Optimal Deceptive and Reference Policies for Supervisory Control

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

***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 linear temporal logic formulae, and study the synthesis of optimal deceptive policies for such agents. We also study the synthesis of optimal reference policies that prevents deceptive strategies of the agent and achieves the supervisor's task with high probability. We show that the synthesis of deceptive policies has a convex optimization problem formulation, while the synthesis of reference policies requires solving a nonconvex optimization problem.**

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 supervisors that try to prevent deception besides achieving a task.

Supervisory control [6] refers to high-level regulation of a low-level controller and has applications including to autonomous vehicles [7], multithreaded software [8], and swarm robotics [9]. In a supervisory control setting, a controlled machine receives instructions from the supervisor level as the process evolves and operates autonomously. The setting described in this paper can be considered as a probabilistic discrete event system under supervisory control where the agent represents the controlled machine and the reference policy represents the instructions of the supervisor. In a broad sense, the reference policy is the expected behavior of the agent by the supervisor.

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. The agent's misleading behavior should have plausibility as misleading acts are plausibly deniable [10]. In detail, the supervisor has an expectation on 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 [11]. KL divergence quantifies the extra information needed to encode a posterior distribution using the information of a given prior distribution. 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 linear temporal logic (LTL) 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 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 does 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, which is not tractable in general [12]. In fact, the supervisor's optimization problem remains nonconvex even when the agent uses a predetermined policy. We give a relaxation of the supervisor's problem that can be modeled with a linear program.

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

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

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

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

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

Similar to our approach, [13] 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 [13]. In the context of transition systems, [14], [15] used the metric proposed in this paper, the KL divergence between distribution 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 [14], [15]. We work in a setting with known transition dynamics and provide a convex optimization problem to synthesize the optimal policy while [14], [15] work with unknown dynamics and use sampling-based gradient descent to synthesize the optimal policy. The entropy maximization for MDPs is discussed in [16], which can be considered as a special case of the synthesis of the optimal deceptive policy where the reference policy follows every possible path with equal probability. For the synthesis of optimal deceptive policies, we use a method similar to [16] in that we represent a path as a collection of transitions between the states. We explore the synthesis of optimal reference policies, which, to the best of our knowledge, has not been discussed before.

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 show that the synthesis of optimal reference policies requires solving a nonconvex optimization problem and give a relaxed problem that relies on a linear program for the synthesis of optimal reference policies[1]. We present numerical examples in Section VI and conclude with suggestions for future work in Section VII.

II. Preliminaries

The indicator function $\mathbb{1}_y(x)$ of a variable y is defined as $\mathbb{1}_y(x) = 1$ if $x = y$ and 0 otherwise. A Bernoulli random variable with parameter p is denoted by $Ber(p)$.

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

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

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

A. Markov Decision Processes

A *Markov decision process* (MDP) is a tuple $\mathcal{M} = (S, A, P, s_0, AP, L)$ 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, s_0 is the initial state, AP is a finite set of atomic proposition, and $L : S \to 2^{AP}$ is a labeling function. $A(s)$ 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 $P(s, a, q) > 0$.

A *policy* for $\mathcal{M}$ is a sequence $\pi = \mu_0\mu_1\ldots$ where each $\mu_t : S \times A \to [0, 1]$ is a function such that $\sum_{a\in A(s)} \mu_t(s, a) = 1$ for every $s \in S$. A *stationary policy* is a sequence $\pi = \mu\mu\ldots$ 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 $\pi = \mu\mu\ldots$, 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 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* $\xi = s_0s_1s_2\ldots$ for an MDP $\mathcal{M}$ is an infinite sequence of states under policy $\pi = \mu_0\mu_1\ldots$ such that $\sum_{a\in A(s_t)} P(s_t, a, s_{t+1})\mu_t(s_t, a) > 0$ for all $t \geq 0$. The probability distribution of paths under policy π is denoted by $\Gamma^\pi_{\mathcal{M}}$.

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

B. Linear Temporal Logic and Deterministic Finite Automata

Linear temporal logic (LTL) [17] is a specification language to describe properties of a system. An LTL formula is constructed using a set AP of atomic propositions, Boolean logic operators $\wedge$ (and), $\vee$ (or), $\neg$ (not), and $\implies$ (if), and temporal connectives $\Box$ (always), $\Diamond$ (eventually), $\bigcirc$ (next) and $\mathcal{U}$ (until). For instance, $\Diamond(a \wedge \Diamond b)$ means "eventually reach a and upon reaching a eventually reach b". We refer interested readers to [18] for further details about LTL.

We use a class of LTL called *co-safe* LTL to describe the tasks of the agent and the supervisor. A co-safe formula is satisfied in finite time, i.e., every sequence that satisfies the co-safe LTL formula has a finite good prefix. A co-safe LTL formula is constructed using the same components of LTL semantics but the operator $\neg$ is only applicable to atomic propositions and the connective $\Box$ is not allowed.

Definition 2. A *deterministic finite automaton* (DFA) is a tuple $\mathcal{L} = (Q, \Sigma, \delta, q_0, Acc)$ where Q is a finite set of states, Σ is an alphabet, $\delta : Q \times \Sigma \to Q$ is the transition function, q_0 is the initial state, and $Acc \subseteq Q$ is the accepting states.

[1]The proofs for the results can be found at `https://arxiv.org/abs/1902.00590`.

Any co-safe LTL formula can be translated into a DFA. We denote the DFA representing a co-safe LTL formula ϕ by $\mathcal{L}_\phi$.

Definition 3. For a DFA $\mathcal{L} = (Q, \Sigma, \delta, q_0, Acc)$ and an MDP $\mathcal{M} = (S, A, P, s_0, AP, L)$, the *product MDP* $\mathcal{M}_p$ is a tuple $\mathcal{M}_p = (S_p, A, P_p, s_{0_p}, Q, L_p)$ where $S_p = S \times Q$, $P((s,q),a,(s',q')) = P(s,a,s')$ if $q' = \delta(q, L(s'))$, 0 otherwise, $s_{0_p} = (s_0, q)$ such that $q = \delta(q_0, L(s_0))$, and $L_p((s,q)) = \{q\}$.

Let $\mathcal{M}_p$ be the product MDP of $\mathcal{M}$ and $\mathcal{L}_\phi$. We say that a state (s,q) is *accepting* on $\mathcal{M}_p$ if and only if $q \in Acc$. In detail, a path $\xi = (s_0,q_0),(s_1,q_1)\ldots$ satisfies the co-safe LTL specification ϕ if there exists a k such that $q_k \in Acc$.

On an MDP $\mathcal{M}$, the probability that a specification ϕ is satisfied under a policy π, is denoted by $\Pr^\pi_\mathcal{M}(\phi)$.

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[2]. The supervisor expects the agent to follow $\pi^\mathcal{S}$ on $\mathcal{M}$, thereby performing a task that is specified by the co-safe LTL formula $\phi^\mathcal{S}$. The agent aims to perform another task that is specified by the co-safe LTL formula $\phi^\mathcal{A}$ 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 $\phi^\mathcal{A}$ with high probability.

We use Kullback-Leibler (KL) divergence to measure the deviation from the supervisor's policy. Recall that $\Gamma^\mathcal{S}_\mathcal{M}$ and $\Gamma^\mathcal{A}_\mathcal{M}$ are the distributions of paths under $\pi^\mathcal{S}$ and $\pi^\mathcal{A}$, respectively. We consider $KL(\Gamma^\mathcal{A}_\mathcal{M}||\Gamma^\mathcal{S}_\mathcal{M})$ 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 that the reference policy lacks while encoding the path distributions of the agent. By limiting the deviations from the reference policy, we aim to make the reference policy lack less information while explaining the agent's behavior. Sanov's theorem [11] 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.

[2]It is often considered that the supervisor provides a nondeterministic policy to the agent by disabling some actions, instead of providing an explicit policy with no nondeterminism. More discussion on deception under nondeterministic reference policies can be found at `https://arxiv.org/abs/1902.00590`.

Formally, let π^* be a solution to $\min_{\pi\in\Pi(\mathcal{M})} \; KL\left(\Gamma^\pi_\mathcal{M}||\Gamma^{\pi^\mathcal{S}}_\mathcal{M}\right)$ subject to $\Pr^\pi_\mathcal{M}(\phi) \geq \nu^\mathcal{A}$. Assume that we simulate n paths under the supervisor's policy. The probability that the observed paths satisfy $\phi^\mathcal{A}$ 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 $\phi^\mathcal{A}$ with a probability higher than $\nu^\mathcal{A}$, the most likely empirical distribution is $\Gamma^{\pi^*}_\mathcal{M}$ [11].

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 co-safe LTL specification $\phi^\mathcal{A}$, a probability threshold $\nu^\mathcal{A}$, and a reference policy $\pi^\mathcal{S}$, solve

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

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

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 $\Pr^{\pi^\mathcal{S}}_\mathcal{M}(\phi^\mathcal{S}) \geq \nu^\mathcal{S}$. We assume that the supervisor has knowledge on the agent's task and formulate the following problem for the synthesis of reference policies for the supervisor.

Problem 2 (Synthesis of Optimal Reference Policies)**.** Given an MDP $\mathcal{M}$, co-safe LTL specifications $\phi^\mathcal{S}$ and $\phi^\mathcal{A}$, probability thresholds $\nu^\mathcal{A}$ and $\nu^\mathcal{S}$, solve

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

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

$$\Pr^{\pi^\mathcal{S}}_\mathcal{M}(\phi^\mathcal{S}) \geq \nu^\mathcal{S}. \tag{3c}$$

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

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

We first consider the synthesis of optimal deceptive policies where the reference policy satisfies $\pi^\mathcal{S}_{s_0,\beta} = 1$. Consider $\phi^\mathcal{A} = \Diamond s_3$ 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,

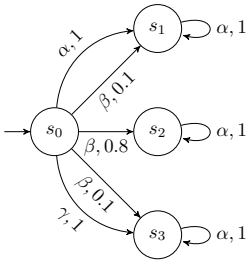


Fig. 1: An MDP with 4 states. A label a, p of a transition refers to the transition that happens with probability p when action a is taken.

the 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 $\phi^{\mathcal{S}} = \Diamond(s_1 \vee s_2)$ and $\nu^{\mathcal{S}} = 0.9$. Consider $\phi^{\mathcal{A}} = \Diamond s_3$ and $\nu^{\mathcal{A}} = 0.1$. Assume that the reference policy has $\pi^{\mathcal{S}}_{s_0,\beta} = 1$. In this case, the agent can directly follow the supervisor's 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 modify the MDP to simplify the problem. Then, we show the existence of an optimal deceptive policy and give an optimization problem to synthesize one.

Recall that $\mathcal{L}_{\phi^{\mathcal{S}}} = (Q_{\mathcal{S}}, \Sigma, \delta_{\mathcal{S}}, q_{0_{\mathcal{S}}}, Acc_{\mathcal{S}})$ and $\mathcal{L}_{\phi^{\mathcal{A}}} = (Q_{\mathcal{A}}, \Sigma, \delta_{\mathcal{A}}, q_{0_{\mathcal{A}}}, Acc_{\mathcal{A}})$ are the equivalent DFAs for the supervisor's specification $\phi^{\mathcal{S}}$ and the agent's specification $\phi^{\mathcal{A}}$. We create a product DFA $\mathcal{L}_p = (Q, \Sigma, \delta, q_0, Acc)$ of $\mathcal{L}_{\phi^{\mathcal{S}}}$ and $\mathcal{L}_{\phi^{\mathcal{A}}}$ to represent the agent's and supervisor's specifications together, where $Q = Q_{\mathcal{A}} \times Q_{\mathcal{S}}$, $\delta((q_1, q_2), a) = (\delta_{\mathcal{S}}(q_1, a), \delta_{\mathcal{A}}(q_2, a))$, $q_0 = (q_{0_{\mathcal{S}}}, q_{0_{\mathcal{A}}})$, and $Acc = Acc_{\mathcal{S}} \times Acc_{\mathcal{A}}$.

We create a product MDP $\mathcal{M}_p = (S_p, A, P, s_{0_p}, L_p)$ of DFA $\mathcal{L}_p$ and MDP $\mathcal{M}$. Let $C_{\mathcal{S}}$ and $C_{\mathcal{A}}$ be the sets of accepting states in $\mathcal{M}_p$ for $\phi^{\mathcal{S}}$ and $\phi^{\mathcal{A}}$, respectively. A state $s_p = (s, q^{\mathcal{S}}, q^{\mathcal{A}})$ belongs to $C_{\mathcal{S}}$, if $q^{\mathcal{S}} \in Acc_{\mathcal{S}}$, and to $C_{\mathcal{A}}$, if $q^{\mathcal{A}} \in Acc_{\mathcal{A}}$. A path $\xi = (s_0, q^{\mathcal{S}}_0, q^{\mathcal{A}}_0), (s_1, q^{\mathcal{S}}_1, q^{\mathcal{A}}_1), \dots$ satisfies $\phi^{\mathcal{S}}$ if there exists an integer k such that $q^{\mathcal{S}}_k \in Acc_{\mathcal{S}}$. Similarly, ξ satisfies $\phi^{\mathcal{A}}$ if there exists an integer k such that $q^{\mathcal{A}}_k \in Acc_{\mathcal{A}}$. On $\mathcal{M}_p$, the agent's specification is $\Diamond C_{\mathcal{A}}$ and the supervisor's specification is $\Diamond C_{\mathcal{S}}$. The reference policy $\pi^{\mathcal{S}}$ induces a policy on $\mathcal{M}_p$. With some abuse of notation, we denote the induced policy also by $\pi^{\mathcal{S}}$. Similarly, $\pi^{\mathcal{A}}$ denotes the induced policy of the agent.

We note that there is a one-to-one correspondence between the path distributions of $\mathcal{M}$ and $\mathcal{M}_p$ [18]. Consequently, the optimization problem given in (2) is equivalent to

$$\inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M}_p)} \quad KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}\right) \tag{4a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}_p}(\Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}. \tag{4b}$$

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 1. *The policy that is induced by the reference policy is stationary for the product MDP $\mathcal{M}_p$.*

In many applications the supervisor aims to achieve the specification with the maximum possible probability. Stationary policies on the product MDP suffice to maximize the probability to satisfy an LTL formula [18].

We assume that the optimal value of Problem 1 is finite. If the KL divergence between the path distributions is finite, the agent's policy cannot differ from the reference policy for some states. The reference policy $\pi^{\mathcal{S}}$ induces a Markov chain $\mathcal{M}^{\mathcal{S}}_p$. A state is recurrent in $\mathcal{M}^{\mathcal{S}}_p$ if it belongs to some closed communicating class. Let C_{cl} be the set of states that belong to a closed communicating class of $\mathcal{M}^{\mathcal{S}}_p$. Assume that under the agent's policy $\pi^{\mathcal{A}}$, there exists a path that visits a state in C_{cl} and leaves C_{cl} 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} must also be closed under $\pi^{\mathcal{A}}$. Furthermore, since the probability of satisfying $\phi^{\mathcal{A}}$ is zero upon entering C_{cl}, the agent should choose the same policy as the supervisor to minimize the KL divergence between the distributions of paths. If a state s is transient in $\mathcal{M}^{\mathcal{S}}_p$, the agent's policy must eventually stop visiting s, since otherwise we have infinite divergence. Furthermore, we have the following property.

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

Also, we remark that the agent's policy should not be different from the supervisor's policy on the states that belong to $C_{\mathcal{A}}$, since the specification of the agent is already satisfied. We denote the set of states for which the agent's policy can differ from the supervisor's policy by $S_d = S_p \setminus (C_{cl} \cup C_{\mathcal{A}})$.

Since the occupation times are bounded for the states that the agent's policy may differ from the supervisor's policy, it is possible to show the sufficiency of stationary policies for the synthesis of optimal deceptive policies [19].

Proposition 2. For any policy $\pi^{\mathcal{A}} \in \Pi(\mathcal{M}_p)$ that satisfies $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}_p}(\Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}$, there exists a stationary policy $\pi^{\mathcal{A},St} \in \Pi(\mathcal{M}_p)$ that satisfies $\Pr^{\pi^{\mathcal{A},St}}_{\mathcal{M}_p}(\Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}$ and $KL\left(\Gamma^{\pi^{\mathcal{A},St}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}\right) \leq KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}\right)$.

We solve the following optimization problem to compute the occupation times of an optimal deceptive policy:

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

$$\text{subject to} \quad x^{\mathcal{A}}_{s,a} \geq 0, \qquad \forall s \in S_d,\ \forall a \in A(s), \tag{5b}$$

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

$$\sum_{q \in C_{\mathcal{A}}} \sum_{s \in S_d} \sum_{a \in A(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} + \mathbb{1}_{s_0}(q) \geq \nu^{\mathcal{A}}, \tag{5d}$$

where $\pi^{\mathcal{S}}_{s,q}$ is the transition probability between 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 (5a) 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. The constraint (5c) encodes the feasible policies and the constraint (5d) represents the task constraint.

Proposition 3. The optimization problem given in (5) is a convex optimization problem that shares the same optimal value with (4). Furthermore, there exists a policy $\pi \in \Pi^{St}(\mathcal{M})$ that attains the optimal value of (5).

The optimization problem given in (5) gives the optimal state-action occupation times 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.

Remark 1. *We assume that the optimal value of Problem 1 is finite. This assumption does not hurt generality of the method. 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 $\mathcal{M}_p$ by removing the actions that assign a positive transition probability to such state-state pairs. If there exists a policy that satisfies the constraint* (4b) *then the value is finite.*

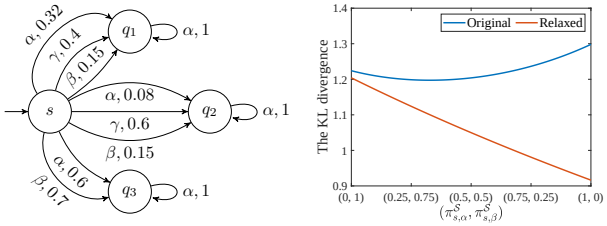
V. Synthesis of Optimal Reference Policies

The optimal reference policy can be computed via dualization. In detail, the optimization problem given in (5) satisfies Slater's conditions and hence the dual problem of (5) shares the same optimal value with (5). One can compute the optimal reference policy by solving the dual problem of (5) with the parameters of the reference policy as the decision variables.

While it is possible to compute the optimal reference policy, in this section we show that to find the optimal solution of Problem 2 one needs to solve a nonconvex optimization problem. We provide a relaxation of the problem which can be modeled with a linear program.

For Problem 2, we observe that it is possible that there are multiple locally optimal reference policies. For example, consider the 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. 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 parametrization to reformulate the optimization problem given in Problem 2. Consider a continuous and bijective transformation from the policy parameters to the new parameters, that makes new parameters to span all stationary policies. After this transformation, an optimal solution to for the space of policy parameters yields an optimal solution in the new parameter space. If the optimization problem formulated with the supervisor's policy parameters has multiple local optima, then any reformulation spanning all stationary policies for the supervisor has multiple optima. Therefore, it is not possible to obtain a convex reformulation.



(a) (b)

Fig. 2: (a) An MDP with 4 states. A label a, p of a transition refers to the transition that happens with probability p when action a is taken. (b) The KL divergence between the path distributions of the agent and the supervisor.

Since it is not possible to obtain a convex reformulation of the optimization problem 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 $\phi^{\mathcal{A}}$ and the paths that do not satisfy $\phi^{\mathcal{A}}$. After this transformation, the probability assigned to the first subset is $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}})$ under policy $\pi^{\mathcal{S}}$ and $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(\phi^{\mathcal{A}})$ under policy $\pi^{\mathcal{A}}$. By the data processing inequality given in (1), this transformation yields a lower bound on the KL divergence between the path distributions: $KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}\right)$ is greater than or equal to

$$KL\left(Ber\left(\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}\left(\phi^{\mathcal{A}}\right)\right) || Ber\left(\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}\left(\phi^{\mathcal{A}}\right)\right)\right). \quad (6)$$

We use this lower bound to construct the relaxed problem

$$\sup_{\pi^{\mathcal{S}} \in \Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M})} \quad (6) \quad (7a)$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(\phi^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \quad (7b)$$

$$\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}. \quad (7c)$$

If $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}}) \geq \nu_{\mathcal{A}}$, the agent may directly use the reference policy. Without loss of generality, assuming that $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}}) < \nu_{\mathcal{A}}$, (6) is decreasing in $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}})$ and increasing in $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(\phi^{\mathcal{A}})$. The problem

$$\sup_{\pi^{\mathcal{S}} \in \Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M})} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(\phi^{\mathcal{A}}) + \Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}}) \quad (8a)$$

$$\text{subject to (7b) and (7c).} \quad (8b)$$

shares the same optimal policies with the problem given in (7). We note that the optimization problem given in (8) 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 $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}})$ subject to $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{S}}) \geq \nu_{\mathcal{S}}$.

The lower bound given in (6) provides a sufficient condition on the optimality of a reference policy for Problem 2. A policy $\pi^{\mathcal{S}}$ satisfying $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{A}}) = 0$ and $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}$ is an optimal reference policy since the optimization problem given in (7) has the optimal value of ∞. However, in general the gap due to the relaxation may get arbitrarily large, and the reference policy synthesized via (7) is not necessarily optimal for Problem 2. For example, consider the MDP given in Figure 2a where the agent's policy again has $\pi^{\mathcal{A}}_{s,\gamma} = 1$. For simplicity, there is no specification for the supervisor, i.e., $\nu^{\mathcal{S}}$ is 0. The policy $\pi^{\mathcal{S}}$ that minimizes $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(\Diamond q_1 \vee \Diamond q_2)$ 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 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.

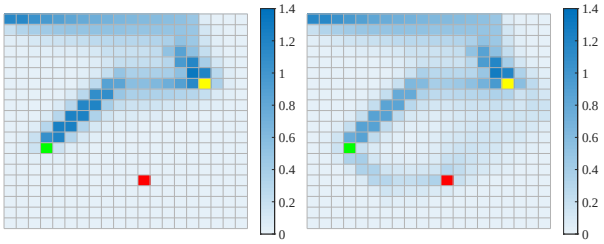
VI. Numerical Examples

In this section, we give numerical examples on the synthesis of optimal deceptive policies. In Section VI-A, we explain some characteristics of the optimal deceptive policies through different scenarios. In Section VI-B, we compare the proposed metric, the KL divergence between the distributions of paths, to some other metrics. We solved the optimization problems with CVX [20] toolbox using MOSEK [21].

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 yellow, green, and red states have labels y, 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. At the green state there is an extra action that allows self transition with probability 1. The initial state is the top-left state.

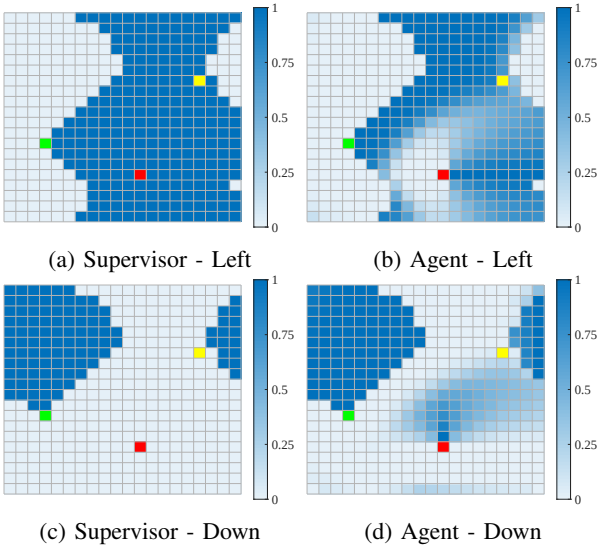
The specification of the supervisor is to first reach the yellow state then reach the green state. The specification is encoded with the co-safe LTL formula $\phi^{\mathcal{S}} = \Diamond(y \wedge \Diamond g)$. The reference policy $\pi^{\mathcal{S}}$ is constructed so that it satisfies $\phi^{\mathcal{S}}$ with probability 1 in minimum expected time. The specification of the agent is to reach the red state. The specification is encoded with the co-safe LTL formula $\phi^{\mathcal{A}} = \Diamond r$. The probability threshold $\nu^{\mathcal{A}}$ for the agent's specification is 0.3.



(a) Reference (b) Deceptive

Fig. 3: Heat maps of the occupation times. The value of a state is the expected number of visits to the state. The deceptive policy follows the same policy until reaching the yellow state. Upon reaching the yellow state, the agent move towards the red state to achieve its objective.

We synthesize the policy of the agent according to Problem 1, which leads to the KL divergence value of 2.975. While the reference policy satisfies $\phi^{\mathcal{A}}$ with probability 3×10^{-5}, the agent's policy satisfies $\phi^{\mathcal{A}}$ with probability 0.3. Until reaching the yellow state, the deceptive policy follows the reference policy since any deviation from the reference policy incurs high divergence. As we see in Figure 4b, upon reaching the yellow state, the reference policy takes action left and the agent's policy takes action down to move to toward the red state. The misleading occurs during this period: while the agent goes down on purpose, he may hold the stochasticity of the environment accountable for this behavior.



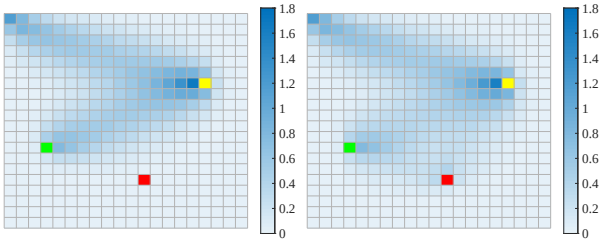
(a) Supervisor - Left (b) Agent - Left

(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 also observe a significant detail in the agent's policy. At the yellow state the reference policy takes action left, on the other hand the agent's policy tries to go right. Note that in the top-right region the reference policy takes action down. The agent wants to drive himself to this region so that he can 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.

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 $\phi^{\mathcal{A}}$ with probability 7×10^{-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 0.899.



(a) Supervisor (b) Agent

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

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 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 one of the transient states.

We randomly generate three different reference policies for the randomly generated MDP. The reference policies satisfy the agent's specification $\phi^{\mathcal{A}}$ with probabilities 0.30, 0.14, and 0.13. For each 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 state-action occupation times for the agent's policy and the reference policies, and by minimizing the L_2-norm between the state-action occupation times 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 policies. 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 [22] 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.

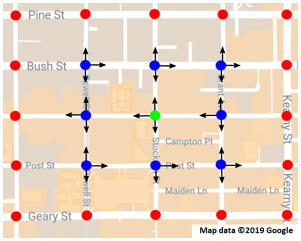
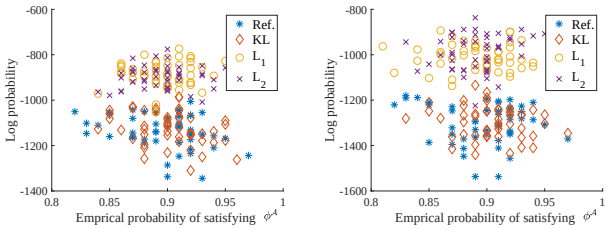


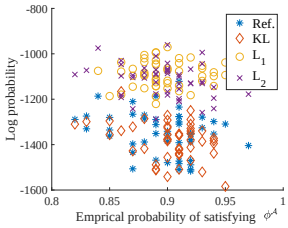
Fig. 6: The map of a region from north east of San Francisco. The green dot indicates the intersection at which the highest number of crimes happened. The data is from [23]. The dots on the map represent the states of the MDP and the arrows represent the available actions. The initial state is chosen uniformly randomly among the blue states and the red states are absorbing. The agent aims to patrol the green state.

For each simulation, we plot the log-probability under the reference policy in Figure 7. From Figure 7, 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

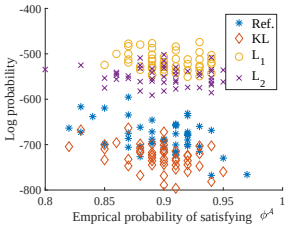


(a) Random MDP – 1

(b) Random MDP – 2



(c) Random MDP – 3



(d) MDP for San Francisco

Fig. 7: The log-likelihoods under the reference policy and the empirical probability of satisfying $\phi^{\mathcal{A}}$. 'Ref.' refers to the rare events of the reference policy that satisfy $\phi^{\mathcal{A}}$ with high probability. 'KL', 'L_1', and 'L_2' refer to the candidate deceptive policies.

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 was explained in Section III.

VII. Conclusion

We considered the problem of deception under a supervisor that provides a reference policy. We modeled the problem using MDPs and co-safe LTL 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 requires solving a nonconvex optimization problem, and provided a linear programming relaxation.

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.

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] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete event processes," *SIAM Journal on Control and Optimization*, vol. 25, no. 1, pp. 206–230, 1987.

[7] M. L. Cummings and S. Guerlain, "Developing operator capacity estimates for supervisory control of autonomous vehicles," *Human Factors*, vol. 49, no. 1, pp. 1–15, 2007.

[8] H. Liao, Y. Wang, J. Stanley, S. Lafortune, S. A. Reveliotis, T. Kelly, and S. A. Mahlke, "Eliminating concurrency bugs in multithreaded software: A new approach based on discrete-event control." *IEEE Transactions on Control Systems and Technology*, vol. 21, no. 6, pp. 2067–2082, 2013.

[9] Y. K. Lopes, S. M. Trenkwalder, A. B. Leal, T. J. Dodd, and R. Groß, "Probabilistic supervisory control theory (pSCT) applied to swarm robotics," in *16th Conference on Autonomous Agents and MultiAgent Systems*, 2017, pp. 1395–1403.

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

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

[12] S. A. Vavasis, "Complexity issues in global optimization: a survey," in *Handbook of Global Optimization*. Springer, 1995, pp. 27–41.

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

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

[15] 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.

[16] Y. Savas, M. Ornik, M. Cubuktepe, and U. Topcu, "Entropy maximization for constrained Markov decision processes," in *56th Annual Allerton Conference on Communication, Control, and Computing*, 2018.

[17] A. Pnueli, "The temporal logic of programs," in *18th Annual Symposium on Foundations of Computer Science*, 1977, pp. 46–57.

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

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

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

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

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

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