Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

On Optimal Control of Discounted Cost Infinite-Horizon Markov Decision Processes Under Local State Information Structures

Guanze Peng * **Veeraruna Kavitha** ** **Quanyan Zhu** ***

* *Department of Electrical and Computer Engineering, New York University, NY, 11201, USA. (E-mail: guanze.peng@nyu.edu)*
** *Industrial Engineering and Operations Research Department, Indian Institute of Technology, Bombay, India. (E-mail: vkavitha@iitb.ac.in)*
*** *Department of Electrical and Computer Engineering, New York University, NY, 11201, USA. (E-mail: quanyan.zhu@nyu.edu).*

Abstract: This paper investigates a class of optimal control problems associated with Markov processes with local state information. The decision-maker has only a local access to a subset of a state vector information as often encountered in decentralized control problems in multi-agent systems. Under this information structure, part of the state vector cannot be observed. We leverage ab initio principles and find a new form of Bellman equations to characterize the optimal policies of the control problem under local information structures. The dynamic programming solutions feature a mixture of dynamics associated unobservable state components and the local state-feedback policy based on the observable local information. We further characterize the optimal local-state feedback policy using linear programming methods. To reduce the computational complexity of the optimal policy, we propose an approximate algorithm based on virtual beliefs to find a sub-optimal policy. We show the performance bounds on the sub-optimal solution and corroborate the results with numerical case studies.

Keywords: Partially Observable Markov Decision Process, Bellman Equation, Linear Programming, Approximate Algorithms, Distributed Control.

1. INTRODUCTION

The subject of Markov decision process (MDP) has been broadly explored in the area of robotics, wireless communication, and economics. In MDPs, the decision-maker is assumed to have complete state information. Notwithstanding, in many real world application, the direct observation of the state is either impossible or difficult to acquire (See Sharma and Sutanto (1997), Peng and Zhu (2019), Huang et al. (2019)). Therefore, partially observable Markov decision process (POMDP) becomes a standard framework where the decision-maker does not have direct access to the state information but indirect observations that are correlated with the true state. A substantial literature has been established over the past few decades, including Puterman (2014); Altman (1999); Krishnamurthy (2016); Sondik (1978).

In standard POMDPs, the state information as a whole is taken as incompletely observable and the observations are statistically dependent on the state. In this work, we consider a class of problems where the state takes the form of a vector and its information can be partitioned into two components. One component contains a subset of states that are completely observable while the other component consists of a subset of states that are completely unobservable. This class of problem often arises from distributed multi-agent control systems, where one agent can only observe his own state while the state information of the others are not observable. We refer this class of problems as MDP under Local State Information or LSI-MDP, in short.

One difference between this class of problems and the classical POMDPs is that decision-maker of LSI-MDP has no information of a subset of states. As a result, the optimal control policy of the decision-maker takes the form of local-state feedback, which depends solely on the observable components of the state vector. We use two examples to motivate the LSI-MDP model as follows.

(1) *Team Optimization Problem and Multiagent System Problems*
In both team optimization problem and multiagent systems, multiple agents make decisions based on their observations to optimize their objective functions, and the decisions can impact the state of the system, which is the aggregation of states of all the agents (See Singh (1994), Gupta et al. (2015)). If their objective functions are fully aligned, the problem becomes a team problem. If their objective functions are partially aligned with each other, the problem becomes a nonzero-sum game problem. Our work studies this problem from the perspective of a single agent in which the agent knows his own state but has no access to the states of other agents.
(2) *Optimal Planning in Robotics*
The robots plan the route or actions based on the

observations or information it acquires (See Kaelbling et al. (1998); Parkan and Wu (1999)). Nevertheless, due to the physical limitation of the sensors, there is no guarantee that the robots are capable of obtaining the complete observation of the state (See Sharma and Sutanto (1997)). Hence, the state can be divided into two parts: one part is observable and the other part is unobservable. As the unobservable part of the state is also influenced by the actions, this scenario coincides with our model.

Specifically, our contributions can be summarized as follows:

- We formulate an LSI-MDP problem and characterize the local-state feedback policy using the principle of optimality. We identify the connections with MDPs and POMDPs.
- We show that the local-state feedback policy is characterized by a mixture of open-loop deterministic nonlinear system dynamics and a feedback solution arising from dynamic programming.
- We develop a method termed as *Virtual Belief Method* to provide a suboptimal stationary local feedback policy. We can show that the worst-case performance degradation is bounded.

This paper is organized as the following. In Section II, we present the problem formulation and identify the relations of our framework with MDPs and POMDPs. In Section III, we use the principle of optimality to establish the associated Bellman-like equation. In Section IV, we propose a method to find suboptimal solutions. In Section V, we study several special cases regarding the structures of the system dynamics, cost function, and transition probabilities. It is shown that under some of the special cases, the method proposed in Section IV can yield optimal solution. In Section VI, we conclude this work and give possible directions of the future work.

2. PROBLEM FORMULATION

In this section, we present the problem formulation of the infinite-horizon discounted cost optimal control problem under local-state information. Let $\mathcal{A}$ be the finite action space and $\mathcal{X}$ be the finite state space. The state of the dynamical system is assumed to be a joint process of two *substates*: *observable substate* and *unobservable substate*. The *observable substate*, which is denoted by $x_{\rm o}$, can be observed to the agent directly and utilized for the decision making. The *unobservable substate*, which is denoted by $x_{\rm u}$, cannot be obtained as observations by the agent. Thus, the state space is the Cartesian product of two state subspaces as follows:

$$\mathcal{X} = \mathcal{X}_{\rm u} \times \mathcal{X}_{\rm o},$$

where $\mathcal{X}_{\rm o}$ contains all the possible observable substates $x_{\rm o}$ and $\mathcal{X}_{\rm u}$ contains all the possible unobservable substates $x_{\rm u}$.

The stage cost function is assumed to be a nonnegative and bounded stationary function

$$c(x,a) \ : \ \mathcal{X} \times \mathcal{A} \ \rightarrow \ \mathbb{R}^+.$$

The transition probability is given by a stationary function

$$p(x'|x,a) \ : \ \mathcal{X}_{\rm o} \times \mathcal{X}_{\rm u} \times \mathcal{A} \ \rightarrow \ [0,1]^{|\mathcal{X}|}.$$

More specifically, it can be written as $p(x'_{\rm o}, x'_{\rm u}|x_{\rm o}, x_{\rm u}, a)$.

In this work, we study the following criteria:

$$V^{\pi}_{\alpha_{\rm u}}(x_{\rm o,0}) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t c(x_{{\rm o},t}, x_{{\rm u},t}, a_t)\middle| x_{\rm o,0}, \right], \qquad (1)$$

where $x_{{\rm o},t}$, $x_{{\rm u},t}$, and a_t are the observable substate, unobservable substate, and action at time t, respectively. We aim to determine the policy which minimizes (1). Here, β is the discount factor and $0 \leq \beta < 1$. The distribution of the initial state $x_{\rm u,0}$ is given by $\alpha_{\rm u}$, π a policy is a collection of the decision rules, and each decision rule is a mapping from the space of the history of states and actions to the action space. The agent only has access to the observable substate $x_{\rm o}$ at each time instant. Therefore, his decision can only be dependent on the observation history formed by $x_{\rm o}$. Formally, denote the state-action history of the original system at time t as

$$h_t = \{x_{{\rm o},t}, x_{{\rm o},t-1}, ..., x_{\rm o,0}, \alpha_{\rm u}, a_{t-1}, a_{t-2}, ..., a_0\},$$

and

$$h_0 = \{x_{\rm o,0}, \alpha_{\rm u}\}.$$

Let $\mathcal{H}_t$ be the space of h_t. By definition, $\pi = \{d_t\}_t$ and $d_t : \mathcal{H}_t \rightarrow \Delta(\mathcal{A})$.

Here, $x_{\rm u}$ can be regarded as unobservable uncertainty in the dynamic system. Thus, to cope with this uncertainty, we have the expectation in (1) that averages out the randomness induced by $x_{\rm u}$.

It is clear that our framework differs from the classical MDPs and POMDPs and there exist close connections between the LSI-MDP framework and these two models. To see this, we let the dynamical system evolve according to the following rule

$$(X'_{\rm o}, X'_{\rm u}) = f(x_{\rm o}, x_{\rm u}, a, \Gamma),$$

where Γ is an exogenous random variable.

Assumption 1. There exists a deterministic function $g(\cdot,\cdot)$ such that at each time instant,

$$X_{{\rm u},t} = g(X_{{\rm o},t}, \tilde{X}_{{\rm u},t}), \qquad (2)$$

where $X_{{\rm o},t}$ and $\tilde{X}_{{\rm u},t}$ are conditionally independent conditioned on a given state-action history.

This assumption means that we can decompose the unobservable into two parts: the first part is correlated with the observable substate and second part is independent of the observable substate. Let $\tilde{\mathcal{X}}_{\rm u}$ be the space of $\tilde{x}_{\rm u}$. Then $g : \mathcal{X}_{\rm o} \times \tilde{\mathcal{X}}_{\rm u} \rightarrow \mathcal{X}_{\rm u}$.

Assumption 2. For every $x_{\rm o} \in \mathcal{X}_{\rm o}$, $g(x_{\rm o}, \tilde{x}_{\rm u})$ is an injective function.

This assumption implies that, given a pair of $(x_{\rm o}, x_{\rm u})$, we can identify the value of $\tilde{x}_{\rm u}$ uniquely.

Next, we use the following theorem to construct a controlled Markov process in which $\{x_{{\rm u},t}\}_t$ is conditionally independent of $\{x_{{\rm o},t}\}_t$.

Theorem 1. Under *Assumptions 1 and 2*, for any given fixed policy, there exists a random process $\{\tilde{X}_{{\rm u},t}\}_t$ which satisfies the following:

a) it evolves (conditionally) independently of $\{X_{{\rm o},t}\}_t$, i.e., such that

$$p(x'_{\rm o}, x'_{\rm u}|x_{\rm o}, x_{\rm u}, a) = p(x'_{\rm o}|x_{\rm o}, a)p(\tilde{x}'_{\rm u}|\tilde{x}_{\rm u}, a),$$

where $x_{\mathrm{u}} = g(x_{\mathrm{o}}, \tilde{x}_{\mathrm{u}})$ and $x'_{\mathrm{u}} = g(x_{\mathrm{o}}, \tilde{x}'_{\mathrm{u}})$; and

b) one can represent the state of the system as $(X_{\mathrm{o},t}, \tilde{X}_{\mathrm{u},t})$ with the same amount of information; at time $t+1$, there exists a deterministic function $\tilde{f}$ such that
$$(X_{\mathrm{o},t+1}, X_{\mathrm{u},t+1}) = \tilde{f}(X_{\mathrm{o},t}, \tilde{X}_{\mathrm{u},t}, a, \Gamma).$$

Proof See Peng et al. (2020). □

In view of the above theorem, in some of the future sections, we focus on the class of MDPs with 'conditionally independent' transition probabilities as in the right hand side of equation (1).

To complete the earlier argument, here we discuss how our model is related to POMDPs and MDPs. In POMDP, the observation and the state are assumed to be statistically correlated. Our system formulation includes the case where $x_{\mathrm{u},t}$ and $x_{\mathrm{o},t}$ are conditionally independent and $x_{\mathrm{u},t}$ provides no information of $x_{\mathrm{u},t}$ at all (once actions are observed). Hence it is a generalization of POMDP. When the relation between x_{o} and x_{u} can be described by a deterministic function, $\bar{g}$, such that $x_{\mathrm{u}} = \bar{g}(x_{\mathrm{o}})$, then our framework reduces to a classical MDP, as x_{o} can represent the system state.

3. DYNAMIC PROGRAMMING WITH BELIEFS

As the substate x_{u} cannot be observed, the agent can form belief over the unobservable state. Denote the belief at time t by $b(x_{\mathrm{u},t})$, which evolves (depending upon observation $x_{o,t+1}$) according to
$$b(x_{\mathrm{u},t+1}) = \frac{\sum_{x_{\mathrm{u},t}} p(x_{\mathrm{o},t+1}, x_{\mathrm{u},t+1} | x_{\mathrm{o},t}, x_{\mathrm{u},t}, a_t) b(x_{\mathrm{u},t})}{\sum_{\hat{x}_{\mathrm{u},t}, \hat{x}_{\mathrm{u},t+1}} p(x_{\mathrm{o},t+1}, \hat{x}_{\mathrm{u},t+1} | x_{\mathrm{o},t}, \hat{x}_{\mathrm{u},t}, a_t) b(\hat{x}_{\mathrm{u},t})}, \tag{3}$$
and
$$b(x_{\mathrm{u},0}) = \alpha_u(x_{\mathrm{u},0}).$$
With a slight abuse of notation, let b_t be the belief vector at time t. Thus, with the state denoted by (x_{o}, b), the system is Markovian. We define the transition function of the belief state as
$$b_{t+1} = T(x_{\mathrm{o},t+1}, x_{\mathrm{o},t}, b_t, a_t). \tag{4}$$
We would like to point out that the belief state acts as a deterministic nonlinear subsystem.

Define
$$\bar{c}(x_{\mathrm{o}}, b, a) = \sum_{x_{\mathrm{u}}} b(x_{\mathrm{u}}) c(x_{\mathrm{o}}, x_{\mathrm{u}}, a).$$
Since x_{u}'s are not observable and we can only form belief over x_{u}. After taking expectation using the belief of x_{u}, we define the new objective function
$$\bar{V}^{\bar{\pi}}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}, b_0) = \mathbb{E}^{\bar{\pi}} \left[\sum_{t=0}^{\infty} \beta^t \bar{c}(x_{\mathrm{o},t}, b_t, a_t) \middle| x_{\mathrm{o},0} \right]. \tag{5}$$
As we have mentioned above, the new system whose state is (x_{o}, b) is Markovian. Denote the set of Markovian deterministic policies in this new system by Π_{MD}. That is, the decision at time t is only dependent on the current state $(x_{\mathrm{o},t}, b_t)$.

For the system whose state is (x_{o}, b), the state-action history at time t is given by
$$\bar{h}_t = \{x_{\mathrm{o},t}, x_{\mathrm{o},t-1}, ..., x_{\mathrm{o},0}, b_t, b_{t-1}, ..., b_0, a_{t-1}, a_{t-2}, ..., a_0\},$$
and
$$\bar{h}_0 = \{x_{\mathrm{o},0}, b_0\}.$$
It is worth noting that $\bar{h}_t$ provides the same information as h_t, as b_t evolves according to the rule (3).

Lemma 1. If a given pair of policies $\pi = \{d_t\}_t$ and $\bar{\pi} = \{\bar{d}_t\}_t$ satisfies that $\bar{d}_t(\bar{h}_t) = d_t(h_t)$, then
$$V^{\pi}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}) = \bar{V}^{\bar{\pi}}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}, b_0).$$

Proof See Peng et al. (2020). □

Theorem 2.
$$\inf_{\pi \in \Pi} V^{\pi}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o}}) = \inf_{\bar{\pi} \in \Pi} \bar{V}^{\bar{\pi}}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}, b_0) = \inf_{\pi_{\mathrm{MD}} \in \Pi_{\mathrm{MD}}} \bar{V}^{\pi_{\mathrm{MD}}}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}, b_0).$$

Proof See Peng et al. (2020). □

Using standard arguments of dynamic programming, we can write down the following Bellman equation:
$$u(x_{\mathrm{o}}, b) = \min_a \Big\{ \bar{c}(x_{\mathrm{o}}, b, a) + \beta \sum_{x'_{\mathrm{o}}, b'} u(x'_{\mathrm{o}}, b') p(x'_{\mathrm{o}}, b' | x_{\mathrm{o}}, b, a) \Big\}. \tag{6}$$

We note that the system can be regarded as a mixture of two subsystems: one is MDP and the other is a nonlinear deterministic system. And the states of these two subsystems are intertwined through the transition probability.

Let $\alpha_{\mathrm{o}}(x_{\mathrm{o}})$ be the distribution of the initial observable substate x_{o}. Moreover, let $\mathcal{B}$ be the *reachable set* of the belief, which contains all the possible belief. If $\mathcal{B}$ is infinite, then the number of constraints of linear programming formed by (6) are also infinite, even for finite state and action spaces. Therefore, solving this optimization problem is challenging using the classical linear programming method.

4. VIRTUAL BELIEF METHOD

In this section, we propose a method called *virtual belief method*, which aims to approximate the system with an MDP and provide a suboptimal solution. We show that this proposed method reduces the complexity and yet guarantees the performance by a bounded term.

In the virtual belief method, the agent is assumed to believe that at each time instant, the system is at x_{u} with probability $b_0(x_{\mathrm{u}})$, which is equal to the distribution of the initial substate $x_{\mathrm{u},0}$. And this belief stays unaltered throughout the whole process.

To proceed, let us formally define the virtual system constructed by this method. The transition probability is in this system given by
$$\tilde{p}(x'_{\mathrm{o}} | x_{\mathrm{o}}, a) = \sum_{x_{\mathrm{u}}} p(x'_{\mathrm{o}} | x_{\mathrm{o}}, x_{\mathrm{u}}, a) b_0(x_{\mathrm{u}}).$$
The new cost function now becomes
$$\tilde{c}(x_{\mathrm{o}}, a) = \sum_{x_{\mathrm{u}}} b_0(x_{\mathrm{u}}) c(x_{\mathrm{o}}, x_{\mathrm{u}}, a).$$
The objective function in the new system is given by
$$\tilde{V}^{\pi}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}) = \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t \tilde{c}(x_{\mathrm{o},t}, a_t) \middle| x_{\mathrm{o},0} \right].$$

It is straightforward to check that the system considered in the virtual belief method is a classical MDP. Following the procedures in Puterman (2014), the Bellman equation associated with the MDP is

$$\tilde{u}(x_{\mathrm{o}}) = \min_a \left\{ \tilde{c}(x_{\mathrm{o}}, a) + \beta \sum_{x'_{\mathrm{o}}} \tilde{u}(x'_{\mathrm{o}}) \tilde{p}(x'_{\mathrm{o}} | x_{\mathrm{o}}, a) \right\}. \quad (7)$$

To solve this MDP, we revisit the method of linear programming. Likewise, we begin with the primal linear programming.

Primal LP′ (Virtual Belief Model)

$$\min_{\tilde{u}(x_{\mathrm{o}})} \sum_{x_{\mathrm{o}}} \tilde{u}\,(x_{\mathrm{o}})\,\alpha_{\mathrm{o}}(x_{\mathrm{o}})$$

$$\text{s.t. } \tilde{c}(x_{\mathrm{o}}, a) + \beta \sum_{x'_{\mathrm{o}}} \tilde{u}(x'_{\mathrm{o}}) \tilde{p}(x'_{\mathrm{o}} | x_{\mathrm{o}}, a) \le \tilde{u}(x_{\mathrm{o}}), \quad \forall\ x_{\mathrm{o}}, a.$$

The corresponding dual LP is given by the following.

Dual LP′

$$\max_{\tilde{y}(x_{\mathrm{o}},a)} \sum_{x_{\mathrm{o}},a} \tilde{y}(x_{\mathrm{o}}, a)\tilde{c}(x_{\mathrm{o}}, a)$$

$$\text{s.t. } \alpha_{\mathrm{o}}(x'_{\mathrm{o}}) + \sum_{x_{\mathrm{o}},a} \beta \tilde{p}(x'_{\mathrm{o}}|x_{\mathrm{o}}, a) y(x_{\mathrm{o}}, a) = \sum_a \tilde{y}(x'_{\mathrm{o}}, a), \quad \forall\ x'_{\mathrm{o}}$$

$$\tilde{y}(x_{\mathrm{o}}, a) \ge 0, \quad \forall\ x_{\mathrm{o}}, a.$$

Both linear programmings above are solvable as they have finite constraints (with finite state and action spaces). To see how the disregard of the evolution of the belief process $\{b_t\}_t$ can deteriorate the optimization performance, we first define the operator acting on $\tilde{u}(x_{\mathrm{o}})$ as follows:

$$\tilde{\mathcal{L}}\tilde{u}(x_{\mathrm{o}}) = \min_a \left\{ \tilde{c}(x_{\mathrm{o}}, a) + \beta \sum_{x'_{\mathrm{o}}, x_{\mathrm{u}}} \tilde{u}(x'_{\mathrm{o}}) \tilde{p}(x'_{\mathrm{o}}|x_{\mathrm{o}}, a) \right\}. \quad (8)$$

To make comparisons, we consider the bellman equation in a full-information setting. In this setting, both x_{o} and x_{u} are available for decision making, resulting in a classical MDP. The objective funtion in this setting is given by

$$V^{\pi}_{\mathrm{f},\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}) = \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t c(x_{\mathrm{o},t}, x_{\mathrm{u},t}, a_t) \middle| x_{\mathrm{o},0} \right].$$

Let the value function in the full-information setting be $v(x_{\mathrm{o}}, x_{\mathrm{u}})$, which satisfies the following fixed-point equation

$$v(x_{\mathrm{o}}, x_{\mathrm{u}}) = \min_a \Big\{ c(x_{\mathrm{o}}, x_{\mathrm{u}}, a) + \beta \sum_{x'_{\mathrm{o}}, x'_{\mathrm{u}}} v(x'_{\mathrm{o}}, x'_{\mathrm{u}}) p(x'_{\mathrm{o}}, x'_{\mathrm{u}} | x_{\mathrm{o}}, x_{\mathrm{u}}, a) \Big\}. \quad (9)$$

Define the operator acting on $\{v(x_{\mathrm{o}}, x_{\mathrm{o}})\}_{x_{\mathrm{o}},x_{\mathrm{u}}}$ as

$$\mathcal{L}v(x_{\mathrm{o}}, x_{\mathrm{u}}) = \min_a \Big\{ c(x_{\mathrm{o}}, x_{\mathrm{u}}, a) + \beta \sum_{x'_{\mathrm{o}}, x'_{\mathrm{u}}} v(x'_{\mathrm{o}}, x'_{\mathrm{u}}) p(x'_{\mathrm{o}}, x'_{\mathrm{u}} | x_{\mathrm{o}}, x_{\mathrm{u}}, a) \Big\}. \quad (10)$$

Also, the fixed-point equation (9) can be transformed to the following linear programming problems.

Primal LP″ (Full information case)

$$\min_v \sum_{x_{\mathrm{o}}} v\,(x_{\mathrm{o}}, x_{\mathrm{u}})\,\alpha_{\mathrm{o}}(x_{\mathrm{o}})\alpha_{\mathrm{u}}(x_{\mathrm{u}})$$

$$\text{s.t. } c(x_{\mathrm{o}}, x_{\mathrm{u}}, a) + \beta \sum_{x'_{\mathrm{o}}, x'_{\mathrm{u}}} v(x'_{\mathrm{o}}, x'_{\mathrm{u}}) p(x'_{\mathrm{o}}, x'_{\mathrm{u}} | x_{\mathrm{o}}, x_{\mathrm{u}}, a) \le v(x_{\mathrm{o}}, x_{\mathrm{u}}), \quad \forall\ x_{\mathrm{o}}, x_{\mathrm{u}}, a.$$

And the corresponding dual LP is given by the following.

Dual LP″

$$\max_{y_{\mathrm{f}}} \sum_{x_{\mathrm{o}},x_{\mathrm{u}},a} y_{\mathrm{f}}(x_{\mathrm{o}}, x_{\mathrm{u}}, , a) c(x_{\mathrm{o}}, x_{\mathrm{u}}, a)$$

$$\text{s.t. } \sum_{x_{\mathrm{o}},x_{\mathrm{u}},a} \beta p(x'_{\mathrm{o}}, x'_{\mathrm{u}} | x_{\mathrm{o}}, x_{\mathrm{u}}, a) y_{\mathrm{f}}(x_{\mathrm{o}}, x_{\mathrm{u}}, a) = \sum_a y_{\mathrm{f}}(x'_{\mathrm{o}}, x'_{\mathrm{u}}, , a) - \alpha_{\mathrm{o}}(x'_{\mathrm{o}})\alpha_{\mathrm{u}}(x'_{\mathrm{u}}),\ \forall\ x'_{\mathrm{o}}, x'_{\mathrm{u}}$$

$$y_{\mathrm{f}}(x_{\mathrm{o}}, x_{\mathrm{u}}, , a) \ge 0, \quad \forall\ x_{\mathrm{o}}, x_{\mathrm{u}}, a.$$

Before we give the main theorem of this section, we present the following two propositions. Puterman (2014).

Proposition 3. The operator defined in (8) is a contraction mapping and it has the following properties:

(1) if $\tilde{u} \ge \tilde{\mathcal{L}}\tilde{u}$, then $\tilde{u} \ge \tilde{u}^*$;
(2) if $\tilde{u} \le \tilde{\mathcal{L}}\tilde{u}$, then $\tilde{u} \le \tilde{u}^*$;
(3) if $\tilde{u} = \tilde{\mathcal{L}}\tilde{u}$, then $\tilde{u} = \tilde{u}^*$.

Proposition 4. The operator defined in (10) is a contraction mapping and it has the following properties:

(1) if $v \ge \mathcal{L}v$, then $v \ge v^*$;
(2) if $v \le \mathcal{L}v$, then $v \le v^*$;
(3) if $v = \mathcal{L}v$, then $v = v^*$.

Theorem 5. (**Comparison between Full information and Virtual Belief models**)
If the transition probability can be decomposed as

$$p(x'_{\mathrm{o}}, x'_{\mathrm{u}} | x_{\mathrm{o}}, x_{\mathrm{u}}, a) = p(x'_{\mathrm{o}}|x_{\mathrm{o}}, a) p(x'_{\mathrm{u}}|x_{\mathrm{u}}, a), \quad (11)$$

then

$$\sup_{x_{\mathrm{o},0}} \left| \inf_{\pi\in\Pi} \tilde{V}^{\pi}_{\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}) - \inf_{\pi\in\Pi} V^{\pi}_{\mathrm{f},\alpha_{\mathrm{u}}}(x_{\mathrm{o},0}) \right|$$

is bounded by

$$C = \max\left\{ \frac{-\underline{C}}{1-\beta}, \frac{\bar{C}}{1-\beta} \right\},$$

where

$$\bar{C} := \max_{x_{\mathrm{o}},x_{\mathrm{u}},x'_{\mathrm{u}},a} \left\{ c(x_{\mathrm{o}}, x_{\mathrm{u}}, a) - c(x, x'_{\mathrm{u}}, a) \right\},$$

and

$$\underline{C} := \min_{x_{\mathrm{o}},x_{\mathrm{u}},x'_{\mathrm{u}},a} \left\{ c(x_{\mathrm{o}}, x_{\mathrm{u}}, a) - c(x, x'_{\mathrm{u}}, a) \right\}.$$

Proof See Peng et al. (2020). □

Remark 1. The theorem above states that, even though the direct observation of x_{u} cannot be obtained, we can still guarantee that the performance is deteriorated at most by a bounded term.

We note that the bound on the difference is dependent on the structure of the cost function with respect to x_u. More explicitly, the bounds depend on the sensitivity of $c(x_\mathrm{o}, x_\mathrm{u}, a)$ with respect to the change in x_u.

Also, we can compare the value function in virtual belief method and the value function define in (5). The comparison results are stated in the following theorem.

Theorem 6. (**Comparison between POMDP and Virtual Belief Models**)
If the transition probability can be decomposed as

$$p(x'_\mathrm{o}, x'_\mathrm{u}|x_\mathrm{o}, x_\mathrm{u}, a) = p(x'_\mathrm{o}|x_\mathrm{o}, a)p(x'_\mathrm{u}|x_\mathrm{u}, a),$$

then

$$\sup_{x_{\mathrm{o},0}} \left| \inf_{\pi\in\Pi} V^\pi_{\alpha_\mathrm{u}}(x_{\mathrm{o},0}) - \inf_{\pi\in\Pi} \tilde{V}^\pi_{\alpha_\mathrm{u}}(x_{\mathrm{o},0}) \right|$$

is bounded by

$$C' = \max\left\{ \frac{-\underline{C}'}{1-\beta}, \frac{\bar{C}'}{1-\beta} \right\},$$

where

$$\bar{C}' := \max_{b,b'\in\mathcal{B},x_\mathrm{o},a} \left\{ \bar{c}(x_\mathrm{o}, b, a) - \bar{c}(x_\mathrm{o}, b', a) \right\},$$

and

$$\underline{C}' := \min_{b,b'\in\mathcal{B},x_\mathrm{o},a} \left\{ \bar{c}(x_\mathrm{o}, b, a) - \bar{c}(x_\mathrm{o}, b', a) \right\}.$$

Proof. The proof of *Theorem 6* largely relies on the proof of *Theorem 5.* □

The results in *Theorem 5* & *6* hold under the assumption that the transition probability can be decomposed. Generally, when this assumption does not hold, even if the cost function does not change significantly with respect to x_u, the results may not hold. In such cases, the update of the belief is required for estimating the evolution of observable part $\{x_{\mathrm{o},t}\}_t$.

5. SPECIAL CASES

In this section, we discuss several special cases concerning the structure of the system dynamics, cost function, and transition probabilities.

5.1 $x_\mathrm{u} = \bar{g}(x_\mathrm{o})$ *or* $\mathcal{X}_\mathrm{u} = \emptyset$

If $x_\mathrm{u} = \bar{g}(x_\mathrm{o})$, the unobservable substate can be fully determined from the observable substate. That is, we can infer the true value of the unobservable state from the observation. Then, the overall state can be fully characterized by x_o. Therefore, x_o is sufficient to represent the overall state of the system. As we mentioned earlier, in this case, the system reduces to MDP. It is straightforward to see that (7) and (9) coincide and thus they yield the same value function. Similar arguments hold for the case where $\mathcal{X}_\mathrm{u} = \emptyset$.

5.2 $\mathcal{X}_\mathrm{o} = \emptyset$

In this case, the system is a deterministic system in which the state can be fully characterized by the belief b. And the approximated optimization faced here is given by

$$\begin{aligned} \min_\pi \ & \sum_{t=0}^{\infty} \beta^t b_t^\mathrm{T} c(a) \\ \text{s.t. } & b_{t+1} = P_\mathrm{u}(a) b_t \end{aligned} \tag{12}$$

Here, with a slight abuse of notation, $c(a) = \{c(:, a)\}$ and $P_\mathrm{u}(a)$ is the transition matrix of x_u for a given action a. The optimization above is a classical nonlinear optimal control problem.

5.3 $c(x_\mathrm{o}, x_\mathrm{u}, a) = c(x_\mathrm{o}, x'_\mathrm{u}, a)$

This case is trivial as the stage cost function is no longer a function of the unobservable substate. Thus, $\bar{C} = \underline{C}$. Here, $\bar{C}$ and $\underline{C}$ are defined in the proof of **Theorem 2**. The two value functions are equal and the virtual belief method loses no performance.

5.4 $p(x'_\mathrm{o}, x'_\mathrm{u}|x_\mathrm{o}, x_\mathrm{u}, a) = p(x'_\mathrm{o}|x_\mathrm{o}, a)p(x'_\mathrm{u}|x_\mathrm{u})$

In this case, the independent random process, $\{x_{\mathrm{u},t}\}_t$, is not controllable and thus evolves independently respect to the actions. By assuming that the transition kernel $p(x'_\mathrm{u}|x_\mathrm{u})$ is ergodic, we denote the stationary measure of x_u by $b_\mathrm{s}(x_\mathrm{u})$ and the corresponding belief vector by b_s. In such that a setting, as there exist stationary measures over x_o and x_u jointly, we can reduce (6) to

$$u_\mathrm{s}(x_\mathrm{o}) = \min_a \left\{ \bar{c}(x_\mathrm{o}, b_\mathrm{s}, a) + \beta \sum_{x'_\mathrm{o}} u_\mathrm{s}(x'_\mathrm{o}) p(x'_\mathrm{o}|x_\mathrm{o}, a) \right\}, \tag{13}$$

which leads to tractable linear programmings as the state space and number of constraints are finite. As for the virtual belief method, if we replace the initial belief vector b_0 with b_s, then it will yield the optimal solution.

6. NUMERICAL EXAMPLE

In this section, we use numerical experiments to demonstrate o results. Consider the following dynamical system:

$$\mathcal{A} = \{0, 1\},\ \mathcal{X}_\mathrm{o} = \{0, 1\},\ \mathcal{X}_\mathrm{u} = \{0, 1\},$$

Let $P_\mathrm{o}(a)$ and $P_\mathrm{u}(a)$ be the transition matrices of the observable substate and unobservable substate, respectively.

$$P_\mathrm{o}(0) = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}, \quad P_\mathrm{o}(1) = \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix},$$

$$P_\mathrm{u}(0) = \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}, \quad P_\mathrm{u}(1) = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}.$$

Let $\mathcal{C}(a) = \{c(x_\mathrm{o}, x_\mathrm{u}, a)\}_{\mathcal{X}_\mathrm{o}\times\mathcal{X}_\mathrm{u}}$ be the cost matrix.

$$\mathcal{C}(0) = \begin{bmatrix} 2 & 0.2 \\ 1 & 0.4 \end{bmatrix}, \quad \text{and} \quad \mathcal{C}(1) = \begin{bmatrix} 0.7 & 0.4 \\ 1 & 0.4 \end{bmatrix}.$$

And the probabilities of $x_{\mathrm{u},0} = 0$ and $x_{\mathrm{o},0} = 0$ are both assumed to be $1/2$. The discount factor is set as $\beta = 1/2$.

First consider the system of full information, i.e., the agent has access to both x_o and x_u. The optimal value found by solving LP is approximately 2.0524. And if x_u cannot be observed, using virtual belief method and we have that the policy $d(x_\mathrm{o}) = 1$, $\forall\ x_\mathrm{o}$. and we obtain the value as 2.3714, which is bounded by C.

Another way to find the find the stationary policy is by solving the following constrained linear programming:

(Constrained) Dual LP*

$$
\begin{aligned}
\min_{y_c} \quad & \sum_{a,x_o,x_u} y_c(x_o,x_u,a)c(x_o,x_u,a) \\
\text{s.t.} \quad & \sum_{a} y_c(x_o',x_u',a) - \alpha_o(x_o')\alpha_u(x_u') \\
& = \sum_{x_o,x_u,a} \beta p(x_o',x_u'|x_o,x_u,a)y_c(x_o,x_u,a), \quad \forall\ x_o',x_u' \\
& y_c(x_o,x_u,a) = y_c(x_o,x_u',a), \quad \forall\ x_o,x_u,x_u',a.
\end{aligned}
$$

Here, $y(x_o, x_u, a)$ is the measure function measuring the frequency of the system visiting the state-action pair (x_o, x_u, a). The constraint arises from the fact that x_u cannot be observed and used in the policy in the system whose state is solely x_o. The corresponding primal LP is given by

Primal LP*

$$
\begin{aligned}
\min_{u_c} & \sum_{x_o,x_u} u_c\left(x_o,x_u\right)\alpha_o(x_o)\alpha_u(x_u) \\
\text{s.t.} & \sum_{x_u} u_c(x_o,x_u) - \sum_{x_u} c(x_o,x_u,a) \\
& \geq \beta \sum_{x_u,x_o',x_u'} u_c(x_o',x_u')p(x_o',x_u'|x_o,x_u,a), \quad \forall\ x_o,a.
\end{aligned}
$$

Note that there exists one-to-one correspondence between the solutions to **Primal LP*** and **Dual LP***. There is no dynamic programming equation associated with **Primal LP***, yet it provides a numerical method to compute the stationary policy. The optimal deterministic stationary found using the constrained LPs is given by $d_{\text{MD}}(x_o) = 1$, $\forall\ x_o$, which yields value 1.8706.

7. CONCLUSIONS AND FUTURE WORKS

7.1 Conclusions

In this paper, we have studied LSI-MDP, which is a Markov decision process with incomplete state information. In this model, the state can be divided into two parts, one of which is observable and the other is unobservable. System of this kind is closely related to MDP and POMDP and we have pointed out their relations. We have shown that directly solving optimal control problem in such systems is challenging using the classical linear programming approach, as the number of decision variables (or constraints) is possibly infinite. We have proposed a new method to tackle this challenge, which provides a suboptimal solution. We have provided bounds on the difference between optimal and sub-optimal solution, under certain separability conditions.

7.2 Future Works

When coping with an optimization problem with uncertainty, the agent can either average the randomness out or be robust to the uncertainty. As a consequence, if we consider the unobservable substate as the source of uncertainty, we can formulate a robust optimal control problem, where the objection is given by the following:

$$
V_R^\pi(x_{o,0}) = \sup_{\mathbf{x}_u} \sum_{t=0}^{\infty} \mathbb{E}^\pi \left[\beta^t c(x_{o,t}, x_{u,t}, a_t) \middle| \{\mathbf{x}_{u,s}\}_{s\leq t} \right].
$$

The agent aims to find an optimal policy while being robust to all the possible trajectories of the unobservable substate, $\mathbf{x}_u$. It is worth noting that the transition probabilities and the cost function share the same uncertainty induced by x_u, which makes the robust problem NP-hard as shown in Mannor et al. (2012); Bagnell et al. (2001).

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers whose comments helped to improve the paper significantly. The research has been partially supported by U.S. National Science Foundation Awards ECCS-1847056, CNS-1544782, and SES-1541164, and grant W911NF-19-1-0041 from U.S. Army Research Office (ARO).

REFERENCES

Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

Bagnell, J.A., Ng, A.Y., and Schneider, J.G. (2001). *Solving uncertain Markov decision processes*. Carnegie Mellon University, the Robotics Institute.

Gupta, A., Yuksel, S., Basar, T., and Langbort, C. (2015). On the existence of optimal policies for a class of static and sequential dynamic teams. *SIAM Journal on Control and Optimization*, 53(3), 1681–1712.

Huang, Y., Kavitha, V., and Zhu, Q. (2019). Continuous-time markov decision processes with controlled observations. *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*.

Kaelbling, L.P., Littman, M.L., and Cassandra, A.R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2), 99–134.

Krishnamurthy, V. (2016). *Partially observed Markov decision processes*. Cambridge University Press.

Mannor, S., Mebel, O., and Xu, H. (2012). Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*.

Parkan, C. and Wu, M.L. (1999). Decision-making and performance measurement models with applications to robot selection. *Computers & Industrial Engineering*, 36(3), 503–523.

Peng, G., Kavitha, V., and Zhu, Q. (2020). On optimal control of discounted cost infinite-horizon markov decision processes under local state information structuress. *arXiv preprint arXiv:2005.03169*.

Peng, G. and Zhu, Q. (2019). Game-theoretic analysis of optimal control and sampling for linear stochastic systems. *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*.

Puterman, M.L. (2014). *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Sharma, R. and Sutanto, H. (1997). A framework for robot motion planning with sensor constraints. *IEEE Transactions on Robotics and Automation*, 13(1), 61–73.

Singh, M.P. (1994). *Multiagent systems*. Springer.

Sondik, E.J. (1978). The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2), 282–304.