Unpredictable Planning Under Partial Observability

Michael Hibbard, Yagiz Savas, Bo Wu, Takashi Tanaka, Ufuk Topcu

Abstract— **We study the problem of synthesizing a controller that maximizes the entropy of a partially observable Markov decision process (POMDP) subject to a constraint on the expected total reward. Such a controller minimizes the predictability of a decision-maker's trajectories while guaranteeing the completion of a task expressed by a reward function. In the first part of the paper, we establish that the problem of synthesizing an entropy-maximizing controller for a POMDP is undecidable. We then show that the maximum entropy of a POMDP is upper bounded by the maximum entropy of its fully observable counterpart. In the second part, we consider the entropy maximization problem over finite-state controllers (FSCs). We first recast the entropy-maximizing FSC synthesis problem for a POMDP as a parameter synthesis problem for a parametric Markov chain (pMC). Then, we show that the maximum entropy of a POMDP is lower bounded by the maximum entropy of this pMC. Finally, we present an algorithm, based on a nonlinear optimization problem, to synthesize an FSC that locally maximizes the entropy of a POMDP over FSCs with the same number of memory states. In numerical examples, we demonstrate the proposed algorithm on motion planning scenarios.**

I. Introduction

A Partially observable Markov decision process (POMDP) model sequential decision-making in stochastic environments with imperfect information and nondeterministic choices [1], [2]. A controller, i.e., a decision rule based on the received imperfect information, resolves the nondeterminism in a POMDP and induces a stochastic process. In this paper, we are interested in synthesizing a controller that induces a stochastic process with maximum entropy among the ones whose realizations accumulate an expected total reward above a given threshold.

Entropy measures the unpredictability of outcomes in a random variable [3]. Following [4], [5], we quantify the unpredictability of realizations in a stochastic process by defining the entropy of the process as the joint entropy of a sequence of random variables. Intuitively, our objective is then to synthesize a controller that induces a process whose realizations accumulate rewards in a way that maximizes the unpredictability to an outside observer.

A controller for a POMDP specifies a probability distribution over available actions for each history of information received from the environment. In the first part of the paper, we investigate whether there exists a controller that induces a stochastic process with maximum entropy, and prove that the existence problem is undecidable. We then show that the maximum entropy of a POMDP is upper bounded by the entropy of its corresponding fully observable counterpart. Since the maximum entropy of a Markov decision process can be computed efficiently [5], the derived inequality provides a practical benchmark to evaluate the performance of a given controller on a POMDP.

A finite-state controller (FSC) for a POMDP specifies a probability distribution over actions for each of its memory states according to the most recent information received from the environment [6]. In this regard, FSCs represent a subset of controllers which may, in general, utilize the whole information history. In the second part of the paper, we investigate the problem of synthesizing an FSC that maximizes the entropy of a POMDP over all FSCs with the same number of memory states. By restricting our attention to FSCs with deterministic memory transitions, we recast the controller synthesis problem for a POMDP as a so-called parameter synthesis problem for a parametric Markov chain (pMC) [7], [8]. We first show that the maximum entropy of a pMC induced from a POMDP by FSCs with deterministic memory transitions is a lower bound on the maximum entropy of the POMDP. We also show that by using a specific memory transition function for FSCs, one can monotonically increase the maximum entropy of the stochastic process induced from a POMDP by increasing the number of memory states in FSCs. Finally, we present an algorithm, based on a nonlinear optimization problem, to synthesize parameters that maximize the entropy of a pMC subject to expected reward constraints.

There are several possible applications of the theoretical framework introduced in this paper. For example, the proposed methods can be used to synthesize a controller for an autonomous agent that carries out a mission in an adversarial environment. In particular, if the agent's sensor measurements are noisy and the mission is defined in terms of a reward function, the synthesized controller leaks the minimum information about the trajectories of the agent to an outside observer while guaranteeing the accumulation of an expected total reward above a desired threshold.

Related Work. A recent study [5] showed that an entropy-maximizing controller for an MDP could be synthesized efficiently by solving a convex optimization problem. In POMDPs, entropy has often been used for active sensing applications [9]–[12], in which an agent seeks to select actions that maximize its information gain from the environment. In object recognition, for example, [12] used entropy as a measure of the certainty a robot had in correctly identifying an object. Although conceptually similar, these applications

All authors are with the Department of Aerospace Engineering and Engineering Mechanics, and the Oden Institute for Computational Engineering and Sciences, University of Texas, Austin, 201 E 24th St, Austin, TX 78712. email: {mwhibbard, yagiz.savas, bwu3, ttanaka, utopcu}@utexas.edu

differ from our own as we seek to maximize the entropy of the trajectories an agent follows rather than maximizing its knowledge of the environment.

In the reinforcement learning literature, the entropy of a controller has been used as a regularization term in an agent's objective to balance the trade-off between exploration and exploitation [13]. As discussed in [14], using a controller with high entropy, an agent can learn a greater variety of admissible methods to complete a task, leading to a greater robustness when subsequently fine-tuned to specific scenarios. In imitation learning [15], a controller with high entropy similarly yields greater robustness when the provided demonstrations are imperfect. Unlike the aforementioned work, here, we aim to synthesize a controller that maximizes the entropy of the induced stochastic process, rather than synthesizing a controller with high entropy.

There is also an extensive amount of research regarding the decidability and computational complexity of synthesizing controllers for POMDPs. Undecidability results for controllers that optimize a variety of value functions are established in [16]. Although obtaining an optimal controller is intractable in general, it has been shown that approximately optimal controllers can be obtained by FSCs [17], [18]. Inspired by these results, in this paper, we also consider FSCs to maximize the entropy of a POMDP. We provide a specific memory transition function that is guaranteed to increase the entropy of an induced stochastic process with increasing number of memory states.

II. Preliminaries

For a set $\mathcal{S}$, we denote its power set and cardinality by $2^{\mathcal{S}}$ and $|\mathcal{S}|$, respectively. The set of all probability distributions on a finite set $\mathcal{S}$, i.e., all functions $f{:}\mathcal{S}{\to}[0,1]$ such that $\sum_{s\in\mathcal{S}} f(s){=}1$, is denoted by $\Delta(\mathcal{S})$. If $\{x_t\}$ is a sequence, a subsequence $(x_k, x_{k+1}, \dots, x_l)$ is denoted by x_k^l. We also write $x^l{:=}(x_1, x_2, \dots, x_l)$. Finally, $\mathbb{N}{=}\{1,2,\ldots\}$, $\mathbb{N}_0{=}\{0,1,2,\ldots\}$ and $\mathbb{R}_{\geq 0}{=}[0,\infty)$.

A. Partially Observable Markov Decision Processes

Definition 1: A *partially observable Markov decision process* (POMDP) is a tuple $\mathcal{M} = (\mathcal{S}, s_I, \mathcal{A}, \mathcal{P}, \mathcal{Z}, \mathcal{O}, \mathcal{R})$ where $\mathcal{S}$ is a finite set of states, $s_I{\in}\mathcal{S}$ is a unique initial state, $\mathcal{A}$ is a finite set of actions, $\mathcal{P}{:}\mathcal{S}{\times}\mathcal{A}{\to}\Delta(\mathcal{S})$ is a transition function, $\mathcal{Z}$ is a finite set of observations, $\mathcal{O}{:}\mathcal{S}{\to}\Delta(\mathcal{Z})$ is an observation function, and $\mathcal{R} : \mathcal{S}{\times}\mathcal{A}{\to}\mathbb{R}_{\geq 0}$ is a reward function.

For simplicity, we assume that all actions $a{\in}\mathcal{A}$ are available in all states $s{\in}\mathcal{S}$. For the ease of notation, we denote the transition probability $\mathcal{P}(s'|s,a)$ and the observation probability $\mathcal{O}(z|s)$ by $\mathcal{P}_{s,a,s'}$ and $\mathcal{O}_{s,z}$, respectively.

For a POMDP $\mathcal{M}$, the *corresponding fully observable* MDP $\mathcal{M}_{fo}$ is obtained by setting $\mathcal{Z}{=}\mathcal{S}$ and $\mathcal{O}_{s,s}{=}1$ for all $s{\in}\mathcal{S}$. A Markov chain (MC) is a fully observable MDP such that $|\mathcal{A}|{=}1$.

A *system history* of length $t{\in}\mathbb{N}$ for a POMDP $\mathcal{M}$ is a sequence $h^t{=}(s_I, a_1, s_2, a_2, s_3, \dots, s_t)$ of states and actions such that $\mathcal{P}_{s_k,a_k,s_{k+1}}{>}0$ for all $k{\geq}1$. We denote the set of all system histories of length t by $\mathcal{H}^t$ and define the set of all system histories as $\mathcal{H}{:=}\cup_{t\in\mathbb{N}}\mathcal{H}^t$.

For any system history $h^t{=}(s_I, a_1, s_2, \dots, s_t)$ of length t, there is an associated *observation history* $o^t{=}(\mathcal{O}(s_I), a_1, \mathcal{O}(s_2), \dots, \mathcal{O}(s_t))$ of length $t{\in}\mathbb{N}$. Note that there are, in general, multiple observation histories that are admissible for a given system history. We denote the collection of all observation histories of length t by $Obs_{\mathcal{M}}^t$ and define the set of all observation histories as $Obs_{\mathcal{M}}{:=}\cup_{t\in\mathbb{N}} Obs_{\mathcal{M}}^t$.

Definition 2: A *controller* π for a POMDP $\mathcal{M}$ is a mapping $\pi : Obs_{\mathcal{M}}{\to}\Delta(\mathcal{A})$. We denote the collection of all controllers by $\Pi(\mathcal{M})$.

The probability that the controller π takes the action $a{\in}\mathcal{A}$ upon receiving the observation history $o^t{\in}Obs_{\mathcal{M}}^t$ is denoted by $\pi(a|o^t)$.

In general, a controller $\pi{\in}\Pi(\mathcal{M})$ may require the use of the entire observation history which can be of an arbitrary length [19]. By restricting controllers to use only the most recent fragment of their observation history, we obtain the special class of controllers known as finite-state controllers [20], [21].

Definition 3: For a POMDP $\mathcal{M}$, a *k-finite-state controller* (k-FSC) is a tuple $\mathcal{C}{=}(Q, q_1, \gamma, \delta)$, where $Q{=}\{q_1, q_2, \dots, q_k\}$ is a finite set of memory states, $q_1{\in}Q$ is the initial memory state, $\gamma{:}Q{\times}\mathcal{Z}{\to}\Delta(\mathcal{A})$ is a decision function and $\delta{:}Q{\times}\mathcal{Z}{\times}\mathcal{A}{\to}\Delta(Q)$ is a memory transition function. We denote the collection of all k-FSCs by $\mathcal{F}_k(\mathcal{M})$.

For a memory state $q{\in}Q$ of a k-FSC $\mathcal{C}$, we denote its set of successor memory states $q'{\in}Q$ by $Succ(q){:=}\{q'{\in}Q | \sum_{z\in\mathcal{Z}} \sum_{a\in\mathcal{A}} \delta(q'|q,z,a){>}0\}$.

Definition 4: A *deterministic k-FSC* $\mathcal{C}{=}(Q, q_0, \gamma, \delta)$ is a k-FSC such that for all $q{\in}Q$, $|\mathrm{Succ}(q)|{=}\ 1$. We denote the collection of all deterministic k-FSCs by $\mathcal{F}_k^{det}(\mathcal{M})$.

An FSC prescribes a probability distribution for both the action selection γ and the memory state update δ based on the most recent observation and the FSC's current memory state. An example of the structure of an FSC is illustrated in Fig. 1 which shows the action selection and memory state update for the initial memory state in a 2-FSC. In Fig. 1, the initial memory state prescribes the distributions for γ and δ according to the two possible observations $\{z_1, z_2\}$ and the two available actions $\{a_1, a_2\}$.

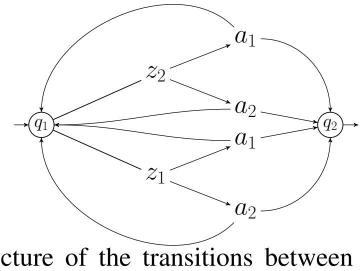

Fig. 1: Structure of the transitions between memory states for the initial memory state of a 2-FSC.

B. Entropy of Stochastic Processes

The *entropy of a random variable* X with a countable support $\mathcal{X}$ and probability mass function (pmf) $p(x)$ is

$$H(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x). \tag{1}$$

We use the convention that 0log0=0. Let (X_1, X_2) be a pair of random variables with the joint pmf $p(x_1, x_2)$ and the support $\mathcal{X}\times\mathcal{X}$. The *joint entropy* of (X_1, X_2) is

$$H(X_1, X_2) := -\sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}} p(x_1, x_2)\log p(x_1, x_2), \tag{2}$$

and the *conditional entropy* of X_2 given X_1 is

$$H(X_2|X_1) := -\sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}} p(x_1, x_2)\log p(x_2|x_1). \tag{3}$$

The definitions of the joint and conditional entropies extend to collections of k random variables as shown in [3]. A discrete *stochastic process* $\mathbb{X}$ is a discrete time-indexed sequence of random variables, i.e., $\mathbb{X}{=}\{X_k{\in}\mathcal{X} : k{\in}\mathbb{N}\}$.

Definition 5: (Entropy of a stochastic process) [22] The *entropy of a stochastic process* $\mathbb{X}$ is defined as

$$H(\mathbb{X}) := \lim_{k\to\infty} H(X^k). \tag{4}$$

Recall that $X^k{:=}(X_1, X_2, \ldots, X_k)$. The above definition is different from the *entropy rate* of a stochastic process, which is defined as $\lim_{k\to\infty}\frac{1}{k}H(X^k)$ when the limit exists [3]. The limit in (4) either converges to a non-negative real number or diverges to positive infinity [22].

For a POMDP $\mathcal{M}$, a controller $\pi{\in}\Pi(\mathcal{M})$ induces a discrete stochastic process $\{X_k{\in}\mathcal{S} : k{\in}\mathbb{N}\}$ in which each X_k is a random variable over the state space $\mathcal{S}$. We denote the entropy of a POMDP $\mathcal{M}$ under a controller $\pi{\in}\Pi(\mathcal{M})$ by $H^{\pi}(\mathcal{M})$.

Definition 6: (Maximum entropy of a POMDP) The *maximum entropy of a POMDP* $\mathcal{M}$ is defined as

$$H(\mathcal{M}) := \sup_{\pi\in\Pi(\mathcal{M})} H^{\pi}(\mathcal{M}). \tag{5}$$

III. Problem Statement and an Upper Bound on Maximum Entropy

We consider an *agent* whose decision-making process is modeled as a POMDP and an *outside observer* whose objective is to infer the states occupied by the agent in the future from the states occupied in the past. Being aware of the observer's objective, the agent aims to synthesize a controller that minimizes the predictability of its future states while ensuring that the expected total reward it collects exceeds a specified threshold.

We measure the predictability of the agent's future states by the entropy of the underlying stochastic process. The rationale behind this choice can be better understood by recalling (see, e.g., Theorem 2.5.1 in [3]) that for any given $n{\in}\mathbb{N}$ and $k{\leq}n$,

$$H(X^n) = H(X_k^n|X^{k-1}) + H(X^{k-1}). \tag{6}$$

Therefore, by maximizing the value of the left hand side of (6), one maximizes the entropy of the all future sequences $(X_n, \ldots, X_k)$ for any history of sequences $(X_{k-1}, \ldots, X_1)$.

We first focus on the problem of maximizing the entropy of a POMDP in the absence of reward constraints.

Problem 1 (Entropy maximization): For a POMDP $\mathcal{M}$, synthesize (if it exists) a controller $\pi^{\star}{\in}\Pi(\mathcal{M})$ such that $\pi^{\star}{\in}\arg\max_{\pi\in\Pi(\mathcal{M})} H^{\pi}(\mathcal{M})$.

For simplicity, we assume that for a given POMDP $\mathcal{M}$, the maximum entropy of the corresponding fully observable MDP $\mathcal{M}_{fo}$ is finite. The validity of this assumption for a given POMDP can be efficiently verified through Algorithm 1 in [5].

Assumption 1: For a POMDP $\mathcal{M}$, corresponding fully observable MDP $\mathcal{M}_{fo}$ satisfies $\sup_{\pi\in\Pi(\mathcal{M}_{fo})} H^{\pi}(\mathcal{M}_{fo}){<}\infty$.

Next, we introduce expected total reward constraints to the framework.

Problem 2 (Constrained entropy maximization): For a POMDP $\mathcal{M}$ and a constant Γ, synthesize a controller $\pi^{\star}{\in}\Pi(\mathcal{M})$ that solves the following problem.

$$\underset{\pi\in\Pi(\mathcal{M})}{\text{maximize}} \quad H^{\pi}(\mathcal{M}) \tag{7a}$$

$$\text{subject to:} \quad \mathbb{E}^{\pi}\Big[\sum_{t=1}^{\infty}\mathcal{R}(S_t, A_t)\Big] \geq \Gamma. \tag{7b}$$

As it is a common practice in the analysis of undiscounted reward models [23], we also assume that the maximum expected total reward that can be collected by the agent is finite even if it has perfect observations.

Assumption 2: For a POMDP $\mathcal{M}$, corresponding fully observable MDP $\mathcal{M}_{fo}$ satisfies $\mathbb{E}^{\pi}\Big[\sum_{t=1}^{\infty}\mathcal{R}(S_t, A_t)\Big]{<}\infty$ for any $\pi{\in}\Pi(\mathcal{M}_{fo})$.

A. Undecidability of the Entropy-Maximizing Controller Synthesis Problem

In this section, we show that the problem of synthesizing a controller that solves the entropy maximization problem is undecidable. Using this result, we then conclude that the constrained entropy maximization problem is also undecidable.

Recall that for a POMDP $\mathcal{M}$, a controller $\pi{\in}\Pi(\mathcal{M})$ induces a stochastic process $\{X_k{\in}\mathcal{S} : k{\in}\mathbb{N}\}$ whose entropy $H^{\pi}(\mathcal{M})$ can be written as

$$H^{\pi}(\mathcal{M}) := \lim_{k\to\infty} H^{\pi}(X^k) = \sum_{t=2}^{\infty} H^{\pi}(X_t|X^{t-1}). \tag{8}$$

Note that since $\mathcal{M}$ has a unique initial state by definition, we have $H(X_1){=}0$. In what follows, we first show that the infinite sum in (8) satisfies a form of recursive Bellman equations. Then, using known results from the literature, we conclude the undecidability of the synthesis problem.

For a given system history $h^t{=}(s_I, a_1, s_2, a_2, s_3, \ldots, s_t)$, let the sequences $s^t{=}(s_I, s_2, s_3, \ldots, s_t)$ and $a^t{=}(a_1, a_2, a_3, \ldots, a_t)$ be the corresponding state and action histories of length t, respectively. We denote the set of all state and action histories of length t by $\mathcal{SH}^t$ and $\mathcal{AH}^t$.

Additionally, we define the set of all possible state and action histories as $\mathcal{SH}{:=}\cup_{t\in\mathbb{N}}\mathcal{SH}^t$ and $\mathcal{AH}{:=}\cup_{t\in\mathbb{N}}\mathcal{AH}^t$.

For a POMDP $\mathcal{M}$ under the controller $\pi{\in}\Pi(\mathcal{M})$, it can be shown that the realization probability $Pr^\pi(s^{t+1}|s^t)$ of the state history $s^{t+1}{\in}\mathcal{SH}^{t+1}$ for a given $s^t{\in}\mathcal{SH}^t$ is

$$Pr^\pi(s^{t+1}|s^t) = \sum_{a^t\in\mathcal{AH}^t}\prod_{k=1}^{t}\mu_k(a_k|h^k)\mathcal{P}_{s_t,a_t,s_{t+1}} \tag{9}$$

where h^k are prefixes of h^t from which the state sequence s^t is obtained, and $\mu_t:\mathcal{H}^t{\to}\Delta(\mathcal{A})$ is a mapping such that

$$\mu_t(a|h^t) := \sum_{o^t\in Obs^t_{\mathcal{M}}}\pi(a|o^t)Pr(o^t|h^t) \tag{10}$$

where the realization probability $Pr(o^t|h^t)$ of the observation history o^t for a given h^t can be recursively written as

$$Pr(o^t|h^t) = \mathcal{O}_{s_t,z_t}\mathcal{P}_{s_{t-1},a_{t-1},s_t}Pr(o^{t-1}|h^{t-1}) \tag{11}$$

for all $t{>}1$ by assuming that $o^1{=}s_I$ with probability 1.

Now, for a given controller $\pi{\in}\Pi(\mathcal{M})$ and a finite constant $T{\in}\mathbb{N}$, let $\mathcal{V}^\pi_{t,T}:\mathcal{SH}^t{\to}\mathbb{R}$ be the *value function* such that

$$\mathcal{V}^\pi_{t,T}(s^t) := \sum_{k=t}^{T}H^\pi(X_{k+1}|X^k_t, X^t=s^t). \tag{12}$$

It is worth noting that

$$\sup_{\pi\in\Pi(\mathcal{M})}H^\pi(\mathcal{M}) = \sup_{\pi\in\Pi(\mathcal{M})}\lim_{T\to\infty}\mathcal{V}^\pi_{1,T}(s_I). \tag{13}$$

Moreover, since $\mathcal{V}^\pi_{t,T}$ is monotonically increasing in T for all $\pi{\in}\Pi(\mathcal{M})$, using Assumption 1, we have

$$\sup_{\pi\in\Pi(\mathcal{M})}\lim_{T\to\infty}\mathcal{V}^\pi_{t,T}(s^t) = \lim_{T\to\infty}\sup_{\pi\in\Pi(\mathcal{M})}\mathcal{V}^\pi_{t,T}(s^t) \tag{14}$$

for all $s^t{\in}\mathcal{SH}^t$.

Lemma 1: For a POMDP $\mathcal{M}$, a controller $\pi{\in}\Pi(\mathcal{M})$ and a finite constant $T{\in}\mathbb{N}$, the value function $\mathcal{V}^\pi_{t,T}$, defined in (12), satisfies the equality

$$\begin{aligned}\mathcal{V}^\pi_{t,T}(s^t) =& H^\pi(X_{t+1}|X^t=s^t)\\ &+\sum_{s^{t+1}\in\mathcal{SH}^{t+1}}Pr^\pi(s^{t+1}|s^t)\mathcal{V}^\pi_{t+1,T}(s^{t+1})\end{aligned} \tag{15}$$

for all $t{<}T$ and $s^t\in\mathcal{SH}^t$.

Proof: See Appendix.□

As a consequence of (14) and Lemma 1, we can now define functions $\mathcal{V}^\star_{t,T}:\mathcal{SH}^t{\to}\mathbb{R}$ for $t{\leq}T$ such that

$$\mathcal{V}^\star_{t,T}(s^t) := \sup_{\pi\in\Pi(\mathcal{M})}\mathcal{V}^\pi_{t,T}(s^t) \tag{16}$$

and conclude that, for all $t{<}T$ and $s^t\in\mathcal{SH}^t$,

$$\begin{aligned}\mathcal{V}^\star_{t,T}(s^t) = \sup_{\pi\in\Pi(\mathcal{M})}\Big[&H^\pi(X_{t+1}|X^t=s^t)\\ &+\sum_{s^{t+1}\in\mathcal{SH}^{t+1}}Pr^\pi(s^{t+1}|s^t)\mathcal{V}^\star_{t+1,T}(s^{t+1})\Big].\end{aligned} \tag{17}$$

By taking the limit of both sides of the above equation as $T{\to}\infty$, and using (13) and (14), we finally conclude that $H(\mathcal{M}){=}\lim_{T\to\infty}\mathcal{V}^\star_{1,T}(s_I)$ satisfies the equations in (17) which are recursive Bellman equations [23]. The following undecidability result is then an immediate consequence of the above derivations.

Theorem 1: Suppose Assumption 1 holds. For a POMDP $\mathcal{M}$, the problem of synthesizing a controller $\pi^\star{\in}\Pi(\mathcal{M})$ such that $\pi^\star{\in}\arg\max_{\pi\in\Pi(\mathcal{M})}H^\pi(\mathcal{M})$ is undecidable.

Proof: Using (17) together with Assumption 1, we can recast the entropy maximization problem as a positive-bounded reward maximization problem where the reward function is $\overline{\mathcal{R}}{:}\mathcal{SH}{\times}\Pi(\mathcal{M}){\to}\mathbb{R}$ such that $\overline{\mathcal{R}}(s^t,\pi){=}H^\pi(X_{t+1}|X^t=s^t)$. Since the problem of synthesizing a controller that maximizes the expected total reward is undecidable by Theorem 4.1 in [16], the result follows.□

Note that Theorem 1 implies that the constrained entropy maximization problem is also undecidable.

B. An Upper Bound on Maximum Entropy

Although we cannot compute the maximum entropy of a POMDP $\mathcal{M}$, we can derive an upper bound on it by relating it to the maximum entropy of the corresponding fully observable MDP $\mathcal{M}_{fo}$.

Recall that for any given controller $\pi{\in}\Pi(\mathcal{M})$ on a POMDP $\mathcal{M}$, we can construct, through (10), a controller $\pi'{\in}\Pi(\mathcal{M}_{fo})$ on the corresponding MDP $\mathcal{M}_{fo}$ which satisfies $Pr^\pi(s^{t+1}|s^t){=}Pr^{\pi'}(s^{t+1}|s^t)$ for all $s^t,s^{t+1}{\in}\mathcal{SH}$. Then, for all $s^t{\in}\mathcal{SH}$, we have

$$\sup_{\pi\in\Pi(\mathcal{M})}H^\pi(X_{t+1}|X^t=s^t) \leq \sup_{\pi\in\Pi(\mathcal{M}_{fo})}H^\pi(X_{t+1}|X^t=s^t).$$

Informally, by having access to the state history s^t, a controller $\pi'{\in}\Pi(\mathcal{M}_{fo})$ can achieve an immediate reward $H^{\pi'}(X_{t+1}|X^t{=}s^t)$ in (17) that is at least as high as the immediate reward achieved by a controller $\pi{\in}\Pi(\mathcal{M})$. Considering the entropy maximization problem as a positive-bounded reward maximization problem, we then conclude the following result.

Theorem 2: Suppose Assumption 1 holds. For a POMDP $\mathcal{M}$ and its corresponding fully observable MDP $\mathcal{M}_{fo}$, we have

$$H(\mathcal{M}) \leq H(\mathcal{M}_{fo}). \tag{18}$$

Proof: See Appendix.□

Whereas the computability of a controller that maximizes the entropy of a POMDP $\mathcal{M}$ is undecidable, we can synthesize a controller $\pi'{\in}\arg\max_{\pi\in\Pi(\mathcal{M}_{fo})}H^\pi(\mathcal{M}_{fo})$ in time polynomial in the size of the corresponding MDP [5]. Therefore, we can efficiently compute $H(\mathcal{M}_{fo})$ to determine an upper bound for the maximum entropy $H(\mathcal{M})$ over all controllers $\pi{\in}\Pi(\mathcal{M})$.

IV. Reformulation using Finite-State Controllers

Since the synthesis problem over general controllers is undecidable, in this section, we consider the entropy maximization problem over deterministic finite-state controllers with fixed numbers of memory states.

Problem 3 (Entropy maximization over FSCs): For a POMDP $\mathcal{M}$, and a constant $k>0$, synthesize (if it exists) a controller $\mathcal{C}^\star \in \mathcal{F}_k^{det}(\mathcal{M})$ such that $\mathcal{C}^\star \in \arg\max_{\mathcal{C}\in\mathcal{F}_k(\mathcal{M})} H^{\mathcal{C}}(\mathcal{M})$.

Problem 4 (Constrained entropy maximization over FSCs): For a POMDP $\mathcal{M}$ and constants $k>0$ and Γ, synthesize (if it exists) a controller $\mathcal{C}^\star \in \mathcal{F}_k^{det}(\mathcal{M})$ that solves the following problem.

$$\underset{\mathcal{C}\in\mathcal{F}_k^{det}(\mathcal{M})}{\text{maximize}} \quad H^{\mathcal{C}}(\mathcal{M}) \tag{19a}$$

$$\text{subject to:} \quad \mathbb{E}^{\mathcal{C}}\Big[\sum_{t=1}^{\infty} \mathcal{R}(S_t, A_t)\Big] \geq \Gamma. \tag{19b}$$

A. A Solution Approach Through Parametric Markov Chains

We develop solution methods to Problems 3 and 4 through the use of parametric Markov chains. Recall that for a POMDP $\mathcal{M}$, a k-FSC $\mathcal{C}\in\mathcal{F}_k(\mathcal{M})$ induces a Markov chain (MC). The collection of all MCs that can be induced from $\mathcal{M}$ by a k-FSC is described by the induced parametric MC which is defined as follows.

Definition 7: For a POMDP $\mathcal{M}$ and a constant $k>0$, the *induced parametric Markov chain* (pMC) is a tuple $\mathcal{D}_{\mathcal{M},k}=(S_{\mathcal{M},k}, s_{I,\mathcal{M},k}, V_{\mathcal{M},k}, P_{\mathcal{M},k})$ where

- $S_{\mathcal{M},k} = \mathcal{S} \times \{1,2,...,k\}$ is the finite set of states,
- $s_{I,\mathcal{M},k} = \langle s_I, 1\rangle$ is the initial state,
- $V_{\mathcal{M},k} = \{\gamma_a^{q,z} | z \in \mathcal{Z}, q \in Q, a \in \mathcal{A}\}$ $\cup \{\delta_{q'}^{q,z,a} | z \in \mathcal{Z}, q, q' \in Q, a \in \mathcal{A}\}$ is the finite set of parameters,
- $P_{\mathcal{M},k} : S_{\mathcal{M},k} \to \Delta(S_{\mathcal{M},k})$ is a transition function such that $P_{\mathcal{M},k}(s'|s) := \sum_{a\in A} \overline{P}(s'|s,a)$ for all $s, s' \in S_{\mathcal{M},k}$ where $\overline{P} : S_{\mathcal{M},k} \times \mathcal{A} \to \Delta(S_{\mathcal{M},k})$ is a mapping such that

$$\overline{P}(\langle s', q'\rangle \mid \langle s, q\rangle, a) := \sum_{z\in\mathcal{Z}} \mathcal{O}_{s,z}\, \mathcal{P}_{s,a,s'}\, \gamma_a^{q,z}\, \delta_{q'}^{q,z,a}. \tag{20}$$

Note that when defining the (parametric) transition probabilities of the induced pMC, we suppose that the observations $\mathcal{O}_{s,z}$ are obtained before selecting actions $a\in\mathcal{A}$. We also remark that different definitions of the induced pMC can be used to reduce the number of parameters in $V_{\mathcal{M},k}$ [21].

Now, an MC can be obtained from the induced pMC by instantiating the parameters $\mathcal{V}_{\mathcal{M},k}$ in a way that the resulting transition function $P_{\mathcal{M},k}$ is well-defined. Formally, let $Z=\{p_1,\dots,p_n\}$ be a finite set of parameters over the domain $\mathbb{R}$, and $\mathbb{Q}[Z]$ be the set of multivariate polynomials over Z. An *instantiation* for Z is a function $u{:}Z\to\mathbb{R}$. Additionally, replacing each parameter p_i in a polynomial $f\in\mathbb{Q}[V]$ by $u(p_i)$ yields $f[u]\in\mathbb{R}$.

Applying an instantiation $u{:}V_{\mathcal{M},k} \to \mathbb{R}$ to the induced pMC $\mathcal{D}_{\mathcal{M},k}$, denoted $\mathcal{D}_{\mathcal{M},k}[u]$, replaces each polynomial $P_{\mathcal{M},k}$ by $P_{\mathcal{M},k}[u]$. An instantiation u is then *well-defined* for $\mathcal{D}_{\mathcal{M},k}$ if the replacement yields probability distributions, i.e., if $\mathcal{D}_{\mathcal{M},k}[u]$ is an MC.

Every well-defined instantiation u describes a k-FSC $\mathcal{C}_u\in\mathcal{F}_k(\mathcal{M})$ [21]. Thus, we can synthesize all admissible MCs that can be induced from a POMDP $\mathcal{M}$ by a k-FSC $\mathcal{C}\in\mathcal{F}_k(\mathcal{M})$ through well-defined instantiations u over $V_{\mathcal{M},k}$. This implies that Problems 3 and 4 can be reduced to a parameter synthesis problem for the induced pMC. In Section V, for a pMC, we present a method to synthesize parameters that induces a stochastic process with maximum entropy whose realizations achieve an expected total reward above a given threshold.

In the next section, we provide two results that allow one to compare the maximum entropy of a POMDP with the maximum entropy of the induced pMC.

B. An Upper Bound and a Monotonocity Result

For a given k-FSC $\mathcal{C}$, let $u_{\mathcal{C}}{:}V_{\mathcal{M},k}\to\mathbb{R}$ be the corresponding instantiation of $\mathcal{D}_{\mathcal{M},k}$ such that $u_{\mathcal{C}}(\gamma_a^{q,z}){:=}\gamma(a|q,z)$ and $u_{\mathcal{C}}(\delta_{q'}^{q,z,a}){:=}\delta(q'|q,z,a)$. Note that $\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}}]$ is a stochastic process. For a given POMDP $\mathcal{M}$ and a constant $k>0$, let the maximum entropy of the induced pMC $\mathcal{D}_{\mathcal{M},k}$ be defined as

$$H(\mathcal{D}_{\mathcal{M},k}) := \sup_{\mathcal{C}\in\mathcal{F}_k^{det}(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}}]). \tag{21}$$

Theorem 3: Suppose Assumption 1 holds. Let $\mathcal{M}$ be a POMDP, $k>0$ be constant, and $\mathcal{D}_{\mathcal{M},k}$ be the induced pMC. Then, the following inequality holds.

$$H(\mathcal{D}_{\mathcal{M},k}) \leq H(\mathcal{M}). \tag{22}$$

Proof: By definition the of $H(\mathcal{D}_{\mathcal{M},k})$, each possible instantiation $u_{\mathcal{C}}$ can only correspond to a deterministic FSC $\mathcal{C}$, i.e., all corresponding FSCs satisfy $|Succ(q)|{=}1$. Then, it can be shown by construction that there is a one-to-one correspondence between the state histories of the instantiated pMC $\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}}]$ and its corresponding POMDP $\mathcal{M}$ under the FSC $\mathcal{C}$. Additionally, for any given instantiation $u_{\mathcal{C}}$, there exists a deterministic FSC $\mathcal{C}\in\mathcal{F}_k^{det}(\mathcal{M})$ which induces the same state history transition function $Pr^{u_{\mathcal{C}}}(s^{t+1}|s^t)$ with $\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}}]$. Therefore, using the result of Lemma 1, we can show that for any instantiated pMC, there exists an FSC that will induce from the POMDP a stochastic process with the same entropy. Because $\mathcal{F}_k^{det}(\mathcal{M}){\subset}\Pi(\mathcal{M})$, it then follows that $H(\mathcal{D}_{\mathcal{M},k}){\leq}H(\mathcal{M})$.□

Theorem 3 implies that by synthesizing a deterministic k-FSC $\mathcal{C}$ such that the instantiation $u_{\mathcal{C}}$ maximizes the entropy of the induced pMC $\mathcal{D}_{\mathcal{M},k}$, we can guarantee that the entropy $\mathcal{H}^{\mathcal{C}}(\mathcal{M})$ of the POMDP $\mathcal{M}$ under the controller $\mathcal{C}$ is at least as high as the entropy of $H(\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}}])$.

We now present a subclass of deterministic k-FSCs, using which we can monotonically increase the maximum entropy of a stochastic process induced from a POMDP by increasing the number of memory states for FSCs.

For a POMDP $\mathcal{M}$, consider a k-FSC $\mathcal{C}=(Q, q_0, \gamma, \delta)$ where the memory transition function $\delta{:}Q\times\mathcal{Z}\times\mathcal{A}\to\Delta(Q)$ is defined as

$$\begin{cases} \delta(q_{i+1}|q_i, z, a) = 1 & \forall z \in \mathcal{Z}, a \in \mathcal{A}, 1 \leq i < k \\ \delta(q_k|q_k, z, a) = 1 & \forall z \in \mathcal{Z}, a \in \mathcal{A} \\ \delta(q_i|q_j, z, a) = 0 & \text{otherwise.} \end{cases} \tag{23}$$

A k-FSC with the memory transition function defined above is shown in Fig. 2. Let $\overline{\mathcal{F}}_k(\mathcal{M}){\subset}\mathcal{F}_k^{det}(\mathcal{M})$ be the set of k-FSCs whose memory transition function is given in (23). Then, we have the following result.

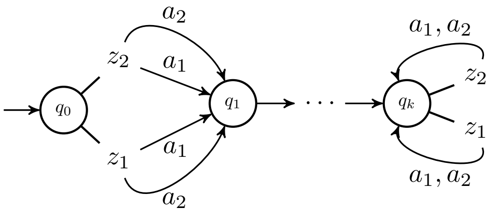

Fig. 2: A deterministic k-FSC example.

Lemma 2: The following inequality holds for all $j{\leq}k$.

$$\sup_{\mathcal{C}\in\overline{\mathcal{F}}_j(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},j}[u_\mathcal{C}]) \leq \sup_{\mathcal{C}\in\overline{\mathcal{F}}_k(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},k}[u_\mathcal{C}]). \quad (24)$$

Proof: See Appendix.□

Based on the result of Lemma 2, we can now set an initial number of memory states for a *deterministic* FSC with the memory transition function (23) and solve Problem 3 to determine the maximum entropy of the induced pMC. Comparing this value to the upper bound given in Theorem 2, we may then iteratively adjust the number of memory states in the FSC to achieve a greater maximum entropy.

V. Finite-State Controller Synthesis

In this section, we present a method to synthesize a deterministic k-FSC that maximizes the entropy of a POMDP over all deterministic k-FSCs whose memory transition function is given in (23).

Recall that for a POMDP $\mathcal{M}$ and a constant $k{>}0$, the induced pMC represents all possible MCs that can be induced from $\mathcal{M}$ by a k-FSC. Additionally, the maximum entropy of the induced pMC provides a lower bound on the maximum entropy of the POMDP due to Theorem 3. Furthermore, by increasing the number of memory states in k-FSCs with transition function given in (23), we can synthesize controllers that improves the entropy of the induced stochastic process.

Using Lemma 1, for a POMDP $\mathcal{M}$ and a constant $k{>}0$, we can write the entropy of an instantiation $u{:}V_{\mathcal{M},k}{\to}\mathbb{R}$ of the induced pMC $\mathcal{D}_{\mathcal{M},k}$, denoted $\mathcal{D}_{\mathcal{M},k}[u]$, as a solution to a form of Bellman equations. Specifically, let $T{\subseteq}S_{\mathcal{M},k}$ be the set of absorbing states in $\mathcal{D}_{\mathcal{M},k}[u]$, i.e., $s{\in}T$ implies that the only successor state of s is itself. Let $P^u_{\mathcal{M},k}{:}S_{\mathcal{M},k}{\to}\Delta(S_{\mathcal{M},k})$ be the transition function of the instantiated pMC such that $P^u_{\mathcal{M},k}(s'|s)$ is defined by replacing parameters $\gamma^{q,z}_a$ and $\delta^{q,z,a}_{q'}$ in (20) with their corresponding instantiations $u(\gamma^{q,z}_a)$ and $u(\delta^{q,z,a}_{q'})$. Additionally, let $L^u{:}S_{\mathcal{M},k}{\to}\mathbb{R}$ be the *local entropy* function such that, for all $s{\in}S_{\mathcal{M},k}$,

$$L^u(s) := -\sum_{s'\in S_{\mathcal{M},k}} P^u_{\mathcal{M},k}(s'|s)\log P^u_{\mathcal{M},k}(s'|s). \quad (25)$$

Using Lemma 1 and Assumption 1, and defining variables $\nu{\in}\mathbb{R}^{|S_{\mathcal{M},k}|}$, it can be shown that the entropy of $\mathcal{D}_{\mathcal{M},k}[u]$ is the unique fixed-point of the system of equations

$$\nu(s) = L^u(s) + \sum_{s'\in S_{\mathcal{M},k}} P^u_{\mathcal{M},k}(s'|s)\nu(s') \;\; s\in S_{\mathcal{M},k}\backslash T \quad (26a)$$

$$\nu(s) = 0 \qquad\qquad s\in T, \quad (26b)$$

such that $H(\mathcal{D}_{\mathcal{M},k}[u]){=}\nu(s_{I,\mathcal{M},k})$. Then, the maximum entropy $H(\mathcal{D}_{\mathcal{M},k})$ of $\mathcal{D}_{\mathcal{M},k}$ can be computed by finding the maximum $\nu(s_{I,\mathcal{M},k})$ that satisfies

$$\nu(s) \leq L^u(s) + \sum_{s'\in S_{\mathcal{M},k}} P^u_{\mathcal{M},k}(s'|s)\nu(s') \;\; s\in S_{\mathcal{M},k}\backslash T \quad (27)$$

together with the condition (26b). Similarly, for the expected total reward constraint, let $\mathcal{R}^u{:}\mathcal{S}_{\mathcal{M},k}{\to}\mathbb{R}$ define the expected immediate rewards on $\mathcal{D}_{\mathcal{M},k}$ such that, for all $s{\in}S_{\mathcal{M},k}$,

$$\mathcal{R}^u(s) := \sum_{s'\in S_{\mathcal{M},k}}\sum_{a\in\mathcal{A}} \overline{P}^u(s'|s,a)\mathcal{R}(s,a) \quad (28)$$

where $\overline{P}^u{:}S_{\mathcal{M},k}\times\mathcal{A}{\to}\Delta(S_{\mathcal{M},k})$ is defined by replacing parameters $\gamma^{q,z}_a$ and $\delta^{q,z,a}_{q'}$ in (20) with their corresponding instantiations $u(\gamma^{q,z}_a)$ and $u(\delta^{q,z,a}_{q'})$. Then, the nonlinear optimization problem to compute the maximum entropy of $\mathcal{D}_{\mathcal{M},k}$ over $\overline{F}_k(\mathcal{M})$ subject to an expected total reward constraint is given as follows.

$$\underset{\nu,u,\eta}{\text{maximize}} \qquad \nu(s_{I,\mathcal{M},k}) \quad (29a)$$

subject to:

$$\nu(s) \leq L^u(s) + \sum_{s'\in S_{\mathcal{M},k}} P^u_{\mathcal{M},k}(s'|s)\nu(s') \;\; \forall\, s\in S_{\mathcal{M},k}\backslash T \quad (29b)$$

$$\nu(s) = 0 \qquad\qquad \forall\, s\in T \quad (29c)$$

$$\eta(s) \leq \mathcal{R}^u(s) + \sum_{s'\in S_{\mathcal{M},k}} P^u_{\mathcal{M},k}(s'|s)\eta(s') \;\; \forall s\in S_{\mathcal{M},k} \quad (29d)$$

$$\eta(s_{I,\mathcal{M},k}) \geq \Gamma \quad (29e)$$

$$\sum_{s'\in S_{\mathcal{M},k}} P^u_{\mathcal{M},k}(s'|s) = 1 \qquad \forall s\in S_{\mathcal{M},k} \quad (29f)$$

$$P^u_{\mathcal{M},k}(s'|s) \geq 0 \qquad \forall s,s'\in S_{\mathcal{M},k} \quad (29g)$$

As previously explained, the constraints in (29b)-(29c) describe a subspace in $\mathbb{R}^{|S_{\mathcal{M},k}|}$ such that the maximum point $\nu(s_{I,\mathcal{M},k})$ of the subspace corresponds to the value of the maximum entropy $H(\mathcal{D}_{\mathcal{M},k})$ of the pMC $\mathcal{D}_{\mathcal{M},k}$. The constraints (29d)-(29e) ensure that the instantiation u satisfies the expected reward constraint given in (19b). Finally, the constraints (29f)-(29g) guarantee that the optimization is performed only over well-defined instantiations u.

Note that in the above optimization problem, $P^u_{\mathcal{M},k}(s'|s)$, $\eta(s')$ and $\nu(s)$ are functions of decision variables. Therefore, the constraints (29b) and (29d) contain bilinear terms. Additionally, $L^u(s)$ is concave in $P^u_{\mathcal{M},k}(s'|s)$, and $\mathcal{R}^u(s)$ is affine in $u(\gamma^{q,z}_a)$ since we consider k-FSCs with fixed memory transitions (23).

To solve the optimization problem (29a)-(29g), we use a variation of convex-concave-procedure (CCP) [24], called *penalty* CCP [25]. In particular, we utilize the parameter synthesis method explained in [7]. Here, we briefly explain the solution approach and refer the reader to [7] for details.

We first represent each bilinear term $f(x)$, e.g., $P^u_{\mathcal{M},k}(s'|s)\nu(s')$, as a difference-of-convex function $f(x){=}f_1(x){-}f_2(x)$ and linearize the concave part $f_2(x)$ around an initial point. This process yields a convex optimization problem. We then introduce nonnegative penalty variables ψ_i to the constraints (29b) and (29d), and replace the objective function with $\nu(s_{I,\mathcal{M},k}) - \tau \sum_i \psi_i$ where τ is a constant regularization parameter. We solve the resulting convex problem and update the initial point with the optimal solution to the convex problem. By iteratively performing the same steps, we obtain, if the procedure converges, a local optimal solution to our original problem (29a)-(29g).

VI. Numerical Examples

In this section, we provide two numerical examples to demonstrate the relation between the maximum entropy of a POMDP, the threshold Γ on the expected total reward, and the number of memory states in FSCs. We use MOSEK [26] solver together with CVX [27] interface to solve the convex optimization problems obtained from the convex-concave procedure. To improve the approximation of exponential cone constraints, we use CVXQUAD [28] package.

A. Relation Between the Maximum Entropy and the Expected Reward Threshold

In the first example, we consider a POMDP with 6 states shown in Fig. 3. There is only one observation $\mathcal{Z} = \{z_1\}$ and therefore the observation function is $\mathcal{O}_{s,z_1}{=}1$ for all states s. We use a deterministic 2-FSC whose memory transition function δ is given in (23). Because there is only one observation, the synthesized controller is an open-loop controller. We suppose that the agent aims to reach state s_4 and encode this objective by defining a reward function $\mathcal{R}$ such that $\mathcal{R}(s_2, a_1){=}\mathcal{R}(s_3, a_1){=}1$ and $\mathcal{R}(s, a){=}0$ otherwise.

We investigate the effect of the threshold Γ in (19b) on the maximum entropy by synthesizing controllers for values between $\Gamma{=}0.5$ and $\Gamma{=}1$. For each value of Γ, we solve the optimization problem given in Section V for 10 times by randomly initializing the convex-concave procedure. For each Γ, we pick the best result of 10 trials, and plot the maximum entropy of the stochastic process induced by the synthesized controllers in Fig. 4. For comparison, we synthesize controllers by solving a feasibility problem given in [7]. We obtain the feasibility problem from (29a)-(29g) by removing the entropy constraint (29b) and replacing the objective function (29a) with a constant value.

In this example, the proposed approach yields the globally optimal controller in all simulations by attaining the tight bound on Γ. Because the feasibility program only seeks to find a feasible instantiation of the parameters that satisfy the expected total reward constraint (29d), the entropy of the stochastic processes it yields is less than the maximum attainable entropy.

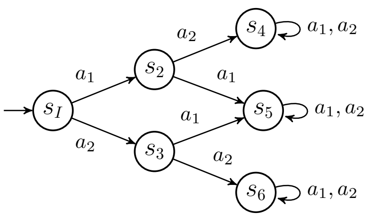

Fig. 3: POMDP illustrating the relation between the maximum entropy and the expected total rewards.

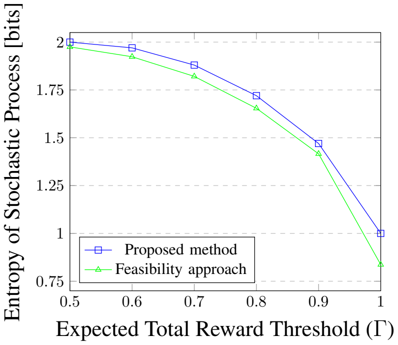

Fig. 4: The trade-off between the maximum entropy and the collected rewards. Feasibility approach synthesizes a feasible FSC that collects expected rewards above the given threshold Γ. Proposed method synthesizes a feasible FSC with maximum entropy.

B. Relation Between the Maximum Entropy and the Number of Memory States

In the second example, we consider a POMDP with 15 states shown in Fig. 5. As in the previous example, there is only a single observation $\mathcal{Z}{=}\{z_1\}$ resulting in $\mathcal{O}_{s,z_1}{=}1$ for all states s. We suppose that the agent aims to reach s_{14} with probability 1. To encode this objective, we set $\Gamma{=}1$ with $\mathcal{R}(s_{10}, a_2){=}1$, $\mathcal{R}(s_{11}, a_2){=}1$, $\mathcal{R}(s_{12}, a_2){=}1$, and $\mathcal{R}(s, a){=}0$ otherwise.

We study the relation between the number of memory states and the maximum entropy of the induced pMC by synthesizing controllers for $k{=}1, \ldots, 6$ memory states. As in the previous example, we run the optimization problem given in Section V for 10 times while randomly initializing the convex-concave procedure. In Fig. 6, we plot the maximum entropy of the stochastic process induced by the controller for each value of k. Furthermore, Fig. 7 shows the entropy-maximizing controller for the POMDP, where edge weights correspond to the probability of action selection.

From Fig. 6, we see that the 1-FSC achieves a maximum entropy of zero. If a 1-FSC selects any action other than a_2, then it cannot reach state s_{14} while collecting an expected total reward of 1. The addition of a memory state allows the agent to randomize its action selection for an additional time step. After 4 memory states, however, additional memory states do not affect the maximum entropy of the induced

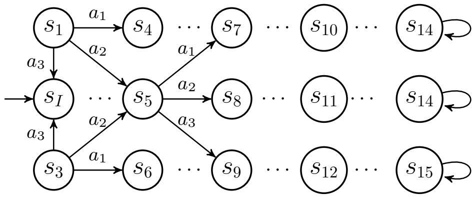

Fig. 5: POMDP illustrating the relation between the maximum entropy and the number of memory states in FSCs.

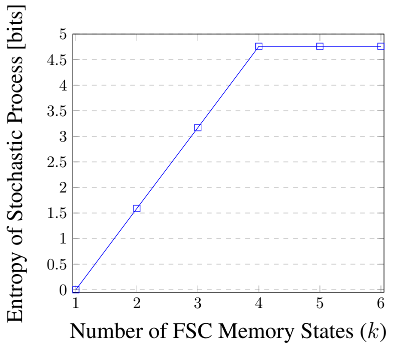

Fig. 6: Comparison between the maximum entropy of the induced stochastic process for varying values of k.

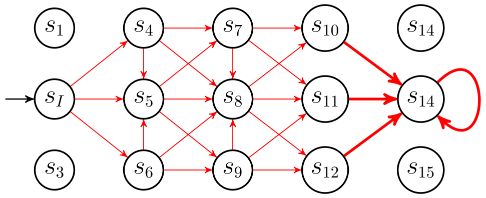

Fig. 7: The trajectories of the synthesized entropy maximizing controller. Edge thicknesses indicate the transition probabilities.

stochastic process. Any controller with at least 4 memory states achieves the globally optimal action distribution shown in Fig. 7. This example demonstrates the monotonicity of the maximum entropy with the number of states in the controller.

VII. Conclusions and Future Extensions

In this paper, we consider an entropy maximization problem in POMDPs subjected to additional expected reward constraints. We first define the entropy in POMDPs and show that general entropy maximization problem is undecidable in POMDPs but the maximal entropy is upper bounded by that of the underlying MDP. Then we consider the entropy maximization problem over deterministic FSC in POMDPs. Such a problem can be translated to parameter synthesis in a parametric Markov chain obtained by the product of the POMDP and FSC. We propose to use penalty CCP to solve such a nonlinear optimization problem. Two examples are presented to show the validity of our proposed approach.

There are several avenues for further research related to this problem. For example, the FSCs we consider are deterministic and assumed to self-loop in the final memory state. An extension of this work would be to optimize over all possible monotonically increasing, deterministic structures for a k-FSC as no structure strictly dominates another in terms of the maximum entropy achievable on the induced pMC. Furthermore, we can also expand our attention to the class of FSCs with nondeterministic transitions between memory states; however, this introduces the complication that the maximum entropy on the pMC is no longer upper bounded by the entropy of the POMDP.

References

[1] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[2] M. J. Kochenderfer, *Decision making under uncertainty: Theory and Application*. MIT press, 2015.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[4] F. Biondi, A. Legay, B. F. Nielsen, and A. Wasowski, "Maximizing entropy over Markov processes," *Journal of Logical and Algebraic Methods in Programming*, vol. 83, no. 5, pp. 384 – 399, 2014.

[5] Y. Savas, M. Ornik, M. Cubuktepe, and U. Topcu, "Entropy maximization for constrained Markov decision processes," in *Allerton Conference on Communication, Control, and Computing*, 2018, pp. 911–918.

[6] P. Poupart and C. Boutilier, "Bounded finite-state controllers," in *Advances in Neural Information Processing Systems*, 2004, pp. 823–830.

[7] M. Cubuktepe, N. Jansen, S. Junges, J.-P. Katoen, and U. Topcu, "Synthesis in pmdps: A tale of 1001 parameters," in *Automated Technology for Verification and Analysis*, S. K. Lahiri and C. Wang, Eds. Cham: Springer International Publishing, 2018, pp. 160–176.

[8] L. Hutschenreiter, C. Baier, and J. Klein, "Parametric Markov chains: PCTL complexity and fraction-free gaussian elimination," *arXiv preprint arXiv:1709.02093*, 2017.

[9] C. Kreucher, K. Kastella, and A. O. Hero Iii, "Sensor management using an active sensing approach," *Signal Processing*, vol. 85, no. 3, pp. 607–624, 2005.

[10] N. Roy, G. Gordon, and S. Thrun, "Finding approximate POMDP solutions through belief compression," *Journal of artificial intelligence research*, vol. 23, pp. 1–40, 2005.

[11] M. Araya, O. Buffet, V. Thomas, and F. Charpillet, "A POMDP extension with belief-dependent rewards," in *Advances in neural information processing systems*, 2010, pp. 64–72.

[12] R. Eidenberger and J. Scharinger, "Active perception and scene modeling by planning with probabilistic 6D object poses," in *International Conference on Intelligent Robots and Systems*, 2010, pp. 1036–1043.

[13] T. Haarnoja, V. Pong, A. Zhou, M. Dalal, P. Abbeel, and S. Levine, "Composable deep reinforcement learning for robotic manipulation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6244–6251.

[14] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1352–1361.

[15] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning." 2008.

[16] O. Madani, S. Hanks, and A. Condon, "On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems," in *AAAI/IAAI*, 1999, pp. 541–548.

[17] C. Amato, D. S. Bernstein, and S. Zilberstein, "Optimizing fixed-size stochastic controllers for pomdps and decentralized pomdps," *Autonomous Agents and Multi-Agent Systems*, vol. 21, no. 3, pp. 293–320, 2010.

[18] K. Chatterjee, L. De Alfaro, and T. A. Henzinger, "Trading memory for randomness," in *International Conference on the Quantitative Evaluation of Systems*, 2004, pp. 206–217.

[19] S. M. Ross, *Introduction to stochastic dynamic programming*. Academic press, 2014.

[20] N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling, "Learning finite-state controllers for partially observable environments," in *Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 427–436.

[21] S. Junges, N. Jansen, R. Wimmer, T. Quatmann, L. Winterer, J.-P. Katoen, and B. Becker, "Permissive finite-state controllers of pomdps using parameter synthesis," *arXiv preprint arXiv:1710.10294*, 2017.

[22] F. Biondi, *Markovian Processes for Quantitative Information Leakage*. PhD thesis, IT University of Copenhagen, 2014.

[23] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[24] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (cccp)," in *Advances in Neural Information Processing Systems*, 2002, pp. 1033–1040.

[25] T. Lipp and S. Boyd, "Variations and extension of the convex-concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, 2016.

[26] M. ApS, *MOSEK Optimizer API for Python. Version 8.1.*, 2019. [Online]. Available: https://docs.mosek.com/8.1/pythonapi/index.html

[27] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[28] H. Fawzi, J. Saunderson, and P. A. Parrilo, "Semidefinite approximations of the matrix logarithm," *Foundations of Computational Mathematics*, 2018, package cvxquad at https://github.com/hfawzi/cvxquad.

VIII. Appendix

Proof of Lemma 1. We prove the claim by strong induction on *t*. For the base case, we have

$$\begin{aligned}
\mathcal{V}^{\pi}_{T-1,T}(s^{T-1}) &= H^{\pi}(X_T|X_{T-1}, X^{T-1}=s^{T-1}) \\
&\quad + H^{\pi}(X_{T+1}|X_T, X^{T-1}=s^{T-1}) && \text{(30a)}\\
&= H^{\pi}(X_T|X^{T-1}=s^{T-1}) \\
&\quad + H^{\pi}(X_{T+1}|X_T, \mathcal{H}^{T-1}=h^{T-1}) && \text{(30b)}\\
&= H^{\pi}(X_T|X^{T-1}=s^{T-1}) \\
&\quad + \sum_{s^T\in\mathcal{SH}^T} Pr^{\pi}(s^T|s^{T-1}) H^{\pi}(X_{T+1}|X^T=s^T) && \text{(30c)}\\
&= H^{\pi}(X_T|X^{T-1}=s^{T-1}) \\
&\quad + \sum_{s^T\in\mathcal{SH}^T} Pr^{\pi}(s^T|s^{T-1}) \mathcal{V}^{\pi}_{T,T}(s^T). && \text{(30d)}
\end{aligned}$$

where (30b) follows from (30a) by the fact that X_{T-1} is a component of S^{T-1}. By the total law of probability and the definition of the state history, we obtain (30c) from (30b). Lastly, (30d) holds by the definition of the value function defined in (15). We now assume that the equality in (15) holds for time steps $T-2, T-3, \dots, t+1$, and show that the equality holds for t.

$$\begin{aligned}
\mathcal{V}^{\pi}_{t,T}(s^t) &= \sum_{k=t}^{T} H^{\pi}(X_{k+1}|X_t^k, X^t=s^t) && \text{(31a)}\\
&= H^{\pi}(X_{t+1}|X_t, X^t=s^t) \\
&\qquad + \sum_{k=t+1}^{T} H^{\pi}(X_{k+1}|X_{t+1}^k, X^t=s^t) && \text{(31b)}\\
&= H^{\pi}(X_{t+1}|X^t=s^t) + \sum_{s^{t+1}\in\mathcal{SH}^t} \sum_{k=t+1}^{T} \dots \\
&\quad H^{\pi} Pr^{\pi}(s^{t+1}|s^t)(X_{k+1}|X_t^k, X^{t+1}=s^{t+1}) && \text{(31c)}\\
&= H^{\pi}(X_{t+1}|X^t=s^t) \\
&\qquad + \sum_{s^{t+1}\in\mathcal{SH}^{t+1}} Pr^{\pi}(s^{t+1}|s^t) \mathcal{V}^{\pi}_{t+1,T}(s^{t+1}). && \text{(31d)}
\end{aligned}$$

As in the base case, (31b) follows from (31a) by the fact that X_t is a component of s^t. We then obtain (31c) from (31b) by the total law of probability and the definition of the state history h^t. Lastly, (31d) holds by the definition of the value function defined in (15). The equality holds for a general t, completing the induction. We may thus write the total expected entropy in this recursive form.

Proof of Theorem 2. We prove the claim by strong induction on t. Denote the value function for $\pi \in \Pi(\mathcal{M})$ as $\mathcal{V}^{\pi}_{t,T}(s^t)$ and the value function for $\pi'\in\Pi(\mathcal{M}_{fo})$ constructed according to (10) as $\mathcal{V}^{\pi'}_{t,T}(s^t)$, respectively. Starting with the base case $t=T$, we have

$$\begin{aligned}
\mathcal{V}^{\pi}_{T,T}(s^T) &= H^{\pi}(X_{T+1}|X^T=s^T) && \text{(32a)}\\
&= H^{\pi'}(X_{T+1}|X^T=s^T) && \text{(32b)}\\
\sup_{\pi\in\Pi(\mathcal{M})} \mathcal{V}^{\pi}_{T,T}(s^T) &\leq \sup_{\pi'\in\Pi(\mathcal{M}_{fo})} H^{\pi'}(X_{T+1}|X^T=s^T) && \text{(32c)}\\
&= \mathcal{V}^{\pi'}_{T,T}(s^T). && \text{(32d)}
\end{aligned}$$

The equality in (32b) follows from the fact that we can construct an equivalent history-dependent controller on the underlying MDP that achieves the same transition probabilities for any observation-based controller. We then obtain (32c) by the fact that $\Pi(\mathcal{M})\subset\Pi(\mathcal{M}_{fo})$. By the definition of the value function in (15), we then obtain (32c).

Now assume that the inequality holds for time steps $T-1, ..., t+1$. We show that it also holds for t as follows. Note first that

$$\begin{aligned}
\mathcal{V}^{\pi}_{t,T}(s^t) &= H^{\pi}(X_{t+1}|X^t=s^t) \\
&\quad + \sum_{s^{t+1}\in\mathcal{SH}^{t+1}} Pr^{\pi}(s^{t+1}|s^t) \mathcal{V}^{\pi}_{t+1,T}(s^{t+1}) && \text{(33a)}\\
&\leq H^{\pi}(X_{t+1}|X^t=s^t) \\
&\quad + \sum_{s^{t+1}\in\mathcal{SH}^{t+1}} Pr^{\pi}(s^{t+1}|s^t) \mathcal{V}^{\pi'}_{t+1,T}(s^{t+1}) && \text{(33b)}\\
&= H^{\pi'}(X_{t+1}|X^t=s^t) \\
&\quad + \sum_{\substack{s^{t+1}\in\\ \mathcal{SH}^{t+1}}} \mathcal{V}^{\pi'}_{t+1,T}(s^{t+1}) Pr^{\pi'}(s^{t+1}|s^{t+1}). && \text{(33c)}
\end{aligned}$$

By Lemma 1, we can write the value function recursively in (33a). The equality in (33b) then follows by the induction hypothesis. By (10), we can construct an equivalent controller on the underlying MDP that has the same transition probabilities. Doing so yields (33c). Then, we have

$$\begin{aligned}
\sup_{\pi\in\Pi(\mathcal{M})} \mathcal{V}^{\pi}_{t,T}(s^t) &\leq \sup_{\pi'\in\Pi(\mathcal{M}_{fo})} H^{\pi'}(X_{t+1}|X^t=s^t) \\
&\qquad + \sum_{\substack{s^{t+1}\in\\ \mathcal{SH}^{t+1})}} Pr^{\pi'}(s^{t+1}|s^t) \mathcal{V}^{\pi'}_{t+1,T}(s^{t+1}) && \text{(34a)}\\
&= \mathcal{V}^{\pi'}_{t,T}(s^t). && \text{(34b)}
\end{aligned}$$

where inequality in (34a) is due to the fact that $\Pi(\mathcal{M})\subset\Pi(\mathcal{M}_{fo})$ and (34b) follows by the definition of the value function in (15). Thus the induction holds for t. Since the claim holds for all t, we have $\mathcal{V}^{\pi}_{1,T}(s_I)\leq\mathcal{V}^{\pi'}_{1,T}(s_I)$. By (8), this implies that $H^{\pi}(X^T)\leq H^{\pi'}(X^T)$ for all T. Taking the limit as $T\rightarrow\infty$ on both sides of the inequality completes the proof.

Proof of Lemma 2. We prove the claim by induction on the number of memory states k. We start with the base case $n{=}1$. Consider an instantiated pMC $\mathcal{D}_{\mathcal{M},1}[u_{\mathcal{C}}]$ for which there exists a corresponding deterministic 1-FSC $\mathcal{C}{\in}\bar{\mathcal{F}}_1(\mathcal{M})$ whose decision function γ satisfies $\gamma(a|q_1,z){=}u_{\mathcal{C}}(\gamma_a^{q_1,z})$. Now, construct a deterministic 2-FSC $\mathcal{C}'$ whose decision function γ' satisfies $\gamma'(a|q_1,z){=}\gamma'(a|q_2,z){=}u_{\mathcal{C}}(\gamma_a^{q_1,z})$. Then, since the memory transitions of both FSCs satisfy (23), there is a one to one correspondence between the state histories of $\mathcal{D}_{\mathcal{M},1}[u_{\mathcal{C}}]$ and $\mathcal{D}_{\mathcal{M},2}[u_{\mathcal{C}'}]$. Using Lemma 1, it can be shown that $H(\mathcal{D}_{\mathcal{M},1}[u_{\mathcal{C}}]){=}H(\mathcal{D}_{\mathcal{M},2}[u_{\mathcal{C}'}])$. Since we choose $\mathcal{C}$ arbitrarily, the maximum entropy of $\mathcal{D}_{\mathcal{M},2}$ cannot be lower than that of $\mathcal{D}_{\mathcal{M},1}$, i.e.,

$$\sup_{\mathcal{C}\in\overline{\mathcal{F}}_1(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},1}[u_{\mathcal{C}}]) \leq \sup_{\mathcal{C}\in\overline{\mathcal{F}}_2(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},2}[u_{\mathcal{C}}]). \tag{35}$$

We assume that the claim holds for $n{=}1,2,...,k-1$, and show that it also holds for $k{=}n$. Consider an instantiated pMC $\mathcal{D}_{\mathcal{M},k-1}[u_{\mathcal{C}}]$ for which there exists a corresponding deterministic $(k-1)$-FSC $\mathcal{C}{\in}\bar{\mathcal{F}}_{k-1}(\mathcal{M})$ whose decision function γ satisfies $\gamma(a|q_i,z){=}u_{\mathcal{C}}(\gamma_a^{q_i,z})$ for $i{=}1,\ldots,k-1$. Then, we can construct an k-FSC $\mathcal{C}'$ whose decision function γ' satisfies $\gamma'(a|q_i,z) := \gamma(a|q_i,z)$ for $i{=}1,\ldots,k-1$, and $\gamma'(a|q_k,z) := \gamma(a|q_{k-1},z)$. Since the memory transitions of both FSCs satisfy (23), there is a one to one correspondence between the state histories of $\mathcal{D}_{\mathcal{M},k-1}[u_{\mathcal{C}}]$ and $\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}'}]$. Using Lemma 1, it can be shown that $H(\mathcal{D}_{\mathcal{M},k-1}[u_{\mathcal{C}}]){=}H(\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}'}])$. Then, since $\mathcal{C}$ is chosen arbitrarily, using the induction hypothesis, we obtain

$$\sup_{\mathcal{C}\in\overline{\mathcal{F}}_j(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},j}[u_{\mathcal{C}}]) \leq \sup_{\mathcal{C}\in\overline{\mathcal{F}}_k(\mathcal{M})} H(\mathcal{D}_{\mathcal{M},k}[u_{\mathcal{C}}]) \tag{36}$$

for all $j{\leq}k$. This completes the proof. □