2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

A Strong Duality Result for Cooperative Decentralized Constrained POMDPs

Nouman Khan, *Student Member, IEEE,* and Vijay Subramanian, *Senior Member, IEEE*

***Abstract*— The work studies cooperative decentralized constrained POMDPs with asymmetric information. Using an extension of Sion's Minimax theorem for functions with positive infinity and results on weak-convergence of measures, strong duality and existence of a saddle point are established for the setting of infinite-horizon expected total discounted costs when the observations lie in a countable space, the actions are chosen from a finite space, the immediate constraint costs are bounded, and the immediate objective cost is bounded from below.**

I. Introduction

Single-Agent Markov Decision Processes (SA-MDPs) [1] and Single-Agent Partially Observable Markov Decision Processes (SA-POMDPs) [2] have long served as the basic building-blocks in the study of sequential decision-making. An SA-MDP is an abstraction in which an agent sequentially interacts with a fully-observable Markovian environment to solve a multi-period optimization problem; in contrast, in SA-POMDP, the agent only gets to observe a noisy or incomplete version of the environment. In 1957, Bellman proposed dynamic-programming as an approach to solve SA-MDPs [1], [3]. This combined with the characterization of SA-POMDP into an equivalent SA-MDP [4]–[6] (in which the agent maintains a belief about the environment's true state) made it possible to extend dynamic-programming results to SA-POMDPs. Reinforcement learning [7] based algorithmic frameworks use data-driven dynamic-programming approaches to solve such single-agent sequential decision-making problems when the environment is unknown.

In many engineering systems, there are multiple decision-makers that collectively solve a sequential decision-making problem but with safety constraints: e.g., a team of robots performing a joint task, a fleet of automated cars navigating a city, multiple traffic-light controllers in a city, etc. Bandwidth constrained communications and communication delays in such systems lead to a decentralized team problem with information asymmetry. In this work, we study a fairly general abstraction of such systems, namely that of a cooperative decentralized constrained POMDP, hereon referred to as Dec-C-POMDP. The special cases of Dec-C-POMDP when there are no constraints, when there is only one agent, or when the environment is fully observable to each agent, are referred to as Dec-POMDP[1], SA-C-POMDP, and Dec-C-MDP, respectively. The relationships among such models are shown in Figure 1.

A. Related Work

1) Single-Agent Settings: Prior work on planning and learning under constraints has primarily focused on single-agent constrained MDP (SA-C-MDP) where unlike in SA-MDPs, the agent solves a constrained optimization problem. For this setup, a number of fundamental results from the planning perspective have been derived – for instance, [9]–[15]; see [16] for details of the convex-analytic approach for SA-C-MDPs. These aforementioned results have led to the development of many algorithms in the learning setting: see [17]–[23]. Unlike SA-C-MDPs, rigorous results for SA-C-POMDPs are limited; few works include [24]–[27].

2) Multi-Agent Settings: Challenges arising from the combination of partial observability of the environment and information-asymmetry[2] have led to difficulties in developing general solutions to Dec-POMDPs: e.g., solving a finite-horizon Dec-POMDP with more than two agents is known to be NEXP-complete [28]. Nevertheless, conceptual approaches exist to establish solution methodologies and structural properties in (finite-horizon) Dec-POMDPs namely: i) the person-by-person approach [29]; ii) the designer's approach [30]; and iii) the common-information (CI) approach [31], [32]. Using a fictitious coordinator that only uses the common information to take actions, the CI approach allows for the transformation of the problem to a SA-POMDP which can be used to solve for an optimal control. The CI approach has also led to the development of a multi-agent reinforcement learning (MARL) framework [33] where agents learn good compressions of common and private information that can suffice for approximate optimality. On the empirical front, worth-mentioning works include [34], [35]. Finally, as far as we know, work on Dec-C-POMDPs is non-existent.

B. Contribution

For Dec-C-POMDPs, the technical challenges increase even more from those of Dec-POMDPs because restriction of the search space to deterministic policy-profiles is no longer an option[3]. Therefore, the coordinator in the equivalent SA-C-POMDP has an uncountable prescription space, which leads to an uncountable state-space in its equivalent SA-C-MDP. This is an issue because most fundamental results in the theory of SA-C-MDPs (largely based on occupation-measures) rely heavily on the state-space being countably-

N. Khan and V. Subramanian are with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, 48109-2122. knouman@umich.edu, vgsubram@umich.edu

[1]For a good introduction to Dec-POMDPs, see [8].

[2]Mismatch in the information of the agents.

[3]Restricting to deterministic policies can be sub-optimal in SA-C-MDPs and SA-C-POMDPs: see [16] and [24].

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383989

infinite; see [16]. Due to these reasons, the study of Dec-C-POMDPs calls for a new methodology—one which avoids this transformation and directly studies the decentralized problem. Our work takes the first steps in this direction and presents a rigorous approach for Dec-C-POMDPs which is based on structural characterization of the set of behavioral policies and their performance measures, and using measure theoretic results. The main result in this paper, namely Theorem 1, establishes strong duality and existence of a saddle-point for Dec-C-POMDPs, thus providing a firm theoretical basis for (future) development of primal-dual type planning and learning algorithms.

C. Organization

The rest of the paper is organized as follows. Mathematical model of (cooperative) Dec-C-POMDP is introduced in Section II. The optimization problem is formulated in Section III. Results on strong duality and existence of a saddle point are then derived in Section IV. Finally, concluding remarks are given in Section V.

D. Notation

Before we present the model, we highlight the key notations in this paper.

- The sets of integers and positive integers are respectively denoted by $\mathbb{Z}$ and $\mathbb{N}$. For integers a and b, $[a,b]_{\mathbb{Z}}$ represents the set $\{a, a+1, \ldots, b\}$ if $a \leq b$ and $\emptyset$ otherwise. The notations [a] and $[a,\infty]_{\mathbb{Z}}$ are used as shorthand for $[1,a]_{\mathbb{Z}}$ and $\{a, a+1, \ldots\}$, respectively.
- For integers $a \leq b$ and $c \leq d$, and a quantity of interest q, $q^{(a:b)}$ is a shorthand for the vector $\left(q^{(a)}, q^{(a+1)}, \ldots, q^{(b)}\right)$ while $q_{c:d}$ is a shorthand for the vector $(q_c, q_{c+1}, \ldots, q_d)$. The combined notation $q_{a:b}^{(c:d)}$ is a shorthand for the vector $(q_i^{(j)} : i \in [a,b]_{\mathbb{Z}}, j \in [c,d]_{\mathbb{Z}})$. The infinite tuples $\left(q^{(a)}, q^{(a+1)}, \ldots,\right)$ and $(q_c, q_{c+1}, \ldots,)$ are respectively denoted by $q^{(a:\infty)}$ and $q_{c:\infty}$.
- For two real-valued vectors v_1 and v_2, the inequalities $v_1 \leq v_2$ and $v_1 < v_2$ are meant to be element-wise inequalities.
- Probability and expectation operators are denoted by $\mathbb{P}$ and $\mathbb{E}$, respectively. Random variables are denoted by upper-case letters and their realizations by the corresponding lower-case letters. At times, we also use the shorthand $\mathbb{E}[\cdot|x] \triangleq \mathbb{E}[\cdot|X=x]$ and $\mathbb{P}(y|x) \triangleq \mathbb{P}(Y=y|X=x)$ for conditional quantities.
- Topological spaces are denoted by upper-case calligraphic letters. For a topological-space $\mathcal{W}$, $\mathcal{B}(\mathcal{W})$ denotes the Borel σ-algebra, measurability is determined with respect to $\mathcal{B}(\mathcal{W})$, and $\mathcal{M}_1(\mathcal{W})$ denotes the set of all probability measures on $\mathcal{B}(\mathcal{W})$ endowed with the topology of weak convergence. Also, unless stated otherwise, "measure" means a non-negative measure.
- Unless otherwise stated, if a set $\mathcal{W}$ is countable, as a topological space it will be assumed to have the discrete topology. Therefore, the corresponding Borel σ-algebra $\mathcal{B}(\mathcal{W})$ will be the power-set $2^{\mathcal{W}}$.

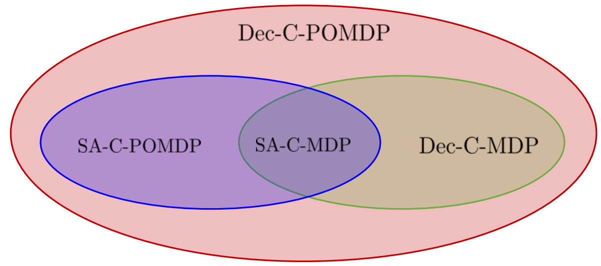

Fig. 1: Relationships between Models of Cooperative Sequential Decision-Making under Constraints.

- Unless stated otherwise, the product of a collection of topological spaces will be assumed to have the product topology.

II. Model

Let $(N, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}_{tr}, (c,d), P_1, \mathcal{U}, \alpha)$ denote a (cooperative) Dec-C-POMDP with N agents, state space $\mathcal{S}$, joint-observation space $\mathcal{O}$, joint-action space $\mathcal{A}$, transition-law $\mathcal{P}_{tr}$, immediate-cost functions c and d, (fixed) initial distribution P_1, space of decentralized policy-profiles $\mathcal{U}$, and discount factor $\alpha \in (0,1)$. The decision problem (to be detailed later on) has the following attributes and notation.

- **State Process**: The state-space $\mathcal{S}$ is some topological space with a Borel σ-algebra $\mathcal{B}(\mathcal{S})$. The state-process is denoted by $\{S_t\}_{t=1}^{\infty}$.
- **Joint-Observation Process**: The joint-observation space $\mathcal{O}$ is a countable discrete space of the form $\mathcal{O} = \prod_{n=0}^{N} \mathcal{O}^{(n)}$, where $\mathcal{O}^{(0)}$ denotes the common observation space of all agents and $\mathcal{O}^{(n)}$ denotes the private observation space of agent $n \in [N]$. The joint-observation process is denoted by $\{O_t\}_{t=1}^{\infty}$ where $O_t = O_t^{(0:N)}$ and is such that at time t, agent $n \in [N]$ observes $O_t^{(0)}$ and $O_t^{(n)}$ only.
- **Joint-Action Process**: The joint-action space $\mathcal{A}$ is a finite discrete space of the form $\mathcal{A} = \prod_{n=1}^{N} \mathcal{A}^{(n)}$, where $\mathcal{A}^{(n)}$ denotes the action space of agent $n \in [N]$. The joint-action process is denoted by $\{A_t\}_{t=1}^{\infty}$ where $A_t = A_t^{(1:N)}$ and $A_t^{(n)}$ denotes the action of agent n at time t. Since all $\mathcal{A}^{(n)}$'s and $\mathcal{A}$ are finite, they are all compact metric spaces.[4]
- **Transition-law**: At time $t \in \mathbb{N}$, given the current state S_t and current joint-action A_t, the next state S_{t+1} and the next joint-observation O_{t+1} are determined in a time-homogeneous manner, independent of all previous states, all previous and current joint-observations, and all previous joint-actions. The transition-law is given by

$$\mathcal{P}_{tr} \triangleq \{P_{saBo} : s \in \mathcal{S}, a \in \mathcal{A}, B \in \mathcal{B}(\mathcal{S}), o \in \mathcal{O}\}, \quad (1)$$

where for all $t \in \mathbb{N}$,

$$\begin{aligned} &\mathbb{P}(S_{t+1} \in B, O_{t+1} = o | S_{1:t-1} = s_{1:t-1}, \\ &\quad O_{1:t} = o_{1:t}, A_{1:t-1} = a_{1:t-1}, S_t = s, A_t = a) \\ &\quad = \mathbb{P}(S_{t+1} \in B, O_{t+1} = o | S_t = s, A_t = a) \quad (2) \\ &\quad \triangleq P_{saBo}. \end{aligned}$$

[4]Hence, also complete and separable.

• **Immediate-costs**: The immediate cost $c : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is a real-valued function whose expected discounted aggregate (to be defined later) we would like to minimize. On the other hand, the immediate cost $d : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^K$ is $\mathbb{R}^K$-valued function whose expected discounted aggregate we would like to keep within a specified threshold. For these reasons, we call c and d as the immediate objective and constraint costs respectively. We shall make use of the following assumption on immediate-costs in Theorem 1.

Assumption 1. *The immediate objective cost is bounded from below and the immediate constraint costs are bounded, i.e., there exist $\underline{c} \in \mathbb{R}$ and $\underline{d}, \overline{d} \in \mathbb{R}^K$ such that*

$$\underline{c} \le c(\cdot, \cdot) \text{ and } \underline{d} \le d(\cdot, \cdot) \le \overline{d}. \tag{3}$$

Let $\bar{\underline{d}} = \|\underline{d}\|_\infty \vee \|\overline{d}\|_\infty$ so that $\|d(\cdot,\cdot)\|_\infty \le \bar{\underline{d}} < \infty$.

• **Initial Distribution**: P_1 is a (fixed) probability measure for the initial state and initial joint-observation, i.e., $P_1 \in \mathcal{M}_1(\mathcal{S} \times \mathcal{O})$ and

$$P_1(B, o) \triangleq \mathbb{P}(S_1 \in B, O_1 = o). \tag{4}$$

• **Space of Policy-Profiles**: At time $t \in \mathbb{N}$, the common history of all agents is defined as all the common observations received thus far, i.e., $H_t^{(0)} \triangleq \left(O_{1:t}^{(0)}\right)$. Similarly, the private history of agent $n \in [N]$ at time t is defined as all observations received and all the actions taken by the agent thus far (except for those that are part of the common information), i.e.,

$$\begin{aligned} &H_1^{(n)} \triangleq O_1^{(n)} \setminus O_1^{(0)}, \text{ and} \\ &H_t^{(n)} \triangleq \left(H_{t-1}^{(n)}, (A_{t-1}^{(n)}, O_t^{(n)}) \setminus O_t^{(0)}\right) \ \forall t \in [2, \infty]_{\mathbb{Z}}. \end{aligned} \tag{5}$$

Finally, the joint history at time t is defined as the tuple of the common history and all the private histories at time t, i.e., $H_t \triangleq H_t^{(0:n)}$.

For $t \in \mathbb{N}$ and $n \in [0, N]_{\mathbb{Z}}$, let $\mathcal{H}_t^{(n)}$ denote the set of all possible realizations of $H_t^{(n)}$. We define a (decentralized) behavioral policy-profile u as a tuple $u^{(1:N)} \in \mathcal{U} \triangleq \prod_{n=1}^N \mathcal{U}^{(n)}$ where $u^{(n)}$ denotes some behavioral policy used by agent n, i.e., $u^{(n)}$ itself is a tuple of the form $u_{1:\infty}^{(n)}$ where $u_t^{(n)}$ maps $\mathcal{H}_t^{(0)} \times \mathcal{H}_t^{(n)}$ to $\mathcal{M}_1\left(\mathcal{A}^{(n)}\right)$, and where agent n uses the distribution $u_t^{(n)}(H_t^{(0)}, H_t^{(n)})$ at time t to choose its action $A_t^{(n)}$. We pause to emphasize that in a (decentralized) behavioral policy, at any time t, each agent randomizes over its action-set independently of all other agents (*no common randomness is used*). Thus, given a joint-history $h_t \in \mathcal{H}_t$ at time t, the probability that joint-action $a_t \in \mathcal{A}$ is taken is given by

$$\begin{aligned} u_t(a_t|h_t) &\triangleq \prod_{n=1}^N u_t^{(n)}\left(h_t^{(0)}, h_t^{(n)}\right)\left(a_t^{(n)}\right) \\ &= \prod_{n=1}^N u_t^{(n)}\left(a_t^{(n)} \middle| h_t^{(0)}, h_t^{(n)}\right). \end{aligned} \tag{6}$$

Remark 1. *With Assumption 1, the conditional expectations $\mathbb{E}_{P_1}[c(S_t, A_t) \mid H_t = h_t, A_t = a_t]$ and $\mathbb{E}_{P_1}[d(S_t, A_t) \mid H_t = h_t, A_t = a_t]$ exist, are unique, and are bounded from below. Furthermore, the latter are element-wise finite.*

• **Decision Process**: Let $\mathbb{P}_{P_1}^{(u)}$ be the probability measure corresponding to policy-profile $u \in \mathcal{U}$ and initial-distribution P_1, and let $\mathbb{E}_{P_1}^{(u)}$ denote the corresponding expectation operator.[5] We define *infinite-horizon expected total discounted costs* $C : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ and $D : \mathcal{U} \to \mathbb{R}^K$ as

$$C(u) = C^{(P_1,\alpha)}(u) \triangleq \mathbb{E}_{P_1}^{(u)}\left[\sum_{t=1}^\infty \alpha^{t-1} c(S_t, A_t)\right], \tag{7}$$

$$\text{and } D(u) = D^{(P_1,\alpha)}(u) \triangleq \mathbb{E}_{P_1}^{(u)}\left[\sum_{t=1}^\infty \alpha^{t-1} d(S_t, A_t)\right]. \tag{8}$$

Remark 2. *Assumption 1 ensures that $C(u) \in \mathbb{R} \cup \{\infty\}$, and $D(u) \in \mathbb{R}^K$ with (absolute) element-wise bound $\bar{\underline{d}}/(1-\alpha)$.*

The decision process proceeds as follows: *i*) At time $t \in \mathbb{N}$, the current state S_t and observations O_t are generated; *ii*) each agent $n \in [N]$ chooses an action $a^{(n)} \in \mathcal{A}^{(n)}$ based on $H_t^{(0)}, H_t^{(n)}$; *iii*) the immediate-costs $c(S_t, A_t), d(S_t, A_t)$ are incurred; *iv*) The system moves to the next state and observations according to the transition-law $\mathcal{P}_{tr}$.

III. Optimization Problem

To formulate the Dec-C-POMDP optimization problem, we first need to give a suitable topology to the space of behavioral policy-profiles, in particular, one in which it is compact and convex.[6] To this end, we use the finiteness of the action space $\mathcal{A}^{(n)}$ and the countability of the joint-observation space $\mathcal{O}$ to associate $\mathcal{U}$ with a product of compact sets that are parameterized by (countable number of) all possible histories. Tychonoff's theorem (see [37][Proposition 4]) then helps achieve compactness under the product topology. (Convexity comes trivially). Thus, the sets

$$\begin{aligned} \mathcal{H}_t &\triangleq \prod_{n=0}^N \mathcal{H}_t^{(n)}, \\ \mathcal{H}^{(n)} &\triangleq \bigcup_{t=1}^\infty \mathcal{H}_t^{(0)} \times \mathcal{H}_t^{(n)}, \text{ and} \\ \mathcal{H} &\triangleq \bigcup_{t=1}^\infty \mathcal{H}_t, \end{aligned} \tag{9}$$

are countable. Here, $\mathcal{H}_t$ is the set of all possible joint-histories at time t, $\mathcal{H}^{(n)}$ is the set of all possible histories of agent n, and $\mathcal{H}$ is the set of all possible joint-histories. With this in mind, one observes that $\mathcal{U}$ is in one-to-one correspondence with the set $\mathcal{X}_{\mathcal{U}} \triangleq \prod_{n=1}^N \mathcal{X}_{\mathcal{U}^{(n)}}$, where

$$\mathcal{X}_{\mathcal{U}^{(n)}} \triangleq \prod_{h \in \mathcal{H}^{(n)}} \mathcal{M}_1\left(\mathcal{A}^{(n)}; h\right), \tag{10}$$

[5] The existence and uniqueness of $\mathbb{P}_{P_1}^{(u)}$ can be ensured by an adaptation of the Ionesca-Tulcea theorem [36].

[6] Convexity is a set property rather than a topological property. In the rest of the paper, by a "convex topological space", we mean convexity of the set on which the topology is defined.

and $\mathcal{M}_1\left(\mathcal{A}^{(n)};h\right)$[7] is a copy of $\mathcal{M}_1\left(\mathcal{A}^{(n)}\right)$ dedicated for agent-n's history h. For example, a given policy u would correspond to a point $x \in \mathcal{X}_{\mathcal{U}}$ such that $x_{n,\left(h_t^{(0)},h_t^{(n)}\right)} = u_t^{(n)}\left(\cdot|h_t^{(0)},h_t^{(n)}\right)$, and similarly, vice versa.

Since $\mathcal{A}^{(n)}$ is a complete separable (compact) metric space, by Prokhorov's Theorem (see [37][Proposition 6]), each $\mathcal{M}_1\left(\mathcal{A}^{(n)};h\right)$ is a compact (and convex[8]) metric space (with the topology of weak-convergence). Therefore, endowing $\mathcal{X}_{\mathcal{U}^{(n)}}$ and $\mathcal{X}_{\mathcal{U}}$ with the product topology makes each a compact (and convex) metric space via Tychonoff's theorem (see [37][Proposition 4]), which is also metrizable (via [37][Proposition 5]). Given the one-to-one correspondence, **from now onward, we assume that $\mathcal{U}^{(n)}$ and $\mathcal{U}$ have the same topology as that of $\mathcal{X}_{\mathcal{U}^{(n)}}$ and $\mathcal{X}_{\mathcal{U}}$ respectively**. Henceforth, we will consider C and D_k's as functions on topological spaces. Furthermore, since $\mathcal{U}$ has been shown to be a compact metric space (hence, also complete and separable), we can also define $\mathcal{B}(\mathcal{U}) = \otimes_{n=1}^{N}\mathcal{B}\left(\mathcal{U}^{(n)}\right)$[9], and $\mathcal{M}_1(\mathcal{U})$, where $\mathcal{M}_1(\mathcal{U})$ is compact (and convex) metrizable space by Prokhorov's theorem (see [37][Proposition 6]).

It will be helpful to work with mixtures of behavioral policy-profiles – wherein the team first uses a measure $\mu \in M_1(\mathcal{U})$ to choose its policy-profile $u \in \mathcal{U}$ and then proceeds with it from time 1 onward. Under this setup, the policy-profile chosen collectively by the agents becomes a random object, and we extend the definitions of C and D to $\widehat{C} : \mathcal{M}_1(\mathcal{U}) \to \mathbb{R}\cup\{\infty\}$ and $\widehat{D} : \mathcal{M}_1(\mathcal{U}) \to \mathbb{R}^K$ as follows:

$$\begin{aligned}\widehat{C}(\mu) &= \widehat{C}^{(P_1,\alpha)}(\mu) \triangleq \mathbb{E}^{(U\sim\mu)}\left[C(U)\right], \text{and}\\ \widehat{D}(\mu) &= \widehat{D}^{(P_1,\alpha)}(\mu) \triangleq \mathbb{E}^{(U\sim\mu)}\left[D(U)\right].\end{aligned} \tag{11}$$

The goal of the agents is to work cooperatively to solve the following constrained optimization problem.

$$\left.\begin{aligned}&\text{minimize } \widehat{C}(\mu)\\ &\text{subject to } \mu \in \mathcal{M}_1(\mathcal{U}) \text{ and } \widehat{D}(\mu) \le \breve{D}.\end{aligned}\right\} \tag{\textit{Dec-C-POMDP}}$$

Here, $\breve{D}$ is a fixed K-dimensional real-valued vector. We refer to the solution of (*Dec-C-POMDP*) as its optimal value and denote it by $\underline{C} = \underline{C}^{(P_1,\alpha)}$. In particular, if the set of feasible mixtures is empty, we set $\underline{C}$ to ∞, and, with slight abuse of terminology, consider any mixture in $\mathcal{M}_1(\mathcal{U})$ to be optimal.

The following assumption about feasibility of (*Dec-C-POMDP*) will be used in one of the parts of Theorem 1.

Assumption 2 (Slater's Condition)**.** *There exists a mixture $\mu \in \mathcal{M}_1(\mathcal{U})$ and $\zeta > 0$ for which*

$$D(\mu) \le \breve{D} - \zeta 1. \tag{12}$$

[7] $M_1(\cdot)$ denotes the set of all probability measures on $\cdot$.
[8] Convexity of $\mathcal{M}_1\left(\mathcal{A}^{(n)}\right)$ is trivial.
[9] For separable metric spaces $\mathcal{W}_1, \mathcal{W}_2, \ldots$, $\mathcal{B}(\mathcal{W}_1 \times \mathcal{W}_2 \times \ldots) = \mathcal{B}(\mathcal{W}_1) \otimes \mathcal{B}(\mathcal{W}_2) \otimes \ldots$. See [38][Lemma 1.2].

IV. Characterization of Strong Duality

To solve (*Dec-C-POMDP*), we define the Lagrangian function $\widehat{L} : \mathcal{M}_1(\mathcal{U}) \times \mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ as follows.

$$\begin{aligned}\widehat{L}(\mu,\lambda) = \widehat{L}^{(P_1,\alpha)}(\mu,\lambda) &\triangleq \widehat{C}(\mu) + \langle\lambda, \widehat{D}(\mu) - \breve{D}\rangle\\ &= \mathbb{E}^{(U\sim\mu)}\underbrace{\left[C(U) + \langle\lambda, D(U) - \breve{D}\rangle\right]}_{\triangleq L^{(P_1,\alpha)}(U,\lambda) = L(U,\lambda)}.\end{aligned} \tag{13}$$

Here, $\mathcal{Y} \triangleq \{\lambda \in \mathbb{R}^K : \lambda \ge 0\}$ is the set of tuples of K non-negative real-numbers, each commonly known as a Lagrange-multiplier. Our main result shows that the the solution $\underline{C}$ satisfies

$$\underline{C} = \inf_{\mu\in\mathcal{M}_1(\mathcal{U})}\sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu,\lambda), \tag{14}$$

and that the inf and sup can be interchanged, i.e.,

$$\underline{C} = \sup_{\lambda\in\mathcal{Y}}\inf_{\mu\in\mathcal{M}_1(\mathcal{U})}\widehat{L}(\mu,\lambda). \tag{15}$$

Theorem 1 (Strong Duality and Existence of Saddle Point)**.** *Under Assumption 1, the following statements hold.*

(a) The optimal value satisfies

$$\underline{C} = \inf_{\mu\in\mathcal{M}_1(\mathcal{U})}\sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu,\lambda). \tag{16}$$

(b) A mixture $\mu^\star \in \mathcal{M}_1(\mathcal{U})$ is optimal if and only if $\underline{C} = \sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu^\star,\lambda)$.

(c) Strong duality holds for (Dec-C-POMDP), i.e.,

$$\underline{C} = \inf_{\mu\in\mathcal{M}_1(\mathcal{U})}\sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu,\lambda) = \sup_{\lambda\in\mathcal{Y}}\inf_{\mu\in\mathcal{M}_1(\mathcal{U})}\widehat{L}(\mu,\lambda). \tag{17}$$

Moreover, there exists a $\mu^\star \in \mathcal{M}_1(\mathcal{U})$ such that $\underline{C} = \sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu^\star,\lambda)$ and $\mu^\star$ is optimal for (Dec-C-POMDP).

(d) If Assumption 2 holds, then there also exists $\lambda^\star \in \mathcal{Y}$ such that the following saddle-point condition holds for all $(\mu,\lambda) \in \mathcal{M}_1(\mathcal{U}) \times \mathcal{Y}$,

$$\widehat{L}(\mu^\star,\lambda) \le \widehat{L}(\mu^\star,\lambda^\star) = \underline{C} \le \widehat{L}(\mu,\lambda^\star). \tag{18}$$

i.e., $\mu^\star$ minimizes $\widehat{L}(\cdot,\lambda^\star)$ and $\lambda^\star$ maximizes $\widehat{L}(\mu^\star,\cdot)$. In addition to this, the primal dual pair $(\mu^\star,\lambda^\star)$ satisfies the complementary-slackness condition:

$$\langle\lambda^\star, \widehat{D}(\mu^\star) - \breve{D}\rangle = 0. \tag{19}$$

Proof. (a) If $\mu \in \mathcal{M}_1(\mathcal{U})$ is feasible (i.e., it satisfies $\widehat{D}(\mu) \le \breve{D}$), then the sup is obtained by choosing $\lambda = 0$, so

$$\sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu,\lambda) = \widehat{C}(\mu). \tag{20}$$

If $\mu \in \mathcal{M}_1(\mathcal{U})$ is not feasible, then

$$\sup_{\lambda\in\mathcal{Y}}\widehat{L}(\mu,\lambda) = \infty. \tag{21}$$

Indeed, suppose WLOG that the k^{th} constraint is violated, i.e., $\widehat{D}_k(\mu) > \breve{D}_k$, then ∞ can be obtained by choosing λ_k arbitrarily large and setting other λ_k's to 0.

From (20), (21), and our convention that $\underline{C} = \infty$ whenever the feasible-set is empty, it follows that

$$\underline{C} = \inf_{\mu \in \mathcal{M}_1(\mathcal{U})} \sup_{\lambda \in \mathcal{Y}} \widehat{L}(\mu, \lambda). \tag{22}$$

(b) By our convention on the value of $\underline{C}$ (when there is no feasible mixture), $\mu^\star$ is optimal if and only if $\widehat{C}(\mu^\star) = \underline{C}$, i.e., $\sup_{\lambda \in \mathcal{Y}} \widehat{L}(\mu^\star, \lambda) = \underline{C}$.
(c) To establish strong duality, we use [37][Proposition 11] which requires $\mathcal{M}_1(\mathcal{U})$ and $\mathcal{Y}$ to be convex topological spaces, with $\mathcal{M}_1(\mathcal{U})$ being compact as well. It is clear that $\mathcal{Y}$ is convex and we can endow it with the usual subspace topology of $\mathbb{R}^K$. Convexity of $\mathcal{M}_1(\mathcal{U})$ is trivial and its compactness has been ensured in Section III. By definition, $\widehat{L}$ is affine and thus trivially concave in λ. [37][Proposition 8] implies that $\widehat{L}$ is convex in μ and Lemma 2 shows that $\widehat{L}$ is lower semi-continuous[10] in μ. From [37][Proposition 11], it then follows that

$$\inf_{\mu \in \mathcal{M}_1(\mathcal{U})} \sup_{\lambda \in \mathcal{Y}} \widehat{L}(\mu, \lambda) = \sup_{\lambda \in \mathcal{Y}} \inf_{\mu \in \mathcal{M}_1(\mathcal{U})} \widehat{L}(\mu, \lambda),$$

and that there exists $\mu^\star \in \mathcal{M}_1(\mathcal{U})$ such that

$$\sup_{\lambda \in \mathcal{Y}} \widehat{L}(\mu^\star, \lambda) = \inf_{\mu \in \mathcal{M}_1(\mathcal{U})} \sup_{\lambda \in \mathcal{Y}} \widehat{L}(\mu, \lambda).$$

The optimality of $\mu^\star$ is implied by parts (b) and (a).
(d) This follows from Lagrange-multiplier theory.

This concludes the proof. □

Lemma 2 (Lower Semi-Continuity of $\widehat{L}$ on $\mathcal{M}_1(\mathcal{U})$)**.** *Under Assumption 1, $\widehat{L}$ is lower semi-continuous on $\mathcal{M}_1(\mathcal{U})$.*

Proof. Fix $\lambda \in \mathcal{Y}$ and $\mu \in \mathcal{M}_1(\mathcal{U})$. Let $\{\mu_i\}_{i=1}^{\infty}$ be a sequence of measures in $\mathcal{M}_1(\mathcal{U})$ that converges to μ. We want to show

$$\liminf_{i\to\infty} \mathbb{E}^{(U\sim\mu_i)}[L(U,\lambda)] \geq \mathbb{E}^{(U\sim\mu)}[L(U,\lambda)].$$

By Lemma 3, L is point-wise lower semi-continuous on $\mathcal{U}$. Therefore, [37][Proposition 9] applies on $\mathcal{M}_1(\mathcal{U})$ and the above inequality follows. □

Lemma 3 (Lower Semi-Continuity of L on $\mathcal{U}$)**.** *Under Assumption 1, the functions C and D_k's are lower semi-continuous on $\mathcal{U}$. Hence, L is lower semi-continuous on $\mathcal{U}$.*

Proof. We will prove the statement for C. The proof of lower semi-continuity of D_k's is similar. For brevity, let

$$p(u,t,h_t,a_t) = p_{P_1}(u,t,h_t,a_t) \triangleq \mathbb{P}_{P_1}^{(u)}(H_t = h_t, A_t = a_t),$$
$$W(u,t,h_t,a_t) = W_{P_1}(u,t,h_t,a_t)$$
$$\triangleq p(u,t,h_t,a_t)\,\mathbb{E}_{P_1}[c(S_t,A_t)\,|H_t = h_t, A_t = a_t],$$

where we use the convention $0 \cdot \infty = 0$. Then,

$$\begin{aligned}
C(u) &= \mathbb{E}_{P_1}^{(u)}\left[\sum_{t=1}^{\infty} \alpha^{t-1} c(S_t,A_t)\right] \\
&= \mathbb{E}_{P_1}^{(u)}\left[\sum_{t=1}^{\infty} \alpha^{t-1}\left(c(S_t,A_t) - \underline{c}\right)\right] + \sum_{t=1}^{\infty} \alpha^{t-1}\underline{c} \\
&\overset{(a)}{=} \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}_{P_1}^{(u)}[c(S_t,A_t) - \underline{c}] + \sum_{t=1}^{\infty} \alpha^{t-1}\underline{c} \\
&\overset{(b)}{=} \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}_{P_1}^{(u)}[\mathbb{E}_{P_1}[c(S_t,A_t)|H_t,A_t]] \\
&= \sum_{t=1}^{\infty} \sum_{h_t \in \mathcal{H}_t} \sum_{a_t \in \mathcal{A}} \alpha^{t-1} W(u,t,h_t,a_t).
\end{aligned}$$

Here, (a) follows from applying the Monotone-Convergence Theorem to the (increasing non-negative) sequence $\{\sum_{t=1}^{i} \alpha^{t-1}(c(S_t,A_t) - \underline{c})\}_{i=1}^{\infty}$ (see [37][Proposition 1]); and (b) uses the tower property of conditional expectation.[11]

Let $\{{}^{i}u\}_{i=1}^{\infty}$ be a sequence in $\mathcal{U}$ that converges to u. By Fatou's Lemma (see [37][Proposition 3]),

$$\liminf_{i\to\infty} C({}^{i}u) \geq \sum_{t=1}^{\infty} \sum_{h_t \in \mathcal{H}_t} \sum_{a_t \in \mathcal{A}} \alpha^{t-1} \liminf_{i\to\infty} W({}^{i}u,t,h_t,a_t). \tag{23}$$

Following Lemma 4, $p({}^{i}u,t,h_t,a_t) \geq 0$ converges to $p(u,t,h_t,a_t)$. Therefore,

$$\liminf_{i\to\infty} W({}^{i}u,t,h_t,a_t) \geq W(u,t,h_t,a_t). \tag{24}$$

Then, (23) and (24) result in $\liminf_{i\to\infty} C({}^{i}u) \geq C(u)$, which establishes the lower semi-continuity of $C(u)$. □

Lemma 4. *[Limit Probabilities for a converging sequence of policy-profiles] Let $\{{}^{i}u\}_{i=1}^{\infty}$ be a sequence in $\mathcal{U}$ that converges to u. Then, for any $t \in \mathbb{N}$, $h_t \in \mathcal{H}_t$, and $a_t \in \mathcal{A}$,*

$$\lim_{i\to\infty} p({}^{i}u,t,h_t,a_t) = p(u,t,h_t,a_t),$$

where $p(\cdot,t,h_t,a_t) = \mathbb{P}_{P_1}^{(\cdot)}(H_t = h_t, A_t = a_t)$. In other words, for every $t \in \mathbb{N}$, the sequence of measures $\{p({}^{i}u,t,\cdot,\cdot)\}_{i=1}^{\infty}$ converges weakly to $p(u,t,\cdot,\cdot)$.

Proof. Given that ${}^{i}u$ converges to u, by the definition of convergence in product topology, for every $n \in [N]$, ${}^{i}u_t^{(n)}(h_t^{(0)}, h_t^{(n)})$ converges weakly to $u_t^{(n)}(h_t^{(0)}, h_t^{(n)})$. Since $\mathcal{A}^n$ is finite, this means that for each $a^{(n)} \in \mathcal{A}^{(n)}$, ${}^{i}u_t^{(n)}(a^{(n)}|h_t^{(0)}, h_t^{(n)})$ converges to $u_t^{(n)}(a^{(n)}|h_t^{(0)}, h_t^{(n)})$, which further implies that for all $a \in \mathcal{A}$, ${}^{i}u_t(a|h_t)$ converges to $u_t(a|h_t)$. Now, we use forward induction to prove the statement.

1) **Base Case**: For time $t = 1$, let $o_1 \in \mathcal{H}_1 = \mathcal{O}$ and $a_1 \in \mathcal{A}$. We have

$$p({}^{i}u,1,o_1,a_1) = P_1(\mathcal{S},o_1)\,{}^{i}u_1(a_1|o_1) \to p(u,1,o_1,a_1).$$

2) **Induction Step**: Assuming that the statement is true for time t, we show that it is true for time $t+1$. Let $h_{t+1} = (o_{1:t+1}, a_{1:t}) = (h_t, a_t, o_{t+1}) \in \mathcal{H}_{t+1}$ and $a_{t+1} \in \mathcal{A}$. We have

$$\begin{aligned}
p({}^{i}u,t+1,h_{t+1},a_{t+1}) &= p({}^{i}u,t,h_t,a_t) \\
&\times {}^{i}u_{t+1}(a_{t+1}|h_{t+1})\,\mathbb{P}_{P_1}(o_{t+1}|h_t,a_t).
\end{aligned}$$

[10]For definition of lower semi-continuity, see [37][Definition 1].

[11]The conditional expectations $\mathbb{E}_{P_1}[c(S_t,A_t)|H_t,A_t]$ exist and are unique because $c(\cdot,\cdot)$ is bounded from below.

By inductive hypothesis, $p\left({}^{i}u, t, h_t, a_t\right)$ converges to $p\left(u, t, h_t, a_t\right)$, and ${}^{i}u_t\left(a_{t+1}|h_{t+1}\right)$ converges to $u_t\left(a_{t+1}|h_{t+1}\right)$ by assumption. We conclude that $p\left({}^{i}u, t+1, h_{t+1}, a_{t+1}\right)$ converges to $p\left(u, t+1, h_{t+1}, a_{t+1}\right)$.

This completes the proof. □

V. Conclusion

In this work, we studied a (cooperative) decentralized constrained POMDP in the setting of infinite-horizon expected total discounted costs. We established strong duality and existence of a saddle point using an extension of Sion's Minimax Theorem which required giving a suitable topology to the space of all possible policy-profiles and then establishing lower semi-continuity of the Lagrangian function. The strong duality result provides a firm theoretical footing for future development of primal-dual type planning and learning algorithms for Dec-C-POMDPs—see [39] for one such algorithm.

VI. ACKNOWLEDGMENTS

This work was funded by NSF via grants ECCS2038416, EPCN1608361, EARS1516075, CNS1955777, CCF2008130, and CMMI2240981 for V. Subramanian, and grants EARS1516075, CNS1955777, CCF2008130, and CMMI2240981 for N. Khan. The authors would also like to thank Dr. Hsu Kao for helpful discussions.

References

[1] R. Bellman, "A Markovian decision process," Journal of Mathematics and Mechanics, vol. 6, no. 5, pp. 679–684, 1957.
[2] K. J. Astrom, "Optimal control of Markov processes with incomplete state information," Journal of Mathematical Analysis and Applications, vol. 10, pp. 174–205, 1965.
[3] R. A. Howard, Dynamic Programming and Markov Processes. Cambridge, MA: MIT Press, 1960.
[4] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," Operations research, vol. 21, no. 5, pp. 1071–1088, 1973.
[5] E. J. Sondik, "The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs," Operations research, vol. 26, no. 2, pp. 282–304, 1978.
[6] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," Artificial Intelligence, vol. 101, no. 1, pp. 99–134, 1998.
[7] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. The MIT Press, second ed., 2018.
[8] F. A. Oliehoek and C. Amato, "A concise introduction to decentralized POMDPs," in SpringerBriefs in Intelligent Systems, 2016.
[9] E. Altman, "Denumerable constrained Markov decision processes and finite approximations," Mathematics of Operations Research, vol. 19, no. 1, pp. 169–191, 1994.
[10] E. Altman, "Constrained Markov decision processes with total cost criteria: Occupation measures and primal lp," Mathematical Methods of Operations Research, vol. 43, pp. 45–72, Feb 1996.
[11] E. A. Feinberg, "Constrained Semi-Markov decision processes with average rewards," Zeitschrift für Operations Research, vol. 39, pp. 257–288, Oct 1994.
[12] E. Feinberg and A. Shwartz, "Constrained discounted dynamic programming," Mathematics of Operations Research, vol. 21, 11 1995.
[13] E. A. Feinberg and A. Shwartz, "Constrained discounted dynamic programming," Mathematics of Operations Research, vol. 21, no. 4, pp. 922–945, 1996.
[14] E. A. Feinberg, "Constrained discounted Markov decision processes and hamiltonian cycles," Mathematics of Operations Research, vol. 25, no. 1, pp. 130–140, 2000.
[15] E. A. Feinberg, A. Jaśkiewicz, and A. S. Nowak, "Constrained discounted Markov decision processes with borel state spaces," Automatica, vol. 111, p. 108582, 2020.
[16] E. Altman, Constrained Markov Decision Processes. Chapman and Hall, 1999.
[17] V. S. Borkar, "An actor-critic algorithm for constrained Markov decision processes," Syst. Control. Lett., vol. 54, pp. 207–213, 2005.
[18] S. Bhatnagar, "An actor-critic algorithm with function approximation for discounted cost constrained Markov decision processes," Syst. Control. Lett., vol. 59, pp. 760–766, 2010.
[19] S. Bhatnagar and K. Lakshmanan, "An online actor–critic algorithm with function approximation for constrained Markov decision processes," Journal of Optimization Theory and Applications, vol. 153, pp. 688 – 708, 2012.
[20] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for infinite-horizon average-reward constrained Markov decision processes," in AAAI Conference on Artificial Intelligence, 2022.
[21] H. Wei, X. Liu, and L. Ying, "Triple-Q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation," in AISTATS, vol. 151, pp. 3274–3307, Mar 2022.
[22] A. Bura, A. HasanzadeZonuzy, D. Kalathil, S. Shakkottai, and J.-F. Chamberland, "DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning." https://arxiv.org/abs/2112.00885?context=cs.AI, 2021.
[23] S. Vaswani, L. Yang, and C. Szepesvari, "Near-optimal sample complexity bounds for constrained MDPs," in NeurIPS, 2022.
[24] D. Kim, J. Lee, K.-E. Kim, and P. Poupart, "Point-Based Value Iteration for Constrained POMDPs," in IJCAI, p. 1968–1974, 2011.
[25] J. Lee, G.-h. Kim, P. Poupart, and K.-E. Kim, "Monte-Carlo tree search for constrained POMDPs," in NeurIPS, vol. 31, 2018.
[26] A. Undurti and J. P. How, "An online algorithm for constrained POMDPs," in ICRA, pp. 3966–3973, 2010.
[27] A. Jamgochian, A. Corso, and M. J. Kochenderfer, "Online planning for constrained POMDPs with continuous spaces through dual ascent." https://arxiv.org/abs/2212.12154, 2022.
[28] D. S. Bernstein, S. Zilberstein, and N. Immerman, "The complexity of decentralized control of markov decision processes," in UAI, p. 32–37, 2000.
[29] H. S. Witsenhausen, "On the structure of real-time source coders," Bell System Technical Journal, vol. 58, no. 6, pp. 1437–1451, 1979.
[30] H. S. Witsenhausen, "A standard form for sequential stochastic control," Mathematical systems theory, vol. 7, no. 1, pp. 5–11, 1973.
[31] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," IEEE Trans. Automatic Control, vol. 58, no. 7, pp. 1644–1658, 2013.
[32] A. Nayyar, A. Mahajan, and D. Teneketzis, The Common-Information Approach to Decentralized Stochastic Control, pp. 123–156. Cham: Springer International Publishing, 2014.
[33] H. Kao and V. Subramanian, "Common information based approximate state representations in multi-agent reinforcement learning," in AISTATS, vol. 151, pp. 6947–6967, PMLR, 28–30 Mar 2022.
[34] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in AAMAS, (Cham), pp. 66–83, Springer International Publishing, 2017.
[35] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning," in NeurIPS, vol. 33, pp. 10199–10210, 2020.
[36] C. T. Ionescu Tulcea, "Mesures dans les espaces produits," Lincei–Rend. Sc. fis. mat. e nat., vol. 7, pp. 208–211, 1949.
[37] N. Khan and V. Subramanian, "A Strong Duality Result for Constrained POMDPs with Multiple Cooperative Agents." https://arxiv.org/abs/2303.14932, 2023.
[38] O. Kallenberg, Foundations of modern probability. Probability and its Applications, Springer-Verlag, New York, second ed., 2002.
[39] N. Khan and V. Subramanian, "Cooperative Multi-Agent Constrained Pomdps: Strong Duality and Primal-Dual Reinforcement Learning with Approximate Information States," 2023.