau} \; \beta^{i-1}_\tau(\varsigma^{i-1}_\tau, u^{i-1}_\tau),$$

where $\beta^{i-1}_\tau \colon (\varsigma^{i-1}_\tau, u^{i-1}_\tau) \mapsto \mathbb{E}_{\varsigma^i_\tau \sim \Pr\{\cdot | \varsigma^{i-1}_\tau, u^{i-1}_\tau\}}\{\alpha^i_\tau(\varsigma^i_\tau)\}$, which proves the statement holds at player $i-1$. Define function $\alpha^{i-1}_\tau \colon \varsigma^{i-1}_\tau \mapsto \max_{u^{i-1}_\tau} \beta^{i-1}_\tau(\varsigma^{i-1}_\tau, u^{i-1}_\tau)$ at player $i-1$. Consequently, the value of the sequential-move subgame $G^{\beta_\tau}_{s_\tau, \varnothing}$ can be rewritten by expanding over the sequential-move nodes $\varsigma^1_\tau \doteq \langle s_\tau, o^1_\tau\rangle$, *i.e.*,

$$V_{\beta_\tau}(s_\tau) = \mathbb{E}_{\varsigma^1_\tau \sim \Pr\{\cdot | s_\tau\}}\{\alpha^1_\tau(\varsigma^1_\tau)\}.$$

The value of a cooperative game being unique, we know the optimal solution for $\bar{G}^{\beta_\tau}_{s_\tau}$ is also an optimal solution for $G^{\beta_\tau}_{s_\tau}$. In demonstrating this statement, we also exhibited Bellman's optimality equations, providing the solution of the perfect-information extensive-form game $\bar{G}^{\beta_\tau}_{s_\tau}$, *i.e.*, at any player i, node ς^i_τ, and action u^i_τ,

$$\beta^{i,*}_\tau(\varsigma^i_\tau, u^i_\tau) = \mathbb{E}_{\varsigma^{i+1}_\tau \sim T(\cdot | \varsigma^i_\tau, u^i_\tau)}\{\max_{u^{i+1}_\tau} \beta^{i+1,*}_\tau(\varsigma^{i+1}_\tau, u^{i+1}_\tau)\},$$

with boundary condition $\beta^{n,*}_\tau \colon (\varsigma^n_\tau, u^n_\tau) \mapsto R(\varsigma^n_\tau, u^n_\tau)$. Which ends the proof. □

Theorem 2.2 introduces Bellman's optimality equations that enable us to find a greedy joint decision at single-stage subgame $G^{\beta_\tau}_{s_\tau}$ by solving the corresponding extensive-form game $\bar{G}^{\beta_\tau}_{s_\tau}$. It proceeds in two phases. From player n at the top of the hierarchy to player 1 at the bottom, a backward pass computes optimal action-values $\beta^{i,*}_\tau(\varsigma^i_\tau, u^i_\tau)$ for each player i, each node ς^i_τ, and each action u^i_τ. Then, from player 1 at the bottom of the hierarchy to player n at the top, a forward pass selects a greedy decision rule independently for each player i, and each node ς^i_τ. This backward induction algorithm requires a linear time complexity with the number of players, nodes, and actions $\mathbf{O}(n|\Sigma||U^*|)$ instead of double exponential $\mathbf{O}(|O^*|^{|U^*|^n})$ where $O^* \doteq \operatorname{argmax}_{O^i} |O^i|$ with O^i being the set of reachable histories of player i in s_τ and $U^* \doteq \operatorname{argmax}_{U^i} |U^i|$. A careful reader would notice that the linearity of β_τ over occupancy states and joint decision rules is key in demonstrating Theorem 2.2.

2.3. Nested-Occupancy States

Upon inspection of perfect-information extensive-form game $\bar{G}^{\beta_\tau}_{s_\tau}$, one can see that despite the polynomial-time complexity of the point-based backup, $\bar{G}^{\beta_\tau}_{s_\tau}$ may contain a significant number of nodes. That is because nodes in $\bar{G}^{\beta_\tau}_{s_\tau}$ provide total information available to the planner at any player i—*i.e.*, $\varsigma^i_\tau \doteq \langle s_\tau, o^i_\tau, u^{:i-1}_\tau\rangle$, which may result in redundant and unnecessary computations. To address this challenge, we propose the introduction of a statistic referred to as *nested-occupancy state* that we shall maintain in place of the total information available to the planner.

At player i, a nested-occupancy state $s^i_\tau \doteq (b^i_\tau, o^i_\tau, u^{:i-1}_\tau)$ consists of a private history o^i_τ of player i, the actions of its subordinates $u^{:i-1}_\tau$, and a nested-belief state b^i_τ. Besides, the nested-belief state b^i_τ at player i is a posterior distribution over histories o^{i+1}_τ and nested-belief states b^{i+1}_τ of the immediate superior player $i+1$. This distribution is conditional on the total data available to the planner at player i, i.e., $b^i_\tau(o^{i+1}_\tau, b^{i+1}_\tau) \doteq \Pr\{o^{i+1}_\tau, b^{i+1}_\tau | \varsigma^i_\tau\}$, for any histories o^{i+1}_τ and nested-belief states b^{i+1}_τ; with boundary condition $b^n_\tau(x_\tau) \doteq \Pr\{x_\tau | \varsigma^n_\tau\}$, for any hidden state x_τ. Interestingly, the nested-occupancy state has many important properties. First, it is a sufficient statistic for optimally solving $\bar{G}^{\beta_\tau}_{s_\tau}$.

Theorem 2.3 (Proof in Appendix B). *At player i, the nested-occupancy state s^i_τ is a sufficient statistic of the total data ς^i_τ available to the planner for optimally solving the perfect-information extensive-form game $\bar{G}^{\beta_\tau}_{s_\tau}$.*

Theorem 2.3 suggests using a nested-occupancy state as an alternative to the total data available to the planner without compromising optimality. This statistic facilitates the aggregation of histories of a player that convey the same information about the game, thus effectively reducing the dimensionality of the game. Prior to delving further, it is necessary to introduce three equivalence relations. First, two nested-occupancy states at player i, represented as $s_\tau^{i,\bullet} \doteq (b_\tau^{i,\bullet}, o_\tau^{i,\bullet}, u_\tau^{:i-1,\bullet})$ and $s_\tau^{i,\circ} \doteq (b_\tau^{i,\circ}, o_\tau^{i,\circ}, u_\tau^{:i-1,\circ})$, are considered $\mathscr{B}_1$-equivalent if they differ only through their histories, *i.e.,* whenever $(b_\tau^{i,\circ}, u_\tau^{:i-1,\circ}) = (b_\tau^{i,\bullet}, u_\tau^{:i-1,\bullet})$ then $s_\tau^{i,\bullet} \sim_{\mathscr{B}_1} s_\tau^{i,\circ}$. Similarly, they are considered $\mathscr{B}_2$-equivalent if they share the same nested-belief state and histories of subordinates, *i.e.,* whenever $(b_\tau^{i,\circ}, o_\tau^{i-1,\circ}) = (b_\tau^{i,\bullet}, o_\tau^{i-1,\bullet})$ then $s_\tau^{i,\bullet} \sim_{\mathscr{B}_2} s_\tau^{i,\circ}$. Last, two private histories are considered $\mathscr{P}$-equivalent if their optimal actions or, more generally, policies are interchangeable.

Theorem 2.4 (Proof in Appendix C.4)**.** *Let $\bar{G}_{s_\tau}^{\beta_\tau}$ be a perfect-information extensive-form game. Let $s_\tau^{i,\bullet}$ and $s_\tau^{i,\circ}$ be two nested-occupancy states induced by occupancy state s_τ at player i. The following properties hold.*

1. *If $s_\tau^{i,\bullet} \sim_{\mathscr{B}_1} s_\tau^{i,\circ}$ then $\beta_\tau^{i,*}(s_\tau^{i,\bullet}, u_\tau^i) = \beta_\tau^{i,*}(s_\tau^{i,\circ}, u_\tau^i)$.*
2. *If $s_\tau^{i,\bullet} \sim_{\mathscr{B}_2} s_\tau^{i,\circ}$ then $o_\tau^{i,\bullet} \sim_{\mathscr{P}} o_\tau^{i,\circ}$.*

Theorem 2.4 establishes that two $\mathscr{B}_1$-equivalent nested-occupancy states have the same optimal actions, resulting in significant computational savings. Moreover, it showcases that two $\mathscr{B}_2$-equivalent nested-occupancy states have their corresponding histories following the same policy. This insight allows for compact occupancy states, wherein only one history per equivalent class is retained, leading to faster estimations. Similarly, Dibangoye et al. (2016) employed compact occupancy states while utilizing the complete distribution over hidden states and histories of all teammates to determine when two histories are equivalent. Our equivalence relations, however, are based on nested belief states that are more concise than occupancy states, resulting in a more aggressive compression. At player n, for instance, the planner groups together histories that share the belief state, and this process continues down the hierarchy.

3. Near-Optimally Solving M' Under HIS

This section adapts the point-based value-iteration (PBVI) algorithm (Pineau et al., 2003) to compute ϵ-optimal joint policy for M' (resp. M) under HIS starting at initial state distribution s_0 for planning horizon ℓ. We chose the PBVI algorithm because it leverages the linear functions β_τ involved in the optimal value function. Besides, it is guaranteed to find near-optimal solutions asymptotically. Notice that algorithms that do not leverage the linear functions β_τ, *e.g.,* feature-based heuristic search value iteration (Dibangoye et al., 2013; 2016), cannot benefit from our findings.

PBVI, *cf.* Algorithm 1 in Appendix A, has two main parts for solving M' (resp. M) under HIS. First, it bounds the size of the value function at each stage τ of the game by representing the value only at a finite, reachable occupancy subset $\tilde{S}_\tau$. Next, it optimizes the value function represented as a collection $\mathcal{V}_\tau$ at each stage τ using point-based backup, *i.e.,* at any stage τ, $\mathcal{V}_\tau = \{\texttt{backup}(s_\tau, \mathcal{V}_{\tau+1})\colon s_\tau \in \tilde{S}_\tau\}$, where backups are executed in no particular order, *i.e.,* $\texttt{backup}(s_\tau, \mathcal{V}_{\tau+1}) = \operatorname{argmax}_{\alpha_\tau^{a_\tau:}\colon\, a_\tau \in A, \alpha_\tau^{a_{\tau+1:}} \in \mathcal{V}_\tau} \; Q_{\beta_\tau^{a_{\tau+1:}}}(s_\tau, a_\tau)$. Each iteration traverses occupancy-state subsets bottom up. This iterative process repeats until convergence or until a budget, *e.g.,* CPU time, memory, or number of iterations, has been exhausted. The algorithm adds supplemental points into occupancy subsets to improve the value functions further. It selects candidate points using a portfolio of exploration strategies, including random explorations and greedy *w.r.t.* underlying (PO)MDP value functions. For every stage τ, the algorithm adds only candidate points beyond a certain distance from the occupancy subset $\tilde{S}_\tau$ to create a new occupancy-state set $\tilde{S}_{\tau+1}$.

For any arbitrary occupancy-state subsets $\tilde{S}_{0:}$, PBVI produces a value $\upsilon_0(s_0)$. The error between $\upsilon_0(s_0)$ and $\upsilon_0^*(s_0)$ is bounded. The bound depends on how $\tilde{S}_{0:}$ samples the entire occupancy-state space; with denser sampling, the estimate $\upsilon_0(s_0)$ converges to $\upsilon_0^*(s_0)$. The remainder of this section states and proves our error bound. It is also shown that the PBVI algorithm under HIS allows an exponential decrease in time complexity over standard versions of PBVI.

Define the density $\delta_{\tilde{S}_{0:}}$ to be the maximum distance from any reachable occupancy state to subsets $\tilde{S}_{0:}$. More precisely, $\delta_{\tilde{S}_{0:}} \doteq \max_{\tau \in [\![0:\ell-1]\!]} \max_{s \in S_\tau} \min_{s' \in \tilde{S}_\tau} \|s - s'\|_1$. Define a positive scalar c such that $\|r(\cdot,\cdot)\|_\infty \leqslant c$.

Theorem 3.1 (Proof in Appendix D)**.** *For any occupancy subsets $\tilde{S}_{0:}$, the error of the PBVI algorithm is bounded by*

$$\upsilon_0^*(s_0) - \upsilon_0(s_0) \leqslant 2c\delta_{\tilde{S}_{0:}} \frac{1 + \ell\gamma^{\ell+1} - (\ell+1)\gamma^\ell}{(1-\gamma)^2}.$$

It is worth noticing that whenever ℓ goes to infinity, our bound meets that from Pineau et al. (2003) for infinite-horizon partially observable Markov decision processes.

Theorem 3.2. *Let $|\tilde{S}^*| = \max_{t \in 0,1,\ldots,\ell-1} |\tilde{S}_t|$ be the maximum size of the selected spaces of occupancy states. The complexity of the PBVI algorithm under HIS is about $\boldsymbol{O}\left(n\ell|\tilde{S}^*|^2|Z^*|^{n\ell}|U^*|^{1+n(\ell+1)}\right)$.*

Proof. As stated in Section 2.2, the complexity of solving a single-stage subgame $G_{s_\tau}^{\beta_\tau}$ is about $\boldsymbol{O}(n|\Sigma||U^*|)$, where $|\Sigma|$ is the size of the extensive-form game. Since the set of reachable histories for each player is bounded in size by $(|Z^*||U^*|)^{n\ell}$, we get $|\Sigma| \leqslant (|Z^*||U^*|)^{n\ell}|U^*|^n$.
At stage τ a subgame is solved for each $s_\tau \in \tilde{S}_\tau$ and each β_τ (obtained from $\tilde{S}_{\tau+1}$). Thus, at most $\ell|\tilde{S}^*|^2$ subgames are solved on the whole horizon.
As a consequence, the total complexity is about $\mathbf{O}(\ell|\tilde{S}^*|^2(|Z^*||U^*|)^{n\ell}|U^*|^n n|U^*|) = \mathbf{O}(n\ell|\tilde{S}^*|^2|Z^*|^{n\ell}|U^*|^{1+n(\ell+1)})$. □

As a comparison, the number of joint decision rules at step-time τ is bounded by $|U^*|^{(|Z^*||U^*|)^{n\ell}}$. Thus, using similar reasoning as in the proof of Theorem 3.2, the complexity of the vanilla PBVI algorithm on the single-agent reformulation is about $\boldsymbol{O}\left(\ell|\tilde{S}^*|^2|U^*|^{(|Z^*||U^*|)^{n\ell}}\right)$

	hPBVI		**PBVI^{milp}**		**PBVI^{enum}**		**A2C**		**IQL**	
tiger(2)	**0.18**	**112.50**	1.63	91.81	OOT		–	95.73	–	80.15
tiger(3)	**1.05**	**262.50**	141.72	218.81	OOT		–	167.16	–	255.99
tiger(4)	**6.28**	**393.75**	OOT		OOT		–	207.70	–	218.47
tiger(6)	**912.63**	**483.78**	OOT		OOT		–	200.96	–	-129.51
recycling(2)	**0.02**	**93.73**	0.78	**93.73**	0.04	**93.73**	–	93.34	–	93.02
recycling(3)	**0.05**	**252.83**	19.28	**252.83**	143.59	247.80	–	142.00	–	129.57
recycling(4)	**0.19**	**310.07**	835.96	283.05	OOT		–	181.25	–	153.03
recycling(6)	**1.91**	**459.78**	OOT		OOT		–	186.11	–	197.93
recycling(8)	**138.28**	**600.00**	OOT		OOT		–	126.19	–	244.02
mabc(2)	0.05	**27.42**	0.15	27.40	**0.04**	**27.42**	–	27.18	–	27.2
mabc(3)	**0.03**	**23.24**	1.21	**23.24**	0.65	**23.24**	–	23.27	–	23.24
mabc(4)	**0.07**	**24.94**	223.26	**24.94**	OOT		–	24.36	–	**24.94**
mabc(7)	**1.66**	**27.25**	OOT		OOT		–	16.72	–	26.82
mabc(10)	**139.82**	**27.75**	OOT		OOT		–	12.25	–	24.84
grid3x3(2)	**0.61**	**24.44**	1329.33	24.33	OOT		–	22.93	–	24.35
grid3x3(3)	**65.43**	**28.16**	OOT		OOT		–	27.92	–	**28.16**
mars(2)	**0.28**	**84.33**	248.61	76.15	OOT		–	43.20	–	52.86
boxpushing(2)	**0.66**	**675.46**	24.58	576.30	OOT		–	180.11	–	614.6

Table 1. Snapshot of empirical results, *cf.* Appendix F. For each game(n) and algorithm, we report time (in seconds) per backup and the best value for horizon $\ell = 30$. OOT means time limit of 30 minutes has been exceeded and '–' is not applicable.

4. Experiments

This section presents the outcomes of our experiments, which were carried out to juxtapose our findings with the leading-edge theory employed in global methods, encompassing the utilization of the PBVI algorithm as a standard algorithmic scheme. Our analysis involves three variants of the PBVI algorithm, namely PBVI^{enum}, PBVI^{milp}, and hPBVI, each employing distinct methods of performing point-based backups. PBVI^{enum} relies on brute-force enumeration of joint decision rules. At the same time, PBVI^{milp} utilizes mixed-integer linear programs (MILPs) for implicit enumeration following the state-of-art approach for general Dec-POMDPs (Dibangoye et al., 2016). In contrast, hPBVI leverages the subgame solving methods described above. We used ILOG CPLEX Optimization Studio to solve the MILPs. Finally, hPBVI incorporates our findings to facilitate point-based backups under hierarchical information sharing. Global methods are not designed to scale up with players. To present a comprehensive view, we have also compared our results against local policy- and value-based methods, *i.e.,* advantage actor-critic (A2C) (Konda & Tsitsiklis, 1999) and independent Q-learning (IQL) (Tan, 1998), respectively. The experiments were executed on an Ubuntu machine with 32GB of available RAM and a 2.5GHz processor, utilizing only one core, with a time limit of 30 minutes.

We have comprehensively assessed various algorithms using several two-player benchmarks sourced from academic literature, available at `masplan.org`. These benchmarks encompass mabc, recycling, grid3x3, boxpushing, mars, and tiger. To enable a comparison of multiple players, we have also introduced the multi-player variants of these benchmarks. Please refer to Appendix E for a detailed definition of these multi-player benchmarks.

Our study aimed to assess the reduction in complexity achieved by point-based backups and its effect on solving larger multi-player games. Our findings show that hPBVI performs point-based backups significantly faster than other methods, which enables it to scale up to larger teams, as illustrated in Table 4. Specifically, hPBVI was able to perform point-based backups for up to 10 players in about 139.82 seconds in `mabc(10)` at $\ell = 30$, while PBVI^{enum} ran out of time for 4 players, and PBVI^{milp} for 5 players. Additionally, hPBVI converges faster than PBVI^{enum} and PBVI^{milp} in 2- to 3-player domains. For example, hPBVI can converge in under 1 second in `grid3x3(2)` at $\ell = 30$, while PBVI^{milp} takes about 1329.33 seconds, not to mention PBVI^{enum}. Our results in Table 4 demonstrate that hPBVI can scale up to larger teams of players where neither PBVI^{milp} nor PBVI^{enum} can. Figure 4 illustrates the capacity of hPBVI to address larger problems when compared to standard PBVI algorithms (a more extensive comparison of computational times is proposed in Appendix F).

Local methods A2C and IQL do scale up to larger teams as expected. Surprisingly, they perform very well on certain domains with weakly coupled players, as shown in `mabc(4)` and `grid3x3(3)`, *cf.* Table 4. However, hPBVI always performs better A2C and IQL on all benchmarks except `mabc(4)` and `grid3x3(3)`, which exhibit local behaviors that are global optimal solutions. Moreover, it converges faster than A2C and IQL on all tested benchmarks, Figure 4 illustrates anytime performances for the recycling problem (Figures 13 to 16 in Appendix F provide more detailed results for each benchmark). Although this observation goes beyond our original goal, it provides encouraging insights when comparing local against global methods over teams of medium sizes. Nonetheless, we caution readers against drawing general conclusions from this observation, as different local methods may yield different local optima and convergence rates.

5. Discussion

This paper presents a point-based value iteration algorithm for near-optimally solving Dec-POMDPs. It exploits a hierarchical information-sharing structure, a dominant management style in our society for corporations, governments, criminal enterprises, armies, and religions. Under this assumption, it shows that point-based backup operations can be solved as perfect-information extensive-form games without compromising optimality. Doing so results in an expo-

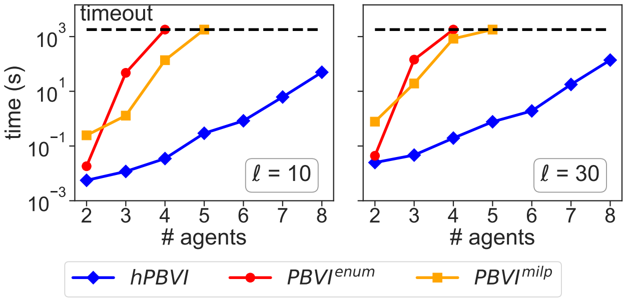

Figure 3. Average backup time for the recycling problem with different numbers of agents.

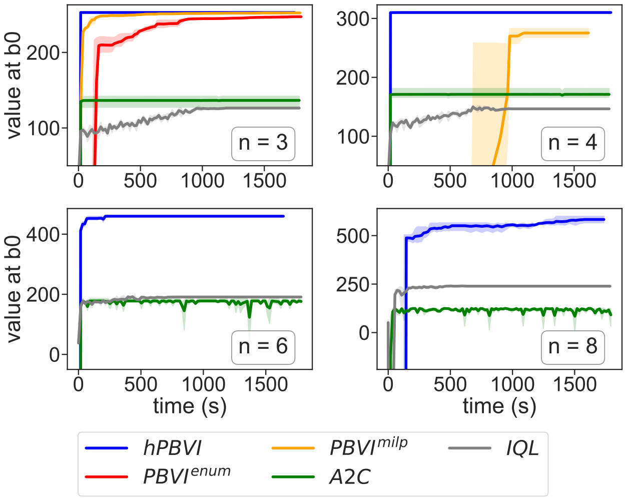

Figure 4. Anytime values for the recycling problem with teams of size $n \in \{3, 4, 6, 8\}$ and planning horizon $\ell = 30$.

nential complexity drop, allowing global methods to scale up to larger teams of players. A thorough empirical analysis reveals that algorithms utilizing our findings can scale up to larger teams of players. In contrast, the state-of-the-art global approaches quickly ran out of resources. Another important empirical finding is that our approach scales to all medium-sized tested domains while providing equal or better performances than a state-of-the-art local method.

Traditionally, global methods have been considered ineffective in games that involve medium to large-sized teams of players. For instance, state-of-the-art Dec-POMDP solvers such as FB-HSVI were only designed for two players (Oliehoek et al., 2010; Dibangoye et al., 2009; 2013; 2016). However, we have presented a paper that puts forth several propositions for developing global methods that possess the scalability of local methods while maintaining global guarantees. In applications where the stakes are high and critical, such as search and rescue, security, and healthcare, scalable global methods with more reliable solutions than those from local methods are essential.

Similarly to Kovařík et al. (2022), our paper demonstrates that simultaneous-move games can be solved sequentially and centrally while allowing each player to act optimally in a decentralized manner. This sequential and centralized training for decentralized execution (SCTDE) approach enables us to leverage private information available to players in a simple manner. Additionally, the SCTDE approach enables us to reason for each player individually in a way that is similar to extensive-form games. This results in a significant reduction in complexity, especially when faced with public observations. This insight also allows us to transfer theories and algorithms from extensive-form games to simultaneous-move games. While our study demonstrates how to optimally solve single-stage games as extensive-form games, the principle we discussed also applies to planning and learning to act in multi-stage general-sum games.

Our study focuses on analyzing the line hierarchical structure. Although previous studies, such as Xie et al. (2020), have successfully applied the SCTDE approach in two-player common-payoff games, our paper extends the research to multiple players. Additionally, our research can be further expanded to consider other structures, such as a tree structure, where players at the same level are independent. In the past, several forms of structure have been investigated such as dynamics independence (Becker et al., 2004), weak-separability (Nair et al., 2005), and delayed information-sharing (Nayyar et al., 2010). However, it is not clear how the hierarchical assumption affects these structures and the corresponding planning and learning theories.

Acknowledgements

This work was supported by ANR project Planning and Learning to Act in Systems of Multiple Agents under Grant ANR-19-CE23-0018, and ANR project Data and Prior, Machine Learning and Control under Grant ANR-19-CE23-0006, and ANR project Multi-Agent Trust Decision Process for the Internet of Things under Grant ANR-21-CE23-0016, all funded by French Agency ANR.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amato, C., Bernstein, D. S., and Zilberstein, S. Optimizing memory-bounded controllers for decentralized pomdps. *arXiv preprint arXiv:1206.5258*, 2012.

Amato, C., Chowdhary, G., Geramifard, A., Üre, N. K., and

Kochenderfer, M. J. Decentralized control of partially observable markov decision processes. In *CDC*, 2013.

Becker, R., Zilberstein, S., Lesser, V. R., and Goldman, C. V. Solving Transition Independent Decentralized Markov Decision Processes. *JAIR*, 22:423–455, 2004.

Bellman, R. E. *Dynamic Programming*. Dover Publications, Incorporated, 1957.

Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of Operations Research*, 27, 2002.

Bono, G., Dibangoye, J. S., Matignon, L., Pereyron, F., and Simonin, O. Cooperative Multi-agent Policy Gradient. In *ECML-PKDD*, pp. 459–476, 2018.

Dibangoye, J. S., Mouaddib, A., and Chaib-draa, B. Point-based incremental pruning heuristic for solving finite-horizon Dec-POMDPs. In *AAMAS*, pp. 569–576, 2009.

Dibangoye, J. S., Amato, C., and Doniec, A. Scaling up decentralized mdps through heuristic search. In de Freitas, N. and Murphy, K. P. (eds.), *UAI*, pp. 217–226, 2012.

Dibangoye, J. S., Amato, C., Buffet, O., and Charpillet, F. Optimally solving Dec-POMDPs as continuous-state MDPs. In *IJCAI*, pp. 90–96, 2013.

Dibangoye, J. S., Amato, C., Buffet, O., and Charpillet, F. Exploiting Separability in Multi-Agent Planning with Continuous-State MDPs. In *AAMAS*, 2014.

Dibangoye, J. S., Amato, C., Buffet, O., and Charpillet, François, D. Optimally solving Dec-POMDPs as continuous-state MDPs. *JAIR*, 2016.

Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. In *NIPS*, 2016.

Hansen, E. A., Bernstein, D. S., and Zilberstein, S. Dynamic Programming for Partially Observable Stochastic Games. In *AAAI*, 2004.

Horák, K. and Bošanskỳ, B. Solving partially observable stochastic games with public observations. In *AAAI*, 2019.

Horák, K., Bošanskỳ, B., and Pěchouček, M. Heuristic search value iteration for one-sided partially observable stochastic games. In *AAAI*, 2017.

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

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Neural Information Processing Systems*, 1999.

Kovařík, V., Schmid, M., Burch, N., Bowling, M., and Lisỳ, V. Rethinking formal models of partially observable multiagent decision making. *Artificial Intelligence*, 2022.

Lowe, R., WU, Y., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NIPS*, volume 30, pp. 6379–6390, 2017.

MacDermed, L. C. and Isbell, C. L. Point based value iteration with optimal belief compression for dec-pomdps. In *NIPS*, 2013.

Malik, D., Palaniappan, M., Fisac, J., Hadfield-Menell, D., Russell, S., and Dragan, A. An efficient, generalized Bellman update for cooperative inverse reinforcement learning. In *ICML*, 2018.

Nair, R., Tambe, M., Yokoo, M., Pynadath, D. V., and Marsella, S. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *International Joint conference on Artificial Intelligence (IJCAI)*, 2003.

Nair, R., Varakantham, P., Tambe, M., and Yokoo, M. Networked Distributed POMDPs: A Synthesis of Distributed Constraint Optimization and POMDPs. In *AAAI*, 2005.

Nayyar, A., Mahajan, A., and Teneketzis, D. Optimal control strategies in delayed sharing information structures. *IEEE Transactions on Automatic Control*, 2010.

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

Oliehoek, F. A. Sufficient plan-time statistics for decentralized pomdps. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Oliehoek, F. A., Spaan, M. T. J., Dibangoye, J. S., and Amato, C. Heuristic search for identical payoff bayesian games. In *AAMAS*, pp. 1115–1122, 2010.

Ooi, J. and Wornell, G. Decentralized control of a multiple access broadcast channel: performance bounds. In *CDC*, 1996.

Peshkin, L., Kim, K.-E., Meuleau, N., and Kaelbling, L. P. Learning to cooperate via policy search. *arXiv preprint cs/0105032*, 2001.

Pineau, J., Gordon, G., Thrun, S., et al. Point-based value iteration: An anytime algorithm for pomdps. In *IJCAI*, 2003.

Rabinovich, Z., Goldman, C. V., and Rosenschein, J. S. The complexity of multiagent systems: The price of silence. In *AAMAS*, 2003.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J. N., and Whiteson, S. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, 2018.

Shoham, Y. and Leyton-Brown, K. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Szer, D. and Charpillet, F. An optimal best-first search algorithm for solving infinite horizon dec-pomdps. In *ECML*, 2005.

Tan, M. Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents. In Huhns, M. N. and Singh, M. P. (eds.), *Readings in Agents*, pp. 487–494. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

Tsitsiklis, J. N. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1984.

Wang, M., Daamen, W., Hoogendoorn, S. P., and van Arem, B. Cooperative car-following control: Distributed algorithm and impact on moving jam features. *IEEE Transactions on Intelligent Transportation Systems*, 2015.

Xie, Y., Dibangoye, J., and Buffet, O. Optimally solving two-agent decentralized POMDPs under one-sided information sharing. In *ICML*, pp. 10473–10482, 2020.

A. The PBVI Algorithm

This section presents a pseudocode for the point-based value iteration algorithm to solve decentralized, partially observable Markov decision processes with hierarchical information sharing near-optimally.

Algorithm 1 PBVI for M' under HIS.

```
function PBVI()
  Initialize $\tilde{S}_{0:}$ and $\mathcal{V}_{0:}$.
  while $\mathcal{V}_{0:}$ has not converged do
    improve($\mathcal{V}_{0:}, \tilde{S}_{0:}$).
    $\tilde{S}_{0:} \leftarrow$ expand($\tilde{S}_{0:}$).
  end while
function improve($\mathcal{V}_{0:}, \tilde{S}_{0:}$)
  for $\tau = \ell - 1$ to 0 do
    for $s_\tau \in \tilde{S}_\tau$ do
      $\mathcal{V}_\tau \leftarrow \mathcal{V}_\tau \cup \{$backup$(s_\tau, \mathcal{V}_{\tau+1})\}$.
    end for
  end for
```

B. Proof of Theorem 2.3

This section shows that the nested-occupancy state is a sufficient statistic for the central planner to optimally solve perfect-information extensive-form game $\bar{G}^{\beta_\tau}_{s_\tau}$ for any player i.

Proof. The nested-occupancy state is a sufficiency statistic of the total data available to the planner at any player i for optimally solving $\bar{G}^{\beta_\tau}_{s_\tau}$, if it is sufficient to predict (1) the next nested-occupancy state and (2) the immediate reward. Let the total data available to the planner at player i be $\varsigma^i_\tau \doteq (s_\tau, o^i_\tau, u^{:i-1}_\tau)$. Let u^i_τ be the action chosen at player i after experiencing total data ς^i_τ. The nested-occupancy state $s^i_\tau \doteq (b^i_\tau, o^i_\tau, u^{:i-1}_\tau)$ summarizing ς^i_τ is sufficient to predict the next nested-occupancy state s^{i+1}_τ, if and only if the following holds $\Pr\{s^{i+1}_\tau \mid \varsigma^i_\tau, u^i_\tau\} = \Pr\{s^{i+1}_\tau \mid s^i_\tau, u^i_\tau\}$. To prove this property, we start with the definition of a nested-belief state b^i_τ associated with nested-occupancy state s^i_τ, *i.e.,*

$$\begin{aligned}
\Pr\{s^{i+1}_\tau \mid \varsigma^i_\tau, u^i_\tau\} &\doteq \Pr\{b^{i+1}_\tau, o^{i+1}_\tau, u^{:i,\bullet}_\tau \mid s_\tau, o^i_\tau, u^{:i,\circ}_\tau\}, && \text{(by definition of } \varsigma^i_\tau \text{ and } s^i_\tau\text{)}\\
&= \Pr\{b^{i+1}_\tau, o^{i+1}_\tau \mid s_\tau, o^i_\tau, u^{:i,\circ}_\tau\} \cdot \Pr\{u^{:i,\bullet}_\tau | s_\tau, o^i_\tau, u^{:i,\circ}_\tau\}, && \text{(by application of the Bayes rule)}\\
&= \Pr\{b^{i+1}_\tau, o^{i+1}_\tau \mid \varsigma^i_\tau\} \cdot \Pr\{u^{:i,\bullet}_\tau | u^{:i,\circ}_\tau\}, && \text{(by checking constraint } u^{:i,\bullet}_\tau = u^{:i,\circ}_\tau\text{)}\\
&= b^i_\tau(b^{i+1}_\tau, o^{i+1}_\tau) \cdot \mathbb{1}\{u^{:i,\circ}_\tau = u^{:i,\bullet}_\tau\}, && \text{(by definition of } b^i_\tau\text{).}
\end{aligned}$$

It will prove useful to define transition rule $\tilde{T}\colon (s^{i+1}_\tau | b^i_\tau, u^{:i}_\tau) \mapsto \Pr\{s^{i+1}_\tau | \, b^i_\tau, u^{:i}_\tau\}$, describing the probability to transitionning into nested-occupancy state s^{i+1}_τ upon talking action u^i_τ in nested-occupancy state s^i_τ. Notice that the transition does not depend on the current history o^i_τ, or it does so only through $(b^i_\tau, u^{:i}_\tau)$. Next, we show the sufficiency of nested-occupancy states to predict immediate rewards. Since rewards occur only at player n, we shall only consider nested occupancy states at that player. The nested-occupancy state $s^n_\tau \doteq (b^n_\tau, o^n_\tau, u^{:n-1}_\tau)$ summarizing ς^n_τ is sufficient to predict the immediate reward upon taking action u^n_τ, if and only if there exists a reward function $(s^n_\tau, u^n_\tau) \mapsto \tilde{R}(s^n_\tau, u^n_\tau)$ such that the following holds: $R(\varsigma^n_\tau, u^n_\tau) = \tilde{R}(s^n_\tau, u^n_\tau)$. To prove this statement, we start with the definition of $R(\varsigma^n_\tau, u^n_\tau)$, *i.e.,*

$$\begin{aligned}
R(\varsigma^n_\tau, u^n_\tau) &\doteq \mathbb{E}_{x \sim \Pr\{\cdot|\varsigma^n_\tau, u^n_\tau\}}\{\beta_\tau(x, o, u)\}, && \text{(by definition of } R(\varsigma^n_\tau, u^n_\tau)\text{)}\\
&= \mathbb{E}_{x \sim b^n_\tau(\cdot)}\{\beta_\tau(x, o, u)\}, && \text{(by definition of } b^n_\tau\text{).}
\end{aligned}$$

If we let $\tilde{R}(s^n_\tau, u^n_\tau) \doteq \mathbb{E}_{x \sim b^n_\tau(\cdot)}\{\beta_\tau(x, o, u)\}$, then the statement holds.

Nested-occupancy states describe a perfect-information extensive-form game $\tilde{G}^{\beta_\tau}_{s_\tau} \doteq \langle n, \tilde{\Sigma}, \tilde{\Psi}, \tilde{T}, \tilde{R}\rangle$ where nodes $\tilde{\Sigma}$ are nested-occupancy states, $\tilde{\Psi}\colon \tilde{\Sigma} \to 2^{(\cup^n_{i=1} U^i)}$ specifies the action set available to each nested-occupancy state, and $\tilde{T}$ and $\tilde{R}$ are already defined. Clearly, any optimal solution for perfect-information extensive-form game $\tilde{G}^{\beta_\tau}_{s_\tau}$ is also optimal for perfect-information extensive-form game $\bar{G}^{\beta_\tau}_{s_\tau}$. To prove this statement, we need to prove that the optimal action-value functions $\tilde{\beta}^{1:n,*}_\tau$ of $\tilde{G}^{\beta_\tau}_{s_\tau}$ such that at any player i and node ς^i_τ (resp. nested-occupancy state s^i_τ) and action u^i_τ the following equality holds: $\beta^{i,*}_\tau(\varsigma^i_\tau, u^i_\tau) = \tilde{\beta}^{i,*}_\tau(s^i_\tau, u^i_\tau)$. We prove the statement by induction. The statement trivially holds at player n since $R(\varsigma^n_\tau, u^n_\tau) = \tilde{R}(s^n_\tau, u^n_\tau)$, then $\beta^{n,*}_\tau(\varsigma^i_\tau, u^i_\tau) \doteq R(\varsigma^n_\tau, u^n_\tau) = \tilde{R}(s^n_\tau, u^n_\tau) \doteq \tilde{\beta}^{i,*}_\tau(s^i_\tau, u^i_\tau)$. Suppose the statement hold at player $i+1$ onward, *i.e.*, $\beta^{i+1,*}_\tau(\varsigma^{i+1}_\tau, u^{i+1}_\tau) = \tilde{\beta}^{i+1,*}_\tau(s^{i+1}_\tau, u^{i+1}_\tau)$. We are now ready to show it also hold at player i. We start with the expression of the optimal action-value function $\beta^{i,*}_\tau$, *i.e.*,

$$\begin{aligned}
\beta^{i,*}_\tau(\varsigma^i_\tau, u^i_\tau) &= \mathbb{E}_{\varsigma^{i+1}_\tau \sim T(\cdot|\varsigma^i_\tau, u^i_\tau)}\{\max\nolimits_{u^{i+1}_\tau} \beta^{i+1,*}_\tau(\varsigma^{i+1}_\tau, u^{i+1}_\tau)\}, && \text{(by Theorem 2.2)}\\
&= \mathbb{E}_{s^{i+1}_\tau \sim \Pr\{\cdot|\varsigma^i_\tau, u^i_\tau\}}\{\max\nolimits_{u^{i+1}_\tau} \tilde{\beta}^{i+1,*}_\tau(s^{i+1}_\tau, u^{i+1}_\tau)\}, && \text{(by induction hypothesis)}\\
&= \mathbb{E}_{s^{i+1}_\tau \sim \tilde{T}(\cdot|b^i_\tau, u^{:i}_\tau)}\{\max\nolimits_{u^{i+1}_\tau} \tilde{\beta}^{i+1,*}_\tau(s^{i+1}_\tau, u^{i+1}_\tau)\}, && \text{(by definition of } \tilde{T}(\cdot|b^i_\tau, u^{:i}_\tau))\\
&= \tilde{\beta}^{i,*}_\tau(s^i_\tau, u^i_\tau), && \text{(by definition of } \tilde{\beta}^{i,*}_\tau).
\end{aligned}$$

The statement holds for player i, thus for any arbitrary player. Consequently, one can use nested-occupancy states instead of total data available to the planner without compromising optimality, which ends the proof. □

C. Equivalence Relations

This section presents crucial properties necessary for grouping private histories that convey the same information about the game. To cluster two private histories of a player and reason similarly for the entire cluster, it is imperative to ensure that making identical immediate and future decisions for all members in the cluster does not compromise optimality. To do so, we need to specify how the information we rely on to make decisions evolves over stages. In particular, we need to exhibit rules for calculating the next-stage nested-occupancy state given the current one and decisions made at the current stage.

C.1. Predicting Next-Stage Observations

This subsection shows that the observation at stage $\tau+1$ and player i can be accurately predicted using only the nested occupancy states at stage τ and player i along with decision rules at stage τ and players i to n.

Lemma C.1. *Let $\varsigma^i_\tau \doteq \langle s_\tau, o^i_\tau, u^{:i-1}_\tau\rangle$ be total data available to the planner at stage τ and player i. Let $a^{i:}_\tau$ be the decision rules for player i to player n at stage τ. Let $s^i_\tau \doteq \langle b^i_\tau, o^i_\tau, u^{:i-1}_\tau\rangle$ be the nested-occupancy state at stage τ and player i summarizing total data ς^i_τ. The probability $\Pr\{z^i_{\tau+1}|\varsigma^i_\tau, a^{i:}_\tau\}$ that the planner receives observation $z^i_{\tau+1}$ on behalf of player i upon acting according to $\langle u^i_\tau, a^{i+1:}_\tau\rangle$ starting in total data ς^i_τ satisfies the following recursion:*

$$\Omega^i(z^i_{\tau+1}|b^i_\tau, u^{:i}_\tau, a^{i+1:}_\tau) = \sum_{s^{i+1}_\tau} \tilde{T}(s^{i+1}_\tau|b^i_\tau, u^{:i}_\tau) \sum_{z^{i+1}_{\tau+1}} \mathbb{1}\{z^i_{\tau+1} \sqsubseteq \zeta^{i+1}(z^{i+1}_{\tau+1})\} \cdot \Omega^{i+1}(z^{i+1}_{\tau+1}|b^{i+1}_\tau, u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau), \quad (1)$$

with boundary condition $\Omega^n(z^n_{\tau+1}|b^n_\tau, u^{:n}_\tau) \doteq \sum_x \sum_y b^n_\tau(x) \cdot p(y, z^n_{\tau+1}|x, u^{:n}_\tau)$.

Proof. Starting from conditional probability distribution $\Pr\{z^i_{\tau+1}|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\}$ and expanding over nested-occupancy states s^{i+1}_τ and observations $z^{i+1}_{\tau+1}$ of player $i+1$

$$\Pr\{z^i_{\tau+1}|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} = \sum_{s^{i+1}_\tau} \sum_{z^{i+1}_{\tau+1}} \Pr\{s^{i+1}_\tau, z^{i+1}_{\tau+1}, z^i_{\tau+1}|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\}.$$

The expansion of the joint probability into the product of conditional probabilities yields the following expression:

$$= \sum_{s^{i+1}_\tau} \sum_{z^{i+1}_{\tau+1}} \Pr\{z^i_{\tau+1}|s^{i+1}_\tau, z^{i+1}_{\tau+1}, \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} \cdot \Pr\{z^{i+1}_{\tau+1}|s^{i+1}_\tau, \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} \cdot \Pr\{s^{i+1}_\tau|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\}. \quad (2)$$

The first factor in (2) depends solely upon $z^{i+1}_{\tau+1}$ and not on the tuple $(s^{i+1}_\tau, \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau)$, *i.e.*,

$$\Pr\{z^i_{\tau+1}|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} = \sum_{s^{i+1}_\tau} \sum_{z^{i+1}_{\tau+1}} \Pr\{z^i_{\tau+1}|z^{i+1}_{\tau+1}\} \cdot \Pr\{z^{i+1}_{\tau+1}|s^{i+1}_\tau, \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} \cdot \Pr\{s^{i+1}_\tau|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\}. \quad (3)$$

The last factor in (3) depends solely upon $\langle b^i_\tau, u^{:i}_\tau\rangle$ and not on tuple $\langle \varsigma^i_\tau, a^{i+1:}_\tau\rangle$, which becomes after re-arranging terms:

$$\Pr\{z^i_{\tau+1}|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} = \sum_{s^{i+1}_\tau} \tilde{T}(s^{i+1}_\tau|b^i_\tau, u^{:i}_\tau) \sum_{z^{i+1}_{\tau+1}} \Pr\{z^i_{\tau+1}|z^{i+1}_{\tau+1}\} \cdot \Pr\{z^{i+1}_{\tau+1}|s^{i+1}_\tau, \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\}. \tag{4}$$

Equation (4) makes it possible to prove the statement, Equation (1), recursively. We start at player $n-1$, *i.e.,*

$$\Pr\{z^{n-1}_{\tau+1}|\varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau\} = \sum_{s^n_\tau} \tilde{T}(s^n_\tau|b^{n-1}_\tau, u^{:n-1}_\tau) \sum_{z^n_{\tau+1}} \Pr\{z^{n-1}_{\tau+1}|z^n_{\tau+1}\} \cdot \Pr\{z^n_{\tau+1}|s^n_\tau, \varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau\}. \tag{5}$$

The boundary condition gives the last factor in (5), *i.e.,* for nested-occupancy state $s^n_\tau \doteq (b^n_\tau, o^n_\tau, u^{:n-1}_\tau)$,

$$\Pr\{z^{n-1}_{\tau+1}|\varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau\} = \sum_{s^n_\tau} \tilde{T}(s^n_\tau|b^{n-1}_\tau, u^{:n-1}_\tau) \sum_{z^n_{\tau+1}} \mathbb{1}\{z^{n-1}_{\tau+1} \sqsubseteq z^n_{\tau+1}\} \cdot \Omega^n(z^n_{\tau+1}|b^n_\tau, \langle u^{:n-1}_\tau, a^n_\tau(o^n_\tau)\rangle). \tag{6}$$

Let $\Omega^{n-1}\colon (z^{n-1}_{\tau+1}|b^{n-1}_\tau, u^{:n-1}_\tau, a^n_\tau) \mapsto \sum_{s^n_\tau} \tilde{T}(s^n_\tau|b^{n-1}_\tau, u^{:n-1}_\tau) \sum_{z^n_{\tau+1}} \mathbb{1}\{z^{n-1}_{\tau+1} \sqsubseteq \zeta^n(z^n_{\tau+1})\} \cdot \Omega^n(z^n_{\tau+1}|b^n_\tau, u^{:n-1}_\tau, a^n_\tau(o^n_\tau))$ be the observation model for predicting next observation at stage τ and player $n-1$. Then, statement (1) holds at stage τ and player $n-1$. Suppose the statement holds for any arbitrary stage τ and player $i+1$. We are ready to prove it also holds at stage τ and player i. Starting at (4), the application of the induction hypothesis yields:

$$\Pr\{z^i_{\tau+1}|\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} = \sum_{s^{i+1}_\tau} \tilde{T}(s^{i+1}_\tau|b^i_\tau, u^{:i}_\tau) \sum_{z^{i+1}_{\tau+1}} \mathbb{1}\{z^i_{\tau+1} \sqsubseteq \zeta^{i+1}(z^{i+1}_{\tau+1})\} \cdot \Omega^{i+1}(z^{i+1}_{\tau+1}|b^{i+1}_\tau, u^{:i}_\tau, a^{i+1:}_\tau). \tag{7}$$

If we let $\Omega^i\colon (z^i_{\tau+1}|b^i_\tau, u^{:i}_\tau, a^{i+1:}_\tau) \mapsto \sum_{s^{i+1}_\tau} \tilde{T}(s^{i+1}_\tau|b^i_\tau, u^{:i}_\tau) \sum_{z^{i+1}_{\tau+1}} \mathbb{1}\{z^i_{\tau+1} \sqsubseteq \zeta^{i+1}(z^{i+1}_{\tau+1})\} \cdot \Omega^{i+1}(z^{i+1}_{\tau+1}|b^{i+1}_\tau, u^{:i}_\tau, a^{i+1:}_\tau)$ be the observation model for predicting next observation at stage τ and player i, then statement (1) holds at stage τ and player i. Thus, the statement holds for any stage and player, which ends the proof. □

C.2. Predicting Next-Stage Nested-Occupancy States

This subsection proves the nested-occupancy states describe a Markovian process, *i.e.,* the next-stage nested-occupancy state depends only upon the current one. Notice that nested-occupancy states have three components. Only the nested belief states are nonobservable and need to be estimated. If we know how to estimate the nested belief state, we can add the history and actions of subordinates, thereby constructing a nested-occupancy state.

Lemma C.2. *Let $\varsigma^i_\tau \doteq \langle s_\tau, o^i_\tau, u^{:i-1}_\tau\rangle$ be total data available to the planner at stage τ and player i. Let $a^{i:}_\tau$ be the decision rules for player i to player n at stage τ. Let $s^i_\tau \doteq \langle b^i_\tau, o^i_\tau, u^{:i-1}_\tau\rangle$ be the nested-occupancy state at stage τ and player i summarizing total data ς^i_τ. The next-stage nested-belief state $b^i_{\tau+1} \doteq T^i(b^i_\tau, \langle u^i_\tau, a^{i+1:}_\tau\rangle, z^i_{\tau+1})$, upon acting according to $\langle u^i_\tau, a^{i+1:}_\tau\rangle$ in nested-occupancy state s^i_τ and receiving observation $z^i_{\tau+1}$, satisfies the following recursion: for any history and nested-belief tuple $(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1})$,*

$$b^i_{\tau+1}(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}) \propto \textstyle\sum_{s^{i+1}_\tau \doteq (b^{i+1}_\tau, o^{i+1}_\tau, u^{:i}_\tau)} \tilde{T}(s^{i+1}_\tau\,|b^i_\tau, u^{:i}_\tau) \sum_{z^{i+1}_{\tau+1}} \mathbb{1}\{z^i_{\tau+1} \sqsubseteq \zeta^{i+1}(z^{i+1}_{\tau+1})\} \cdot \delta^{o^{i+1}_{\tau+1}}_{\langle o^{i+1}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), z^{i+1}_{\tau+1}\rangle}$$
$$\delta^{b^{i+1}_{\tau+1}}_{T^{i+1}(b^{i+1}_\tau, \langle u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau\rangle, z^{i+1}_{\tau+1})} \cdot \Omega^{i+1}(z^{i+1}_{\tau+1}|b^{i+1}_\tau, \langle u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau\rangle).$$

with boundary condition $b^n_{\tau+1} \doteq T^n(b^n_\tau, u^{:n}_\tau, z^n_{\tau+1})$ where $b^n_{\tau+1}(y) \propto \sum_x b^n_\tau(x) \cdot p(y, z^n_{\tau+1}|x, u^{:n}_\tau)$ for any hidden state y.

Proof. The proof proceeds by induction. Starting with player $n-1$, we define the nested-belief state $b^{n-1}_{\tau+1}$ at stage $\tau+1$ and player $n-1$ upon acting according to $\langle u^{n-1}_\tau, a^n_\tau\rangle$ in total data available to the planner $\varsigma^{n-1}_\tau \doteq \langle s_\tau, o^{n-1}_\tau, u^{:n-2}_\tau\rangle$ and receiving observation $z^{n-1}_{\tau+1}$, as follows: for any history and nested-belief tuple $(o^n_{\tau+1}, b^n_{\tau+1})$,

$$b^{n-1}_{\tau+1}(o^n_{\tau+1}, b^n_{\tau+1}) \doteq \Pr\{o^n_{\tau+1}, b^n_{\tau+1}|\varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau, z^{n-1}_{\tau+1}\}. \tag{8}$$

The expansion of (8) over nested-occupancy states s^n_τ at stage τ and player n and histories $z^n_{\tau+1}$ at stage $\tau+1$ and player n, result in the following expression:

$$b^{n-1}_{\tau+1}(o^n_{\tau+1}, b^n_{\tau+1}) = \sum_{s^n_\tau \doteq (b^n_\tau, o^n_\tau, u^{:n-1}_\tau)} \sum_{z^n_{\tau+1}} \Pr\{s^n_\tau, z^n_{\tau+1}, o^n_{\tau+1}, b^n_{\tau+1}|\varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau, z^{n-1}_{\tau+1}\}. \tag{9}$$

The application of Bayes' rule in expression (9) yields the following expression:

$$b^{n-1}_{\tau+1}(o^n_{\tau+1}, b^n_{\tau+1}) = \sum_{s^n_\tau \doteq (b^n_\tau, o^n_\tau, u^{:n-1}_\tau)} \sum_{z^n_{\tau+1}} \frac{\Pr\{s^n_\tau, z^n_{\tau+1}, o^n_{\tau+1}, b^n_{\tau+1}, \varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau, z^{n-1}_{\tau+1}\}}{\Pr\{\varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau, z^{n-1}_{\tau+1}\}}. \tag{10}$$

The expansion of the joint probability into the product of conditional probabilities and the application of Lemma C.1 yield the following expression:

$$\begin{aligned} b^{n-1}_{\tau+1}(o^n_{\tau+1}, b^n_{\tau+1}) = {} & \textstyle\sum_{s^n_\tau \doteq (b^n_\tau, o^n_\tau, u^{:n-1}_\tau)} \sum_{z^n_{\tau+1}} \Pr\{z^{n-1}_{\tau+1} | z^n_{\tau+1}\} \cdot \Pr\{o^n_{\tau+1} | o^n_\tau, a^n_\tau(o^n_\tau), z^n_{\tau+1}\} \cdot \Pr\{b^n_{\tau+1} | b^n_\tau, \langle u^{:n-1}_\tau, a^n_\tau(o^n_\tau)\rangle, z^n_{\tau+1}\} \cdot \\ & \Pr\{z^n_{\tau+1} | b^n_\tau, \langle u^{:n-1}_\tau, a^n_\tau(o^n_\tau)\rangle\} \cdot \Pr\{s^n_\tau\ | \varsigma^{n-1}_\tau, u^{n-1}_\tau\} / \Pr\{z^{n-1}_{\tau+1} | \varsigma^{n-1}_\tau, u^{n-1}_\tau, a^n_\tau\} \end{aligned} \tag{11}$$

Using the boundary condition, we obtain the following expression, *i.e.,*

$$\begin{aligned} b^{n-1}_{\tau+1}(o^n_{\tau+1}, b^n_{\tau+1}) \propto {} & \textstyle\sum_{s^n_\tau \doteq (b^n_\tau, o^n_\tau, u^{:n-1}_\tau)} \sum_{z^n_{\tau+1}} \mathbb{1}\{z^{n-1}_{\tau+1} \sqsubseteq \zeta^n(z^n_{\tau+1})\} \cdot \delta^{o^n_{\tau+1}}_{\langle o^n_\tau, a^n_\tau(o^n_\tau), z^n_{\tau+1}\rangle} \\ & \delta^{b^n_{\tau+1}}_{T^n(b^n_\tau, \langle u^{:n-1}_\tau, a^n_\tau(o^n_\tau)\rangle, z^n_{\tau+1})} \cdot \Omega^n(z^n_{\tau+1} | b^n_\tau, \langle u^{:n-1}_\tau, a^n_\tau(o^n_\tau)\rangle) \cdot \tilde{T}(s^n_\tau\ | b^{n-1}_\tau, u^{:n-1}_\tau) \end{aligned} \tag{12}$$

Hence, the statement holds at stage τ and player $n-1$. Suppose it holds at stage τ and player $i+1$. We are now ready to show the statement also holds at stage τ and player i. We start with the definition the nested-belief state $b^i_{\tau+1}$ at stage $\tau+1$ and player i upon acting according to $\langle u^i_\tau, a^{i+1:}_\tau\rangle$ in total data available to the planner $\varsigma^i_\tau \doteq \langle s_\tau, o^i_\tau, u^{:i-1}_\tau\rangle$ and receiving observation $z^i_{\tau+1}$. The proof proceeds similarly to that of player $n-1$, *i.e.,*

$$b^i_{\tau+1}(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}) \doteq \Pr\{o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1} | \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau, z^i_{\tau+1}\}. \tag{13}$$

The expansion of (13) over nested-occupancy states s^{i+1}_τ at stage τ and player $i+1$ and histories $z^{i+1}_{\tau+1}$ at stage $\tau+1$ and player $i+1$, result in the following expression:

$$b^i_{\tau+1}(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}) = \textstyle\sum_{s^{i+1}_\tau \doteq (b^{i+1}_\tau, o^{i+1}_\tau, u^{:i}_\tau)} \sum_{z^{i+1}_{\tau+1}} \Pr\{s^{i+1}_\tau, z^{i+1}_{\tau+1}, o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1} | \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau, z^i_{\tau+1}\}. \tag{14}$$

The application of Bayes' rule in expression (14) yields the following expression: for any pairs $(o^{1:i+1}_{\tau+1}, b^{i+1}_{\tau+1})$,

$$b^i_{\tau+1}(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}) = \sum_{s^{i+1}_\tau \doteq (b^{i+1}_\tau, o^{i+1}_\tau, u^{:i}_\tau)} \sum_{z^{i+1}_{\tau+1}} \Pr\{s^{i+1}_\tau, z^{i+1}_{\tau+1}, o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}, \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau, z^i_{\tau+1}\} / \Pr\{\varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau, z^i_{\tau+1}\}. \tag{15}$$

The expansion of the joint probability into the product of conditional probabilities and the application of Lemma C.1 yield the following expression:

$$\begin{aligned} b^i_{\tau+1}(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}) = {} & \textstyle\sum_{s^{i+1}_\tau \doteq (b^{i+1}_\tau, o^{i+1}_\tau, u^{:i}_\tau)} \sum_{z^{i+1}_{\tau+1}} \Pr\{z^i_{\tau+1} | z^{i+1}_{\tau+1}\} \cdot \Pr\{o^{i+1}_{\tau+1} | o^{i+1}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), z^{i+1}_{\tau+1}\} \cdot \\ & \Pr\{b^{i+1}_{\tau+1} | b^{i+1}_\tau, \langle u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau\rangle, z^{i+1}_{\tau+1}\} \cdot \Pr\{z^{i+1}_{\tau+1} | b^{i+1}_\tau, \langle u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau\rangle\} \cdot \\ & \Pr\{s^{i+1}_\tau\ | \varsigma^i_\tau, u^i_\tau\} / \Pr\{z^i_{\tau+1} | \varsigma^i_\tau, u^i_\tau, a^{i+1:}_\tau\} \end{aligned} \tag{16}$$

Using the boundary condition and the induction hypothesis, we obtain the following expression, *i.e.,*

$$\begin{aligned} b^i_{\tau+1}(o^{i+1}_{\tau+1}, b^{i+1}_{\tau+1}) \propto {} & \textstyle\sum_{s^{i+1}_\tau \doteq (b^{i+1}_\tau, o^{i+1}_\tau, u^{:i}_\tau)} \tilde{T}(s^{i+1}_\tau\ | b^i_\tau, u^{:i}_\tau) \sum_{z^{i+1}_{\tau+1}} \mathbb{1}\{z^i_{\tau+1} \sqsubseteq \zeta^{i+1}(z^{i+1}_{\tau+1})\} \cdot \delta^{o^{i+1}_{\tau+1}}_{\langle o^{i+1}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), z^{i+1}_{\tau+1}\rangle} \\ & \delta^{b^{i+1}_{\tau+1}}_{T^{i+1}(b^{i+1}_\tau, \langle u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau\rangle, z^{i+1}_{\tau+1})} \cdot \Omega^{i+1}(z^{i+1}_{\tau+1} | b^{i+1}_\tau, \langle u^{:i}_\tau, a^{i+1}_\tau(o^{i+1}_\tau), a^{i+2:}_\tau\rangle). \end{aligned} \tag{17}$$

Hence, the statement holds at any stage τ and player i, which ends the proof. □

C.3. Nested Belief States, Policies and Action-Value Functions At Player n

This section establishes many important properties regarding player n. First, it establishes that for any given stage and player n, the planner can make decisions based on belief states instead of histories. To prove this statement, one must demonstrate that belief states are capable of predicting (1) the next observation for the subsequent stage and player n; (2) the next belief state for the subsequent stage and player n; and (3) the immediate reward. Next, it shows that belief-dependent policies are optimal at player n. Finally, it describes the action-value functions under a history-dependent policy of player 1 to $n-1$ and a belief-dependent policy of player n.

Lemma C.3. *Let $a_\tau^{:n-1}$ be a joint decision rule of player 1 to $n-1$. Let b_τ^n be a nested belief state, $o_\tau^{:n-1}$ be a joint history of player 1 to $n-1$, and $a_\tau^{:n-1}(o_\tau^{:n-1})$ be a joint action of player 1 to $n-1$ when following joint policy $a_\tau^{:n-1}$. Let $s_\tau^n \doteq (b_\tau^n, o_\tau^n, a_\tau^{:n-1}(o_\tau^{:n-1}))$ be a nested-occupancy state at stage τ and player n. Then, the following propositions hold for any nested-occupancy state s_τ^n at stage τ and player n.*

1. *For any action u_τ^n and observation $z_{\tau+1}^n$, we have $\Pr\{z_{\tau+1}^n | s_\tau^n, u_\tau^n\} = \Omega^n(z_{\tau+1}^n | b_\tau^n, \langle u_\tau^n, a_\tau^{:n-1}(o_\tau^{:n-1})\rangle)$.*

2. *For any action u_τ^n and observation $z_{\tau+1}^n$, we have $\Pr\{b_{\tau+1}^n | s_\tau^n, u_\tau^n, z_{\tau+1}^n\} = \delta_{T^n(b_\tau^n, \langle u_\tau^n, a_\tau^{:n-1}(o_\tau^{:n-1})\rangle, z_{\tau+1}^n)}^{b_{\tau+1}^n}$.*

3. *For any action u_τ^n, we have $\mathbb{E}_{(x_\tau, u_\tau)\sim \Pr\{\cdot\ | s_\tau^n, u_\tau^n\}}\{r(x_\tau, u_\tau)\} = \mathbb{E}_{(x_\tau, u_\tau)\sim \Pr\{\cdot\ | b_\tau^n, u_\tau^n, a_\tau^{:n-1}(o_\tau^{:n-1})\}}\{r(x_\tau, u_\tau)\}$.*

Proof. The two first propositions hold directly from Lemmas C.1 and C.2. The last proposition holds because the immediate rewards depend on the histories of player n only through the corresponding belief states. Which ends the proof. □

Lemma C.4. *The optimal policy of player n depends only upon the belief state not on histories.*

Proof. The proof proceeds by induction. Let $a_{0:}^{:n-1}$ be the joint policy of player 1 to $n-1$. The best-response decision rule of player n at stage $\ell-1$ is written as follows: for any history $o_{\ell-1}^n$,

$$\begin{aligned} a_{\ell-1}^n(o_{\ell-1}^n) &\in \operatorname{argmax}_{u_{\ell-1}^n}\ \mathbb{E}_{(x_{\ell-1}, u_{\ell-1})\sim \Pr\{\cdot\ | b_{\ell-1}^n, o_{\ell-1}^n, u_{\ell-1}^n, a_{\ell-1}^{:n-1}(o_{\ell-1}^{:n-1})\}}\{r(x_{\ell-1}, u_{\ell-1})\} && \text{(by Definition)} \\ &\in \operatorname{argmax}_{u_{\ell-1}^n}\ \mathbb{E}_{(x_{\ell-1}, u_{\ell-1})\sim \Pr\{\cdot\ | b_{\ell-1}^n, u_{\ell-1}^n, a_{\ell-1}^{:n-1}(o_{\ell-1}^{:n-1})\}}\{r(x_{\ell-1}, u_{\ell-1})\}. && \text{(by Lemma C.3)} \end{aligned}$$

The statement holds at stage $\ell-1$. Define the value function $\tilde{\alpha}_{\ell-1}^n$ under the joint policy $a_{0:}^{:n-1}$ of player 1 to $n-1$,

$$\bar{\alpha}_{\ell-1}^n \colon (b_{\ell-1}^n, o_{\ell-1}^{:n-1}) \mapsto \max_{u_{\ell-1}^n}\ \mathbb{E}_{(x_{\ell-1}, u_{\ell-1})\sim \Pr\{\cdot\ | b_{\ell-1}^n, u_{\ell-1}^n, a_{\ell-1}^{:n-1}(o_{\ell-1}^{:n-1})\}}\{r(x_{\ell-1}, u_{\ell-1})\}.$$

Define the value function $\bar{\beta}_{\ell-2}^n$ under the joint policy $a_{0:}^{:n-1}$ of player 1 to $n-1$,

$$\bar{\beta}_{\ell-2}^n \colon (b_{\ell-2}^n, o_{\ell-2}^{:n-1}, u_{\ell-2}) \mapsto \mathbb{E}_{(x_{\ell-2}, b_{\ell-1}^n, o_{\ell-1}^{:n-1})\sim \Pr\{\cdot\ | b_{\ell-2}^n, u_{\ell-2}, o_{\ell-2}^{:n-1}\}}\{r(x_{\ell-2}, u_{\ell-2}) + \gamma \bar{\alpha}_{\ell-1}^n(b_{\ell-1}^n, o_{\ell-1}^{:n-1})\}.$$

The best-response decision rule of player n at stage $\ell-2$ is written as follows: for any history $o_{\ell-2}^n$,

$$a_{\ell-2}^n(o_{\ell-2}^n) \in \operatorname{argmax}_{u_{\ell-2}^n}\ \bar{\beta}_{\ell-2}^n(b_{\ell-2}^n, o_{\ell-2}^{:n-1}, \langle u_{\ell-2}^n, a_{\ell-2}^{:n-1}(o_{\ell-2}^{:n-1})\rangle).$$

Consequently, the statement holds for stages $\ell-1$ and $\ell-2$. Suppose the statement holds for stage $\tau+1$, that is there exists an action-value function $\bar{\beta}_\tau^n$ under the joint policy $a_{0:}^{:n-1}$ of player 1 to $n-1$,

$$\bar{\beta}_\tau^n \colon (b_\tau^n, o_\tau^{:n-1}, u_\tau) \mapsto \mathbb{E}_{(x_\tau, b_{\tau+1}^n, o_{\tau+1}^{:n-1})\sim \Pr\{\cdot\ | b_\tau^n, u_\tau, o_\tau^{:n-1}\}}\{r(x_\tau, u_\tau) + \gamma \bar{\alpha}_{\tau+1}^n(b_{\tau+1}^n, o_{\tau+1}^{:n-1})\}.$$

We are now ready to show the statement also holds at stage τ. The best-response decision rule of player n at stage τ is written as follows: for any history o_τ^n,

$$a_\tau^n(o_\tau^n) \in \operatorname{argmax}_{u_\tau^n}\ \bar{\beta}_\tau^n(b_\tau^n, o_\tau^{:n-1}, \langle u_\tau^n, a_\tau^{:n-1}(o_\tau^{:n-1})\rangle).$$

This proves the statement holds for stage τ, ending the proof. □

Lemma C.4 shows that the HIS assumption allows player n to act based solely upon belief states instead of histories optimally. In other words, a belief-dependent policy exists as good or better than any history-dependent policy of player n. The subsequent lemma shows how the use of belief-dependent policies for player n affects the description of the action-value function under a joint history-dependent policy of player 1 to $n-1$ and a belief-dependent policy for player n.

Lemma C.5. *Let $a_{0:}^{:n-1}$ be the joint history-dependent policy of player 1 to $n-1$, $\tilde{a}_{0:}^{n}$ be the belief-dependent policy of player n. The action-value function under joint policy $\langle a_{0:}^{:n-1}, \tilde{a}_{0:}^{n} \rangle$ is given as follows:*

$$\bar{\beta}_\tau^n \colon (b_\tau^n, o_\tau^{:n-1}, u_\tau^{:n}) \mapsto \mathbb{E}_{(x_\tau, b_{\tau+1}^n, o_{\tau+1}^{:n-1}) \sim \Pr\{\cdot\ |b_\tau^n, u_\tau, o_\tau^{:n-1}\}}\{r(x_\tau, u_\tau) + \gamma \bar{\beta}_{\tau+1}^n (b_{\tau+1}^n, o_{\tau+1}^{:n-1}, \langle a_{\tau+1}^{:n-1}(o_{\tau+1}^{:n-1}), \tilde{a}_{\tau+1}^n (b_{\tau+1}^n)\rangle)\}$$

with boundary condition $\bar{\beta}_\ell^n(\cdot,\cdot,\cdot) \doteq 0$.

Proof. The proof follows directly from the proof of Lemma C.4. □

C.4. Proof of Theorem 2.4

Proof. We shall treat each proposition separately.

Statement 1. We prove the first statement by induction. We begin the proof by demonstrating that the statement holds at player n. Let $s_\tau^{n,\circ} \doteq (b_\tau^{n,\circ}, o_\tau^{n,\circ}, u_\tau^{:n-1,\circ})$ and $s_\tau^{n,\bullet} \doteq (b_\tau^{n,\bullet}, o_\tau^{n,\bullet}, u_\tau^{:n-1,\bullet})$ be two nested-occupancy states. Suppose $s_\tau^{n,\circ} \sim_{\mathscr{B}_1} s_\tau^{n,\bullet}$, that is $(b_\tau^{n,\circ}, u_\tau^{:n-1,\circ}) = (b_\tau^{n,\bullet}, u_\tau^{:n-1,\bullet})$. Consider the action-value function $\tilde{\beta}_\tau^{n,*}$ at stage τ, player n, nested-occupancy state $s_\tau^{n,\circ}$ and action u_τ^n, *i.e.,*

$$\begin{aligned}\tilde{\beta}_\tau^{n,*}(s_\tau^{n,\circ}, u_\tau^n) &\doteq \mathbb{E}_{x\sim b_\tau^{n,\circ}(\cdot)}\{\beta_\tau(x, o_\tau^{n,\circ}, \langle u_\tau^{:n-1,\circ}, u_\tau^n\rangle)\}\\ &= \bar{\beta}_\tau^n(b_\tau^{n,\circ}, o_\tau^{:n-1,\circ}, \langle u_\tau^{:n-1,\circ}, u_\tau^n\rangle) \qquad \text{(by Lemma C.5).}\end{aligned}$$

Since the action-value function $\bar{\beta}_\tau^n$ depends on the nested-occupancy state only through the belief state $b_\tau^{n,\circ}$, joint history $o_\tau^{:n-1,\circ}$, and joint action $\langle u_\tau^{:n-1,\circ}, u_\tau^n\rangle$, not upon joint history $o_\tau^{n,\circ}$, thus does the action-value function $\tilde{\beta}_\tau^{n,*}$. Hence, the first statement holds at player n. Suppose the statement holds for player $i+1$. We are now ready to show it also holds at player i. We start with the expression of optimal action-value $\tilde{\beta}_\tau^{i,*}(s_\tau^{i,\circ}, u_\tau^i)$ for nested-occupancy state $s_\tau^{i,\circ}$ and action u_τ^i, *i.e.,*

$$\tilde{\beta}_\tau^{i,*}(s_\tau^{i,\circ}, u_\tau^i) = \mathbb{E}_{s_\tau^{i+1,\circ}\sim \tilde{T}(\cdot|b_\tau^{i,\circ}, \langle u_\tau^i, u_\tau^{:i-1,\circ}\rangle)}\{\max\nolimits_{u_\tau^{i+1}} \tilde{\beta}_\tau^{i+1,*}(s_\tau^{i+1,\circ}, u_\tau^{i+1})\}.$$

An inspection of the transition function $\tilde{T}(\cdot|b_\tau^{i,\circ}, u_\tau^{:i})$ reveals that it depends on nested-occupancy state $s_\tau^{i,\circ}$ only though nested-belief state $b_\tau^{i,\circ}$ and joint action $u_\tau^{:i}$. Consequently, if we let $s_\tau^{i,\circ} \sim_{\mathscr{B}_1} s_\tau^{i,\bullet}$ then we know $(b_\tau^{i,\circ}, u_\tau^{:i,\circ}) = (b_\tau^{i,\bullet}, u_\tau^{:i,\bullet})$, which leads to the statement:

$$\begin{aligned}\tilde{\beta}_\tau^{i,*}(s_\tau^{i,\circ}, u_\tau^i) &= \mathbb{E}_{s_\tau^{i+1,\bullet}\sim \tilde{T}(\cdot|b_\tau^{i,\bullet}, \langle u_\tau^i, u_\tau^{:i-1,\bullet}\rangle)}\{\max\nolimits_{u_\tau^{i+1}} \tilde{\beta}_\tau^{i+1,*}(s_\tau^{i+1,\bullet}, u_\tau^{i+1})\}\\ &= \tilde{\beta}_\tau^{i,*}(s_\tau^{i,\bullet}, u_\tau^i).\end{aligned}$$

This expression proves the first statement at any stage τ and player i.

Statement 2. To prove the second statement, we build upon the first statement. If we let $s_\tau^{i,\circ} \sim_{\mathscr{B}_2} s_\tau^{i,\bullet}$, then for any arbitrary joint action $u_\tau^{:i-1}$ we know that $(b_\tau^{i,\circ}, o_\tau^{i,\circ}, u_\tau^{:i-1}) \sim_{\mathscr{B}_1} (b_\tau^{i,\bullet}, o_\tau^{i,\bullet}, u_\tau^{:i-1})$. If $s_\tau^{i,\circ} \sim_{\mathscr{B}_2} s_\tau^{i,\bullet}$, we know that histories of subordinates are identical $o_\tau^{:i-1,\circ} = o_\tau^{:i-1,\bullet}$, then the following holds $u_\tau^{:i-1} = a_\tau^{:i-1,*}(o_\tau^{:i-1,\bullet}) = a_\tau^{:i-1,*}(o_\tau^{:i-1,\circ})$. Consequently, by the application of the first statement, we have for any arbitrary action u_τ^i,

$$\tilde{\beta}_\tau^{i,*}(\langle b_\tau^{i,\circ}, o_\tau^{i,\circ}, a_\tau^{:i-1,*}(o_\tau^{:i-1,\circ})\rangle, u_\tau^i) = \tilde{\beta}_\tau^{i,*}(\langle b_\tau^{i,\bullet}, o_\tau^{i,\bullet}, a_\tau^{:i-1,*}(o_\tau^{:i-1,\bullet})\rangle, u_\tau^i).$$

Consequently, the sets of optimal actions $A_\tau^{i,*}(o_\tau^{i,\circ}) \doteq \operatorname{argmax}_{u_\tau^i} \tilde{\beta}_\tau^{i,*}(\langle b_\tau^{i,\circ}, o_\tau^{i,\circ}, a_\tau^{:i-1,*}(o_\tau^{:i-1,\circ})\rangle, u_\tau^i)$ and $A_\tau^{i,*}(o_\tau^{i,\bullet}) \doteq \operatorname{argmax}_{u_\tau^i} \tilde{\beta}_\tau^{i,*}(\langle b_\tau^{i,\bullet}, o_\tau^{i,\bullet}, a_\tau^{:i-1,*}(o_\tau^{:i-1,\bullet})\rangle, u_\tau^i)$ at histories $o_\tau^{i,\circ}$ and $o_\tau^{i,\bullet}$, respectively, are equivalent, *i.e.,* $A_\tau^{i,*}(o_\tau^{i,\circ}) = A_\tau^{i,*}(o_\tau^{i,\bullet})$. Since $a_\tau^{i,*}(o_\tau^{i,\bullet})$ and $a_\tau^{i,*}(o_\tau^{i,\circ})$ belong to the same set $A_\tau^{i,*}(o_\tau^{i,\circ}) = A_\tau^{i,*}(o_\tau^{i,\bullet})$, they are interchangeable. In other words, the optimal action for history $o_\tau^{i,\circ}$ is also optimal for history $o_\tau^{i,\bullet}$ and vice versa. Interestingly, one can show that the expansions $\langle o_\tau^{i,\circ}, u_\tau^i, z_{\tau+1}^i\rangle$ and $\langle o_\tau^{i,\bullet}, u_\tau^i, z_{\tau+1}^i\rangle$ of histories $o_\tau^{i,\circ}$ and $o_\tau^{i,\bullet}$ upon taking the same action u_τ^i and receiving the same observation $z_{\tau+1}^i$, respectively, will also have equivalent optimal actions. Hence, essentially providing that the optimal policy for history $o_\tau^{i,\circ}$ is also optimal for history $o_\tau^{i,\bullet}$ and vice versa. To show this statement, first notice that both histories $\langle o_\tau^{i,\circ}, u_\tau^i, z_{\tau+1}^i \rangle$ and $\langle o_\tau^{i,\bullet}, u_\tau^i, z_{\tau+1}^i \rangle$ will have the same histories of subordinates because original histories $o_\tau^{i,\circ}$ and $o_\tau^{i,\bullet}$ had the same histories of subordinates and original histories $o_\tau^{i,\circ}$ and $o_\tau^{i,\bullet}$ were expanded using the same action and observation. Next, we need to show that the nested-belief states associated with the expanded histories $\langle o_\tau^{i,\circ}, u_\tau^i, z_{\tau+1}^i \rangle$ and $\langle o_\tau^{i,\bullet}, u_\tau^i, z_{\tau+1}^i \rangle$ are also equivalent. The proof of this statement follows directly from the fact that the transition function from one stage to the next one depends on $\tilde{T}$, $T^{\cdot}$ and $\Omega^{\cdot}$. A careful inspection of these functions reveals that they depend on nested-occupancy states at player i only through nested-belief states at player i, joint histories of superiors of player i, and actions of subordinates as demonstrated in Lemmas C.2 and C.1. Histories of the current player are only used to select the action for that player. However, our histories of interest have the same optimal action set. So, assuming these histories take the same action does not hurt. Consequently, if we let $s_\tau^{i,\circ} \sim_{\mathscr{B}_2} s_\tau^{i,\bullet}$ then we know that $o_\tau^{i,\circ} \sim_{\mathscr{P}} o_\tau^{i,\bullet}$, which ends the proof. Which ends the proof for both propositions. □

D. Proof of Theorem 3.1

Proof. For simplicity, throughout the proof, we assume with no loss of generality that the central planner does not rely on public observations, so transition function $\boldsymbol{T}$ is deterministic. Let $a_{0:}^*$ be an optimal joint policy with value functions $\upsilon_{0:}^*$. Let s_τ^* be the occupancy state generated under joint policy $a_{0:}^*$, with boundary condition $s_0^* \doteq s_0$. Let $\upsilon_{0:}$ be the value function that the PBVI algorithm produced over occupancy subsets $\tilde{S}_{0:}$. Then, it follows that:

$$\begin{aligned}\upsilon_0^*(s_0^*) - \upsilon_0(s_0^*) &= (\textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \boldsymbol{R}(s_\tau^*, a_\tau^*)) - \upsilon_0(s_0^*), \quad \text{(definition of } \upsilon_0^*(s_0^*)\text{)}\\ &= (\textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \boldsymbol{R}(s_\tau^*, a_\tau^*)) - \sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot (\upsilon_\tau(s_\tau^*) - \upsilon_\tau(s_\tau^*)) - \upsilon_0(s_0^*), \quad \text{(adding zero).}\end{aligned}$$

Next, we use the fact that $\upsilon_\ell(\cdot) \doteq 0$ to re-arrange terms:

$$\begin{aligned}&= \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \boldsymbol{R}(s_\tau^*, a_\tau^*) + \left(\gamma^\ell \cdot \upsilon_\ell(s_\ell^*) + \sum_{\tau=1}^{\ell-1} \gamma^\tau \cdot \upsilon_\tau(s_\tau^*)\right) - \left(\gamma^0 \cdot \upsilon_0(s_0^*) + \sum_{\tau=1}^{\ell-1} \gamma^\tau \cdot \upsilon_\tau(s_\tau^*)\right),\\ &= \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \boldsymbol{R}(s_\tau^*, a_\tau^*) + \sum_{\tau=0}^{\ell-1} \gamma^{\tau+1} \cdot \upsilon_{\tau+1}(s_{\tau+1}^*) - \sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \upsilon_\tau(s_\tau^*),\\ &= \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \left(\boldsymbol{R}(s_\tau^*, a_\tau^*) + \gamma \cdot \upsilon_{\tau+1}(s_{\tau+1}^*) - \upsilon_\tau(s_\tau^*)\right).\end{aligned}$$

Define $\upsilon_\tau^{\cdot}(s_\tau)\colon a_\tau \mapsto \boldsymbol{R}(s_\tau, a_\tau) + \gamma \cdot \upsilon_{\tau+1}(\boldsymbol{T}(s_\tau, a_\tau))$. It follows that

$$\upsilon_0^*(s_0^*) - \upsilon_0(s_0^*) = \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \left(\upsilon_\tau^{a_\tau^*}(s_\tau^*) - \upsilon_\tau(s_\tau^*)\right).$$

If we fix $s_\tau \doteq \operatorname{argmin}_{\tilde{s}_\tau \in \tilde{S}_\tau} \|s_\tau^* - \tilde{s}_\tau\|_1$, then we know that $\|s_\tau^* - s_\tau\|_1 \leqslant \delta_{\tilde{S}_{0:}}$ by definition of $\delta_{\tilde{S}_{0:}}$. Using action-values $\upsilon_\tau^{a_\tau^*}(s_\tau)$ to add zero into the previous error bound results in:

$$\upsilon_0^*(s_0^*) - \upsilon_0(s_0^*) = \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \left(\upsilon_\tau^{a_\tau^*}(s_\tau^*)) - \upsilon_\tau^{a_\tau^*}(s_\tau) + \upsilon_\tau^{a_\tau^*}(s_\tau) - \upsilon_\tau(s_\tau^*)\right).$$

Taking the best joint decision rule for $\upsilon_\tau^{\cdot}(s_\tau)$ results in value $\upsilon_\tau(s_\tau)$ greater or equal to $\upsilon_\tau^{a_\tau^*}(s_\tau)$, which leads to

$$\upsilon_0^*(s_0^*) - \upsilon_0(s_0^*) \leqslant \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \left(\upsilon_\tau^{a_\tau^*}(s_\tau^*) - \upsilon_\tau^{a_\tau^*}(s_\tau) + \upsilon_\tau(s_\tau) - \upsilon_\tau(s_\tau^*)\right).$$

Recall that under a fixed joint policy, value functions are linear functions of occupancy states, which allows us to re-arrange terms as follows:

$$\begin{aligned}\upsilon_0^*(s_0^*) - \upsilon_0(s_0^*) &\leqslant \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \left(\upsilon_\tau^{a_\tau^*}(s_\tau^*) - \upsilon_\tau(s_\tau^*) + \upsilon_\tau(s_\tau) - \upsilon_\tau^{a_\tau^*}(s_\tau)\right)\\ &= \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot (\upsilon_\tau^{a_\tau^*} - \upsilon_\tau) \cdot (s_\tau^* - s_\tau).\end{aligned}$$

The application of the Hölder inegality, the use of the definition of $\delta_{\tilde{S}_{0:}}$, and the use of the bounded reward function $r(\cdot,\cdot)$, permit us to conclude:

$$\begin{aligned}
\upsilon_0^*(s_0^*) - \upsilon_0(s_0^*) &\leqslant \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \|\upsilon_\tau^{a_\tau^*} - \upsilon_\tau\|_\infty \cdot \|s_\tau^* - s_\tau\|_1 \\
&= \delta_{\tilde{S}_{0:}} \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \cdot \|\upsilon_\tau^{a_\tau^*} - \upsilon_\tau\|_\infty \\
&\leqslant 2c\delta_{\tilde{S}_{0:}} \textstyle\sum_{\tau=0}^{\ell-1} \gamma^\tau \sum_{t=\tau}^{\ell-1} \gamma^{t-\tau} \\
&= 2c\delta_{\tilde{S}_{0:}} \textstyle\sum_{\tau=0}^{\ell-1} \sum_{t=\tau}^{\ell-1} \gamma^t \\
&= 2c\delta_{\tilde{S}_{0:}} \textstyle\sum_{\tau=0}^{\ell-1} \frac{\gamma^\tau - \gamma^\ell}{1-\gamma} \\
&= 2c\delta_{\tilde{S}_{0:}} \textstyle\frac{1}{1-\gamma} \sum_{\tau=0}^{\ell-1} (\gamma^\tau - \gamma^\ell) \\
&= 2c\delta_{\tilde{S}_{0:}} \textstyle\frac{1}{1-\gamma} \sum_{\tau=0}^{\ell-1} \frac{1+\ell\gamma^{\ell+1}-(\ell+1)\gamma^\ell}{1-\gamma} \\
&= 2c\delta_{\tilde{S}_{0:}} \textstyle\frac{1+\ell\gamma^{\ell+1}-(\ell+1)\gamma^\ell}{(1-\gamma)^2}.
\end{aligned}$$

Which ends the proof. □

E. Multi-Player Benchmarks

Multi-player Tiger. The 1-player tiger problem was first introduced by Kaelbling et al. (1998) and was later generalized to a 2-player version by Nair et al. (2003). This game describes a scenario where players face two closed doors, one of which conceals a treasure while the other hides a dangerous tiger. Neither player knows which door leads to the treasure and which one to the tiger, but they can receive partial and noisy information about the tiger's location by listening. At any given time, each player can choose to open either the left or right door, which will either reveal the treasure or the tiger, and reset the game. To gain more information about the tiger's location, players can listen to hear the tiger on the left or right side, but with uncertain accuracy.
We have extended this problem to an n-player version by incorporating hierarchical information-sharing and modifying the transition, observation, and reward models following Nair et al. (2003), while ensuring that the original 2-player problem can still be recovered. In this n-player version, only the reward function is not straightforwardly adapted. Listening still costs 1 per player, as in the original problem, while the penalty for opening the wrong door is now set to $-100/n_w$ (with n_w the number of players opening the bad door) and the reward for opening the good door is 10 per player.

Multi-player Recycling Robot. The recycling robot task was first introduced by Sutton & Barto (2018) as a single-player problem. Later on, Amato et al. (2012) generalized it to a two-player version. The multi-player formulation requires robots to work together to recycle soda cans. In this problem, both robots have a battery level, which can be either high or low. They have to choose between collecting small or big cans and recharging their own battery level. Collecting small or big cans can decrease the robot's battery level, with a higher probability when collecting the big can. When a robot's battery is completely exhausted, it needs to be picked up and placed onto a recharging spot, which results in a negative reward. The coordination problem arises since robots cannot pick up a big can independently. In our n-player version of the problem, picking up small cans still rewards 2 per agent. A reward of 5 per agent is given if all agents synchronize to carry a big can, while a penalty of 10 is given if some agents (but not all) try to carry a big can.

Multi-player Broadcast Channel. In 1996, Ooi & Wornell (1996) introduced a scenario in which a unique channel is shared by n players, who aim at transmitting packets. The time is discretized, and only one packet can be transmitted at each time step. If two or more players attempt to send a packet at the same time, the transmission fails due to a collision. In 2004, Hansen et al. extended this problem to a partially observable one, focusing on two players (Hansen et al., 2004). We used similar adaptations to define a partially observable version of the original n-player broadcast channel.

Multi-player Grid3x3. This problem was first introduced by Bernstein et al. (2002). It involves two players who want to meet each other as soon as possible on a two-dimensional grid. Each player has five possible actions: moving north, south, west, east, or staying in place. To simulate an uncertain environment, each player's action has a fixed probability of being successful. Additionally, each player can only sense their own location and has no knowledge of the other player's location. To adapt the problem for multiple players, we placed M players on the grid, each with the same actions and perceptions as described above. The reward has been redefined as the largest number of players minus one present at one of the two meeting points. This way, the original problem can be retrieved for two players.

F. Experiments

We conducted three sets of experiments to assess our findings:

1. To assess the exponential drop in time complexity of backups with respect to an increasing number of players, we maintain the average time required to perform a single backup, *cf.* Section F.1 – Average Backup Time for Increasing Players.
2. To assess the exponential drop in time complexity of backups with respect to increasing horizons, we maintain the average time required to perform a single backup, *cf.* Section F.2 – Average Backup Time for Increasing Horizon.
3. To assess the superiority of our findings with respect to the state-of-the-art approach to solve general decentralized partially Markov decision processes near-optimally, *cf.* Section F.3 – Against State-Of-The-Art Solvers.

F.1. Average Backup Time for Increasing Players

This section investigates the average computational time required to perform a single backup for increasing players, *cf.* Figures F.1,F.1,F.1, and F.1. The experiments show that on all tested benchmarks, hPBVI exhibits a reduction in time complexity compared to the other variants. Moreover, hPBVI can handle a larger number of agents (up to 9 for the Tiger, MABC, and Recycling) compared to the other variants, which are limited to a maximum of 5 agents. This time-complexity reduction in hPBVI is the result of our findings providing the ability to fully exploit the hierarchical information-sharing structure.

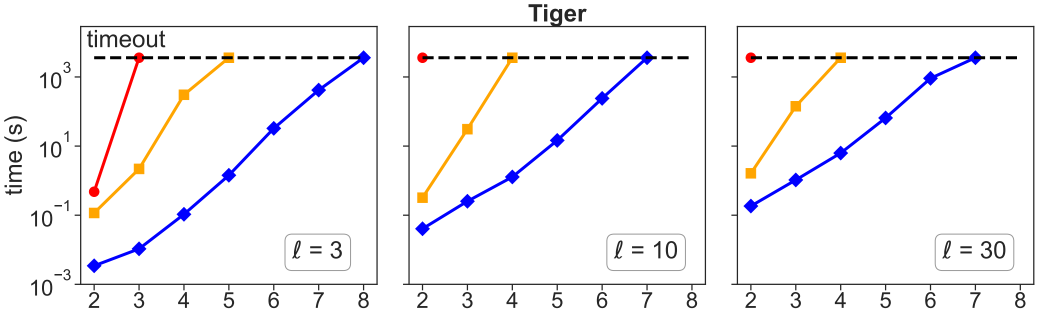

Figure 5. Average Backup Time for the tiger problem and different numbers of players.

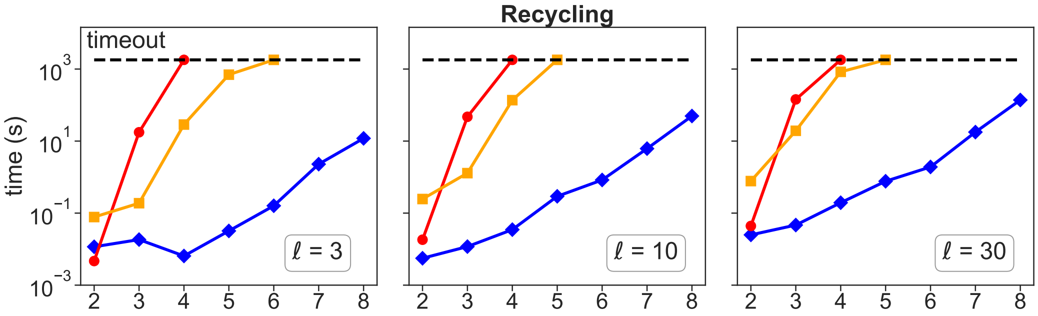

Figure 6. Average Backup Time for the recycling problem and different numbers of players.

F.2. Average Backup Time for Increasing Horizons

This section investigates the average computational time required to perform a single backup for increasing horizons, *cf.* Figures F.2, F.2, 11, and 12. The experiments show once again that on all tested benchmarks, hPBVI exhibits an exponential

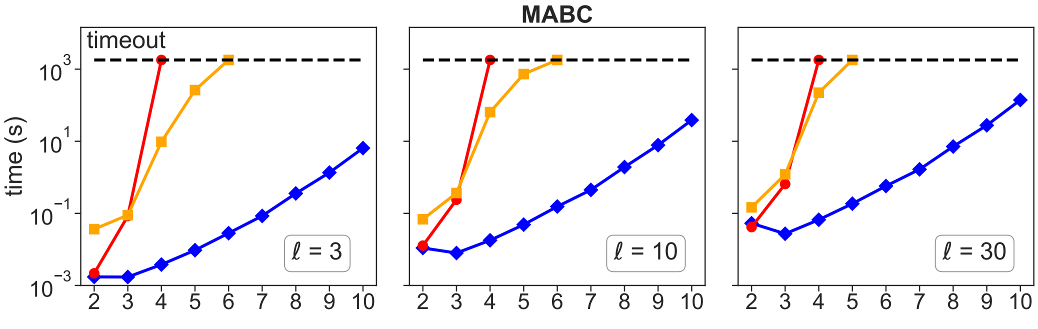

Figure 7. Average Backup Time for the mabc problem and different numbers of players.

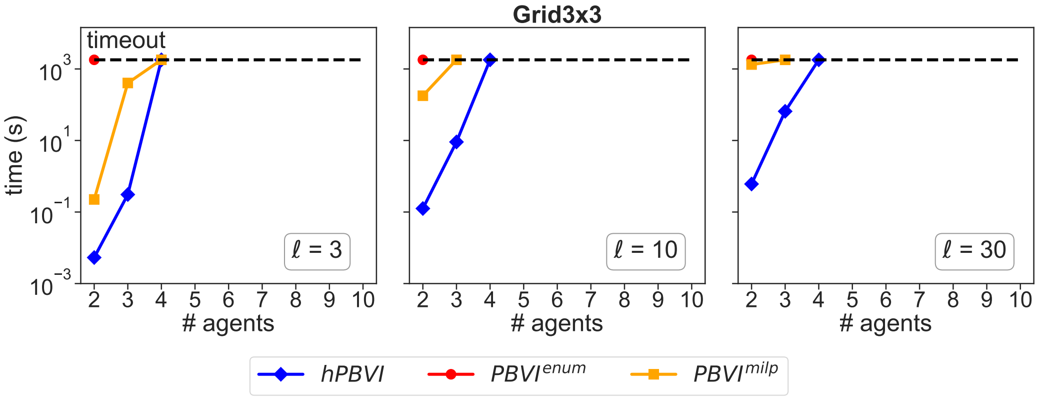

Figure 8. Average Backup Time for the grid3x3 problem and different numbers of players.

drop in time complexity compared to the other variants. However, all three variants of the PBVI algorithm exhibit an increase in time complexity with respect to the planning horizon. This increase in time complexity is expected since, as time goes the number of backups also increases.

F.3. Against State-Of-The-Art Solvers

In this section, we compare our PBVI algorithm variants with two local algorithms, namely A2C and IQL, which are state-of-the-art and can handle a large number of players, as shown in Figures 13, 14, 15, and 15. However, these algorithms prioritize scalability over optimality and may get stuck in local optima. Our experiments demonstrate that hPBVI consistently outperforms all competitors in nearly all tested benchmarks in terms of convergence time and the value of the solution found within 30 minutes. In some weakly coupled domains, A2C and IQL find nearly optimal solutions close to those found by hPBVI.

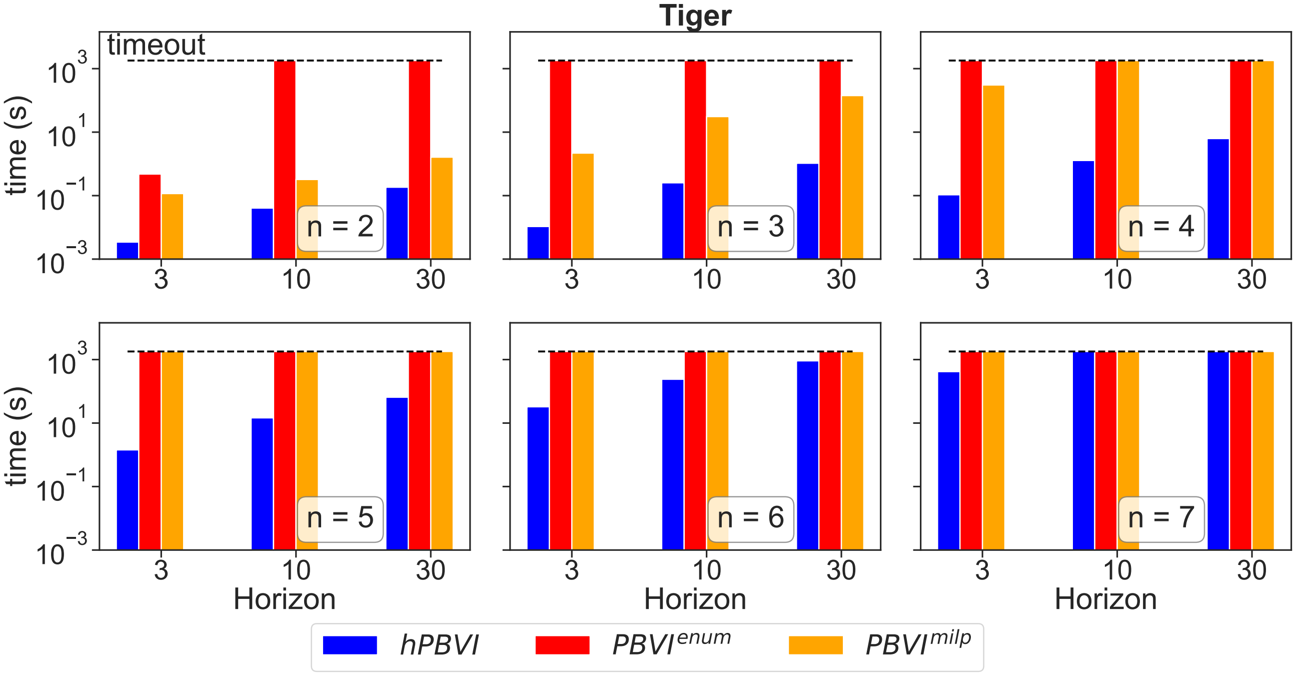

Figure 9. Average backup time as a function of planning horizons for Tiger.

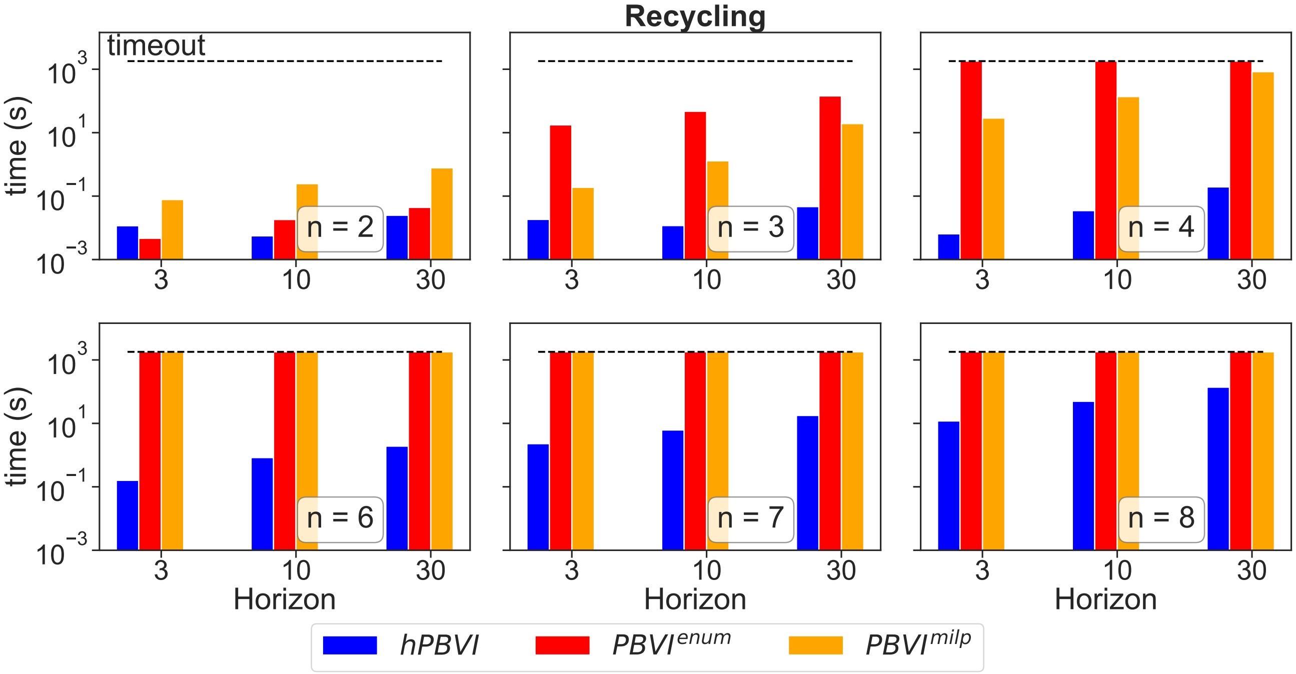

Figure 10. Average backup time as a function of planning horizons for Recycling.

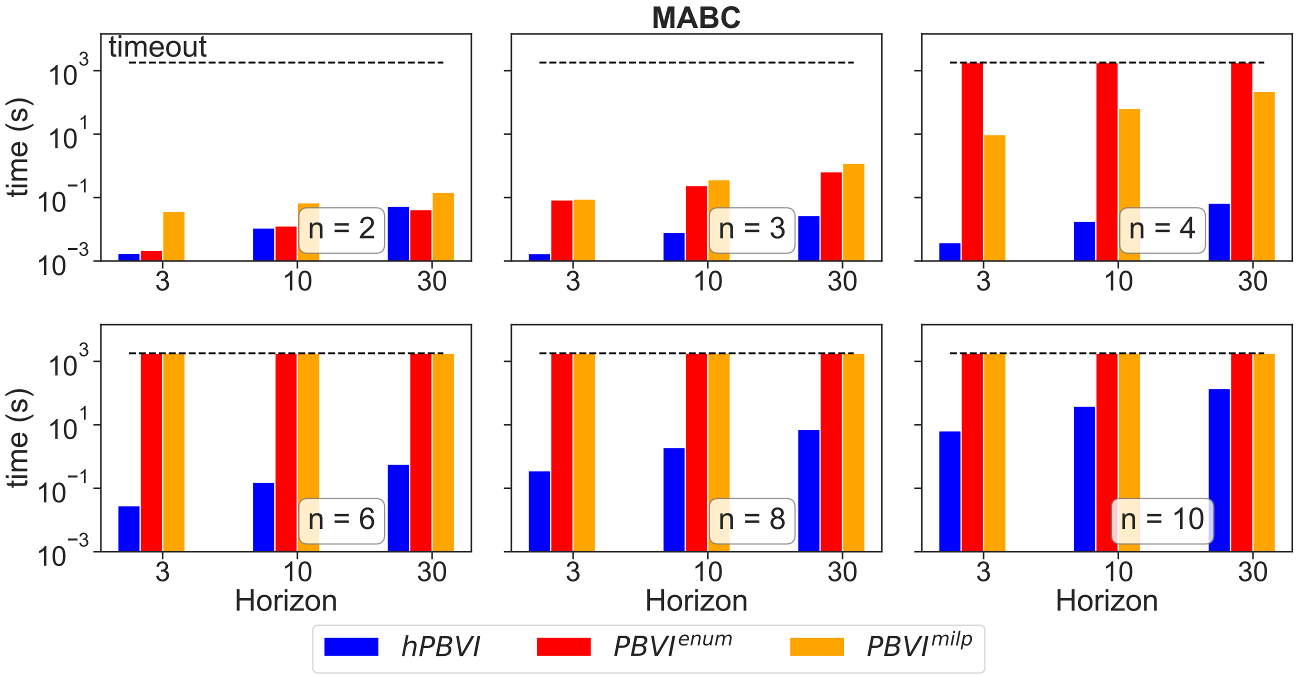

Figure 11. Average backup time as a function of planning horizons for MABC.

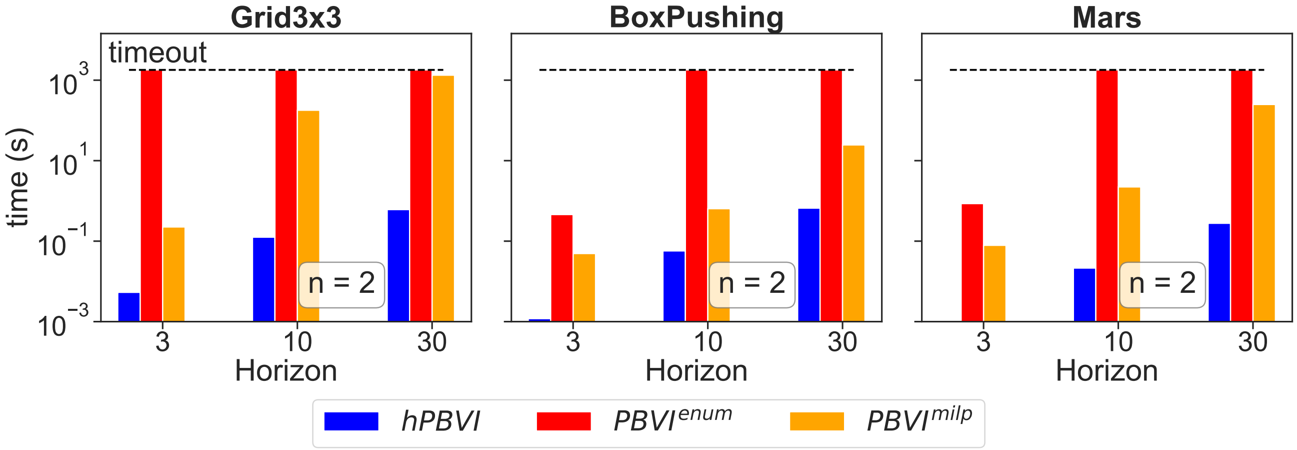

Figure 12. Average backup time as a function of planning horizons for Grid3x3, BoxPushing and Mars.

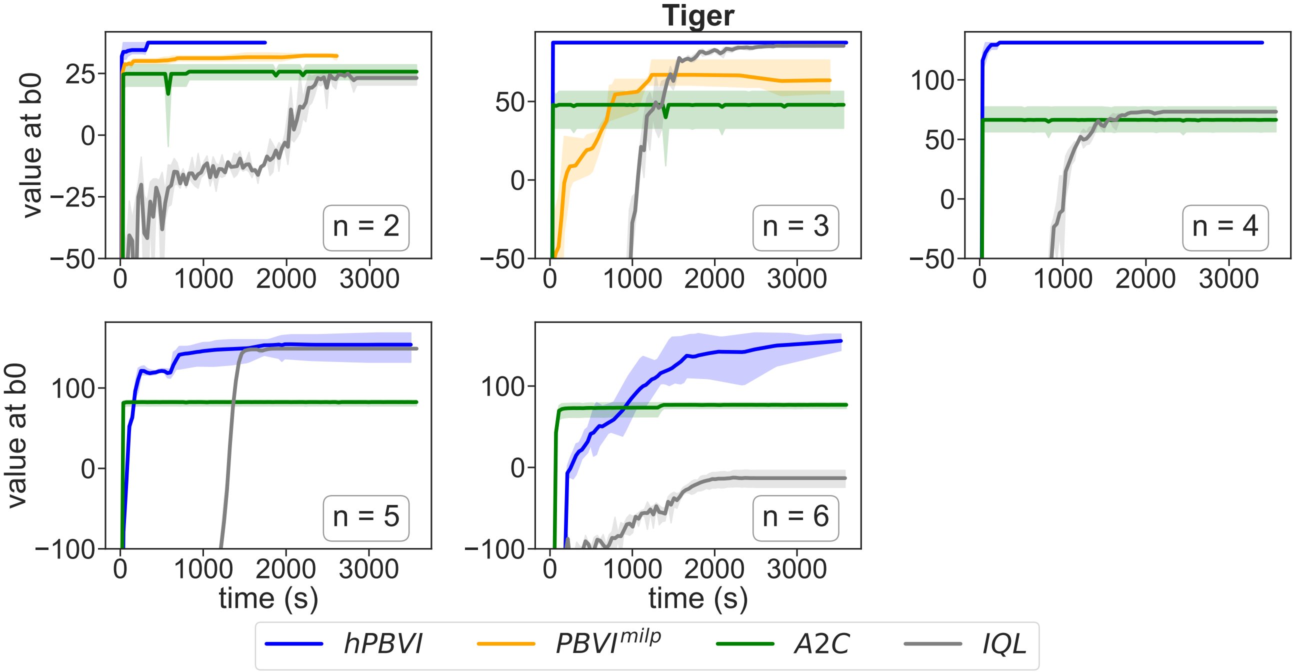

Figure 13. Anytime values for Tiger and $\ell = 30$.

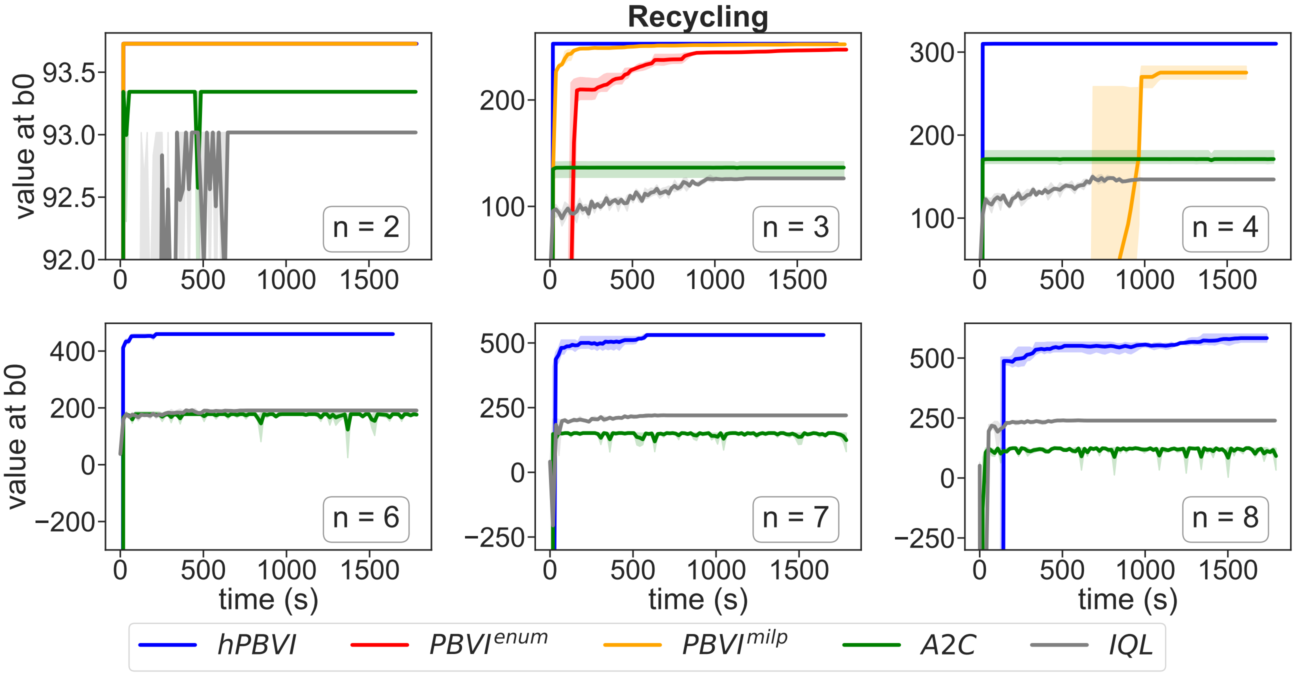

Figure 14. Anytime values for Recycling and $\ell = 30$.

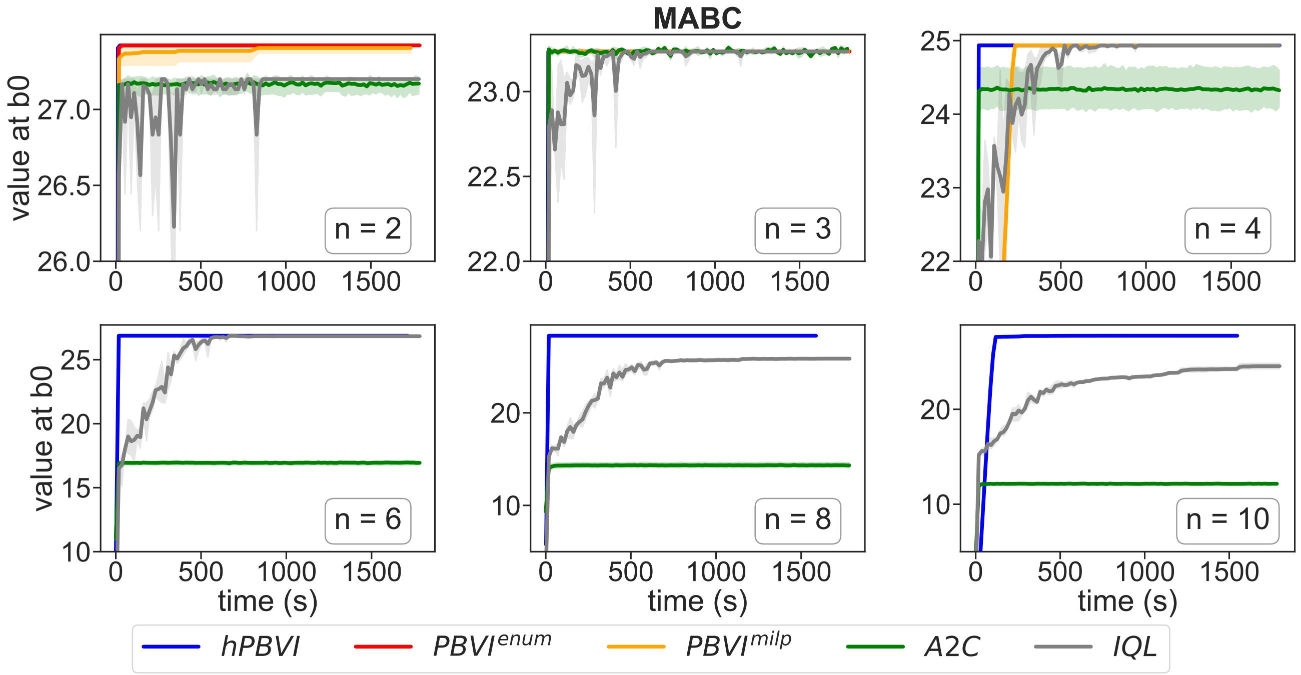

Figure 15. Anytime values for Multi-agent broadcast channel and $\ell = 30$.

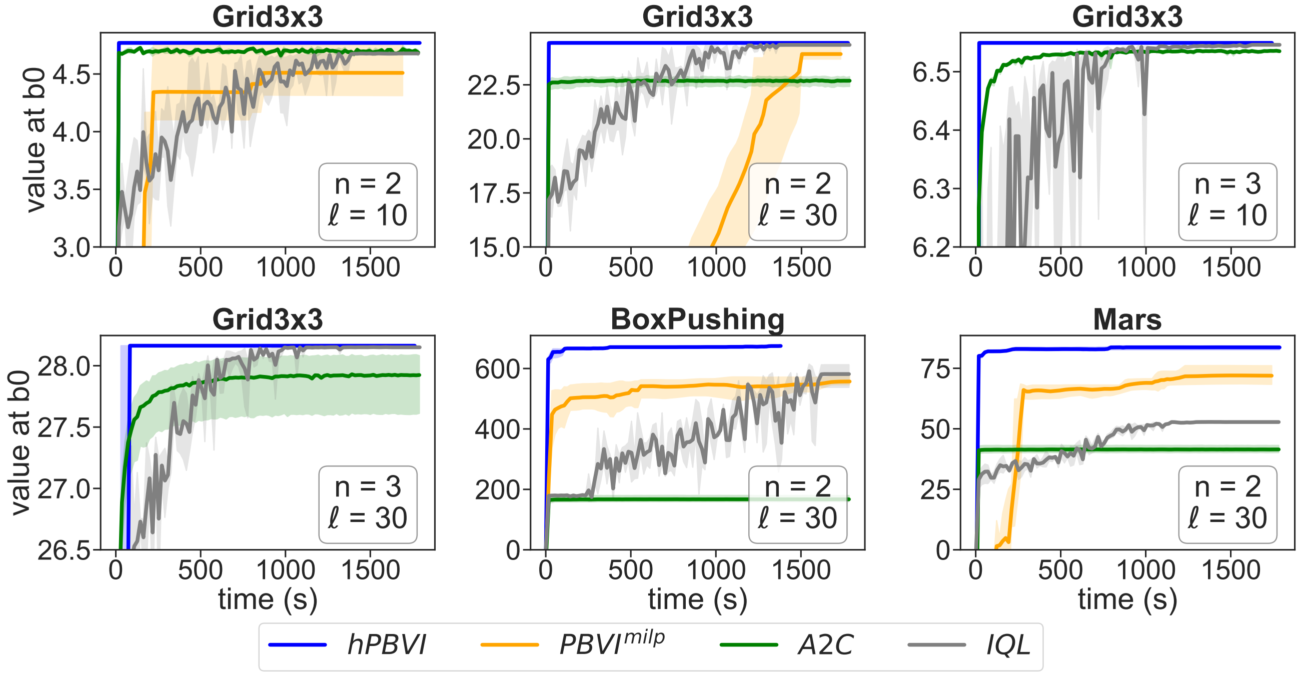

Figure 16. Anytime values for Grid3x3, BoxPushing and Mars.