Deception in Supervisory Control

Mustafa O. Karabag, Melkior Ornik, and Ufuk Topcu

***Abstract*— The use of deceptive strategies is important for an agent that attempts not to reveal his intentions in an adversarial environment. We consider a setting in which a supervisor provides a reference policy and expects an agent to follow the reference policy and perform a task. The agent may instead follow a different, deceptive policy to achieve a different task. We model the environment and the behavior of the agent with a Markov decision process, represent the tasks of the agent and the supervisor with reachability specifications, and study the synthesis of optimal deceptive policies for such agents. We also study the synthesis of optimal reference policies that prevent deceptive strategies of the agent and achieve the supervisor's task with high probability. We show that the synthesis of optimal deceptive policies has a convex optimization problem formulation, while the synthesis of optimal reference policies requires solving a nonconvex optimization problem. We also show that the synthesis of optimal reference policies is NP-hard.**

***Index Terms*— Markov decision processes, deception, supervisory control, computational complexity.**

I. Introduction

DECEPTION is present in many fields that involve two parties, at least one of which is performing a task that is undesirable to the other party. The examples include cyber systems [1], [2], autonomous vehicles [3], warfare strategy [4], and robotics [5]. We consider a setting with a supervisor and an agent where the supervisor provides a reference policy to the agent and expects the agent to achieve a task by following the reference policy. However, the agent aims to achieve another task that is potentially malicious towards the supervisor and follows a different, deceptive policy. We study the synthesis of deceptive policies for such agents and the synthesis of reference policies for such supervisors that try to prevent deception besides achieving a task.

In the described supervisory control setting, the agent's deceptive policy is misleading in the sense that the agent follows his own policy, but convinces the supervisor that he follows the reference policy. Misleading acts result in plausibly deniable outcomes [6]. Hence, the agent's misleading behavior should have plausible outcomes for the supervisor. In detail, the supervisor has an expectation of the probabilities of the possible events. The agent should manipulate these probabilities such that he achieves his task while closely adhering to the supervisor's expectations.

We measure the closeness between the reference policy and the agent's policy by Kullback–Leibler (KL) divergence. KL divergence, also called relative entropy, is a measure of dissimilarity between two probability distributions [7]. KL divergence quantifies the extra information needed to encode a posterior distribution using the information of a given prior distribution. We remark that this interpretation matches the definition of plausibility: The posterior distribution is plausible if the KL divergence between the distributions is low.

We use a Markov decision process (MDP) to represent the stochastic environment and reachability specifications to represent the supervisor's and the agent's tasks. We formulate the synthesis of optimal deceptive policies as an optimization problem that minimizes the KL divergence between the distributions of paths under the agent's policy and reference policy subject to the agent's task specification. In order to preempt the agent's deceptive policies, the supervisor may aim to design its reference policy such that any deviations from the reference policy that achieves some malicious task do not have a plausible explanation. We formulate the synthesis of optimal reference policies as a maximin optimization problem where the supervisor's optimal policy is the one that maximizes the KL divergence between itself and the agent's deceptive policy subject to the supervisor's task constraints.

The agent's problem, the synthesis of optimal deceptive policies, and the supervisor's problem, the synthesis of optimal reference policies, lead to the following questions: Is it computationally tractable to synthesize an optimal deceptive policy? Is it computationally tractable to synthesize an optimal reference policy? We show that given the supervisor's policy, the agent's problem reduces to a convex optimization problem, which can be solved efficiently. On the other hand, the supervisor's problem results in a nonconvex optimization problem even when the agent uses a predetermined policy. We show that the supervisor's problem is NP-hard. We propose the duality approach and alternating direction method of multipliers (ADMM) [8] to locally solve the supervisor's optimization problem. We also give a relaxation of the problem that is a linear program.

The setting described in this paper can be considered as a probabilistic discrete event system under probabilistic supervisory control [9], [10]. The probabilistic supervisor induces an explicit probability distribution over the language generated

This work was supported in part by AFRL FA9550-19-1-0169, AFOSR FA9550-19-1-0005, and DARPA D19AP00004.

M. O. Karabag is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78705 USA (e-mail:karabag@utexas.edu).

M. Ornik was with The University of Texas at Austin, Austin, TX 78705 USA. He is now with the Department of Aerospace Engineering and the Coordinated Science Laboratory, The University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mail:mornik@illinois.edu).

U. Topcu is with the Department of Aerospace Engineering and Engineering Mechanics and the Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78705 USA (e-mail:utopcu@utexas.edu).

by the system by random disablement of the events. The supervisory control model considered in this paper is similar in that the reference policy induces an explicit probability distribution over the paths of the MDP. Different from [9], [10], we consider that the random disablement is done by the agent, and the supervisor is only responsible for providing the explicit random disablement strategy.

Similar to our approach, [11] used KL divergence as a proxy for the plausibility of messages in broadcast channels. While we use the KL divergence for the same purpose, the context of this paper differs from [11]. In the context of transition systems, [12], [13] used the metric proposed in this paper, the KL divergence between the distributions of paths under the agent's policy and the reference policy, for inverse reinforcement learning. In addition to the contextual difference, the proposed method of this paper differs from [12], [13]. We work in a setting with known transition dynamics and provide a convex optimization problem to synthesize the optimal policy while [12], [13] work with unknown dynamics and use sampling-based gradient descent to synthesize the optimal policy. Entropy maximization for MDPs [14] is a special case of the deception problem where the reference policy follows every possible path with equal probability. One can synthesize optimal deceptive policies by maximizing the entropy of the agent's path distribution minus the cross-entropy of the supervisor's path distribution relative to the agent's. For the synthesis of optimal deceptive policies, we use a method similar to [14] as we represent the objective function over transition probabilities. However, our proofs for the existence and synthesis of the optimal deceptive policies significantly differ from the results of [14]. In particular, [14] restricts attention to stationary policies without optimality guarantees whereas we prove the optimality of stationary policies for the deception problem. A related concept to deception is probabilistic system opacity, which was introduced in [15]. Two hidden Markov models (HMMs) are pairwise probabilistically opaque if the likelihood ratio between the HMMs is a non-zero finite number for every infinite observation sequence. This paper is related to [15] in that two HMMs are not pairwise probabilistically opaque if the KL divergence between the distribution of observation sequences is infinite. While [15] provides a method to check whether two HMMs with irreducible Markov chains are pairwise probabilistically opaque, we propose an optimization problem for MDPs that quantifies the deceptiveness of the induced system by the agent. Deception is also interpreted as the exploitation of an adversary's inaccurate beliefs on the agent's behavior [16], [17]. The work [16] focuses on generating unexpected behavior conflicting with the beliefs of the adversary, and [17] focuses on generating noninferable behavior leading to inaccurate belief distributions. On the other hand, the deceptive policy that we present generates behavior that is closest to the beliefs of the other party in order to hide the agent's malicious intentions.

We explore the synthesis of optimal reference policies, which, to the best of our knowledge, has not been discussed before. We propose to use ADMM to synthesize the optimal reference policies. Similarly, [18] also used ADMM for the synthesis of optimal policies for MDPs. While we use the same method, the objective functions of these papers differ since [18] is concerned with the average reward case, whereas we use ADMM to optimize the KL divergence between the distributions of paths. In addition, we use the ADMM to solve a decomposable minimax problem, which, to the best of our knowledge, has not been explored before.

We remark that a preliminary conference version [19] of this paper focuses on the synthesis of deceptive policies. In addition to contents of [19], this version contains the NP-hardness result on the synthesis of optimal reference policies, duality and ADMM methods for the synthesis of locally optimal reference policies, and an additional numerical example (Sections V-A, V-B, V-C, and VI-C). We also provide proofs of our results, which were omitted from [19].

The rest of the paper is organized as follows. Section II provides necessary theoretical background. In Section III, the agent's and the supervisor's problems are presented. Section IV explains the synthesis of optimal deceptive policies. In Section V, we give the NP-hardness result on the synthesis of optimal reference policies. We derive the optimization problem to synthesize the optimal reference policy and give the ADMM algorithm to solve the optimization problem. In this section, we also give a relaxed problem that relies on a linear program for the synthesis of optimal reference policies. We present numerical examples in Section VI and conclude with suggestions for future work in Section VII. We discuss the optimal deceptive policies under nondeterministic reference policies in Appendix I. We provide the proofs for the technical results in Appendix II.

II. Preliminaries

The set $\{x = (x_1, \dots, x_n) | x_i \geq 0\}$ is denoted by $\mathbb{R}^n_+$. The set $\{1, \dots, n\}$ is denoted by $[n]$. The indicator function $\mathbb{1}_y(x)$ of an element y is defined as $\mathbb{1}_y(x) = 1$ if $x = y$ and 0 otherwise. The characteristic function $\mathcal{I}_C(x)$ of a set C is defined as $\mathcal{I}_C(x) = 0$ if $x \in C$ and ∞ otherwise. The projection $Proj_C(x)$ of a variable x to a set C is equal to $\arg\min_{y \in C} \|x - y\|_2^2$. A Bernoulli random variable with parameter p is denoted by $Ber(p)$.

The set $\mathcal{K}$ is a convex cone, if for all $x, y \in \mathcal{K}$ and $a, b \geq 0$, we have $ax + by \in \mathcal{K}$. For the convex cone $\mathcal{K}$, $\mathcal{K}^* = \{y | y^T x \geq 0, \forall x \in \mathcal{K}\}$ denotes the dual cone. The exponential cone is denoted by $\mathcal{K}_{\text{exp}} = \{(x_1, x_2, x_3) | x_2 \exp(x_1/x_2) \leq x_3,\ x_2 > 0\} \cup \{(x_1, 0, x_3) | x_1 \leq 0,\ x_3 \geq 0\}$ and it can be shown that $\mathcal{K}^*_{\text{exp}} = \{(x_1, x_2, x_3) | -x_1 \exp(x_2/x_1 - 1) \leq x_3,\ x_1 < 0\} \cup \{(0, x_2, x_3) | x_2 \geq 0,\ x_3 \geq 0\}$.

Definition 1. *Let Q_1 and Q_2 be discrete probability distributions with a countable support $\mathcal{X}$. The Kullback–Leibler divergence between Q_1 and Q_2 is*

$$KL(Q_1 || Q_2) = \sum_{x \in \mathcal{X}} Q_1(x) \log \left(\frac{Q_1(x)}{Q_2(x)} \right).$$

We define $Q_1(x) \log \left(\frac{Q_1(x)}{Q_2(x)} \right)$ to be 0 if $Q_1(x) = 0$, and ∞ if $Q_1(x) > 0$ and $Q_2(x) = 0$. Data processing inequality states that any transformation $T : \mathcal{X} \to \mathcal{Y}$ satisfies

$$KL(Q_1 || Q_2) \geq KL(T(Q_1) || T(Q_2)). \tag{1}$$

Remark 1. *KL divergence is frequently defined with logarithm to base 2 in information theory. However, we use natural logarithm for the clarity of representation in the optimization problems. The base change does not change the results.*

A. Markov Decision Processes

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

The *history* h_t at time t is a sequence of states and actions such that $h_t = s_0 a_0 s_1 \dots s_{t-1} a_{t-1} s_t$. The set of all histories at time t is $\mathcal{H}_t$. A *policy* for $\mathcal{M}$ is a sequence $\pi = \mu_0 \mu_1 \dots$ where each $\mu_t : \mathcal{H}_t \times A \to [0, 1]$ is a function such that $\sum_{a \in A(s_t)} \mu_t(h_t, a) = 1$ for all $h_t \in \mathcal{H}_t$. A *stationary policy* is a sequence $\pi = \mu\mu \dots$ where $\mu : S \times A \to [0, 1]$ is a function such that $\sum_{a \in A(s)} \mu(s, a) = 1$ for every $s \in S$. The set of all policies for $\mathcal{M}$ is denoted by $\Pi(\mathcal{M})$ and the set of all stationary policies for $\mathcal{M}$ is denoted by $\Pi^{St}(\mathcal{M})$. For notational simplicity, we use $P_{s,a,q}$ for $P(s, a, q)$ and $\pi_{s,a}$ for $\mu(s, a)$ if $\pi = \mu\mu \dots$, i.e., π is stationary.

A stationary policy π for $\mathcal{M}$ induces a Markov chain $\mathcal{M}^\pi = (S, P^\pi)$ where S is the finite set of states and $P^\pi : S \times S \to [0, 1]$ is the transition probability function such that $P^\pi(s, q) = \sum_{a \in A(s)} P(s, a, q)\pi(s, a)$ for all $s, q \in S$. A state q is *accessible* from a state s if there exists an $n \geq 0$ such that the probability of reaching q from s in n steps is greater than 0 . A set C of states is a *communicating class* if q is accessible from s, and s is accessible from q for all $s, q \in C$. A communicating class C is *closed* if q is not accessible from s for all $s \in C$ and $q \in S \setminus C$.

A *path* $\xi = s_0 s_1 s_2 \dots$ for an MDP $\mathcal{M}$ is an infinite sequence of states under policy $\pi = \mu_0 \mu_1 \dots$ such that $\sum_{a \in A(s_t)} P(s_t, a, s_{t+1})\mu_t(h_t, a) > 0$ for all $t \geq 0$. The distribution of paths for $\mathcal{M}$ under policy π is denoted by $\Gamma^\pi_\mathcal{M}$.

For an MDP $\mathcal{M}$ and a policy π, the *state-action occupation measure* at state s and action a is defined by $x^\pi_{s,a} := \sum_{t=0}^{\infty} \Pr(s_t = s|s_0)\mu_t(s_t, a)$. If π is stationary, the state-action occupation measures satisfy $x^\pi_{s,a} = \pi_{s,a} \sum_{a' \in A(s)} x^\pi_{s,a'}$ for all s with finite occupation measures. The state-action occupation measure of a state-action pair is the expected number of times that the action is taken at the state over a path. We use x^π_s for the vector of the state-action occupation measures at state s under policy π and x^π for the vector of all state-action occupation measures.

We use $\Diamond R$ to denote the reachability specification to set R. A path $\xi = s_0 s_1 s_2 \dots$ satisfies $\Diamond R$ if and only if there exists i such that $s_i \in R$. On an MDP $\mathcal{M}$, the probability that a specification $\Diamond R$ is satisfied under a policy π, is denoted by $\Pr^\pi_\mathcal{M}(s_0 \models \Diamond R)$.

III. Problem Statement

We consider a setting in which an agent operates in a discrete stochastic environment modeled with an MDP $\mathcal{M}$, and a supervisor provides a reference policy $\pi^\mathcal{S}$ to the agent. The supervisor expects the agent to follow $\pi^\mathcal{S}$ on $\mathcal{M}$, thereby performing $K^\mathcal{S}$ tasks that are specified by reachability specifications $\Diamond R^\mathcal{S}_i$ for all $i \in [K^\mathcal{S}]$. The agent aims to perform another task that is specified by the reachability specification $\Diamond R^\mathcal{A}$ and may deviate from the reference policy to follow a different policy $\pi^\mathcal{A}$. In this setting, both the agent and the supervisor know the environment, i.e., the components of $\mathcal{M}$.

While the agent operates in $\mathcal{M}$, the supervisor observes the transitions, but not the actions of the agent, to detect any deviations from the reference policy. An agent that does not want to be detected must use a deceptive policy $\pi^\mathcal{A}$ that limits the amount of deviations from reference policy $\pi^\mathcal{S}$ and achieves $\Diamond R^\mathcal{A}$ with high probability.

We use Kullback-Leibler (KL) divergence to measure the deviation from the supervisor's policy. Recall that $\Gamma^{\pi^\mathcal{S}}_\mathcal{M}$ and $\Gamma^{\pi^\mathcal{A}}_\mathcal{M}$ are the distributions of paths under $\pi^\mathcal{S}$ and $\pi^\mathcal{A}$, respectively. We consider $KL(\Gamma^{\pi^\mathcal{A}}_\mathcal{M} || \Gamma^{\pi^\mathcal{S}}_\mathcal{M})$ as a proxy for the agent's deviations from the reference policy.

The perspective of information theory provides two motivations for the choice of KL divergence. The obvious motivation is that this value corresponds to the amount of information bits that the reference policy lacks while encoding the agent's path distribution. By limiting the deviations from the reference policy, we aim to make the agent's behavior easily explainable by the reference policy. Sanov's theorem [7] provides the second motivation. We note that satisfying the agent's objective with high probability is a rare event under the supervisor's policy. By minimizing the KL divergence between the policies, we make the agent's policy mimic the rare event that satisfies the agent's objective and is most probable under the supervisor's policy. Formally, let π^* be a solution to

$$\begin{aligned} \inf_{\pi \in \Pi(\mathcal{M})} \quad & KL\left(\Gamma^\pi_\mathcal{M} || \Gamma^{\pi^\mathcal{S}}_\mathcal{M}\right) \\ \text{subject to} \quad & \Pr^\pi_\mathcal{M}(s_0 \models \Diamond R^\mathcal{A}) \geq \nu^\mathcal{A}. \end{aligned}$$

Assume that we simulate n paths under the supervisor's policy. The probability that the observed paths satisfy $\Diamond R^\mathcal{A}$ with probability higher than $\nu^\mathcal{A}$ is approximately equal to $\exp(-nKL(\Gamma^{\pi^*}_\mathcal{M} || \Gamma^{\pi^\mathcal{S}}_\mathcal{M}))$. Furthermore, given that the observed path distribution satisfies $\Diamond R^\mathcal{A}$ with a probability higher than $\nu^\mathcal{A}$, the most likely distribution is $\Gamma^{\pi^*}_\mathcal{M}$ [7].

The choice of KL divergence is also justified from the perspective of statistical hypothesis testing. Likelihood-ratio test, the most powerful test for a given significance level [20], is a hypothesis test method to compare two models according to their goodness in terms of fitting the data. Assume that the supervisor uses likelihood-ratio test to decide whether the agent followed the reference policy or a different policy. Let $\xi_1, \dots, \xi_n$ be the agent's paths that are observed by the supervisor. We denote the probability of $\xi_1, \dots, \xi_n$ under $\pi^\mathcal{S}$ and $\pi^\mathcal{A}$ by $\Pr(\xi_1, \dots, \xi_n|\pi^\mathcal{S})$ and $\Pr(\xi_1, \dots, \xi_n|\pi^\mathcal{A})$, respectively. By the likelihood-ratio test, the supervisor decides that the agent uses

$\pi^{\mathcal{A}}$ if $\log(\Pr(\xi_1,\dots,\xi_n|\pi^{\mathcal{A}})) - \log(\Pr(\xi_1,\dots,\xi_n|\pi^{\mathcal{S}})) \geq c$, and the agent uses $\pi^{\mathcal{S}}$ otherwise where c is a constant. We note that π^* minimizes $nKL(\Gamma^{\pi}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}) = \mathbb{E}\left[\log(\Pr(\xi_1,\dots,\xi_n|\pi^{\mathcal{A}})) - \log(\Pr(\xi_1,\dots,\xi_n|\pi^{\mathcal{S}}))\right]$ subject to $\Pr^{\pi}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}$. Therefore, in expectation π^* is most likely policy to be not detected by the supervisor among the policies that satisfy $\Pr^{\pi}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}$.

We propose the following problem for the synthesis of deceptive policies for the agents.

Problem 1 (Synthesis of Optimal Deceptive Policies)**.** *Given an MDP $\mathcal{M}$, a reachability specification $\Diamond R^{\mathcal{A}}$, a probability threshold $\nu^{\mathcal{A}}$, and a reference policy $\pi^{\mathcal{S}}$, solve*

$$\inf_{\pi^{\mathcal{A}}\in\Pi(\mathcal{M})} \quad KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}\right) \tag{2a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}. \tag{2b}$$

If the optimal value is attainable, find a policy $\pi^{\mathcal{A}}$ that is a solution to (2)*.*

In order to preempt the possibility of that the agent uses a policy $\pi^{\mathcal{A}}$ that is the best deceptive policy against $\pi^{\mathcal{S}}$, the supervisor aims to find a reference policy $\pi^{\mathcal{S}}$ that maximizes the divergence between $\pi^{\mathcal{A}}$ and $\pi^{\mathcal{S}}$ subject to $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{S}}_i) \geq \nu^{\mathcal{S}}_i$ for all $i \in [K^{\mathcal{S}}]$. We assume that the supervisor knows the agent's task and propose the following problem for the synthesis of reference policies for the supervisor.

Problem 2 (Synthesis of Optimal Reference Policies)**.** *Given an MDP $\mathcal{M}$, reachability specifications $\Diamond R^{\mathcal{A}}$ and $\Diamond R^{\mathcal{S}}_i$ for all $i \in [K^{\mathcal{S}}]$, probability thresholds $\nu^{\mathcal{A}}$ and $\nu^{\mathcal{S}}_i$ for all $i \in [K^{\mathcal{S}}]$, solve*

$$\sup_{\pi^{\mathcal{S}}\in\Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}}\in\Pi(\mathcal{M})} \quad KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}\right) \tag{3a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{3b}$$

$$\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{S}}_i) \geq \nu^{\mathcal{S}}_i, \quad \forall i \in [K^{\mathcal{S}}]. \tag{3c}$$

If the supremum is attainable, find a policy $\pi^{\mathcal{S}}$ that is a solution to (3)*.*

Example 1: We explain the synthesis of optimal deceptive policies and reference policies through the MDP $\mathcal{M}$ given in Figure 1. Note that the policies for $\mathcal{M}$ may vary only at s_0 since it is the only state with more than one action.

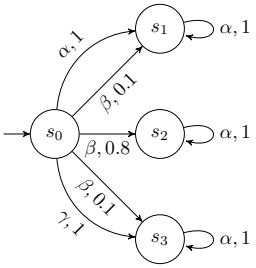


Fig. 1. An MDP with 4 states. A label $\boldsymbol{a}, \boldsymbol{p}$ of a transition refers to the transition that happens with probability $\boldsymbol{p}$ when action $\boldsymbol{a}$ is taken.

We first consider the synthesis of optimal deceptive policies where the reference policy satisfies $\pi^{\mathcal{S}}_{s_0,\beta} = 1$. Consider $\Diamond R^{\mathcal{A}} = \Diamond\{s_3\}$ and $\nu^{\mathcal{A}} = 0.2$. Assume that the agent's policy has $\pi^{\mathcal{A}}_{s,\gamma} = 1$. The value of the KL divergence is 2.30. However, note that as $\pi^{\mathcal{A}}_{s,\beta}$ increases, the KL divergence decreases. In this case, the optimal policy satisfies $\pi^{\mathcal{A}}_{s,\beta} = 0.89$ and $\pi^{\mathcal{A}}_{s,\gamma} = 0.11$ and the optimal value for the KL divergence is 0.04.

We now consider the synthesis of optimal reference policies where the supervisor has a single specification $\Diamond R^{\mathcal{S}} = \Diamond\{s_1, s_2\}$ and $\nu^{\mathcal{S}} = 0.9$. Consider $\Diamond R^{\mathcal{A}} = \Diamond\{s_3\}$ and $\nu^{\mathcal{A}} = 0.1$. Assume that we have $\pi^{\mathcal{S}}_{s_0,\beta} = 1$. In this case, the agent can directly follow the reference policy and make the KL divergence zero. This reference policy is not optimal; the supervisor, knowing the malicious objective of the agent, can choose the reference policy with $\pi^{\mathcal{S}}_{s_0,\alpha} = 1$, which does not allow any deviations and makes the KL divergence infinite.

IV. Synthesis of Optimal Deceptive Policies

In this section, we explain the synthesis of optimal deceptive policies. Before proceeding to the synthesis step, we make assumptions to simplify the problem. Then, we show the existence of an optimal deceptive policy and give an optimization problem to synthesize one.

Without loss of generality, we make the following assumption on the target states of the agent and the supervisor for the clarity of representation. This assumption ensures that the probability of completing a task is constant, either 0 or 1, upon reaching a target state.

Assumption 1. *Every $s \in R^{\mathcal{A}} \cup R^{\mathcal{S}}_1 \cup \ldots \cup R^{\mathcal{S}}_{K^{\mathcal{S}}}$ is absorbing.*

We remark that in the absence of Assumption 1, one can still find the optimal deceptive policy by constructing a product MDP that encodes both the state of the original MDP and the statuses of the tasks. In detail, we need to construct a joint deterministic finite automaton whose states encode the statuses of the specifications for the agent and the supervisor. After creating the joint deterministic finite automaton (DFA), we construct a product MDP by combining the original MDP and the joint DFA and synthesize a policy on the product state space. Since there is a one-to-one mapping between the paths of the original MDP and the product MDP, the synthesized policy for the product MDP can be translated into a policy for the original MDP [21].

If the reference policy is not stationary, we may need to compute the optimal deceptive policy by considering the parameters of the reference policy at different time steps. Such computation leads to a state explosion, which we avoid by adopting the following assumption.

Assumption 2. *The reference policy $\pi^{\mathcal{S}}$ is stationary on $\mathcal{M}$.*

In many applications the supervisor aims to achieve the specifications with the maximum possible probabilities. Under Assumption 1, stationary policies suffice to achieve the Pareto optimal curve for maximizing the probabilities of multiple reachability specifications [22].

Without loss of generality, we assume that the optimal value of Problem 1 is finite. One can easily check whether the optimal value is finite in the following way. Assume that the transition probability between a pair of states is zero under the reference policy. One can create a modified MDP from $\mathcal{M}$ by removing the actions that assign a positive value to such state-state pairs. If there exists a policy that satisfies the constraint (2b) then the value is finite.

Given that the optimal value of Problem 1 is finite, we first identify the three sets of states where the agent should follow the reference policy. Firstly, the agent's policy should not be different from the supervisor's policy on the states that belong to $R^{\mathcal{A}}$, since the specification of the agent is already satisfied. Secondly, the agent should follow the reference policy at states that are recurrent under the reference policy. Formally, the reference policy $\pi^{\mathcal{S}}$ induces a Markov chain $\mathcal{M}^{\mathcal{S}}$. A state is recurrent in $\mathcal{M}^{\mathcal{S}}$ if it belongs to some closed communicating class. The agent should follow the reference policy if a state is recurrent in $\mathcal{M}^{\mathcal{S}}$.

For the second claim, we first remark that every closed communicating class $C \subset S$ of $\mathcal{M}^{\mathcal{S}}$ satisfy either 1) $C \cap (S \setminus R^{\mathcal{A}}) \neq \emptyset$ and $C \cap R^{\mathcal{A}} = \emptyset$, or 2) $C \cap (S \setminus R^{\mathcal{A}}) = \emptyset$ and $C \cap R^{\mathcal{A}} \neq \emptyset$. This is due to the fact that $R^{\mathcal{A}}$ is a closed set, i.e., a state in $R^{\mathcal{A}}$ is reached and the states in $S \setminus R^{\mathcal{A}}$ are not accessible. Hence, there cannot be a closed communicating class of $\mathcal{M}^{\mathcal{S}}$ that has states in both $R^{\mathcal{A}}$ and $S \setminus R^{\mathcal{A}}$. Let C^{cl} be the union of all closed communicating classes of $\mathcal{M}^{\mathcal{S}}$, i.e., the recurrent states of $\mathcal{M}^{\mathcal{S}}$. Note that $C^{cl} \setminus R^{\mathcal{A}}$ is a closed set in $\mathcal{M}^{\mathcal{S}}$ and the states in $R^{\mathcal{A}}$ are not accessible from $C^{cl} \setminus R^{\mathcal{A}}$ in $\mathcal{M}^{\mathcal{S}}$ due to the above discussion.

Assume that under the agent's policy $\pi^{\mathcal{A}}$, there exists a path that visits a state in $C^{cl} \setminus R^{\mathcal{A}}$ and leaves $C^{cl} \setminus R^{\mathcal{A}}$ with positive probability. In this case, the KL divergence is infinite since an event that happens with probability zero under the supervisor's policy happens with a positive probability under the agent's policy. Hence, $C^{cl} \setminus R^{\mathcal{A}}$ must also be a closed set under $\pi^{\mathcal{A}}$. Furthermore, since the agent cannot leave $C^{cl} \setminus R^{\mathcal{A}}$, and the probability of satisfying $\Diamond R^{\mathcal{A}}$ is zero upon entering $C^{cl} \setminus R^{\mathcal{A}}$, the agent should choose the same policy as the supervisor to minimize the KL divergence between the distributions of paths. Note that for the recurrent states in $R^{\mathcal{A}}$, i.e., $C^{cl} \cap R^{\mathcal{A}}$, the second claim is trivially satisfied by the first claim.

For all $s \in S \setminus (C^{cl} \cup R^{\mathcal{A}})$, s is transient in $\mathcal{M}^{\mathcal{S}}$, and the agent's policy must eventually stop visiting s, since otherwise we have infinite divergence. Furthermore, we have the following proposition.

Proposition 1. *If the optimal value of Problem 1 is finite and the optimal policy is $\pi^{\mathcal{A}}$, the state-action occupation measure $x^{\pi^{\mathcal{A}}}_{s,a}$ is finite for all $s \in S \setminus (C^{cl} \cup R^{\mathcal{A}})$ and $a \in A(s)$.*

The occupation measures are bounded for the states that the agent's policy may differ from the supervisor's policy. Since the occupation measures are bounded, the stationary policies suffice for the synthesis of optimal deceptive policies [23].

Proposition 2. *For any policy $\pi^{\mathcal{A}} \in \Pi(\mathcal{M})$ that satisfies $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}$, there exists a stationary policy $\pi^{\mathcal{A},St} \in \Pi(\mathcal{M})$ that satisfies $\Pr^{\pi^{\mathcal{A},St}}_{\mathcal{M}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}$ and*

$$KL\left(\Gamma^{\pi^{\mathcal{A},St}}_{\mathcal{M}} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}\right) \leq KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}\right).$$

Sketch of Proof for Proposition 2. Assume that the KL divergence between the path distributions is finite. Note that the occupation measures of $\pi^{\mathcal{A}}$ are finite for all $s \in S_d = S \setminus (C^{cl} \cup R^{\mathcal{A}})$.

When the reference policy is stationary, we may transform $\mathcal{M}$ into a *semi-infinite MDP*. The semi-infinite MDP shares the same states with $\mathcal{M}$, but has continuous action space such that for all states every randomized action of $\mathcal{M}$ is an action of the semi-infinite MDP. Also the states belong to $R^{\mathcal{A}}$ and C^{cl} are absorbing in the semi-infinite MDP.

Let $X^{\mathcal{S}}_s$ be the successor state distribution at state s under the reference policy in the semi-infinite MDP. At state $s \in S_d$, an action a with successor state distribution $X_{s,a}$ has cost $KL(X_{s,a}||X^{\mathcal{S}}_s)$. The cost is 0 for the other states that do not belong to S_d. Consider an optimization problem that minimizes the expected cost subject to reaching $R^{\mathcal{A}}$ with probability at least $\nu^{\mathcal{A}}$. The result of this optimization problem shares the same value with the result of Problem 1. This problem is a constrained cost minimization for an MDP where the only decision variables are the state-action occupation measures. An optimal policy can be characterized by the state-action occupation measures.

The occupation measures must be finite for all $s \in S_d$ as we showed in Proposition 1. Since every finite occupation measure vector of S_d can also be achieved by a stationary policy, there exists a stationary policy which shares the same occupation measures with an optimal policy [23]. Hence, this stationary policy is also optimal.

Now assume that the stationary optimal policy π^* is randomized. Let π^*_s be the action distribution and $X^{\pi^*}_s$ be the successor state distribution at state s under π^*. Note that at state s there exists an action a^* that has $P(s,a^*,q) = X^{\pi}_s(q)$ since the action space is convex for the semi-infinite MDP. Also due to the convexity of KL divergence we have $\int_{\Delta^{|A(s)|}} KL(X_{s,a}||X^{\mathcal{S}}_s) d\pi^*_s(a) \geq KL(X^{\pi^*}_s||X^{\mathcal{S}}_s)$ where $\Delta^{|A(s)|}$ is $|A(s)|$-dimensional probability simplex. Hence, deterministically taking action a^* is optimal for state s. By generalizing this argument to all $s \in S_s$, we conclude that there exists an optimal stationary deterministic policy for the semi-infinite MDP. Without loss of generality we assume π^* is stationary deterministic.

We note that the stationary deterministic policy π^* of the semi-infinite MDP corresponds to a stationary randomized policy for the original MDP $\mathcal{M}$. Hence the proposition holds. ■

We denote the set of states for which the agent's policy can differ from the supervisor's policy by $S_d = S \setminus (C^{cl} \cup R^{\mathcal{A}})$. We solve the following optimization problem to compute the occupation measures of an optimal deceptive policy:

$$\inf \sum_{s \in S_d} \sum_{a \in A(s)} \sum_{q \in Succ(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} \log\left(\frac{\sum_{a' \in A(s)} x^{\mathcal{A}}_{s,a'} P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q} \sum_{a' \in A(s)} x^{\mathcal{A}}_{s,a'}}\right) \tag{4a}$$

subject to

$$x^{\mathcal{A}}_{s,a} \geq 0, \qquad \forall s \in S_d,\ \forall a \in A(s), \tag{4b}$$

$$\sum_{a \in A(s)} x^{\mathcal{A}}_{s,a} - \sum_{q \in S_d} \sum_{a \in A(q)} x^{\mathcal{A}}_{q,a} P_{q,a,s} = \mathbb{1}_{s_0}(s), \forall s \in S_d, \tag{4c}$$

$$\sum_{q \in R^{\mathcal{A}}} \sum_{s \in S_d} \sum_{a \in A(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} + \mathbb{1}_{s_0}(q) \geq \nu^{\mathcal{A}} \tag{4d}$$

where $\pi^{\mathcal{S}}_{s,q}$ is the transition probability from s to q under $\pi^{\mathcal{S}}$

and the decision variables are $x^{\mathcal{A}}_{s,a}$ for all $s \in S_d$ and $a \in A(s)$. The objective function (4a) is obtained by reformulating the KL divergence between the path distributions as the sum of the KL divergences between the successor state distributions for every time step (See Lemma 3 in Appendix II). The constraint (4c) encodes the feasible policies and the constraint (4d) represents the task constraint.

Proposition 3. *The optimization problem given in* (4) *is a convex optimization problem that shares the same optimal value with* (1)*. Furthermore, there exists a policy* $\pi \in \Pi^{St}(\mathcal{M})$ *that attains the optimal value of* (4)*.*

The optimization problem given in (4) gives the optimal state-action occupation measures for the agent. One can synthesize the optimal deceptive policy $\pi^{\mathcal{A}}$ using the relationship $x^{\mathcal{A}}_{s,a} = \pi^{\mathcal{A}}_{s,a} \sum_{a' \in A(s)} x^{\pi}_{s,a'}$ for all $s \in S_d$ and $\pi^{\mathcal{A}}_{s,a} = \pi^{\mathcal{S}}_{s,a}$ for the other states.

The optimization problem given in (4) can be considered as a constrained MDP problem with an infinite action space [23] and a nonlinear cost function. This equivalence follows from that there exists a deterministic policy that incurs the same cost on the infinite action MDP for every randomized policy for $\mathcal{M}$. Since there exists a deterministic optimal policy for the infinite MDP, we can represent the objective function and constraints of Problem 1 with the occupancy measures. However, we remark that (4) is a convex nonlinear optimization problem whereas the constrained MDPs are often modeled with a linear cost function and solved using linear optimization methods.

Remark 2. *The methods provided in this section can be generalized to task constraints that are co-safe LTL specifications. In detail, every co-safe LTL can be translated into a DFA [24]. By combining the MDP and the DFA, we get the product MDP. Since the co-safe LTL specifications translates into reachability specifications on the product MDP and there is a one-to-one mapping between the paths of the original MDP and the product MDP, we can apply the methods described in this section to compute an optimal deceptive policy.*

V. Synthesis of Optimal Reference Policies

In this section, we prove the hardness of Problem 2. We give an optimization problem based on dualization approach to synthesize locally optimal reference policies. We provide a distributed optimization algorithm based on ADMM for synthesis of locally optimal reference policies. We also derive a lower bound on the objective function and give a linear programming relaxation of Problem 2.

The optimization problem given in (4) has the supervisor's policy parameters as constants. We want to solve the optimization problem given in (4) to formulate the synthesis of optimal reference policies by adding the supervisor's policy parameters as additional decision variables. The set C^{cl} is the set of states that belong to a closed communicating class of $\mathcal{M}^{\mathcal{S}}$. In (4), C^{cl} is a constant set for a given reference policy, but it may vary under different reference policies. We make the following assumption to prevent set C^{cl} from varying under different reference policies.

Assumption 3. *The set* C^{cl} *is the same for all reference policies considered in Problem 2.*

Remark 3. *Assumption 3 is made for the clarity of representation. In the absence of Assumption 3, one can to compute the optimal reference policy for different values of* C^{cl}*. However, we remark that since, in general,* C^{cl} *can have* $\mathcal{O}(2^{|S|})$ *values, computing the optimal reference policy for different values of* C^{cl} *may have exponential complexity in* $|S|$*.*

Under Assumptions 2 and 3, the optimal value of Problem 2 is equal to the optimal value of the following optimization problem:

$$\sup_{x^{\mathcal{S}}_{s,a}} \inf_{x^{\mathcal{A}}_{s,a}} \sum_{s \in S_d} \sum_{a \in A(s)} \sum_{q \in Succ(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} \tag{5a}$$

$$\log \left(\frac{\sum_{a' \in A(s)} x^{\mathcal{A}}_{s,a'} P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q} \sum_{a' \in A(s)} x^{\mathcal{A}}_{s,a'}} \right) \tag{5b}$$

subject to

(4b) − (4d)

$$\pi^{\mathcal{S}}_{s,q} = \sum_{a \in A(s)} P_{s,a,q} \frac{x^{\mathcal{S}}_{s,a}}{\sum_{a' \in A(s)} x^{\mathcal{S}}_{s,a'}}, \ \forall s \in S_d, \ \forall q \in S, \tag{5c}$$

$$x^{\mathcal{S}}_{s,a} \geq 0, \qquad \forall s \in S_d, \ \forall a \in A(s), \tag{5d}$$

$$\sum_{a \in A(s)} x^{\mathcal{S}}_{s,a} - \sum_{q \in S_d} \sum_{a \in A(q)} x^{\mathcal{S}}_{q,a} P_{q,a,s} = \mathbb{1}_{s_0}(s), \forall s \in S_d, \tag{5e}$$

$$\sum_{q \in R^{\mathcal{S}}_i} \sum_{s \in S_d \setminus C_{\mathcal{S}}} \sum_{a \in A(s)} x^{\mathcal{S}}_{s,a} P_{s,a,q} + \mathbb{1}_{s_0}(q) \geq \nu^{\mathcal{S}}_i, \ \forall i \in [K^{\mathcal{S}}] \tag{5f}$$

where $x^{\mathcal{S}}_{s,a}$ variables are the decision variables for the supervisor and $x^{\mathcal{A}}_{s,a}$ variables are the decision variables for the agent.

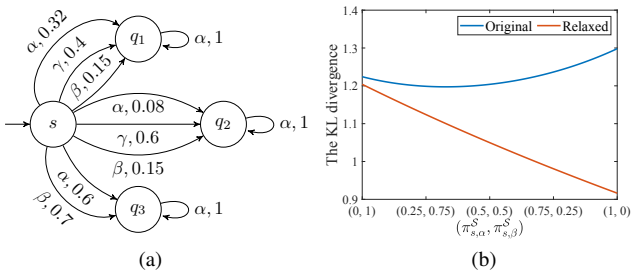
Remark 4. *The optimization problem given in* (5) *has undefined points due to the denominators in* (5b) *and* (5c)*, that are ignored in the above optimization problem for the clarity of representation. If* $\sum_{a \in A(s)} x^{\mathcal{S}}_{s,a} = 0$*, then the state* s *is unreachable and if the KL divergence between the policies is finite, the state must be unreachable also under* $\pi^{\mathcal{A}}$*. Hence there is no divergence at state* s*. If* $\pi^{\mathcal{S}}_{s,q} = 0$ *and if the KL divergence between the policies is finite,* $x^{\mathcal{A}}_{s,q}$ *must be* 0*. Hence there is no divergence for state* s *and successor state* q*.*

We can show the existence of an optimal reference policy if the condition given in Proposition 4 is satisfied. This condition ensures that the objective function of the problem in (5) is finite for all pairs of the supervisor's and the agent's policies.

Proposition 4. *If* $P_{s,a,q} > 0$ *for all* $s \in S_d$*,* $a \in A(s)$*, and* $q \in Succ(s)$*, then there exists a policy* $\pi^{\mathcal{S}}$ *that attains the optimal value of the optimization problem given in* (5)*.*

We note that the optimization problem given in (5) is nonconvex. One might wonder whether there exists a problem formulation that yields a convex optimization problem. We show that it is not possible to obtain a convex reformulation of the optimization problem given in (5).

We first observe that it is possible that there are multiple locally optimal reference policies. For example, consider the



(a) (b)

Fig. 2. (a) An MDP with 4 states. A label $\boldsymbol{a}, \boldsymbol{p}$ of a transition refers to the transition that happens with probability $\boldsymbol{p}$ when action $\boldsymbol{a}$ is taken. (b) The KL divergence between the path distributions of the agent and the supervisor for different reference policies. Note that there are two local optima that maximizes the KL divergence.

MDP given in Figure 2a where the specification of the agent is $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s \models \Diamond q_1 \vee \Diamond q_2) = 1$. Regardless of the reference policy, the agent's policy must have $\pi^{\mathcal{A}}_{s,\gamma} = 1$ due to his specification. For simplicity, there is no specification for the supervisor, i.e., $\nu^{\mathcal{S}}$ is 0. The optimal reference policy maximizes $0.4\log(0.4/(0.32x^{\mathcal{S}}_{s_0,\alpha} + 0.15x^{\mathcal{S}}_{s_0,\beta} + 0.4x^{\mathcal{S}}_{s_0,\gamma})) + 0.6\log(0.6/(0.08x^{\mathcal{S}}_{s_0,\alpha} + 0.15x^{\mathcal{S}}_{s_0,\beta} + 0.6x^{\mathcal{S}}_{s_0,\gamma}))$, which is a convex function of $x^{\mathcal{S}}_{s_0,\alpha}$, $x^{\mathcal{S}}_{s_0,\beta}$, and $x^{\mathcal{S}}_{s_0,\gamma}$. There are two locally optimal reference policies for Problem 2: the policy that satisfies $\pi^{\mathcal{S}}_{s,\alpha} = 1$ and the policy that satisfies $\pi^{\mathcal{S}}_{s,\beta} = 1$. Hence, the problem is not only nonconvex but also possibly multimodal.

We consider a new parametrization to reformulate the optimization problem given in (5). Consider a continuous and bijective transformation from the occupation measures to the new parameters, that makes new parameters to span all stationary policies. After this transformation, an optimal solution to (5) yields an optimal solution in the new parameter space. If the optimization problem given in (5) has multiple local optima, then any reformulation spanning all stationary policies for the supervisor has multiple local optima. Therefore, it is not possible to obtain a convex reformulation.

In Section V-A, we show that Problem 2 is a provably hard problem. In Section V-B, we describe dualization-based procedure to locally solve the optimization problem given in (5). As an alternative to solving the dual problem, we give an algorithm based on alternating direction method of multipliers (ADMM) in Section V-C. Finally, we present a relaxation of the problem in Section V-D that relies on solving a linear program.

A. The Complexity of the Synthesis of Optimal Reference Policies

In this section, we show that the synthesis of an optimal reference policy is NP-hard whereas the feasibility problem for reference policies can be solved in polynomial time.

Finding a feasible policy under multiple reachability constraints has polynomial complexity in the number states and actions for a given MDP. When the target states are absorbing, the complexity of the problem is also polynomial in the number of constraints [22]. This result follows from that a feasible policy can be synthesized with a linear program where the numbers of variables and constraints are polynomial in the number of states, actions, and task constraints.

Matsui transformed the set partition problem to the decision version of an instance of linear multiplicative programming and proved the NP-hardness of linear multiplicative programming [25]. In the proof of Proposition 5, we give an instance of Problem 2 whose decision problem reduces into the decision problem of the instance of linear multiplicative programming that Matsui provided[1].

While a feasible reference policy can be synthesized in polynomial time, the complexity of finding an optimal reference policy is NP-hard even when the target states are absorbing. The hardness proof follows from a reduction of Problem 2 to an instance of linear multiplicative programming which minimizes the multiplication of two variables subject to linear inequality constraints. Formally we have the following result.

Proposition 5. *Problem 2 is NP-hard even under Assumption 1.*

Sketch of Proof for Proposition 5. Problem 2 can be reduced to linear multiplicative programming. Linear multiplicative program can be reduced to the set partition problem. Since the set partition problem is NP-hard, Problem 2 is NP-hard.

In more detail, the set partition problem [26], [27] is NP-hard and is the following:

Instance: An $m \times n$ $0-1$ matrix M satisfying $n > m$.

Question: Is there a $0-1$ vector x satisfying $\sum_{\substack{j=1 \\ M_{ij}=1}}^{n} x_j = 1$ for all $i \in [n]$.

Linear multiplicative programming minimizes the product of two variables subject to linear inequality constraints and is NP-hard [25] . Let M be an $m \times n$ $0-1$ matrix with $n \geq m$ and $n \geq 5$, and $p = n^{n^4}$. The problem

$$\begin{aligned}
\min \quad & (2p^{4n} - p + 2p^{2n}x_0 + y_0)(2p^{4n} - p - 2p^{2n}x_0 + y_0) \\
\text{subject to} \quad & x_0 = \sum_{i=1}^{n} p^i x_i && \text{(6a)} \\
& y_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} p^{i+j} y_{ij} && \text{(6b)} \\
\forall i \in [n], \quad & 0 \leq x_i \leq 1, y_{ii} = x_i, && \text{(6c)} \\
\forall i, j \in [n] \quad & 0 \leq y_{ij} \leq 1, && \text{(6d)} \\
\forall_{\substack{i,j \in [n] \\ i \neq j}}, \quad & x_i \geq y_{ij}, \quad x_j \geq y_{ij}, \quad y_{ij} \geq x_i + x_j - 1, && \text{(6e)} \\
\forall i \in [m], \quad & \sum_{\substack{j=1 \\ M_{ij}=1}}^{n} x_j = 1, && \text{(6f)}
\end{aligned}$$

where the decision variables are x_i for all $i \in [n]$ and y_{ij} for all $i, j \in [n]$, is NP-hard. In detail, [25] proved that the optimal value of (6) is less than or equal to $4p^{8n}$ if and only if there exists a $0-1$ solution for $x_1, \ldots, x_n$ satisfying (6f). Since the decision problem of (6) correspond to solving the set partition problem, (6) is NP-hard.

We can construct an MDP with a size polynomial in n and choose polynomial number of specifications in n such that the

[1]The complete proof is available at https://arxiv.org/abs/1902.00590.

optimal value of Problem 2 is

$$\max \frac{1}{2}\log\frac{1}{(2p^{4n}-p+2p^{2n}x_0+y_0)(2p^{4n}-p-2p^{2n}x_0+y_0)} + \frac{1}{2}\log\left(4C^2(n^2+n+1)^2\right)$$
$$\text{subject to} \quad \text{(6a)}-\text{(6f)}$$

where C is a constant depending on n. Due to the result given in [25], the optimal value of (7) is greater than or equal to $-\log(4p^{8n})/2+\log\left(4C^2(n^2+n+1)^2\right)/2$ if and only if there exists a $0-1$ solution for $x_1,\ldots,x_n$ satisfying (6f). Since the decision problem of (7) correspond to solving the set partition problem, (7) is NP-hard.

Since the number of states, actions, and the task constraints is polynomial in n and (7) synthesizes an optimal reference policy, the synthesis of optimal reference policies is NP-hard. ■

We remark that the hardness of Problem 2 is due to the nonconvexity of the KL objective function since the feasibility problem can be solved in polynomial time.

B. Dualization-based Approach for the Synthesis of Optimal Reference Policies

Observing that Slater's condition [28] is satisfied, and the strong duality holds for the optimization problem given in (4), to find the optimal value of (5) one may consider solving the dual of (4) with $x^{\mathcal{S}}_{s,a}$ as additional decision variables and (5c)-(5f) as additional constraints. In this section, we describe the dualization-based approach for the synthesis of locally optimal reference policies.

The optimization problem given in (4) has the following conic optimization representation:

$$\min_{y} \; c^T y \tag{8a}$$
$$\text{subject to} \quad [G|-I]y = h, \tag{8b}$$
$$y \in \mathcal{K}. \tag{8c}$$

We construct the parameters of the above optimization problem as follows. Define the variable $r_{(s,q)}$ for all $s\in S_d$ and $q\in Succ(s)$. Let r be the $M\times 1$ vector of $r_{(s,q)}$ variables where $r_{(s,q)}$ has the index (s,q). The conic optimization problem has the objective function $\sum_{s\in S_d}\sum_{q\in Succ(s)} r_{(s,q)}$ and the constraint

$$r_{(s,q)} \geq \sum_{a\in A(s)} x^{\mathcal{A}}_{s,a}P_{s,a,q}\log\left(\frac{\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}}\right) \tag{9}$$

for all $s\in S_d$ and $q\in Succ(s)$. The $N\times 1$ vector of $x^{\mathcal{A}}_{s,a}$ variables is $x^{\mathcal{A}}$ where $x^{\mathcal{A}}_{s,a}$ has index s,a. Define $y=[x^{\mathcal{A}},r]^T$. We encode constraint (4c) with $G_{eq}y=h_{eq}$ where G_{eq} is a $N\times(N+M)$ matrix with $(s,(q,a))$-th entry $\mathbb{1}_s(q)-P_{q,a,s}$, and s-th entry of h is $\mathbb{1}_{s_0}(s)$. The constraint (4b) is encoded by $G_+y\geq 0$ where $G_+:=[I_{N\times N}|0_{N\times M}]$. The additional constraint given in (9) is encoded by $G_{(s,q)}y\in K_{\exp}$ where $G_{(s,q)}$ is a $3\times(N+M)$ matrix with $(1,N+(s,q))$-th entry -1, $(2,(s,a))$-th entry $P_{s,a,q}$ for all $a\in A(s)$, $(3,(s,a))$-th entry $\pi^{\mathcal{S}}_{s,q}$ for all $a\in A(s)$. The constraint (4d) is encoded by $G_{\mathcal{A}}y\geq\nu^{\mathcal{A}}$ where $G_{\mathcal{A}}$ is a $1\times(N+M)$ matrix where $(1,(s,a))$-th entry is $\mathbb{1}_{S_d\setminus R^{\mathcal{A}}}(s)\sum_{q\in R^{\mathcal{A}}}P_{s,a,q}$. Finally, $\mathcal{K}=\mathbb{R}^{N+M}\times\{0\}^{|S_d|}\times\mathbb{R}^N_+\times\mathcal{K}_{\exp}\times\ldots\times\mathcal{K}_{\exp}\times\mathbb{R}_+$, $G=[G_{eq},G_+,G_{(1,1)},\ldots,G_{(|S_d|,|S|)},G_{\mathcal{A}}]^T$, $h=[h_{eq},0,\ldots,0,\nu^{\mathcal{A}}]$, and $c=[0_{N\times 1},1_{M\times 1}]$.

The dual of the optimization problem in (8) is

$$\max_{u,w} \; h^T u \tag{10a}$$
$$\text{subject to} \quad \begin{bmatrix}G^T\\-I^T\end{bmatrix}u+w=c, \tag{10b}$$
$$w\in\mathcal{K}^*, \tag{10c}$$

where the decision variables are u and w, and $\mathcal{K}^*=\{0\}^{N+M}\times\mathbb{R}^{|S_d|}\times\mathbb{R}^N_+\times\mathcal{K}^*_{\exp}\times\ldots\times\mathcal{K}^*_{\exp}\times\mathbb{R}_+$.

By combining the optimization problem in (10) and the constraints in (5c)-(5f), and adding $x^{\mathcal{S}}_{s,a}$ as decision variables, we get an optimization problem that shares the same optimal value with (5). However, we remark that this problem is nonconvex because of the constraint (5c) and the bilinear constraints that are due to $\pi^{\mathcal{S}}_{s,q}$ parameter introduced in the construction of $G_{(s,q)}$.

C. Alternating Direction Method of Multipliers (ADMM)-based Approach for the Synthesis of Optimal Reference Policies

The alternating direction method of multipliers (ADMM) [8] is an algorithm to solve decomposable optimization problems by solving smaller pieces of the problem. We use the ADMM to locally solve the optimization problem given in (5). The objective function of (5) is decomposable since it is a sum across S_d where each summand consists of different variables. We exploit this feature to reduce the problem size via the ADMM.

For every state $s\in S_d$, we introduce $z^{\mathcal{A}}_s$ and $z^{\mathcal{S}}_s$ such that $z^{\mathcal{A}}_s=x^{\mathcal{A}}_s$ and $z^{\mathcal{S}}_s=x^{\mathcal{S}}_s$. With these extra variables, the augmented Lagrangian of (5) is

$$\begin{aligned}
&L(x^{\mathcal{S}},x^{\mathcal{A}},z^{\mathcal{S}},z^{\mathcal{A}},\lambda^{\mathcal{S}},\lambda^{\mathcal{A}})\\
&=\sum_{s\in S_d}\sum_{a\in A(s)}\sum_{q\in S}\left(x^{\mathcal{A}}_{s,a}P_{s,a,q}\log\left(\frac{\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}}\right)\right.\\
&-\mathcal{I}_{\mathbb{R}^{|A(s)|}_{\geq 0}}(x^{\mathcal{S}}_s)+\mathcal{I}_{\mathbb{R}^{|A(s)|}_{\geq 0}}(x^{\mathcal{A}}_s)-\rho^{\mathcal{S}}(x^{\mathcal{S}}_s-z^{\mathcal{S}}_s)^T\lambda^{\mathcal{S}}_s\\
&\left.+\rho^{\mathcal{A}}(x^{\mathcal{A}}_s-z^{\mathcal{A}}_s)^T\lambda^{\mathcal{A}}_s-\frac{\rho^{\mathcal{S}}}{2}\|x^{\mathcal{S}}_s-z^{\mathcal{S}}_s\|^2_2+\frac{\rho^{\mathcal{A}}}{2}\|x^{\mathcal{A}}_s-z^{\mathcal{A}}_s\|^2_2\right)\\
&-\mathcal{I}_{X^{\mathcal{S}}}(z^{\mathcal{S}})+\mathcal{I}_{X^{\mathcal{A}}}(z^{\mathcal{A}}),
\end{aligned}$$

where $\rho^{\mathcal{S}}$ and $\rho^{\mathcal{A}}$ are positive constants, $\lambda^{\mathcal{S}}$ and $\lambda^{\mathcal{A}}$ are the dual parameters, $X^{\mathcal{A}}$ is the set of occupation measures of the agent that satisfy (4c) and (4d), $X^{\mathcal{S}}$ is the set of occupation measures of the supervisor that satisfy (5e) and (5f), and $\pi^{\mathcal{S}}_{s,q}=\sum_{a\in A(s)}P_{s,a,q}x^{\mathcal{A}}_{s,a}/(\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'})$ for all $s\in S_d$ and $a\in A(s)$. In Algorithm 1 which is a modified version of the classical ADMM, we give the ADMM for the synthesis of reference policies. Note that we optimize $x^{\mathcal{S}}$ and $x^{\mathcal{A}}$ together to capture the characteristics of the maximin problem.

We remark that Algorithm 1 still requires solving a maximin optimization problem (see line 8). However, the maximin

Algorithm 1 The ADMM for the synthesis of reference policies

Input: An MDP $\mathcal{M}$, reachability specifications $\Diamond R_i^{\mathcal{S}}$ for all $i \in K^{\mathcal{S}}$ and $\Diamond R^{\mathcal{A}}$, probability thresholds $\nu_i^{\mathcal{S}}$ for all $i \in [K^{\mathcal{S}}]$ and $\nu^{\mathcal{A}}$.
Output: A reference policy $\pi^{\mathcal{S}}$.
Set $x^{\mathcal{S},0}$ and $z^{\mathcal{S},0}$ arbitrarily from $X^{\mathcal{S}}$.
Set $x^{\mathcal{A},0}$ and $z^{\mathcal{A},0}$ arbitrarily from $X^{\mathcal{A}}$.
Set $\lambda^{\mathcal{S},0}$ and $\lambda^{\mathcal{A},0}$ to 0.
$k = 0$.
while stopping criteria are not satisfied **do**
  Set $x^{\mathcal{S},k+1}$ and $x^{\mathcal{A},k+1}$ as the solution of $\max_{x^{\mathcal{S}}} \min_{x^{\mathcal{A}}} L(x^{\mathcal{S}}, x^{\mathcal{A}}, z^{\mathcal{S},k}, z^{\mathcal{A},k}, \lambda^{\mathcal{S},k}, \lambda^{\mathcal{S},k})$.
  $z^{\mathcal{S},k+1} := Proj_{X^{\mathcal{S}}}(x^{\mathcal{S},k+1} + \lambda^{\mathcal{S},k})$.
  $z^{\mathcal{A},k+1} := Proj_{X^{\mathcal{A}}}(x^{\mathcal{A},k+1} + \lambda^{\mathcal{A},k})$.
  $\lambda^{\mathcal{S},k+1} := \lambda^{\mathcal{S},k} + x^{\mathcal{S},k+1} - z^{\mathcal{S},k+1}$.
  $\lambda^{\mathcal{A},k+1} := \lambda^{\mathcal{A},k} + x^{\mathcal{A},k+1} - z^{\mathcal{A},k+1}$.
  $k := k + 1$.
end while
Compute $\pi^{\mathcal{S}}$ using $z^{\mathcal{S},k}$ as the occupation measures.

optimization problem in Algorithm 1 can be solved as a local maximin problem separately for each state since $x_s^{\mathcal{S}}$ and $x_s^{\mathcal{A}}$ are decoupled from $x_q^{\mathcal{S}}$ and $x_q^{\mathcal{A}}$ for all $s \neq q \in S_d$. While the number of variables for the problem obtained via dualization-based approach is $\mathcal{O}(|S||A|)$, it is $\mathcal{O}(|A|)$ for the local problems in the ADMM algorithm.

Since the strong duality holds, one can use a dualization-based approach as shown in Section V-B to solve the local maximin problems. We remark that after dualization, the resulting optimization problems are nonconvex similar to the optimization problem obtained via dualization-based approach.

Remark 5. *Convergence of ADMM for particular nonconvex optimization problems has been studied [29], [30]. To the best of our knowledge, the method based on the ADMM for the optimization problem given in* (5) *has no convergence guarantees and does not match with the any of the existing convergence results.*

D. A Linear Programming Relaxation for the Synthesis of Reference Policies

Since it is not possible to obtain a convex reformulation of the optimization problem given in (5) via a transformation, we give a convex relaxation of the problem. Intuitively, synthesizing a policy that minimizes the probability of satisfying the agent's specification is a good way to increase the KL divergence between the distributions of paths. Formally, consider a transformation of the path distributions that groups paths of $\mathcal{M}$ into two subsets: the paths that satisfy $\Diamond R^{\mathcal{A}}$ and the paths that do not satisfy $\Diamond R^{\mathcal{A}}$. After this transformation, the probability assigned to the first subset is $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}})$ under policy $\pi^{\mathcal{S}}$ and $\Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond R^{\mathcal{A}})$ under policy $\Diamond R^{\mathcal{A}}$. By the data processing inequality given in (1), this transformation yields a lower bound on the KL divergence between the path distributions: $KL\left(\Gamma_{\mathcal{M}}^{\pi^{\mathcal{A}}} || \Gamma_{\mathcal{M}}^{\pi^{\mathcal{S}}}\right)$ is greater than or equal to

$$KL\left(Ber\left(\Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}\left(s_0 \models \Diamond R^{\mathcal{A}}\right)\right) || Ber\left(\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}\left(s_0 \models \Diamond R^{\mathcal{A}}\right)\right)\right). \tag{11}$$

We use this lower bound to construct the relaxed problem

$$\sup_{\pi^{\mathcal{S}} \in \Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M})} \quad (11) \tag{12a}$$

$$\text{subject to} \quad \Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{12b}$$

$$\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R_i^{\mathcal{S}}) \geq \nu_i^{\mathcal{S}}, \ i \in [K^{\mathcal{S}}]. \tag{12c}$$

If $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}$, the agent may directly use the reference policy. Without loss of generality, assuming that $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}}) < \nu^{\mathcal{A}}$, the objective function of above optimization problem is decreasing in $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}})$ and increasing in $\Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond R^{\mathcal{A}})$. Hence, the problem

$$\sup_{\pi^{\mathcal{S}} \in \Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M})} \quad \Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond R^{\mathcal{A}}) - \Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}}) \tag{13a}$$

$$\text{subject to} \quad \Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{13b}$$

$$\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R_i^{\mathcal{S}}) \geq \nu_i^{\mathcal{S}}, \ i \in [K^{\mathcal{S}}]. \tag{13c}$$

shares the same optimal policies with the problem given in (12). We note that the optimization problem given in (13) can be solved separately for the supervisor's and the agent's parameters where both of the problems are linear optimization problems. The optimal reference policy for the relaxed problem is the policy that minimizes $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}})$ subject to $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R_i^{\mathcal{S}}) \geq \nu_i^{\mathcal{S}}$ for all $i \in [K^{\mathcal{S}}]$.

The lower bound given in (11) provides a sufficient condition on the optimality of a reference policy for Problem 2. A policy $\pi^{\mathcal{S}}$ satisfying $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R^{\mathcal{A}}) = 0$ and $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond R_i^{\mathcal{S}}) \geq \nu_i^{\mathcal{S}}$ for all $i \in [K^{\mathcal{S}}]$ is an optimal reference policy since the optimization problem given in (12) has the optimal value of ∞. However, in general the gap due to the relaxation may get arbitrarily large, and the reference policy synthesized via (12) is not necessarily optimal for Problem 2. For example, consider the MDP given in Figure 2a where the agent's policy again has $\pi_{s,\gamma}^{\mathcal{A}} = 1$. For simplicity, there is no specification for the supervisor, i.e., $\nu^{\mathcal{S}}$ is 0. The policy $\pi^{\mathcal{S}}$ that minimizes $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s \models \Diamond q_1 \vee \Diamond q_2)$ chooses action β at state s. This policy has a KL divergence value of 1.22. On the other hand, a policy that chooses action α is optimal and it has a KL divergence value of 1.30 even though it does not minimize the probability of satisfying $\Diamond q_1 \vee \Diamond q_2$. The gap of the lower bound may get arbitrarily large as P_{s,α,q_2} decreases. Furthermore, the policy synthesized via the relaxed problem may not even be locally optimal as P_{s,α,q_2} decreases.

The relaxed problem focuses on only one event, achieving the malicious objective, and fails to capture all transitions of the agent. On the other hand, the objective function of Problem 2, the KL divergence between the path distributions, captures all transitions of the agent rather than a single event. In particular, to detect the deviations the optimal deceptive

policy assigns a low probability to the transition from s to q_2 which inevitably happens with high probability for the agent. However, the policy synthesized via the relaxed problem fails to capture that the agent have to assign high probability to the transition from s to q_2.

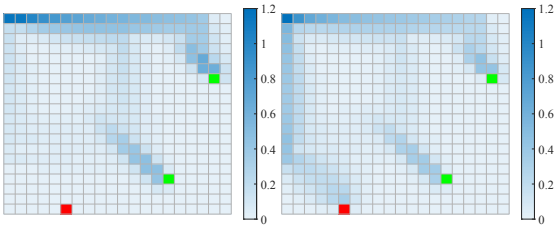
VI. Numerical Examples

In this section we give numerical examples on the synthesis of optimal deceptive policies and optimal reference policies. In Section VI-A we explain some characteristics of the optimal deceptive policies through different scenarios. In the second example given in Section VI-B, we compare the proposed metric, the KL divergence between the distributions of paths, to some other metrics. We demonstrate the ADMM-based algorithm with the example given in Section VI-C.

We solved the convex optimization problems with CVX [31] toolbox using MOSEK [32] and the nonconvex optimization problems using IPOPT [33].

A. Some Characteristics of Deceptive Policies

The first example demonstrates some of the characteristics of the optimal deceptive policies. The environment is a 20×20 grid world given in Figure 3. The green and red states are denoted with sets g and r, respectively. At every state, there are 4 available actions, namely, up, down, left, and right. When the agent takes an action the transition happens into the target direction with probability 0.7 and in the other directions uniformly randomly with probability 0.3. If a direction is out of the grid, the transition probability of that direction is proportionally distributed to the other directions. The green and red states are absorbing. The initial state is the top-left state.

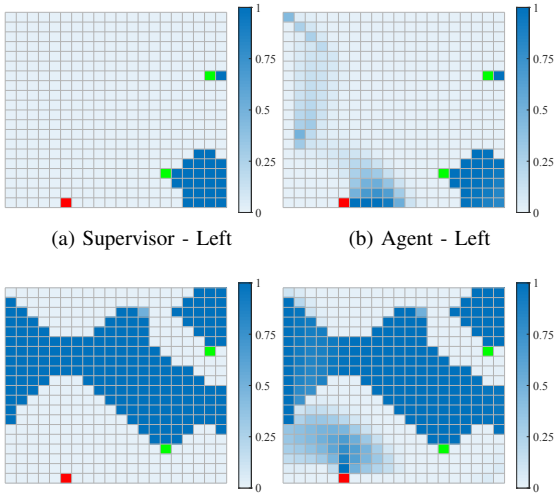


(a) Reference (b) Deceptive

Fig. 3. Heat maps of the occupation measures. The value of a state is the expected number of visits to the state. The deceptive policy makes the agent move towards the red state to achieve the malicious objective.

The specification of the supervisor is to reach each of the green states with probability at least 0.4. The reference policy is constructed so that it reaches the green states with probability at least $1 - 10^{-4}$ in the minimum expected time. The specification of the agent is to reach the red state. The specification is encoded with $\Diamond r$. The probability threshold $\nu^{\mathcal{A}}$ for the agent's specification is 0.3. We synthesize the policy of the agent according to Problem 1, which leads to the KL divergence value of 2.662. While the reference policy satisfies $\Diamond r$ with probability 10^{-4}, the agent's policy satisfies $\Diamond r$ with probability 0.3.

In Figure 3b, we observe that if the deceptive agent is close to the green states, it does not try to reach the red state since deviations from the reference policy in these regions incur high divergence. Instead, as we see in Figure 4, the deceptive policy makes the agent move towards left in the first steps and reach the red state by going down. The misleading occurs during this period: while the agent goes left on purpose, it may hold the stochasticity of the environment accountable for this behavior. We also observe a significant detail in the agent's deceptive policy. The deceptive policy aims to reach the left border since the reference policy takes action down in this region. The agent wants to drive himself to this region to directly follow the reference policy without any divergence. Thus the agent deviates from the reference policy at a particular state to be close to the reference policy as much as possible in the rest of the path. Once the agent is close to the red state, it again deviates from the reference policy and takes action down with a high probability to reach the red state.



(a) Supervisor - Left (b) Agent - Left

(c) Supervisor - Down (d) Agent - Down

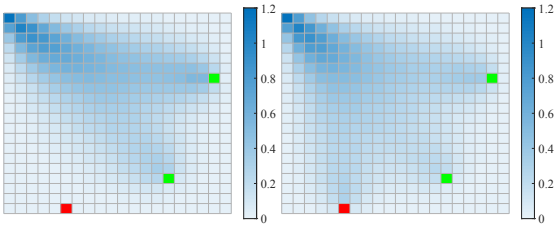
Fig. 4. The assigned probabilities to the actions when the yellow state was visited, but the red state was not visited.

We note that the reference policy is restrictive in this case; as can be seen in Figure 3a, it follows almost a deterministic path. Under such a reference policy, even the policy that is synthesized via Problem 1 is easy to detect. To observe the effect of the reference policy on the deceptive policy, we consider a different reference policy as shown in Figure 5a, which satisfies $\Diamond r$ with probability 10^{-3}. When the reference policy is not as restrictive, the deceptive policy becomes hard to detect. Formally, the value of the KL divergence reduces to 1.462.

B. Detection of a Deceptive Agent

In this example, by comparing KL divergence with some common metrics to synthesize the deceptive policies, we show how the choice of KL divergence helps with preventing detection. We compare the metrics using a randomly generated MDP and an MDP modeling a region from San Francisco.

The randomly generated MDP consists of 21 states. In particular, there are 20 transient states with 4 actions and an



(a) Supervisor (b) Agent

Fig. 5. Heatmaps of the occupation measures. The deceptive policy is hard to detect under a reference policy that is not restrictive.

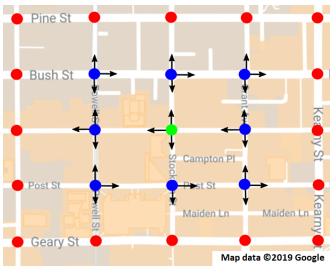


Fig. 6. The map of a region from north east of San Francisco. The green dot indicates the intersection at which the highest number of crimes happened. The data is from [34]. The dots on the map represent the states of the MDP and the arrows represent the available actions. The initial state is chosen uniformly randomly among the blue states and the red states are absorbing. The agent aims to patrol the green state.

absorbing state with 1 action. For the transient states, each action has a successor state that is chosen uniformly randomly among the transient states. In addition to these actions, every transient state has an action that has the absorbing state as the successor state. At every transient state, the reference policy goes to the absorbing state with probability 0.15 and the other successor states with probability 0.85. The agent's specification $\phi^{\mathcal{A}}$ is to reach to a specific transient state.

We randomly generate a reference policy for the randomly generated MDP. The reference policy satisfies the agent's specification with probability 0.30. For the reference policy, we synthesize three candidate policies for deception: by minimizing the KL divergence between the path distributions of the agent's policy and the reference policies, by minimizing the L_1-norm between the occupation measures of the state-action pairs for the agent's policy and the reference policies, and by minimizing the L_2-norm between the occupation measures of the state-action pairs for the agent's policy and the reference policies. The candidate policies are constructed so that they satisfy the agent's specification $\phi^{\mathcal{A}}$ with probability 0.9. For each candidate policy, we run 100 simulations each of which consists of 100 independently sampled paths.

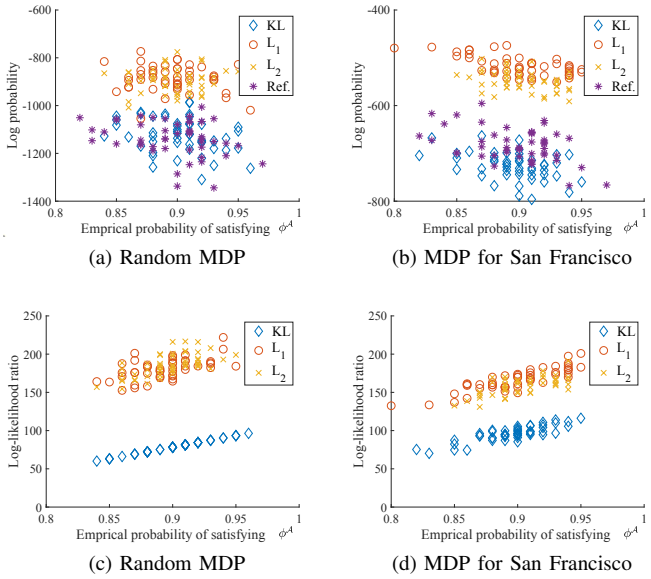
We also simulate the agent's trajectories under the reference policies. In particular, we aim to observe the case where the empirical probability of satisfying $\phi^{\mathcal{A}}$ is approximately 0.9. Note that this is a rare event under the reference policy. We simulate this rare event in the following way. Let $\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}$ be the probability distribution of paths under the reference policy. We create two conditional probability distributions $\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M},+}$ and $\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M},-}$ which are the distribution of paths under the reference policy given that the paths satisfy $\phi^{\mathcal{A}}$ and do not satisfy $\phi^{\mathcal{A}}$, respectively. We sample from $\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M},-}$ with probability 0.9 and $\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M},-}$ with probability 0.1.

In addition to the randomly generated MDP, we use a different MDP to show that the deceptive policy can help patrolling without being detected. The MDP models a region in the north east of San Francisco. The map of the region is given in Figure 6 where each intersection is represented with a state and each road is represented with an action. We design the reference policy to represent the average driver behavior. We obtain the traffic density data from Google Maps [35] and synthesize the reference policy by fitting a stationary policy to the data. The aim of the agent is to patrol the intersection at which the highest number of crimes happens. Formally, the agent's policy reaches the intersection with probability at least 0.9 whereas the reference policy reaches the intersection with probability 0.28. For the simulation, we use the steps as in the randomly generated MDP.

For each simulation, we plot the log-probability under the reference policy and the log-likelihood ratio compared to the reference policy in Figure 7. In Figures 7a and 7b, we observe that the metrics based on L_1 and L_2 norms generate paths that are more probable under the reference policy compared to KL divergence metric. However, we also observe that the probabilities for the KL divergence metric looks closer to the probabilities of the simulations of the reference policy that are satisfying $\phi^{\mathcal{A}}$ with high probability. The paths generated via the KL divergence coincides with the typical set of this rare event. In other words, the other metrics provide paths that have higher probabilities under the reference policy. However, given that the empirical distribution of the paths satisfy $\phi^{\mathcal{A}}$ with high probability, the likelihood of the path generated by the KL divergence is higher and the supervisor expects to see paths that are close to the paths generated by the KL divergence metric. This phenomenon confirms the motivation provided by Sanov's theorem which is explained in Section III. In Figures 7c and 7d, we observe that the paths generated under the KL divergence metric has a lower log-likelihood ratio compared to the other metrics as explained in Section III. This result shows that compared to the other candidate deceptive policies, the deceptive policy generated with the KL divergence is the least likely to be detected under the likelihood-ratio test.

C. Optimal Reference Policies

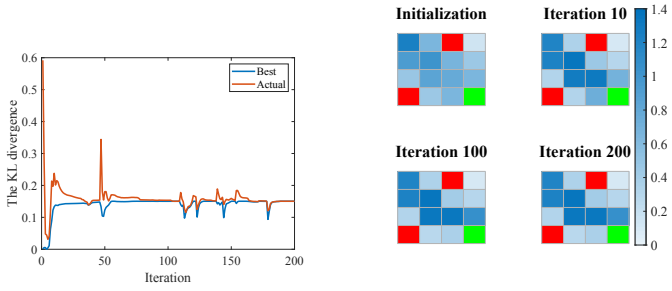
We present an example of synthesis of optimal reference policies. The environment is a 4×4 grid world given in Figure 8b and is similar to the environment described in the example for the characteristics of deceptive policies. The green and red states are denoted with sets g and r, respectively. At every state, there are 4 available 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 0.7 and in the other directions uniformly randomly with probability 0.3. If a direction is out of the grid the transition probability to that direction is proportionally distributed to the other directions. The green state is absorbing and the initial state is the top-left state.



(a) Random MDP (b) MDP for San Francisco

(c) Random MDP (d) MDP for San Francisco

Fig. 7. (a)-(b) The log-likelihoods under the reference policy. 'Ref.' refers to the rare events of the reference policy that satisfies $\phi^{\mathcal{A}}$ with high probability. 'KL', '$\mathbf{L_1}$', and '$\mathbf{L_2}$' refer to the candidate deceptive policies. (c)-(d) The log-likelihood ratios compared with the reference policy.



(a) (b)

Fig. 8. (a) The KL divergence between the agent's policy and the reference policy. The curve "Best" refers to the case that the agent's policy is the best deceptive policy against the reference policy synthesized during the ADMM algorithm. The curve "Actual" refers to the case that the agent's policy is the policy synthesized during the ADMM algorithm. (b) Heatmaps of the occupation measures for the reference policy, i.e., $z^{\mathcal{S},k}$ parameters of the Algorithm 1. The value of a state is the expected number of visits to the state.

The specification of the supervisor is to reach the green state, i.e., $\Diamond g$. Note that the specification of the supervisor is satisfied with probability 1 under any policy. The specification of the agent is to reach one of the red states, i.e., $\Diamond r$. The probability threshold for the agent's task is 0.3.

We synthesize the reference policy via Algorithm 1 given in Section V-C. In Algorithm 1, $z^{\mathcal{S},k}$ represents the reference policy synthesized at iteration k. Similarly, $z^{\mathcal{A},k}$ represents the deceptive policy synthesized at iteration k. We plot the values of the KL divergences between these policies in Figure 8a and give the heatmaps for the occupation measures in Figure 8b. After few tens of iterations of the ADMM algorithm, the KL divergence value is near to the limit value which is 0.150.

In Figure 8a, we also note that if the actual KL divergence value increases suddenly, the best response KL divergence value decreases. The reference policy tries to exploit suboptimal deceptive policies. While this exploitation increases the actual value, it causes suboptimality for the reference policy against the best deceptive policy.

The reference policy gradually gets away from the red states as shown in Figure 8b. Based on this observation, we expect that the relaxed problem given in Section V-D provides useful reference policies for the original problem. This expectation is indeed verified numerically: The reference policy synthesized via the relaxed problem, has a KL divergence of 0.150, which is equal to the limit value of the ADMM algorithm.

VII. Conclusion

We considered the problem of deception under a supervisor that provides a reference policy. We modeled the problem using MDPs and reachability specifications and proposed to use KL divergence for the synthesis of optimal deceptive policies. We showed that an optimal deceptive policy is stationary and its synthesis requires solving a convex optimization problem. We also considered the synthesis of optimal reference policies that easily prevent deception. We showed that this problem is NP-hard. We proposed a method based on the ADMM to compute a locally optimal solution and provided an approximation that can be modeled as a linear program.

In subsequent work we aim to extend the deception problem to a multi-agent settings where multiple malicious agents need to cooperate. Furthermore, it would be interesting to consider the case where a malicious agent first needs to detect the other malicious agents before cooperation. We also aim to study the scenario where the supervisor needs to learn the specification of the agent for the synthesis of the reference policy.

Appendix I
Synthesis of Optimal Deceptive Policies under Nondeterministic Reference Policies

In Problem 1, we assume that the supervisor provides an explicit (possibly probabilistic) reference policy to the agent. It is possible that the reference policy is nondeterministic such that the supervisor disallows some actions at every state and the agent is allowed to take the other actions. We refer the interested readers to [36] for the formal definition of the nondeterministic policies. A *permissible policy* is a policy that takes an allowed action at every time step. The nondeterministic reference policy represents a set $\Pi^{\mathcal{S}}$ of policies that are permissible. The supervisor is indifferent between the policies in $\Pi^{\mathcal{S}}$. As in Section III, we assume that the supervisor observes the transitions, but not the actions of the agent.

We define the optimal deceptive policy as the policy that minimizes the KL divergence to any permissible policy subject to the task constraint of the agent. Under this definition, the optimal deceptive policy mimics the rare event that satisfies the agent's task, and that is most probable under one of the permissible policies. Formally, we solve the following problem for the synthesis of optimal deceptive policies under nondeterministic reference policies.

Problem 3 (Synthesis of Optimal Deceptive Policies under Nondeterministic Reference Policies)**.** *Given an MDP $\mathcal{M}$, a reachability specification $\Diamond R^{\mathcal{A}}$, a probability threshold $\nu^{\mathcal{A}}$, and a set $\Pi^{\mathcal{S}}$ of permissible policies, solve*

$$\inf_{\substack{\pi^{\mathcal{A}}\in\Pi(\mathcal{M})\\ \pi^{\mathcal{S}}\in\Pi^{\mathcal{S}}}} \quad KL\left(\Gamma_{\mathcal{M}}^{\pi^{\mathcal{A}}}||\Gamma_{\mathcal{M}}^{\pi^{\mathcal{S}}}\right) \tag{14a}$$

$$\text{subject to} \quad \Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond R^{\mathcal{A}}) \geq \nu^{\mathcal{A}}. \tag{14b}$$

If the optimal value is attainable, find a policy $\pi^{\mathcal{A}}$ that is a solution to (14)*.*

Similar to Assumption 2, we assume that for every state in MDP $\mathcal{M}$, the set of allowed actions is fixed. For state $s \in S$, we denote the set of allowed actions with $A^{\mathcal{S}}(s)$ and the possible successor state distributions with $\Lambda^{\mathcal{S}}(s)$. Under this assumption, we can identify the maximal end components of $\mathcal{M}$ for the permissible policies. If a maximal end component is closed, i.e., the maximum probability leaving the end component under the permissible policies is 0, then the states of the maximal end component belongs to C^{cl}. If a maximal end component is open, i.e., the maximum probability leaving the end component under the permissible policies is 1, then there exists an optimal policy that eventually leaves the end component since the agent can guarantee the same objective value by following a policy that leaves the end component and takes permissible actions after leaving the end component.

While there exists an optimal policy that leaves the open end components eventually, the optimal policy may have infinite occupation measures at these states. We have the following assumption to ensure the boundedness of the occupation measures for these states.

Assumption 4. *For all $s \in S \setminus (C^{cl} \cup R^{\mathcal{A}})$ and $a \in A(s)$, the occupation measure $x_{s,a}^{\mathcal{A}}$ is upper bounded by θ.*

As in the proof of Proposition 2, we can define a semi-infinite MDP with an optimal cost equal to the optimal value of Problem 3. In detail, at state $s \in S_d$, an action a with successor state distribution $X_{s,a}$ has cost $\min_{X_s^{\mathcal{S}}\in\Lambda_s^{\mathcal{S}}} KL(X_{s,a}||X_s^{\mathcal{S}})$ where $X_s^{\mathcal{S}}$ is the distribution of successor states under the reference policy. We note that $\min_{X_s^{\mathcal{S}}\in\Lambda_s^{\mathcal{S}}} KL(X_{s,a}||X_s^{\mathcal{S}})$ is a convex function of $X_{s,a}$ since the Kullback-Leibler divergence is jointly convex in its arguments [28]. Since the occupation measures are bounded for all states in $S_d = S \setminus (C_{cl} \cup C_{\mathcal{A}})$ due to Assumption 4 and the costs are convex functions of the policy parameters, there exists a stationary deterministic optimal policy for the semi-infinite MDP [23]. Consequently, there exists a stationary randomized optimal policy for Problem 3 under Assumption 4.

Given that the optimal policy is stationary on $\mathcal{M}$, we compute the state-action occupation measures of the optimal policy by solving the following optimization problem:

$$\inf \sum_{s\in S_d}\sum_{a\in A(s)}\sum_{q\in Succ(s)} x_{s,a}^{\mathcal{A}} P_{s,a,q}$$

$$\log\left(\frac{\sum_{a'\in A(s)} x_{s,a'}^{\mathcal{A}} P_{s,a',q}}{\left(\sum_{a\in A^{\mathcal{S}}(s)} \pi_{s,a} P_{s,a,q}\right)\left(\sum_{a'\in A(s)} x_{s,a'}^{\mathcal{A}}\right)}\right) \tag{15a}$$

subject to

(4b) − (4d)

$$\sum_{s\in S_d}\sum_{a\in A(s)} x_{s,a}^{\mathcal{A}} \leq \theta, \tag{15b}$$

$$\sum_{a\in A^{\mathcal{S}}(s)} \pi_{s,a}^{\mathcal{S}} = 1, \qquad \forall s \in S_d \tag{15c}$$

where the decision variables are $x_{s,a}^{\mathcal{A}}$ for all $s \in S_d$ and $a \in A(s)$ and $\pi_{s,a}$ for all $s \in S_d$ and $a \in A^S(s)$.

We remark that the objective function of (15) is a jointly convex function of $x_{s,a}^{\mathcal{A}}$ and $\pi_{s,a}$ parameters. Having computed a set of optimal occupation measures, one can compute a stationary optimal deceptive policy.

APPENDIX II
PROOFS FOR THE TECHNICAL RESULTS

We use the following definition and lemmas in the proof of Proposition 1. We use $\Pr_{\mathcal{M}}^{\pi}(s \models \bigcirc\Diamond s)$ to denote the probability that s is visited again from initial state s under the stationary policy π .

Definition 2. *Let Q be a probability distribution with a countable support $\mathcal{X}$. The entropy of Q is $H(Q) = -\sum_{x\in\mathcal{X}} Q(x)\log(Q(x))$.*

Lemma 1 (Theorem 5.7 of [37])**.** *Let $\mathcal{D}$ be the set of a distributions with support $\{1, 2, \ldots\}$ and the expected value of c. A random variable $X^* \sim Geo(1/c)$ maximizes $H(X)$ subject to $X \in \mathcal{D}$ where $H(X^*) = c\left(-\frac{1}{c}\log\left(\frac{1}{c}\right) - \left(1-\frac{1}{c}\right)\log\left(1-\frac{1}{c}\right)\right) = cH\left(Ber\left(\frac{1}{c}\right)\right)$.*

Lemma 2. *Consider an MDP $\mathcal{M} = (S, A, P, AP, L)$. Let N_s^{π} denote the number of visits to the state s under a stationary policy π such that $\mathbb{E}[N_s^{\pi}] < \infty$. N_s^{π} satisfies $\Pr(N_s^{\pi} = 0) = \Pr_{\mathcal{M}}^{\pi}(s_0 \not\models \Diamond s)$ and $\Pr(N_s^{\pi} = i) = \Pr_{\mathcal{M}}^{\pi}(s_0 \models \Diamond s)\Pr_{\mathcal{M}}^{\pi}(s \models \bigcirc\Diamond s)^{i-1}\Pr_{\mathcal{M}}^{\pi}(s \not\models \bigcirc\Diamond s)$.*

Proof of Proposition 1. We prove this proposition by contradiction. We first provide a lower bound for the objective function of Problem 1. Then, we show that as the state-action occupation measures approach to infinity, the lower bound approaches to infinity. Hence, the state-action occupation measures must be bounded in order to have a finite value for the objective function of Problem 1.

Let d^* be the optimal value of Problem 1. For a state $s \in S \setminus C^{cl}$, first consider the case $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond s) = 0$, i.e., s is unreachable under $\pi^{\mathcal{S}}$. In this case, the agent's policy $\pi^{\mathcal{A}}$ must satisfy $\Pr_{\mathcal{M}}^{\pi^{\mathcal{A}}}(s_0 \models \Diamond s) = 0$, i.e., s must be unreachable under $\pi^{\mathcal{A}}$, otherwise the KL divergence is infinite. Hence the occupation measure is zero in this case.

Consider $\Pr_{\mathcal{M}}^{\pi^{\mathcal{S}}}(s_0 \models \Diamond s) > 0$. For this case, we will show that if the occupation measure is greater than some finite

value, then the KL divergence between the path distributions is greater than d^*. Denote the number visits to s with $N_s^{\pi^{\mathcal{A}}}$ and $N_s^{\pi^{\mathcal{S}}}$ under $\pi^{\mathcal{A}}$ and $\pi^{\mathcal{S}}$, respectively. We have the following claim: Given $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \Diamond s) > 0$, $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s \models \bigcirc\Diamond s) \in [0,1)$, and $d^* > 0$, there exists an M_s such that for all $\pi^{\mathcal{A}}$ that satisfies $\mathbb{E}[N_s^{\pi^{\mathcal{A}}}] > M_s$, we have $KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}) > d^*$.

We consider a partitioning of paths according to the number of times s appears in a path. By the data processing inequality given in (1), we have that $KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}) \geq KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}})$, i.e., the KL divergence between the path distributions is lower bounded by the KL divergence between the distributions of number visits to s. Therefore it suffices to prove the following claim: Given $\Pr^{\mathcal{S}}_{\mathcal{M}}(s_0 \models \Diamond s) > 0$, $\Pr^{\mathcal{S}}_{\mathcal{M}}(s \models \bigcirc\Diamond s) \in [0,1)$, and $d^* > 0$, there exists an M_s such that for all $\pi^{\mathcal{A}}$ that satisfies $\mathbb{E}[N_s^{\pi^{\mathcal{A}}}] > M_s$, we have $KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}}) > d^*$.

Define a random variable $\hat{N}_s^{\pi^{\mathcal{A}}}$ such that $\Pr(\hat{N}_s^{\pi^{\mathcal{A}}} = i) = \Pr(N_s^{\pi^{\mathcal{A}}} = i | N_s^{\pi^{\mathcal{A}}} > 0)$. For notational convenience denote $r^{\mathcal{S}} = 1 - \Pr^{\mathcal{S}}_{\mathcal{M}}(s_0 \models \Diamond s)$, $l^{\mathcal{S}} = \Pr^{\mathcal{S}}_{\mathcal{M}}(s \models \bigcirc\Diamond s)$, $p_i = \Pr(N_s^{\pi^{\mathcal{A}}} = i)$ and $\hat{p}_i = \Pr(\hat{N}_s^{\pi^{\mathcal{A}}} = i)$. Also define $M_s^{\mathcal{A}} := \mathbb{E}[N_s^{\pi^{\mathcal{A}}}]$, $\hat{M}_s^{\mathcal{A}} := \mathbb{E}[\hat{N}_s^{\pi^{\mathcal{A}}}] = \frac{M_s^{\mathcal{A}}}{1-p_0}$, and $M_s^{\mathcal{S}} := \mathbb{E}[N_s^{\pi^{\mathcal{S}}}]$.

We want to show that $M_s^{\mathcal{A}}$ is bounded for a finite d^*. Assume that $M_s^{\mathcal{A}} \leq M_s^{\mathcal{S}}$. In this case the $M_s^{\mathcal{A}}$ is finite since $M_s^{\mathcal{S}}$ is finite. If $M_s^{\mathcal{A}} > M_s^{\mathcal{S}}$, we have

$$
\begin{aligned}
&KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}}) && \text{(16a)}\\
&= p_0 \log\left(\frac{p_0}{r^{\mathcal{S}}}\right) + \sum_{i=1}^{\infty} p_i \log\left(\frac{p_i}{(1-r^{\mathcal{S}})(l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})}\right) && \text{(16b)}\\
&= p_0 \log\left(\frac{p_0}{r^{\mathcal{S}}}\right) + \sum_{i=1}^{\infty} (1-p_0)\hat{p}_i \log\left(\frac{1-p_0}{1-r^{\mathcal{S}}}\right)\\
&\quad + \sum_{i=1}^{\infty} (1-p_0)\hat{p}_i \log\left(\frac{\hat{p}_i}{(l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})}\right) && \text{(16c)}\\
&= p_0 \log\left(\frac{p_0}{r^{\mathcal{S}}}\right) + (1-p_0)\log\left(\frac{1-p_0}{1-r^{\mathcal{S}}}\right)\\
&\quad + \sum_{i=1}^{\infty} (1-p_0)\hat{p}_i \log\left(\frac{\hat{p}_i}{(l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})}\right) && \text{(16d)}\\
&\geq (1-p_0)\sum_{i=1}^{\infty} \hat{p}_i \log\left(\frac{\hat{p}_i}{(l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})}\right) && \text{(16e)}\\
&= (1-p_0)\sum_{i=1}^{\infty} \hat{p}_i \log\left(\hat{p}_i\right)\\
&\quad - (1-p_0)\sum_{i=1}^{\infty} \hat{p}_i \log\left((l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})\right) && \text{(16f)}\\
&= -(1-p_0)H(\hat{N}_s^{\pi^{\mathcal{A}}}) - (1-p_0)\sum_{i=1}^{\infty} \hat{p}_i \log\left((l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})\right) && \text{(16g)}
\end{aligned}
$$

where the equality (16b) follows from Lemma 2. The inequality in (16e) holds since the removed terms correspond to $KL(Ber(p_0)||Ber(r^{\mathcal{S}}))$ which is nonnegative.

By using Lemma 1 to upper bound $H(\hat{N}_s^{\pi^{\mathcal{A}}})$ and the definitions we have the following inequality.

$$
\begin{aligned}
&KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}})\\
&\geq (p_0-1)\left(H(\hat{N}_s^{\pi^{\mathcal{A}}}) + \sum_{i=1}^{\infty} \hat{p}_i \log\left((l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})\right)\right) && \text{(17a)}\\
&\geq (p_0-1)\left(\hat{M}_s^{\mathcal{A}} H\left(Ber\left(\frac{1}{\hat{M}_s^{\mathcal{A}}}\right)\right)\right.\\
&\quad \left. + \sum_{i=1}^{\infty} \hat{p}_i \log\left((l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})\right)\right) && \text{(17b)}\\
&= (p_0-1)\left(\hat{M}_s^{\mathcal{A}} H\left(Ber\left(\frac{1}{\hat{M}_s^{\mathcal{A}}}\right)\right)\right.\\
&\quad \left. + \sum_{i=1}^{\infty} \hat{p}_i \left((i-1)\log(l^{\mathcal{S}}) + \log(1-l^{\mathcal{S}})\right)\right) && \text{(17c)}\\
&= (p_0-1)\left(\hat{M}_s^{\mathcal{A}} H\left(Ber\left(\frac{1}{\hat{M}_s^{\mathcal{A}}}\right)\right)\right.\\
&\quad \left. + \left(\log(1-l^{\mathcal{S}}) + \left(\hat{M}_s^{\mathcal{A}} - 1\right)\log(l^{\mathcal{S}})\right)\right) && \text{(17d)}\\
&= (1-p_0)\left(-\hat{M}_s^{\mathcal{A}} H\left(Ber\left(\frac{1}{\hat{M}_s^{\mathcal{A}}}\right)\right)\right.\\
&= M_s^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1}{\hat{M}_s^{\mathcal{A}}}\right)||Ber(1-l^{\mathcal{S}})\right)\right) && \text{(17e)}
\end{aligned}
$$

Now assume that $M_s^{\mathcal{A}} \geq \frac{c}{1-l^{\mathcal{S}}}$ where $c > 1$ is a constant. In this case, we have

$$
\begin{aligned}
&KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}}) && \text{(18a)}\\
&\geq M_s^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1}{\hat{M}_s^{\mathcal{A}}}\right)||Ber(1-l^{\mathcal{S}})\right)\right) && \text{(18b)}\\
&\geq M_s^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1}{M_s^{\mathcal{A}}}\right)||Ber\left(1-l^{\mathcal{S}}\right)\right)\right) && \text{(18c)}\\
&\geq M_s^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1-l^{\mathcal{S}}}{c}\right)||Ber\left(1-l^{\mathcal{S}}\right)\right)\right) && \text{(18d)}
\end{aligned}
$$

since $\hat{M}_s^{\mathcal{A}} > M_s^{\mathcal{A}}$ and for a variable x such that $x \geq \frac{1}{1-l^{\mathcal{S}}}$, the value of $KL(Ber(\frac{1}{x})||Ber(1-l^{\mathcal{S}}))$ is increasing in x.

Note that $KL\left(Ber\left(\frac{1-l^{\mathcal{S}}}{c}\right)||Ber\left(1-l^{\mathcal{S}}\right)\right)$ is a positive constant. We can easily see that there exists an M_s such that $KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}}) > d^*$ if $M_s^{\mathcal{A}} > M_s$.

We proved that for a given constant, for every transient state of the supervisor the occupancy measure under the agent's policy must be bounded by some constant otherwise the KL divergence between distributions for the number of states to this state is greater than the constant. Since the KL divergence between the path distributions is lower bounded by the KL divergence for states, the finiteness of the KL divergence between the path distributions implies that the occupancy measure under the agent's policy for every transient state of the supervisor.

Thus, if the optimal value of Problem 1 is finite, the occupation measures under $\pi^{\mathcal{A}}$ must be bounded by some $M_s < \infty$ for all $s \in S \setminus C^{cl}$. ■

We use the following definition in the proof of Lemma 3. We remark that the proof of Lemma 3 is fairly similar with the proof of Lemma 2 from [14].

Definition 3. *A k-length path fragment $\xi = s_0 s_1 \dots s_k$ for an MDP $\mathcal{M}$ is a sequence of states under policy $\pi = \mu_0 \mu_1 \dots$*

such that $\sum_{a\in A(s_t)} P(s_t,a,s_{t+1})\mu_t(s_t,a) > 0$ *for all* $k > t \geq 0$. *The distribution of k-length path fragments for* $\mathcal{M}$ *under policy* π *is denoted by* $\Gamma^{\pi}_{\mathcal{M},k}$.

Lemma 3. *The KL divergence* $KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M},k}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M},k})$ *between the distributions of k-length path fragments for stationary policies* $\pi^{\mathcal{A}}$ *and* $\pi^{\mathcal{S}}$ *is equal to the expected sum of KL divergences between the successor state distributions of* $\pi^{\mathcal{A}}$ *and* $\pi^{\mathcal{S}}$ *that is*

$$\sum_{t=0}^{k-1}\sum_{s\in S_d}\Pr^{\pi^{\mathcal{A}}}(s_t=s)\sum_{q\in Succ(s)}\sum_{a\in A(s)}P_{s,a,q}\pi^{\mathcal{A}}_{s,a}\log\left(\frac{\sum_{a'\in A(s)}P_{s,a',q}\pi^{\mathcal{A}}_{s,a'}}{\sum_{a'\in A(s)}P_{s,a',q}\pi^{\mathcal{S}}_{s,a}}\right).$$

Furthermore, if $KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}})$ *is finite, it is equal to*

$$\sum_{s\in S_d}\sum_{q\in S_d}\sum_{a\in A(s)}P_{s,a,q}x^{\mathcal{A}}_{s,a}\log\left(\frac{\sum_{a'\in A(s)}P_{s,a',q}x^{\mathcal{A}}_{s,a'}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}}\right)$$

Proof of Lemma 3. For MDP $\mathcal{M}$, denote the set of k-length path fragments by Ξ_k and the probability of the k-length path fragment $\xi_k = s_0 s_1 \dots s_k$ under the stationary policy π by $\Pr^{\pi}(\xi_k)$. We have $\Pr^{\pi}(\xi_k) = \prod_{t=0}^{k-1}\sum_{a\in A(s_t)}P_{s_t,a,s_{t+1}}\pi_{s_t,a}$. Consequently, we have

$$\begin{aligned}
KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M},k}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M},k}) &= \sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(\xi_k)\log\left(\frac{\Pr^{\pi^{\mathcal{A}}}(\xi_k)}{\Pr^{\pi^{\mathcal{S}}}(\xi_k)}\right)\\
&= \sum_{t=0}^{k-1}\sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(\xi_k)\log\left(\frac{\sum_{a'\in A(s_t)}P_{s_t,a',s_{t+1}}\pi^{\mathcal{A}}_{s_t,a'}}{\sum_{a'\in A(s_t)}P_{s_t,a',s_{t+1}}\pi^{\mathcal{S}}_{s_t,a'}}\right)\\
&= \sum_{t=0}^{k-1}\sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(\xi_k)\sum_{s\in S_d}\mathbb{1}_s(s_t)\\
&\quad\sum_{q\in Succ(s)}\mathbb{1}_q(s_{t+1}|s_t=s)\log\left(\frac{\sum_{a'\in A(s_t)}P_{s,a',q}\pi^{\mathcal{A}}_{s,a'}}{\sum_{a'\in A(s_t)}P_{s,a',q}\pi^{\mathcal{S}}_{s,a'}}\right)\\
&= \sum_{t=0}^{k-1}\sum_{s\in S_d}\Pr^{\pi^{\mathcal{A}}}(s_t=s)\sum_{q\in Succ(s)}\sum_{a\in A(s_t)}\\
&\quad P_{s,a,q}\pi^{\mathcal{A}}_{s,a'}\log\left(\frac{\sum_{a'\in A(s_t)}P_{s,a',q}\pi^{\mathcal{A}}_{s,a'}}{\sum_{a'\in A(s_t)}P_{s,a',q}\pi^{\mathcal{S}}_{s,a'}}\right)
\end{aligned}$$

If $KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}})$ is finite, we have

$$\begin{aligned}
KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}) &= \lim_{k\to\infty}KL(\Gamma^{\pi^{\mathcal{A}}}_{k}||\Gamma^{\pi^{\mathcal{S}}}_{k})\\
&= \lim_{k\to\infty}\sum_{s\in S_d}\sum_{q\in Succ(s)}\sum_{a\in A(s_t)}\sum_{t=0}^{k-1}\\
&\quad\Pr^{\pi^{\mathcal{A}}}(s_t=s)P_{s,a,q}\pi^{\mathcal{A}}_{s,q}\log\left(\frac{\sum_{a'\in A(s_t)}P_{s,a',q}\pi^{\mathcal{A}}_{s,q}}{\sum_{a'\in A(s_t)}P_{s,a',q}\pi^{\mathcal{S}}_{s,q}}\right)\\
&= \sum_{s\in S_d}\sum_{q\in Succ(s)}\sum_{a\in A(s)}\\
&\quad P_{s,a,q}x^{\mathcal{A}}_{s,a}\log\left(\frac{\sum_{a'\in A(s)}P_{s,a',q}x^{\mathcal{A}}_{s,a'}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}}\right).
\end{aligned}$$

Finally, since $P_{s,a,q}$ is zero for all $q \notin Succ(s)$ and we defined $0\log 0 = 0$, we can safely replace $Succ(s)$ with S. ■

Proof of Proposition 3. Assume that $KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}})$ is finite under the stationary policies $\pi^{\mathcal{A}}$ and $\pi^{\mathcal{S}}$. The objective function of the problem given in (2) is equal to

$$\sum_{s\in S_d}\sum_{q\in Succ(s)}\sum_{a\in A(s)}P_{s,a,q}x^{\mathcal{A}}_{s,a}\log\left(\frac{\sum_{a'\in A(s)}P_{s,a',q}x^{\mathcal{A}}_{s,a'}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}}\right)$$

due to Lemma 3. The constraints (4b)-(4c) define the stationary policies that make the states in S_d have valid and finite occupation measures and the constraint (4d) encodes the reachability constraint.

Note that

$$\sum_{q\in S}\sum_{a\in A(s)}P_{s,a,q}x^{\mathcal{A}}_{s,a}\log\left(\frac{\sum_{a'\in A(s)}P_{s,a',q}x^{\mathcal{A}}_{s,a'}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}}\right)$$

is the KL divergence between $\left[\sum_{a\in A(s)}P_{s,a,q}x^{\mathcal{A}}_{s,a}\right]_{q\in Succ(s)}$ and $\left[\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)}x^{\mathcal{A}}_{s,a'}\right]_{q\in Succ(s)}$, which is convex in $x^{\mathcal{A}}_{s,a}$ variables. Since the objective function of (4) is a sum of convex functions and the constraints are affine, (4) is a convex optimization problem.

We now show that there exists a stationary policy on $\mathcal{M}$ that achieves the optimal value of (1). By Proposition 1, we have that for all $s \in S_d$, the occupation measures must be bounded. We may apply the constraints $x^{\mathcal{A}}_{s,a} \leq M_s$ for all s in S_d and a in $A(s)$ without changing the optimal value of (4). After this modification, since the objective function is a continuous function of $x^{\mathcal{A}}_{s,a}$ values and the feasible space is compact, there exists a set of occupation measure values, and consequently a stationary policy that achieves the optimal value of (4). ■

Proof of Proposition 4. The condition $P_{s,a,q} > 0$ for all $s \in S_d$, $a \in A(s)$, and $q \in Succ(s)$ implies that $\sum_{a\in A(s)}x^{\mathcal{S}}_{s,a}P_{s,a,q}$ is strictly positive for all $q \in Succ(s)$. Note that for the states $q \notin Succ(s)$, we have $\sum_{a\in A(s)}x^{\mathcal{A}}_{s,a}P(s,a,q) = 0$. We also note that by Assumption 3, the occupation measures are bounded for all $s \in S_d$ under $\pi^{\mathcal{S}}$. Hence, the objective function of (5) is bounded and jointly continuous in $x^{\mathcal{S}}_{s,a}$ and $x^{\mathcal{A}}_{s,a}$.

Since in we showed that there exists a policy that attains the optimal value of Problem 1, we may represent the optimization problem given in (5) as

$$\sup_{x^{\mathcal{S}}}\min_{x^{\mathcal{A}}} f(x^{\mathcal{S}},x^{\mathcal{A}})$$

subject to $x^{\mathcal{S}} \in X^{\mathcal{S}}$ and $x^{\mathcal{A}} \in X^{\mathcal{A}}$. Note that $X^{\mathcal{S}}$ and $X^{\mathcal{A}}$ are compact spaces, since the occupation measures are bounded for all state-action pairs. Given that $X^{\mathcal{A}}$ is a compact space, the function $f'(x^{\mathcal{S}}) = \min_{x^{\mathcal{A}}} f(x^{\mathcal{S}},x^{\mathcal{A}})$ is a continuous function of $x^{\mathcal{S}}$ [38]. The optimal value of $\sup_{x^{\mathcal{S}}} f'(x^{\mathcal{S}})$ is attained. Consequently, there exists a policy $\pi^{\mathcal{S}}$ that achieves the optimal value of (5). ■

REFERENCES

[1] T. E. Carroll and D. Grosu, "A game theoretic investigation of deception in network security," *Security and Communication Networks*, vol. 4, no. 10, pp. 1162–1172, 2011.

[2] M. H. Almeshekah and E. H. Spafford, "Cyber security deception," in *Cyber Deception*. Springer, 2016, pp. 23–50.

[3] W. McEneaney and R. Singh, "Deception in autonomous vehicle decision making in an adversarial environment," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2005, p. 6152.

[4] M. Lloyd, *The Art of Military Deception*. Pen and Sword, 2003.

[5] J. Shim and R. C. Arkin, "A taxonomy of robot deception and its benefits in HRI," in *International Conference on Systems, Man, and Cybernetics*, 2013, pp. 2328–2335.

[6] R. Doody, "Lying and denying," 2018, Preprint available at http://www.mit.edu/%7Erdoody/LyingMisleading.pdf.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[9] V. Pantelic, S. M. Postma, and M. Lawford, "Probabilistic supervisory control of probabilistic discrete event systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 8, pp. 2013–2018, 2009.

[10] M. Lawford and W. Wonham, "Equivalence preserving transformations for timed transition models," *IEEE Transactions on Automatic Control*, vol. 40, no. 7, pp. 1167–1179, 1995.

[11] M. Bakshi and V. M. Prabhakaran, "Plausible deniability over broadcast channels," *IEEE Transactions on Information Theory*, vol. 64, no. 12, pp. 7883–7902, 2018.

[12] A. Boularias, J. Kober, and J. Peters, "Relative entropy inverse reinforcement learning," in *14th International Conference on Artificial Intelligence and Statistics*, 2011, pp. 182–189.

[13] S. Levine and P. Abbeel, "Learning neural network policies with guided policy search under unknown dynamics," in *Advances in Neural Information Processing Systems*, 2014, pp. 1071–1079.

[14] Y. Savas, M. Ornik, M. Cubuktepe, M. O. Karabag, and U. Topcu, "Entropy maximization for Markov decision processes under temporal logic constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 4, pp. 1552–1567, 2019.

[15] C. Keroglou and C. N. Hadjicostis, "Probabilistic system opacity in discrete event systems," *Discrete Event Dynamic Systems*, vol. 28, no. 2, pp. 289–314, 2018.

[16] M. Ornik and U. Topcu, "Deception in optimal control," in *56th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2018, pp. 821–828.

[17] M. O. Karabag, M. Ornik, and U. Topcu, "Least inferable policies for Markov decision processes," in *2019 American Control Conference*. IEEE, 2019, pp. 1224–1231.

[18] J. Fu, S. Han, and U. Topcu, "Optimal control in Markov decision processes via distributed optimization," in *54th Annual Conference on Decision and Control*, 2015, pp. 7462–7469.

[19] M. O. Karabag, M. Ornik, and U. Topcu, "Optimal deceptive and reference policies for supervisory control," in *58th Conference on Decision and Control*, 2019, pp. 1323–1330.

[20] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[21] C. Baier and J.-P. Katoen, *Principles of Model Checking*. MIT Press, 2008.

[22] 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*. Springer, 2007, pp. 50–65.

[23] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999.

[24] O. Kupferman and R. Lampert, "On the construction of fine automata for safety properties," in *International Symposium on Automated Technology for Verification and Analysis*. Springer, 2006, pp. 110–124.

[25] T. Matsui, "NP-hardness of linear multiplicative programming and related problems," *Journal of Global Optimization*, vol. 9, no. 2, pp. 113–119, 1996.

[26] M. R. Garey and D. S. Johnson, *Computers and intractability: A Guide to the Theory of NP-Completeness*, 1979.

[27] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*. Springer, 1972, pp. 85–103.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[29] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.

[30] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

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

[32] MOSEK ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 8.1.*, 2017. [Online]. Available: http://docs.mosek.com/8.1/toolbox/index.html

[33] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[34] S. Alamdari, E. Fata, and S. L. Smith, "Persistent monitoring in discrete environments: Minimizing the maximum weighted latency between observations," *International Journal of Robotics Research*, vol. 33, no. 1, pp. 138–154, 2014.

[35] Google, "Map of San Francisco," https://www.google.com/maps/@37.789463,-122.4068681,16.98z, accessed: Jan. 25, 2019.

[36] M. M. Fard and J. Pineau, "Non-deterministic policies in Markovian decision processes," *Journal of Artificial Intelligence Research*, vol. 40, pp. 1–24, 2011.

[37] K. Conrad, "Probability distributions and maximum entropy," *Entropy*, vol. 6, no. 452, pp. 1–10, 2004.

[38] F. H. Clarke, "Generalized gradients and applications," *Transactions of the American Mathematical Society*, vol. 205, pp. 247–262, 1975.

Mustafa O. Karabag joined the Department of Electrical and Computer Engineering at the University of Texas at Austin as a Ph.D. student in Fall 2017. He received his B.S. degree in Electrical and Electronics Engineering from Bogazici University in 2017. His research focuses on developing theory and algorithms for non-inferable, deceptive planning in adversarial environments.

Melkior Ornik is an assistant professor in the Department of Aerospace Engineering and the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign. He received his Ph.D. degree from the University of Toronto in 2017. His research focuses on developing theory and algorithms for learning and planning of autonomous systems operating in uncertain, complex and changing environments, as well as in scenarios where only limited knowledge of the system is available.

Ufuk Topcu is an associate professor at The University of Texas at Austin. He received his Ph.D. from the University of California, Berkeley in 2008. Ufuk held a postdoctoral research position at California Institute of Technology until 2012 and was a research assistant professor at the University of Pennsylvania until 2015. He is the recipient of the Antonio Ruberti Young Researcher Prize, the NSF CAREER Award and the Air Force Young Investigator Award. His research focuses on the design and verification of autonomous systems through novel connections between formal methods, learning theory and controls.