Least Inferable Policies for Markov Decision Processes

Mustafa O. Karabag[1], Melkior Ornik[2*], Ufuk Topcu[3]

Abstract— **In a variety of applications, an agent's success depends on the knowledge that an adversarial observer has or can gather about the agent's decisions. It is therefore desirable for the agent to achieve a task while reducing the ability of an observer to infer the agent's policy. We consider the task of the agent as a reachability problem in a Markov decision process and study the synthesis of policies that minimize the observer's ability to infer the transition probabilities of the agent between the states of the Markov decision process. We introduce a metric that is based on the Fisher information as a proxy for the information leaked to the observer and using this metric formulate a problem that minimizes expected total information subject to the reachability constraint. We proceed to solve the problem using convex optimization methods. To verify the proposed method, we analyze the relationship between the expected total information and the estimation error of the observer, and show that, for a particular class of Markov decision processes, these two values are inversely proportional.**

I. Introduction

We consider a setting in which an agent is supposed to accomplish a task in a stochastic environment while an observer that is potentially adversarial tries to infer the characteristics of the agent's behavior. For a scenario where predictable behaviors may put the success of the task at risk, it is crucial for an agent to conceal its strategy. In this paper, we study the synthesis of policies that enable an agent to achieve its task while limiting the ability of the observer to infer.

We model the behavior of the agent by a Markov decision process (MDP). The agent follows a policy to achieve its objective, for example, reaching a set of target states with high probability. This policy determines the transition probabilities of the agent between the states of the MDP. The observer can observe the transitions of the agent at a subset of the states and, solely based on the observed transitions, infers the agent's transition probabilities at these states. As a counter objective, the agent aims to choose its policy such that it limits the ability of the observer to infer the transition probabilities in addition to achieving the agent's task with high probability.

A policy can limit the ability of the observer to infer by minimizing the amount of information on the transition probabilities that the observer can gather from each observed transition. We introduce a metric, *transition information*, to measure the amount of information that a single transition leaks to the observer. This metric is related to the Fisher information which measures the amount of information that a random variable has on a parameter [1]. An observer that is trying to estimate the parameter would have high expected estimation error if the random variable has low Fisher information on the parameter. The notion of transition information generalizes the Fisher information by providing a scalar value describing the information leaked for the agent's transition that is parametrized by the transition probabilities.

While the notion of transition information is appropriate for a single observed transition, we also need to consider the effect of the number of observed transitions on the ability of the observer to infer. A policy that solely minimizes the transition information for each observed state adjusts the transition probabilities to the successor states as close to each other as possible since the uniform distribution of the successor states minimizes the transition information for a state. However, this approach might increase the number of observed transitions, and the observer may be able to infer the transition probabilities due to high number of observed transitions. Hence a policy that minimizes the ability of the observer to infer the transition probabilities must also take into account the number of observed transitions and balance the number of observed transitions and the transition information of each observed transition.

We account for the two quantities of interest, the number of observed transitions and the transition information of each observed transition, through a unified notion of *expected total information* — the expected sum of transition informations over a path generated by the agent's policy. We propose to compute a policy that has the minimum expected total information subject to the constraint that the task of the agent is completed with high probability.

To the best of our knowledge, the proposed method is the first policy synthesis method that uses the Fisher information for planning in MDPs against an adversary. The method introduced in [2] uses the Fisher information for learning and control in unknown systems that are modeled by MDPs. However, in contrast to our approach, [2] aims to increase the information gathered from transitions. A-optimality cri-

[1]Department of Electrical and Computer Engineering, University of Texas at Austin, USA. e-mail: karabag@utexas.edu

[2]Department of Aerospace Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. e-mail:mornik@illinois.edu

[3]Department of Aerospace Engineering and Engineering Mechanics and Institute for Computational Engineering and Sciences, University of Texas at Austin. e-mail:utopcu@utexas.edu

*This work was performed while Melkior Ornik was with the Institute for Computational Engineering and Sciences, University of Texas at Austin.

This work was supported in part by DARPA W911NF-16-1-0001, DARPA D19AP00004, and NSF 1652113.

terion [3] for experiment design aims to minimize the total variance of estimators by minimizing the trace of the inverse Fisher information matrix. The transition information is the reciprocal of the trace of inverse Fisher information matrix. In contrast, by minimizing the transition information, we aim to maximize the total variance of estimators unlike A-optimality criterion. In terms of the use of Fisher information, the closest works to the method proposed in this paper are [4] and [5]. The methods introduced in [4] and [5] use the Fisher information to preserve privacy for database systems and smart meters, respectively, and they do not deal with MDPs. Planning in stochastic control settings in the presence of an adversary has been substantially explored previously; the works closest to our paper are [6]–[8]. The reference [7] provides a method for multi-agent perimeter patrolling scenarios and is not applicable to MDPs in general. Papers [6], [8] propose to randomize the policy of an agent by maximizing the entropy of the induced stochastic process. While, for an MDP, increasing the entropy of a process increases randomness of the paths, it does not necessarily limit the ability of an observer to infer the transition probabilities.

The rest of the paper is organized as follows. Section II provides necessary background on the proposed method. In Section III, the definition of information and the problem formulation are presented. Section IV includes the methodology to synthesize the policy that has minimum expected total information subject to a reachability constraint by convex optimization problems. In Section V, we show the relationship between considered problems and estimation errors of the observer. We present numerical examples in Section VI and conclude with suggestions for the future work in Section VII. We provide the proofs for the technical results in [9].

II. Preliminaries

In this section, we present some of the concepts and notation used in the rest of the paper.

We use $[n]$ for the set $\{1,\ldots,n\}$. For a finite set D, we denote the power set with 2^D and cardinality with $|D|$. $\mathbb{E}[\Theta]$ denotes the expectation of the random variable Θ and $\mathrm{Var}(\Theta)$ denotes the variance of Θ which is $\mathbb{E}\left[(\Theta-\mathbb{E}[\Theta])(\Theta-\mathbb{E}[\Theta])^T\right]$. For a discrete random variable X, $Supp(X)$ denotes the set of values for which the probability distribution of X is nonzero. We use $\mathbb{1}_D$ for the indicator function of a set D where $\mathbb{1}_D(x)=1$ if $x\in D$ and $\mathbb{1}_D(x)=0$ otherwise.

A. Markov Decision Processes

A *Markov decision process* is a tuple $\mathcal{M}=(S,\mathcal{A},\mathcal{P},s_0)$ where S is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{P}:S\times\mathcal{A}\times S\to[0,1]$ is the transition probability function, and s_0 is the initial state. We denote $\mathcal{P}(s,a,q)$ by $\mathcal{P}_{s,a,q}$. $\mathcal{A}(s)$ denotes the set of available actions at state s where $\sum_{q\in S}\mathcal{P}_{s,a,q}=1$ for all $a\in\mathcal{A}(s)$. We denote the successor states of state s by $Succ(s)$ such that a state $q\in S$ if and only if there exists an action a such that $\mathcal{P}_{s,a,q}>0$. A state s is *absorbing* if it has only a single successor state that is itself, i.e., $\mathcal{P}_{s,a,s}=1$ for all $a\in\mathcal{A}(s)$.

A *sub-MDP* (C,D) of $\mathcal{M}$ is a pair where $C\subseteq S$ is non-empty and $D:C\to 2^{\mathcal{A}}$ is a function such that $a\ inD(s)$ only if $\mathcal{P}_{s,a,q}=0$ for all $q\notin C$. An *end component* is a sub-MDP (C,D) of $\mathcal{M}$ such that the digraph induced by (C,D) is strongly connected. An end component (C,D) is *closed* if, for all $s\in C$, $Succ(s)\setminus C=\emptyset$. A *maximal end component* (C,D) is an end component where there is no end component (C',D') such that $(C,D)\neq(C',D')$, $C\subseteq C'$, and $D\subseteq D'$.

A *policy* is a sequence $\pi=[\mu_0,\mu_1,\ldots]$ where each $\mu_t: S\times\mathcal{A}\to[0,1]$ is a function such that $\sum_{a\in\mathcal{A}(s)}\mu_t(s,a)=1$ for every $s\in S$. A *stationary policy* π is a a sequence $\pi=[\mu,\mu,\ldots]$ where $\mu: S\times\mathcal{A}\to[0,1]$ is a function such that $\sum_{a\in\mathcal{A}(s)}\mu(s,a)=1$ for every $s\in S$. We denote the set of all policies by $\Pi(\mathcal{M})$ and the set of stationary policies by $\Pi^{St}(\mathcal{M})$. For a stationary policy π, we denote $\mu(s,a)$ by $\pi_{s,a}$. A stationary policy π induces *a Markov chain* $\mathcal{M}^\pi=(S,\mathcal{P}^\pi,s_0)$ from $\mathcal{M}$ where $\mathcal{P}^\pi: S\times S\to[0,1]$ is the transition probability function such that

$$\mathcal{P}^\pi(s,q):=\sum_{a\in\mathcal{A}(s)}\pi_{s,a}\ \mathcal{P}_{s,a,q}$$

for all $s,q\in S$. We denote $\mathcal{P}^\pi(s,q)$ by $\mathcal{P}^\pi_{s,q}$.

A *path* $\xi=s_0s_1s_2\ldots$ is an infinite sequence of states under policy $\pi=[\mu_0,\mu_1,\ldots]$ such that $\sum_{a\in\mathcal{A}(s_t)}\mathcal{P}_{s_t,a,s_{t+1}}\mu_t(s_t,a)>0$ for all $t\geq 0$. The set of paths for $\mathcal{M}$ under policy π is denoted by $Paths(\mathcal{M}^\pi)$.

The reachability probability to the set B of states, i.e., the probability of reaching a state $b\in B$ under policy π, is denoted by $\Pr^\pi(Reach[B])$.

The *expected state residence time* at state s is defined by

$$x^\pi_s:=\sum_{t=0}^{\infty}\Pr(s_t=s|s_0),$$

where s_t is the state at time t. The expected state residence time x^π_s is also equal to $\mathbb{E}[N_{s,\xi}]$ where $N_{s,\xi}$ is the number of appearances of s in the random path ξ that is generated by the policy π. The *expected state-action residence time* at state s and action a is defined by

$$x^\pi_{s,a}:=\sum_{t=0}^{\infty}\Pr(s_t=s|s_0)\mu_t(s_t,a).$$

The expected state-action residence time of a state and an action is the expected number of times that the action is taken at the state. For a stationary policy $\pi\in\Pi^{St}(\mathcal{M})$ and a transient state s, the following equality $x^\pi_{s,a}=\pi_{s,a}\ x^\pi_s$ holds.

B. The Fisher Information and the Cramér-Rao Bound

Let the random variable X represent the observed data from a random variable that is parametrized by $\Theta\in\mathbb{R}^n$. An *estimator* is a function $\hat{\Theta}: X\to\mathbb{R}^n$ that estimates Θ based on observed data. The estimator $\hat{\Theta}$ is an *unbiased estimator* of Θ if $\mathbb{E}[\hat{\Theta}]-\Theta=0$.

The *precision* of a random variable is the reciprocal of the variance of the random variable. For an unbiased estimator, its precision is the reciprocal of the mean squared error (MSE) of the estimator.

The *Fisher information* [10] $I_X(\theta)$ of a discrete random variable X parametrized by $\theta \in \mathbb{R}$ is

$$I_X(\theta) := -\sum_{x \in Supp(X)} \frac{\partial^2 \log(\Pr(X = x|\theta))}{\partial \theta^2} \Pr(X = x|\theta).$$

An important property of the Fisher information is additivity, that is, when the samples are drawn from i.i.d. random variables, the Fisher information based on n samples $I_{X^n}(\theta)$ satisfies $I_{X^n}(\theta) = nI_X(\theta)$ where $I_X(\theta)$ is the Fisher information of one sample.

The *Cramér-Rao inequality* [10] defines a relationship between the variance of an unbiased estimator of parameter θ and the Fisher information on the parameter θ. The inequality is stated as

$$\mathrm{Var}(\hat{\theta}) \geq I_X(\theta)^{-1}$$

where $\hat{\theta}$ is any unbiased estimator of θ.

An unbiased estimator is *efficient* if it achieves the Cramér-Rao bound.

III. Problem Statement

Consider an agent whose behavior is governed by a Markov decision process (MDP) $\mathcal{M} = (S, \mathcal{A}, \mathcal{P}, s_0)$ where a stationary policy followed by the agent π implemented on this MDP induces a Markov chain. An adversary which we call *observer* observes the transitions and tries to infer the transition probabilities for a set W of states in the induced Markov chain. We assume that the observer can only observe the transitions at the states in W which we call *observed states*, and has no side information.

The problem we study is the synthesis of a policy for the agent with two objectives: (i) reach a set C_{reach} of states with probability higher than a given threshold $0 \leq \nu_{reach} \leq 1$ and (ii) minimize the amount of information leaked to the observer.

For the first objective, we assume that the transitions of the agent after reaching C_{reach} are irrelevant, i.e., every state in C_{reach} is absorbing and is not observed.

For the second objective, we define the notion of *transition information* to measure the amount of information leaked to the observer due to a transition.

Definition 1. The *transition information* of a state s is defined by

$$\iota_s^\pi := \frac{1}{\sum_{q \in Succ(s)} I_Q(\mathcal{P}_{s,q}^\pi)^{-1}}$$

where Q is the random variable that is the successor state of state s.

We remark that the Fisher information and the transition information are analogous:

- The reciprocal of the Fisher information is a lower bound on the variance of an unbiased estimator for a single parameter.
- The reciprocal of the transition information is a lower bound on the variance of an unbiased estimator for many parameters.

Under policy π, for a state s, consider an unbiased estimator $\hat{\mathcal{P}}_s$ of transition probabilities. The reciprocal of the transition information ι_s^π is a lower bound on the variance of $\hat{\mathcal{P}}_s$:

$$\mathrm{Var}(\hat{\mathcal{P}}_s) \geq \frac{1}{\iota_s^\pi}.$$

We use the transition information to define the *total information* of a path. The total information of a path $\xi = s_0 s_1 s_2 \ldots$ is defined as the sum of each observed transition's transition information such that

$$\iota_{W,\xi}^\pi := \sum_{t=0}^{\infty} \mathbb{1}_W(s_t) \iota_{s_t}^\pi.$$

We then state the synthesis problem formally as follows:

Problem 1 (Synthesis of Minimum-Information Admissible Policies)**.** Given an MDP $\mathcal{M} = (S, \mathcal{A}, \mathcal{P}, s_0)$, a set C_{reach} of states, a probability threshold ν_{reach}, and the set W of observed states, compute

$$\inf_{\pi \in \Pi^{St}(\mathcal{M})} \quad \mathbb{E}[\iota_{W,\xi}^\pi], \tag{1a}$$

$$\text{s. t.} \quad \Pr{}^\pi(Reach[C_{reach}]) \geq \nu_{reach} \tag{1b}$$

where ξ is a random path generated under policy π. If the optimal value is attainable, compute the optimal policy π^*.

Hereafter we call the policies that satisfy the reachability constraint *admissible policies* and an optimal policy for Problem 1 a *minimum-information admissible policy*.

Example 1. We explain the characteristics of a minimum-information admissible policy through the MDP given in Figure 1 with $\nu_{reach} = 0$ for simplicity.

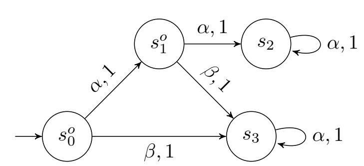


Fig. 1: An MDP with 4 states. A label a, p of a transition refers to the transition that happens with probability p when action a is taken. The states marked with the superscript o are observed.

The goal of the agent is to find a policy that minimizes the expected total information. Consider the policy at state s_1 and note that the policy decision at s_1 does not affect information leaked from state s_0 since it does not change the expected residence time at s_0. Hence we may only consider the information leaked from s_1. If the agent chooses a deterministic policy, the observer can estimate the transition probabilities with no error even after observing a single transition, which means infinite leaked information. Therefore, it is expected that the agent randomizes the transition probabilities. Formally, we explain the reasoning by the fact that the Fisher information is minimized for a $Ber(p)$ random variable with $p = 0.5$. Similarly at state s_0, the agent randomizes the transition probabilities. However, unlike s_1, the policy at s_0 affects the information leaked from s_1. As the agent decreases the probability of taking action α at state s_0, the expected number of visits to state s_1 decreases and consequently information leaked from s_1 decreases. The agent must take the action β with a greater probability than the action α. On the other hand, taking the action β with high probability increases the information leaked from s_0. We expect that, under this trade-off, the agent must choose a policy that takes both actions, but the action β more likely. Numerically, the optimal policy is $\pi_{s_0,\alpha} = 0.38$, $\pi_{s_0,\beta} = 0.62$, $\pi_{s_1,\alpha} = 0.5$, and $\pi_{s_1,\beta} = 0.5$.

Remark. *Note that if the transition probabilities are not constant and change between observations, measurement of inference with a transition information is not meaningful since we assume underlying probability distribution is constant. To have a well-defined problem, we only focus on agents that have to follow stationary policies and we search the optimal policies only in the stationary policies.*

IV. Synthesis of Minimum-Information Admissible Policies

For an MDP $\mathcal{M} = (S, \mathcal{A}, \mathcal{P}, s_0)$, we aim to find a minimum-information admissible policy π that minimizes the expected total information of a path subject to the reachability constraint $\Pr^\pi(Reach[C_{reach}]) \geq \nu_{reach}$ where the set of observed states is W. In this section, we represent the expected total transition transition information in terms of expected state-action residence times, show the existence of a minimum-information admissible policy, and give an optimization problem whose solution is a minimum-information admissible policy. We also show that the proposed optimization problem is convex in the expected state-action residence time parameters and hence can be solved using off-the-shelf convex optimization tools.

Note that the Fisher information for a parameter is well-defined if the regularity conditions are satisfied. These conditions require that the distributions depending on the parameter have a common support that is independent of the parameter [10]. For a random variable $P \sim Ber(p)$, the Fisher information is not defined when $p = 0$ or $p = 1$ since the probability distribution of P does not have a common support. However, such a case practically corresponds to infinite Fisher information, which means that the value of the parameter is estimated exactly even after a single observation. We assume that the Cramér-Rao lower bound is zero if the Fisher information is infinite.

Consider a state $w \in W$ whose successor state is denoted by the random variable Q. For each $q \in Succ(w)$, we have

$$I_Q(\mathcal{P}^\pi_{w,q}) = I_{\mathbb{1}_q(Q)}(\mathcal{P}^\pi_{w,q}) = \left(\mathcal{P}^\pi_{w,q}(1 - \mathcal{P}^\pi_{w,q})\right)^{-1}$$

where $\mathbb{1}_q(Q)$ is a $Ber(\mathcal{P}^\pi_{w,q})$ random variable. The transition information of a state w is equal to

$$\iota^\pi_w = \left(\sum_{q \in Succ(w)} \mathcal{P}^\pi_{w,q}(1 - \mathcal{P}^\pi_{w,q})\right)^{-1} \tag{2}$$

under policy π.

Remark. *The categorical random variable Q has the distribution $\mathcal{P}^\pi_{w,q}$ where $q \in Succ(w)$. The covariance matrix Σ of Q has diagonal entries $\mathcal{P}^\pi_{w,q}(1 - \mathcal{P}^\pi_{w,q})$. The transition information of state w given in (2) is also equal to $\mathrm{tr}(\Sigma)^{-1}$. Since Q is a categorical random variable, a sample mean estimator achieves the Cramér-Rao bound for a single transition. However, since the observed data consists of transitions from a path and the transitions are not independent in general, a sample mean estimator is not necessarily unbiased and efficient.*

We now construct the optimization problem whose solution gives the expected state-action residence times for a minimum-information admissible policy. First, we rewrite (2) as

$$\iota^\pi_w = \left(\sum_{q \in Succ(w)} \left(\sum_{a \in \mathcal{A}(w)} \frac{x^\pi_{w,a}}{\sum_{a' \in \mathcal{A}(w)} x^\pi_{w,a'}} \mathcal{P}_{w,a,q}\right) \left(1 - \sum_{a \in \mathcal{A}(w)} \frac{x^\pi_{w,a}}{\sum_{a' \in \mathcal{A}(w)} x^\pi_{w,a'}} \mathcal{P}_{w,a,q}\right)\right)^{-1}$$

using the definitions of the induced Markov chain and expected state-action residence times.

We assume that the optimal value of Problem 1 is finite. If the optimal value is infinite any admissible policy is a minimum-information admissible policy.

Proposition 1. For an MDP $\mathcal{M}$, if $\mathbb{E}[\iota^\pi_{W,\xi}]$ is finite where ξ is a path generated randomly under a policy $\pi \in \Pi^{St}(\mathcal{M})$, then

$$\mathbb{E}[\iota^\pi_{W,\xi}] = \sum_{w \in W} x^\pi_w \iota^\pi_w.$$

Note that the expected total information $x^\pi_w \iota^\pi_w$ of a state w has some undefined points in $x^\pi_{w,a} \geq 0$ where $a \in \mathcal{A}(w)$. We define the function at such points as follows:

- If the expected state residence time is zero, i.e., $x^\pi_w = \sum_{a \in \mathcal{A}(w)} x^\pi_{w,a} = 0$, then $x^\pi_w \iota^\pi_w := 0$. Since the state will never be visited, the observer cannot get information on the transition probabilities.
- If w deterministically transitions to one of the successor states and expected residence time is greater than zero, i.e., there exists a state $q \in Succ(w)$ such that $\sum_{a \in \mathcal{A}(w)} x^\pi_{w,a} \mathcal{P}_{w,a,q} > 0$ and $\sum_{a \in \mathcal{A}(w)} x^\pi_{w,a} \mathcal{P}_{w,a,q'} = 0$ for all $q' \in Succ(w) \setminus q$, then $x^\pi_w \iota^\pi_w := \infty$. Since the observer can estimate the transition probabilities even after a single observation and there is a positive probability that the state will be visited, the expected total information is infinite.
- If the expected state residence time at w is infinite, i.e., $x^\pi_w = \sum_{a \in \mathcal{A}(w)} x^\pi_{w,a} = \infty$, then $x^\pi_w \iota^\pi_w := \infty$. Since

the observed distribution of transitions converges to the transition probabilities, the expected total information is infinite.

We represent the stationary policies of the agent with a set of constraints which use the expected state-action residence times. A stationary policy makes each state either recurrent or transient. We need to identify the states that can be reachable and recurrent. If a policy leaks finite information, a set of states can be reachable and recurrent if and only if they belong to an end component and are not observed since the recurrence of a reachable observed state results in infinite expected total information.

Definition 2. An *unobserved end component* (UEC) is a sub-MDP (C, D) such that the digraph induced by (C, D) is strongly connected and $C \cap W = \emptyset$. An *unobserved maximal end component* (UMEC) (C, D) is a UEC where $C \subset S$ and there is no UEC (C', D') such that $(C, D) \neq (C', D')$, $C \subseteq C'$, and $D \subseteq D'$.

We denote the set of states that belong to some UMEC by C_{end}. After reaching C_{end}, the agent can follow a stationary policy that always stays in the UMEC and leaks no more information. For example, s_2 is a UMEC state in Figure 2. However, due to the reachability constraints the agent might need to follow a policy that leaves a UMEC. We disallow such cases and make the following assumption to ensure that the agent does not leave UMECs.

Assumption 1. *All unobserved maximal end components are closed.*

Remark. *In the absence of Assumption 1, to find the optimal stationary policy, one needs to check every UEC to determine whether the agents needs to stay or leave the UEC. Such a check increases computational complexity of finding a minimum-information admissible policy. For clarity of presentation, we here adopt Assumption 1. We refer the interested readers to [9] for the more general problem without Assumption 1.*

The optimal value of Problem 1 is

$$\inf \sum_{w \in W} x_w^\pi \iota_w^\pi \tag{4a}$$

$$\text{s. t. } \; x_s^\pi = \sum_{a \in \mathcal{A}(s)} x_{s,a}^\pi, \quad \forall s \in S \setminus C_{end} \tag{4b}$$

$$x_{s,a}^\pi \geq 0, \quad \forall s \in S \setminus C_{end}, \; \forall a \in \mathcal{A}(s) \tag{4c}$$

$$\sum_{a \in \mathcal{A}(s)} x_{s,a}^\pi - \sum_{q \in S} \sum_{a \in \mathcal{A}(q)} x_{q,a}^\pi \mathcal{P}_{q,a,s} = \mathbb{1}_{s_0}(s), \quad \forall s \in S \setminus C_{end} \tag{4d}$$

$$\sum_{q \in C_{reach}} \sum_{s \in S \setminus C_{end}} \sum_{a \in \mathcal{A}(s)} x_{s,a}^\pi \mathcal{P}_{s,a,q} + \mathbb{1}_{s_0}(q) \geq \nu_{reach}, \tag{4e}$$

where the decision variables are $x_{s,a}^\pi$ for all $s \in S \setminus C_{end}$ and $a \in \mathcal{A}(s)$. The objective function (4a) follows from Proposition 1 and the constraints (4b)-(4c) follow from definitions of expected residence times. The constraint (4d) is the flow equation indicating that the expected number of arrivals into a state, i.e., the inflow, is equal to the expected number of departures from the state, i.e., the outflow. These equations ensure that there exists a policy that gives the computed expected state-action residence times [11]. The reachability constraint in (1b) is equivalent to (4e).

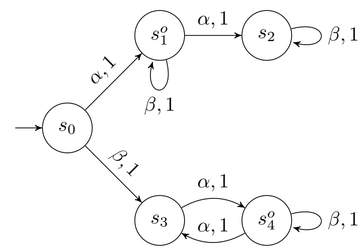


Fig. 2: An MDP with 5 states. A label a, p of a transition refers to the transition that happens with probability p when action a is taken. The states marked with the superscript o are observed.

Note that some stationary admissible policies are infeasible for the optimization problem given in (4). In detail, the stationary policies that eventually always stay in an end component and visit an observed state infinitely often are infeasible. For instance, consider a policy π such that $\Pr^\pi(Reach[s_2]) = 0.5$ for the MDP given in Figure 2 with the reachability constraint $\Pr(Reach[s_2] \geq 0.5)$. While π leads to infinite expected total information and satisfies the reachability constraint, it is not feasible for the problem in (4). One can easily check the existence of a policy that satisfies the reachability constraint via model checking tools such as [12]. If there exists a policy that satisfies the task constraints, but the optimization problem given in (4) is infeasible, we can say that the minimum-information admissible policy leaks infinite information.

Proposition 2. If there exists a policy $\pi \in \Pi^{St}(\mathcal{M})$ that satisfies the reachability constraint given in (1b), then there exists a policy $\pi^* \in \Pi^{St}(\mathcal{M})$ that attains the optimal value of the optimization problem given in (4).

Proposition 3. The optimization problem given in (4) is a convex optimization problem.

Remark. *The convexity of the optimization problem given in* (4) *is due to the fact that the transition information is a convex cost function of the policy parameters. After a preprocessing step that has polynomial-time complexity in the size of $\mathcal{M}$, the optimization problem can be formulated as a conic optimization problem which can be solved using interior-point methods* [13] *in polynomial-time in the size of $\mathcal{M}$.*

After computing the optimal expected state-action residence times by the optimization problem in (4), a stationary, minimum-information admissible policy can be synthesized using the relationship $x_{s,a}^\pi = \pi_{s,a} \; x_s^\pi$.

V. Bounds on the Estimation Error

In this section, we consider estimators for the transition probabilities at the observed states and derive the bounds on the expected estimation error in terms of MSE. Define σ_w as the MSE of an unbiased estimator at a state w. We assume that, for the estimator at state w, the observed data are the whole path of the agent and the transition probabilities for the set $S \setminus \{w\}$ of states are known.

Proposition 4. For an MDP $\mathcal{M}$ and a policy $\pi \in \Pi^{St}(\mathcal{M})$,

$$\sigma_w \geq \frac{\Pr^{\pi}(\mathit{Reach}[w])^2}{x_w^{\pi} \iota_w^{\pi}}$$

for every state $w \in W$.

Corollary 5. *For an MDP $\mathcal{M}$ and a policy $\pi \in \Pi^{St}(\mathcal{M})$, the total MSE $\sum_{w\in W} \sigma_w$ satisfies*

$$\sum_{w \in W} \sigma_w \geq \frac{\min_{w \in W} \Pr^{\pi}(\mathit{Reach}[w])^2 |W|^2}{\mathbb{E}[\iota^{\pi}_{W,\xi}]}.$$

Consequently, if $\Pr^{\pi}(\mathit{Reach}[w]) = 1$ for every $\pi \in \Pi^{St}(\mathcal{M})$ and for all $w \in W$, then $\frac{|W|^2}{\mathbb{E}_{\xi}[\iota^{\pi}_{W,\xi}]}$ is a lower bound on the total MSE.

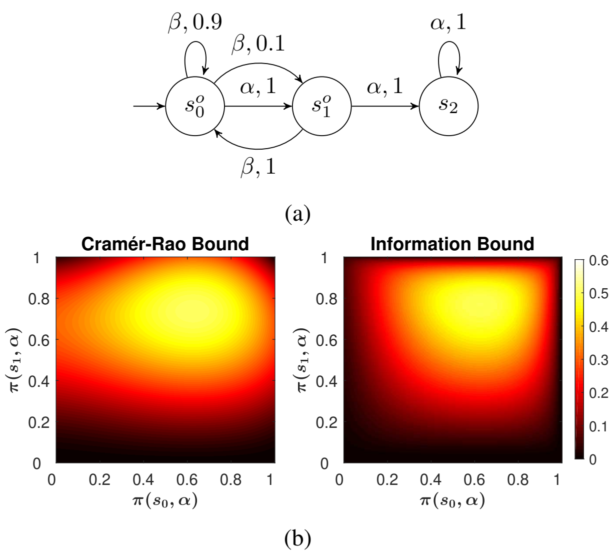


Fig. 3: (a) An MDP with 3 states. A label a, p of a transition refers to the transition that happens with probability p when action a is taken. The states marked with the superscript o are observed. (b) The Cramér-Rao bound on the total MSE of the estimators and the error bound given in Corollary 5.

An example of the bound given in Corollary 5 is illustrated in Figure 3. Both of the observed states are visited under any stationary policy and the reciprocal of the expected total information is directly a lower bound on the total MSE of the estimators. One who wants to maximize the total MSE of the estimators may prefer to optimize over the expected total information instead of the Cramér-Rao bound since the Cramér-Rao is not a convex or concave function of the expected residence time parameters while the minimum-information admissible policy can be computed via a convex optimization problem.

VI. Numerical Examples

In this section, we illustrate the proposed method through two numerical examples. We solved the optimization problems using CVX toolbox [14] with MOSEK [15] on a computer with an Intel Core i7-8550u 1.8 GHz CPU and 8 GB of RAM.

A. Partly Hidden Agent

In this example, we explain the characteristics of the minimum-information admissible policy through different scenarios.

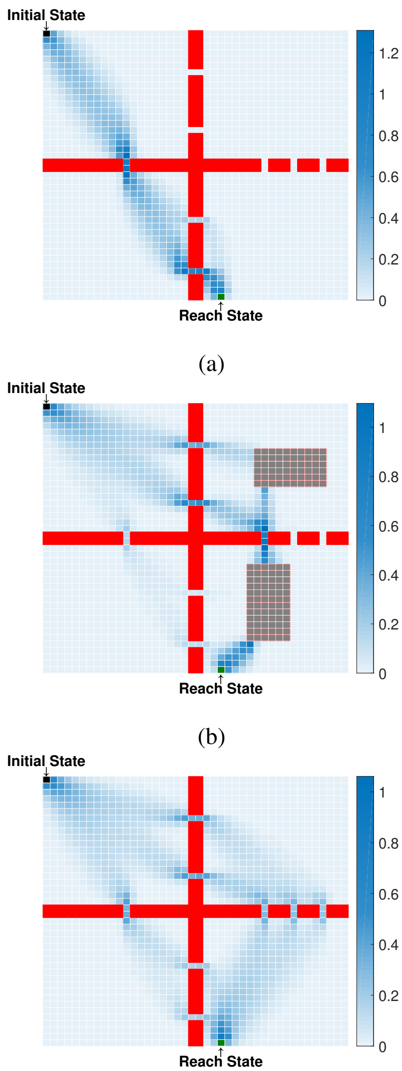
The environment which is given in Figure 4 consists of 4 regions that are separated with walls and connected to each other with bridges. Each region is a 20×20 grid world and each tile in these regions represents a state. Except for the reach state, the agent has 4 actions, namely, up, down, left, and right, at every state. When the agent takes an action the transition happens into the target direction with probability 0.8 and in the other directions uniformly randomly with probability 0.2. If a direction is out of the grid the transition probability to that direction is proportionally distributed to the other directions.

The initial state is the black top-left corner tile and the reach state is the green bottom-middle tile. The task of the agent is to reach the reach state with probability 1. While the agent is in the gray tiles, the observer cannot observe the transitions of the agent.

In the first scenario (see Figure 4a) all states are observed except the reach state and the bridge states. The agent completes the task with a low number of observed transitions (See Table I) with randomized transitions. Note that the randomization only happens between the states that are in the direction of the reach state since further randomization leads to more observations. When the unobserved regions are present in the environment (see Figure 4b), the policy generates paths that pass through the unobserved regions to reduce the number of observations. However, the unobserved regions are not always utilized. For example, in the top-right region if the agent is already away from the unobserved region, it directly goes to the bottom-right region. Although, no information leaks in the unobserved regions, the agent leaks information during the process of reaching those states.

We remark that the minimum-information admissible policy minimizes only the information of transitions from the observed states. While this approach reduces the amount of leaked information in the local sense, i.e., the transitions between the states, the global behavior, i.e., the transitions between the regions, might be easily inferred. We observe such a phenomenon for the scenario given in Figure 4a; the agent leaves the regions using the same bridge. This behavior may be risky if there is an adversary that is interested in the information of which bridge is used. To avoid this behavior, we add a weighted penalty, *exit information*, for each region.

The exit information of a region has the same form with the transition information and consists of the expected state residence times of the bridges. With the exit information (see Figure 4c) the agent randomizes its exit bridge from the regions compared to the initial case (see Figure 4a).



(a)

(b)

(c)

Fig. 4: The heatmaps of expected state residence times for partly hidden agent example. For the scenario given in Figure 4b the environment has some unobserved regions while every state is observed for the scenario given in Figure 4b. The scenario given in Figure 4c considers exit information of in addition to the transition information.

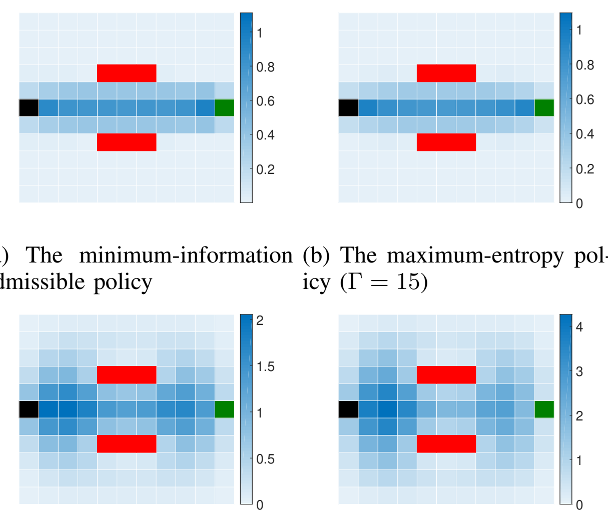
TABLE I: Numerical values for partly hidden agent example.

Scenario	Expected Total Information	Expected Number of Observations	Solving Time
Figure 4a	152.20	81.87	0.52
Figure 4b	146.00	78.89	0.38
Figure 4c	179.64	98.43	0.58

B. *Inference of Local Behavior*

We explain the difference between the proposed method and the policy synthesis via entropy maximization through this example. The environment is a 11×11 grid world given in Figure 5 where each tile represents a state. The black tile is the initial state, the green tile is the reach state, and the red tiles are the absorbing states. Except for the absorbing states and the reach state the agent can transition to 4 directions, namely, up, down, left, and right, at every state. When the agent takes an action, the transition happens in the target direction with probability 1. If a direction is out of the grid the action is not allowed. The task of the agent is to reach the reach state with probability 1.

We compare the policies in terms of their estimation error, which is calculated for different number of sample paths. The observer gets sample paths and estimates the transition probabilities at the observed states using a sample mean estimator. We measure the estimation error for a state by the mean squared error (MSE) between the observed and actual transition distributions at the observed states. The total error is the sum of MSE for each state. If there is no observation sample from a state, we set the MSE for that state.For the weighted MSE error, the weight of a state is ratio between the number of observations from the state and the total number of observations.



(a) The minimum-information admissible policy

(b) The maximum-entropy policy ($\Gamma = 15$)

(c) The maximum-entropy policy ($\Gamma = 60$)

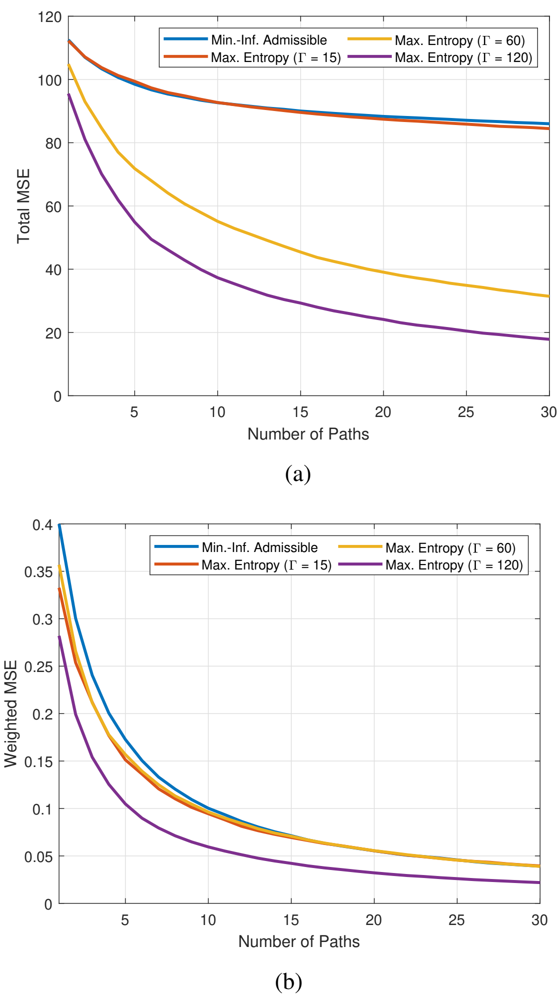
(d) The maximum-entropy policy ($\Gamma = 120$)

Fig. 5: The expected state residence times for inference of local behavior example.

Maximizing the entropy of an MDP is equivalent to maximizing the entropy of the possible paths, and a high entropy value leads to unpredictable paths. Under the reachability constraint, the maximum entropy of the MDP given in Figure 5 is unbounded. For policy synthesis, we follow the procedure given in [6] and impose an upper bound on the expected total state residence time Γ. As the bound increases, the maximum entropy value of the MDP increases.

We synthesize three policies that maximizes the entropy of MDP with different values for $\Gamma = 15, 60,$ and 120.

For low values of Γ such as 15, the minimum-information admissible and the maximum-entropy policies show similar behavior. However, for the high values of Γ, the difference between the minimum-information admissible policy and the maximum-entropy policy becomes clear. The minimum-information admissible policy completes the task with a low number of non-informative observations. On the other hand, the maximum-entropy policy visits the observed states more to explore more paths and randomize the probabilities of paths. While the agent follows different paths, the expected residence times at the observed states increases and observer gets more samples. Although the policy is randomized and samples are less informative, transition probabilities are inferred due to the high number of observations. The result suggests that the unpredictability of the paths does not imply the limitation of inference for the transitions between states. Hence, the minimum-information admissible policy and the maximum-entropy policy serve different purposes.



(a)

(b)

Fig. 6: The expected estimation errors. The curves are averaged over 100 experiments.

VII. Conclusion

We focus on policy synthesis for an agent whose behavior is inferred by an outside adversarial observer. Such an agent must as less informative observations as possible to the observer while completing its task. Based on this criterion, we introduced transition information which is based on the Fisher information and measures the amount of information leaked to the observer from a transition. Then, we formulated a problem that minimizes the expected total information leaked to the observer and showed the existence of such a policy. The significant feature of the proposed method is that it balances a possible trade-off between the number of observations and the informativeness of each observation.

The proposed method relies on the assumption that the agent follows a stationary policy on the observed states. A history dependent planning method may deceive the observer by actively changing the policy. We aim to remove this assumption and design a policy that takes the past transitions into account.

References

[1] B. R. Frieden. *Science from Fisher information: A Unification*. Cambridge University Press, 2004.

[2] T. Alpcan and I. Shames. An information-based learning approach to dual control. *IEEE Transactions on Neural Networks and Learning Systems*, 26(11):2736–2748, 2015.

[3] A. F. Emery and A. V. Nenarokomov. Optimal experiment design. *Measurement Science and Technology*, 9(6):864–876, 1998.

[4] F. Farokhi and H. Sandberg. Optimal privacy-preserving policy using constrained additive noise to minimize the Fisher information. In *56th IEEE Conference on Decision and Control*, pages 2692–2697, 2017.

[5] F. Farokhi and H. Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2018.

[6] Y. Savas, M. Ornik, M. Cubuktepe, and U. Topcu, Entropy maximization for constrained Markov decision processes. In *56th Annual Allerton Conference on Communication, Control, and Computing*, pages 911-918, 2018.

[7] N. Agmon, S. Kraus, and G. A. Kaminka. Multi-robot perimeter patrol in adversarial settings. In *IEEE International Conference on Robotics and Automation*, pages 2339–2345, 2008.

[8] P. Paruchuri, M. Tambe, F. Ordóñez, and S. Kraus. Security in multiagent systems by policy randomization. In *Joint Conference on Autonomous Agents and Multiagent Systems*, pages 273–280, 2006.

[9] M. O. Karabag, M. Ornik, and U. Topcu. Least Inferable Policies for Markov Decision Processes. *arXiv preprint arXiv:1809.06482 [math.OC]*, 2018.

[10] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

[11] K. Etessami, M. Kwiatkowska, M. Y. Vardi, and M. Yannakakis. Multi-objective model checking of Markov decision processes. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 50–65, 2007.

[12] M. Kwiatkowska, G. Norman, and D. Parker. PRISM: Probabilistic symbolic model checker. In *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, pages 200–204, 2002.

[13] Yu. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

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

[15] MOSEK ApS. The MOSEK optimization toolbox for matlab manual, version 8.1. `http://docs.mosek.com/8.1/toolbox/index.html`, 2017.