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 such 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. 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 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

[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.

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 propose alternating direction method of multipliers (ADMM) [13] to locally solve the supervisor's optimization problem. We also give a relaxation of the problem that can be modeled with a linear program.

Similar to our approach, [14] 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 [14]. In the context of transition systems, [15], [16] 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 [15], [16]. We work in a setting with known transition dynamics and provide a convex optimization problem to synthesize of the optimal policy while [15], [16] work with unknown dynamics and use sampling-based gradient descent to synthesize the optimal policy. The entropy maximization for MDPs is discussed in [17], 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 [17] 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. We propose to use ADMM to synthesize the optimal reference policies. Similarly, [18] also used ADMM for the synthesis 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.

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 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 give the proofs for the results in the Appendix.

II. Preliminaries

The set $\{x = (x_1, \dots, x_n) | x_i \geq 0\}$ is denoted by $\mathbb{R}^n_+$. 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. 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 $\mathcal{K}_{\exp} = \{(x_1, x_2, x_3) | x_2 \exp(x_1/x_2) \leq x_3,\ x_2 > 0\} \cup \{(x_1, 0, x_3) | x_1 \leq 0,\ x_3 \geq 0\}$ and it can be shown that $\mathcal{K}^*_{\exp} = \{(x_1, x_2, x_3) | -x_1 \exp(x_2/x_1 - 1) \leq x_3,\ x_1 < 0\} \cup \{(0, x_2, x_3) | x_2 \geq 0,\ x_3 \geq 0\}$.

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 defined with logarithm to base 2 in information theory. However, we use natural logarithm for the clarity of representation in the optimization problems. We remark that the base change does not change the results.*

A. Markov Decision Processes

A *Markov decision process* (MDP) is a tuple $\mathcal{M} = (S, A, P, s_0, AP, L)$ 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 $S \to 2^{AP}$ 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 $\pi = \mu_0 \mu_1 \dots$ 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 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 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 $\sum_{a \in A(s_t)} P(s_t, a, s_{t+1})\mu_t(s_t, a) > 0$ 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 *expected state-action residence time* at state s and action a is defined by

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

If π is stationary, the expected state-action residence times satisfy $x^\pi_{s,a} = \pi_{s,a} \sum_{a' \in A(s)} x^\pi_{s,a'}$ for all s with finite expected residence times. The expected state-action residence time 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 expected state-action residence times at state s under policy π and x^π for the vector of all expected state-action residence times.

B. Linear Temporal Logic and Deterministic Finite Automata

Linear temporal logic (LTL) [19] 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 $\wedge$, $\vee$, $\neg$, and $\implies$, 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 [20] 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 connective $\Box$ and the operator $\neg$ are only applicable to atomic propositions.

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.

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

- $S_p = S \times Q$,
- $P((s, q), a, (s', q')) = \begin{cases} P(s, a, s') & \text{if } q' = \delta(q, L(s')) \\ 0 & \text{otherwise,} \end{cases}$
- $s_{0_p} = (s_0, q)$ such that $q = \delta(q_0, L(s_0))$,
- $L_p((s, q)) = \{q\}$.

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 $\xi = (s_0, q_0), (s_1, q_1) \dots$ satisfies the co-safe LTL specification 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 $\Pr^\pi_\mathcal{M}(s_0 \models \phi)$.

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. 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 $KL(\Gamma^\mathcal{S}_\mathcal{M} || \Gamma^\mathcal{A}_\mathcal{M})$ 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 agent's policy lack less information. 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. Formally, let π^* be a solution to

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

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 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*

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

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

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

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 $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}$. We assume that the supervisor has knowledge on 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 $\mathcal{M}$, co-safe LTL specifications $\phi^{\mathcal{S}}$ and $\phi^{\mathcal{A}}$, probability thresholds $\nu^{\mathcal{A}}$ and $\nu^{\mathcal{S}}$, solve*

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

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{4b}$$

$$\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}. \tag{4c}$$

If the supremum is attainable, find a policy $\pi^{\mathcal{S}}$ that is a solution to (4)*.*

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.

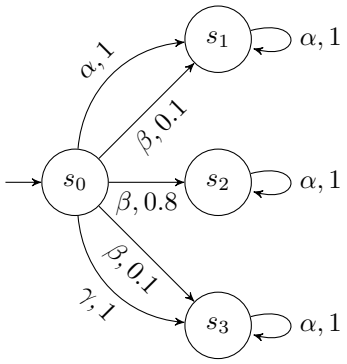


Figure 1: An MDP with 4 states. A label a, p of a transition refers to the transition that happens with probability p when action a is taken.

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, 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 the product MDP $\mathcal{M}_p$, 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, we will use $\pi^{\mathcal{A}}$ for the induced policy of the agent.

We note that there is a one-to-one correspondence between the path distributions of $\mathcal{M}$ and $\mathcal{M}_p$ [20]. Consequently, the optimization problem given in (3) is equivalent to

$$\inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M}_p)} \quad KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}\right) \tag{5a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}_p}(s_{0_p} \models \Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}. \tag{5b}$$

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 [20].

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 $\mathcal{M}_p$ by removing the actions that assign a positive value to such state-state pairs. If there exists a policy that satisfies the constraint (5b) then the value 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 expected state-action residence 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.

Since we know that the expected residence 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.

Proposition 2. *For any policy $\pi^{\mathcal{A}} \in \Pi(\mathcal{M}_p)$ that satisfies $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}_p}(s_{0_p} \models \Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}$, there exists a stationary policy $\pi^{\mathcal{A},St} \in \Pi(\mathcal{M}_p)$ that satisfies $\Pr^{\pi^{\mathcal{A},St}}_{\mathcal{M}_p}(s_{0_p} \models \Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}$ 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 denote the set of states for which the agent's policy can differ from the supervisor's policy by $S_d = S_p \setminus (C_{\mathcal{S}} \cup C_{\mathcal{A}})$. We solve the following optimization problem to compute the expected residence time parameters of the optimal deceptive policy.

$$\inf \quad \sum_{s\in S_d}\sum_{a\in A(s)}\sum_{q\in Succ(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} \log\left(\frac{\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'} P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'}}\right) \tag{6a}$$

$$\text{subject to} \quad x^{\mathcal{A}}_{s,a} \geq 0, \qquad \forall s\in S_d,\ \forall a\in A(s) \tag{6b}$$

$$\sum_{a\in A(s)} x^{\mathcal{A}}_{s,a} - \sum_{q\in S_d}\sum_{a\in A(q)} x^{\mathcal{A}}_{q,a} P_{q,a,s} = \mathbb{1}_{s_0}(s), \qquad \forall s\in S_d \tag{6c}$$

$$\sum_{q\in C_{\mathcal{A}}}\sum_{s\in S_d}\sum_{a\in A(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} + \mathbb{1}_{s_0}(q) \geq \nu^{\mathcal{A}}, \tag{6d}$$

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 (6a) 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). The constraint (6c) encodes the feasible policies and the constraint (6d) represents the task constraint.

Proposition 3. *The optimization problem given in* (6) *is a convex optimization problem that shares the same optimal value with* (5)*. Furthermore, there exists a policy $\pi\in\Pi^{St}(\mathcal{M})$ that attains the optimal value of* (6)*.*

The optimization problem given in (6) gives the optimal expected state-action residence 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.

V. Synthesis of Optimal Reference Policies

In this section, we give an optimization problem to synthesize optimal reference policies. With the modification step described in Section IV, Problem 2 is equivalent to

$$\sup_{\pi^{\mathcal{S}}\in\Pi(\mathcal{M}_p)}\inf_{\pi^{\mathcal{A}}\in\Pi(\mathcal{M}_p)} \quad KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}_p} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}\right) \tag{7a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}_p}(s_{0_p} \models \Diamond C_{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{7b}$$

$$\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}_p}(s_{0_p} \models \Diamond C_{\mathcal{S}}) \geq \nu^{\mathcal{S}}. \tag{7c}$$

The optimization problem given in (6) has the supervisor's policy parameters as constants. We want to solve the optimization problem given in (6) to formulate the synthesis of optimal reference policies by adding the supervisor's policy parameters as additional decision variables. Remember that the set C_{cl} is the set of states that belong to a closed communicating class of $\mathcal{M}^{\mathcal{S}}_p$. In the optimization problem given in (6), 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 2. *The set C_{cl} is the same for all reference policies considered in Problem 2.*

Remark 2. *In the absence of Assumption 2, one needs to compute the optimal reference policy for different values of C_{cl}. Thus, Assumption 2 is made just for the clarity of representation.*

Under Assumptions 1 and 2, the optimal value of Problem 2 is equal to the optimal value of the following optimization problem:

$$\sup_{x^{\mathcal{S}}_{s,a}}\inf_{x^{\mathcal{A}}_{s,a}} \quad \sum_{s\in S_d}\sum_{a\in A(s)}\sum_{q\in Succ(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} \log\left(\frac{\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'} P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'}}\right) \tag{8a}$$

$$\text{subject to} \quad \text{(6b)} - \text{(6d)}$$

$$\pi^{\mathcal{S}}_{s,q} = \sum_{a\in A(s)} P_{s,a,q}\frac{x^{\mathcal{S}}_{s,a}}{\sum_{a'\in A(s)} x^{\mathcal{S}}_{s,a'}}, \qquad \forall s\in S_d,\ \forall q\in S \tag{8b}$$

$$x^{\mathcal{S}}_{s,a} \geq 0, \qquad \forall s\in S_d,\ \forall a\in A(s) \tag{8c}$$

$$\sum_{a\in A(s)} x^{\mathcal{S}}_{s,a} - \sum_{q\in S_d}\sum_{a\in A(q)} x^{\mathcal{S}}_{q,a} P_{q,a,s} = \mathbb{1}_{s_0}(s), \qquad \forall s\in S_d \tag{8d}$$

$$\sum_{q\in C_{\mathcal{S}}}\sum_{s\in S_d\setminus C_{\mathcal{S}}}\sum_{a\in A(s)} x^{\mathcal{S}}_{s,a} P_{s,a,q} + \mathbb{1}_{s_0}(q) \geq \nu^{\mathcal{S}}, \tag{8e}$$

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 3. *The optimization problem given in* (8) *has undefined points due to the denominators in* (8a) *and* (8b)*, 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,* $x^{\mathcal{A}}_{s,q} = 0$ *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 (8) 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* (8)*.*

In Section V-A, we describe dualization-based procedure to solve the optimization problem given in (8). As an alternative to solving the dual problem, we give an algorithm based on alternating direction method of multipliers (ADMM) in Section V-B. Since both approaches, the dualization-based and the ADMM-based, require solving nonconvex optimization problems, we present a relaxation of the problem in Section V-C that relies on solving a linear program. Finally, we investigate the case when the agent uses a fixed policy in Section V-D. In this case we show that one can find a locally optimal reference policy using a convex-concave procedure.

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

Observing that Slater's condition [21] is satisfied and the strong duality holds for the optimization problem given in (6), to find the optimal value of (8) one may consider solving the dual of (6) with $x^{\mathcal{S}}_{s,a}$ as additional decision variables and (8b)-(8e) as additional constraints. In this section, we show that such an approach yields to a nonconvex optimization problem.

The optimization problem given in (6) has the following conic optimization representation:

$$\min_{y} \; c^T y \tag{9a}$$

$$\text{subject to} \;\; [G| - I]y = h, \tag{9b}$$

$$y \in \mathcal{K}. \tag{9c}$$

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

$$r_{(s,q)} \geq \sum_{a\in A(s)} x^{\mathcal{A}}_{s,a} P_{s,a,q} \log\left(\frac{\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'} P_{s,a',q}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'}}\right) \tag{10}$$

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 (6c) 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 (6b) is encoded by $G_+ y \geq 0$ where $G_+ := [I_{N\times N}|0_{N\times M}]$. The additional constraint given in (10) 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 (6d) 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 $\mathbb{1}_{S_d\setminus C_{\mathcal{A}}}(s)\sum_{q\in C_{\mathcal{A}}} P_{s,a,q}$. Finally, $\mathcal{K} = \mathbb{R}^{N+M} \times \{0\}^{|S_d|} \times \mathbb{R}^N_+ \times \mathcal{K}_{\exp} \times \ldots \times \mathcal{K}_{\exp} \times \mathbb{R}_+$,

$$G = \begin{bmatrix} G_{eq} \\ G_+ \\ G_{(1,1)} \\ \vdots \\ G_{(|S_d|,|S|)} \\ G_{\mathcal{A}} \end{bmatrix}, h = \begin{bmatrix} h_{eq} \\ 0 \\ \vdots \\ 0 \\ \nu^{\mathcal{A}} \end{bmatrix}, c = \begin{bmatrix} 0_{N\times 1} \\ 1_{M\times 1} \end{bmatrix}.$$

The dual of the optimization problem in (9) is

$$\max_{u,w} \; h^T u \tag{11a}$$

$$\text{subject to} \; \begin{bmatrix} G^T \\ -I^T \end{bmatrix} u + w = c, \tag{11b}$$

$$w \in \mathcal{K}^*, \tag{11c}$$

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 (11) and the constraints in (8b)-(8e), and adding $x_{s,a}^{\mathcal{S}}$ as decision variables, we get an optimization problem that shares the same optimal value with (8). However, we remark that this problem is nonconvex because of the constraint (8b) and the bilinear constraints that are due to $\pi_{s,q}^{\mathcal{S}}$ parameter introduced in the construction of $G_{(s,q)}$. In general, non-convex optimization problems are intractable [12].

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

The alternating direction method of multipliers (ADMM) [13] 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 (8). The objective function of (8) 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 $z_s^{\mathcal{A}}$ and $z_s^{\mathcal{S}}$ such that $z_s^{\mathcal{A}} = x_s^{\mathcal{A}}$ and $z_s^{\mathcal{S}} = x_s^{\mathcal{S}}$. With these extra variables, the augmented Lagrangian of (8) is

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

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 expected residence time variables of the agent that satisfy (6c) and (6d), $X^{\mathcal{S}}$ is the set of expected residence time variables of the supervisor that satisfy (8d) and (8e), and

$$
\pi_{s,q}^{\mathcal{S}} = \sum_{a \in A(s)} P_{s,a,q} \frac{x_{s,a}^{\mathcal{A}}}{\sum_{a' \in A(s)} x_{s,a'}^{\mathcal{A}}}
$$

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.

Algorithm 1 The ADMM for the synthesis of reference policies

1: **Input:** An MDP $\mathcal{M}$, specifications $\phi^{\mathcal{S}}$ and $\phi^{\mathcal{A}}$, probability thresholds $\nu^{\mathcal{S}}$ and $\nu^{\mathcal{A}}$
2: **Output:** A reference policy $\pi^{\mathcal{S}}$.
3: Set $x^{\mathcal{S},0}$ and $z^{\mathcal{S},0}$ arbitrarily from $X^{\mathcal{S}}$.
4: Set $x^{\mathcal{A},0}$ and $z^{\mathcal{A},0}$ arbitrarily from $X^{\mathcal{A}}$.
5: Set $\lambda^{\mathcal{S},0}$ and $\lambda^{\mathcal{A},0}$ to 0.
6: $k = 0$.
7: **while** stopping criteria are not satisfied **do**
8: Set $x^{\mathcal{S},k+1}$ and $x^{\mathcal{A},k+1}$ as the solution of $\max_{x^{\mathcal{S}}} \min_{x^{\mathcal{A}}} L(x^{\mathcal{S}}, x^{\mathcal{A}}, z^{\mathcal{S},k}, z^{\mathcal{A},k}, \lambda^{\mathcal{S},k}, \lambda^{\mathcal{S},k})$.
9: $z^{\mathcal{S},k+1} := Proj_{X^{\mathcal{S}}}(x^{\mathcal{S},k+1} + \lambda^{\mathcal{S},k})$.
10: $z^{\mathcal{A},k+1} := Proj_{X^{\mathcal{A}}}(x^{\mathcal{A},k+1} + \lambda^{\mathcal{A},k})$.
11: $\lambda^{\mathcal{S},k+1} := \lambda^{\mathcal{S},k} + x^{\mathcal{S},k+1} - z^{\mathcal{S},k+1}$.
12: $\lambda^{\mathcal{A},k+1} := \lambda^{\mathcal{A},k} + x^{\mathcal{A},k+1} - z^{\mathcal{A},k+1}$.
13: $k := k + 1$.
14: **end while**
15: Compute $\pi^{\mathcal{S}}$ using $z^{\mathcal{S},k}$ as the expected residence times.

We remark that Algorithm 1 still requires solving a maximin optimization problem (see line 8). However, the maximin 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 $\mathcal{O}(|S_p||A|)$, it is $\mathcal{O}(|A|)$ for the local problems in the ADMM algorithm.

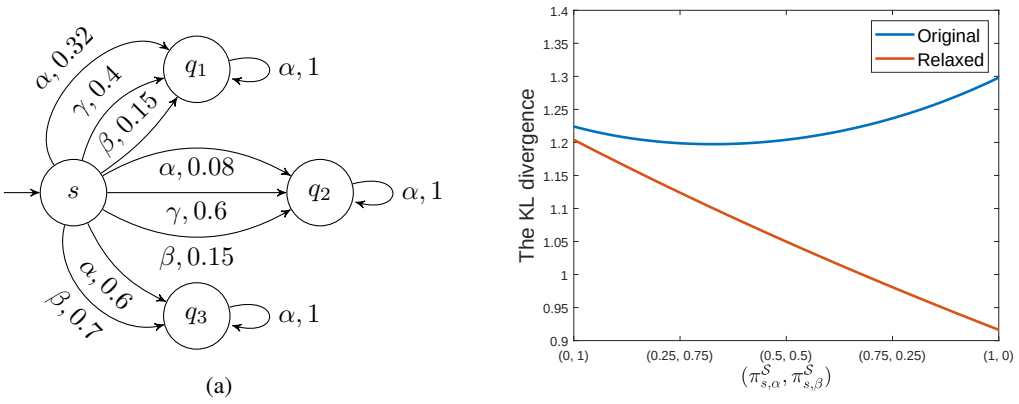
Since the strong duality holds, one can use a dualization-based approach as shown in Section V-A to solve the local maximin problems. We remark that after dualization, the resulting optimization problems are nonconvex similar to the optimization problem obtained via dualization-based approach.

Remark 4. *Convergence of ADMM for particular nonconvex optimization problems has been studied [22], [23]. To the best of our knowledge, the method based on the ADMM for the optimization problem given in* (8) *has no convergence guarantees and does not match with the any of the existing convergence results.*

C. Nonconvexity of the Synthesis of Optimal Deceptive Policies and a Linear Programming Relaxation

Based on the nonconvexity of the optimization problem given in (8), one might wonder whether there exists a problem formulation that yields a convex optimization problem. In this section, we show that it is not possible to obtain a convex reformulation of the optimization problem given in (8) and give a convex relaxation of Problem 2.

We first 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.



(a) (b)

Figure 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 for different reference policies. Note that there are two local optima that maximizes the KL divergence.

We consider a new parametrization to reformulate the optimization problem given in (8). Consider a continuous and bijective transformation from the expected residence time parameters to the new parameters, that makes new parameters to span all stationary policies. After this transformation, an optimal solution to (8) yields an optimal solution in the new parameter space. If the optimization problem given in (8) 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.

Since it is not possible to obtain a convex reformulation of the optimization problem given in (8) 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 $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}})$ under policy $\pi^{\mathcal{S}}$ and $\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}})$ 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:

$$KL\left(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}} || \Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}\right) \geq KL\left(Ber\left(\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}\left(s_0 \models \phi^{\mathcal{A}}\right)\right) || Ber\left(\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}\left(s_0 \models \phi^{\mathcal{A}}\right)\right)\right). \tag{13}$$

We use this lower bound to construct the relaxed problem

$$\sup_{\pi^{\mathcal{S}} \in \Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M})} \quad KL\left(Ber\left(\Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}\left(s_0 \models \phi^{\mathcal{A}}\right)\right) || Ber\left(\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}\left(s_0 \models \phi^{\mathcal{A}}\right)\right)\right) \tag{14a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{14b}$$

$$\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}. \tag{14c}$$

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

$$\sup_{\pi^{\mathcal{S}} \in \Pi(\mathcal{M})} \inf_{\pi^{\mathcal{A}} \in \Pi(\mathcal{M})} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}}) + \Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}}) \tag{15a}$$

$$\text{subject to} \quad \Pr^{\pi^{\mathcal{A}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}}) \geq \nu^{\mathcal{A}}, \tag{15b}$$

$$\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}, \tag{15c}$$

shares the same optimal policies with the problem given in (14). We note that the optimization problem given in (15) 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 $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}})$ subject to $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{S}}) \geq \nu_{\mathcal{S}}$.

The lower bound given in (13) provides a sufficient condition on the optimality of a reference policy for Problem 2. A policy $\pi^{\mathcal{S}}$ satisfying $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{A}}) = 0$ and $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s_0 \models \phi^{\mathcal{S}}) \geq \nu^{\mathcal{S}}$ is an optimal reference policy since the optimization problem given in (14) has the optimal value of ∞. However, in general the gap due to the relaxation may get arbitrarily large, and the reference policy synthesized via (14) 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 $\Pr^{\pi^{\mathcal{S}}}_{\mathcal{M}}(s \models \Diamond q_1 \vee \Diamond q_2)$ 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 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.

D. Synthesis of the Optimal Reference Policy Under a Fixed Deceptive Policy

Given the observation that the synthesis of an optimal reference policy requires solving a nonconvex optimization problem, it is meaningful to consider the problem of synthesizing the optimal policy when the agent's policy is fixed, i.e., $x^{\mathcal{A}}_{s,a}$ variables are fixed. Unfortunately, this problem still remains nonconvex. For instance, consider the MDP given in Figure 2a where the agent's policy has $\pi^{\mathcal{A}}_{s_0,\gamma} = 1$ and the supervisor has no specifications, i.e., $\nu^{\mathcal{S}} = 0$. 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 $x^{\mathcal{S}}_{s_0,\gamma}$.

While computing the globally optimal policy is still hard when the policy of the agent is fixed, computing a locally optimal policy is possible applying the concave-convex procedure [24]. By plugging constraint (8b) into the objective function in (8a), we obtain

$$\sum_{s \in S_d} \sum_{a \in A(s)} \sum_{q \in S} x^{\mathcal{A}}_{s,a} P_{s,a,q} \left(\log \left(\frac{\sum_{a' \in A(s)} x^{\mathcal{A}}_{s,a'} P_{s,a',q}}{\sum_{a' \in A(s)} x^{\mathcal{A}}_{s,a'}} \right) + \log \left(\sum_{a' \in A(s)} x^{\mathcal{S}}_{s,a'} \right) - \log \left(\sum_{a' \in A(s)} x^{\mathcal{S}}_{s,a'} P_{s,a',q} \right) \right). \tag{16}$$

Note that when $x^{\mathcal{A}}_{s,a}$ parameters are fixed, the first $\log$ term is constant, the second $\log$ term is a concave function of $x^{\mathcal{S}}_{s,a}$ parameters, and the last $\log$ term is a convex function of $x^{\mathcal{S}}_{s,a}$ parameters. Based on this observation, we may apply the concave-convex procedure given in Algorithm 2 to get a locally optimal solution.

Algorithm 2 Convex-concave procedure to find a locally optimal reference policy when the agent's policy is fixed

Input: An MDP $\mathcal{M}$, a co-safe LTL specification $\phi^{\mathcal{S}}$, a probability threshold $\nu^{\mathcal{S}}$, and expected residence times of the agent, i.e., $x^{\mathcal{A}}_{s,a}$.
Output: A reference policy $\pi^{\mathcal{S}}$.
Set $\pi^{\mathcal{S},0}$ arbitrarily.
Set $k = 0$
while stopping criteria are not satisfied **do**
    Linearize $-\log\left(\sum_{a\in A(s)} x^{\mathcal{S}}_{s,a} P_{s,a,q}\right)$ at $x^{\mathcal{S},k}_{s,a}$ and find a concave approximation of (16).
    For approximated objective function, $\phi^{\mathcal{S}}$, and $\nu^{\mathcal{S}}$, find the optimal expected residence times $x^{\mathcal{S},k+1}$.
    $k := k + 1$.
end while
Compute $\pi^{\mathcal{S}}$ using $x^{\mathcal{S},k}$ as the expected residence times.

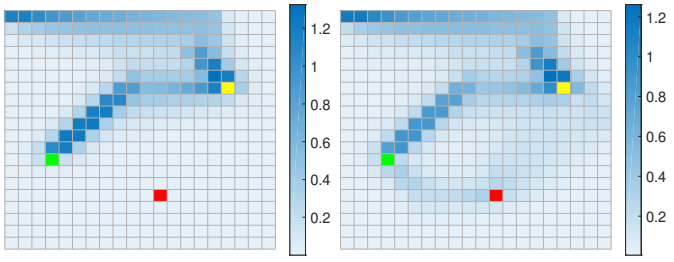
VI. Numerical Examples

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 [25] toolbox using MOSEK [26] and the nonconvex optimization problems using IPOPT [27].

A. Some Characteristics of Deceptive Policies

The first example demonstrates some of the characteristics of the optimal deceptive 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.



(a) Reference (b) Deceptive

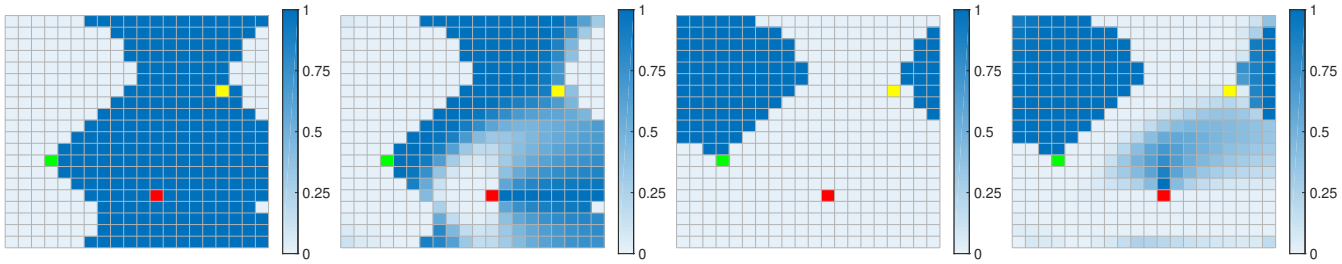
Figure 3: Heat maps of the expected residence times. The value of a state is the expected number of visits to the state. The deceptive policy follows the same policy until reaching the yellow state. Upon reaching the yellow state, the deceptive policy makes the agent move towards the red state to achieve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

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 $\phi^{\mathcal{S}} = \Diamond(y \wedge \Diamond r)$. 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.

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.

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

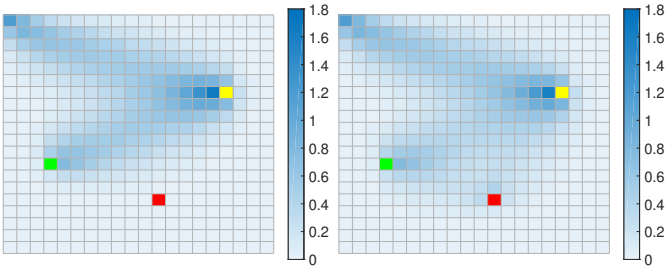
agent wants to drive himself to this region so that he can directly follow the reference policy without any divergence. Thus the agent deviates from the reference policy at a particular state to be close to the reference policy as much as possible in the rest of the path.



(a) Supervisor - Left (b) Agent - Left (c) Supervisor - Down (d) Agent - Down

Figure 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 $\phi^{\mathcal{A}}$ 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

Figure 5: Heatmaps of the expected number residence 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 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 expected residence times of the state-action pairs for the agent's policy and the reference policies, and by minimizing the L_2-norm between the expected residence times 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 policies. We simulate this rare event in the following way. Let $\Gamma^{\pi^S}_{\mathcal{M}}$ 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 [28] 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.

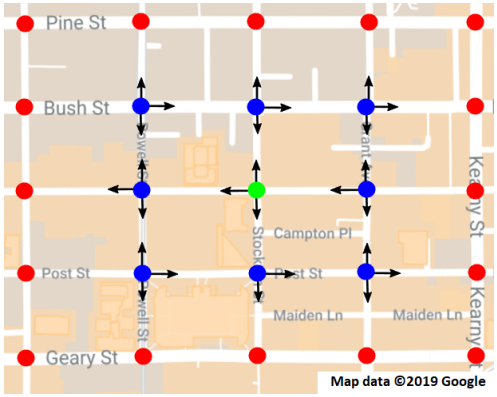
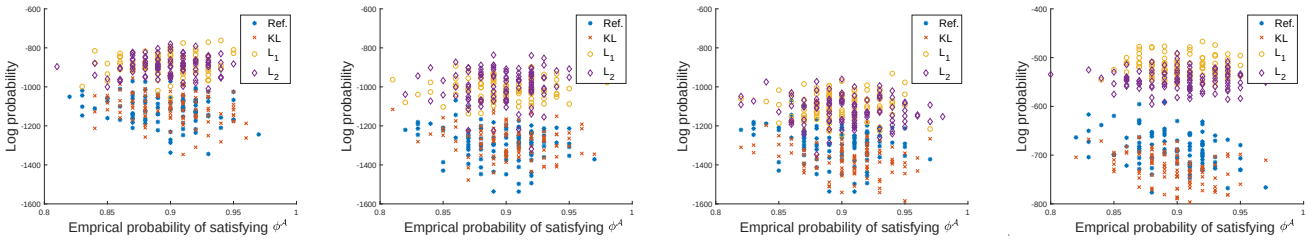


Figure 6: The map of a region from north east of San Francisco. The green dot indicates the intersection at which the highest number of crimes happened. The data is from [29]. The dots on the map represent the states of the MDP and the arrows represent the available actions. The initial state is chosen 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 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.



(a) Random MDP – 1 (b) Random MDP – 2 (c) Random MDP – 3 (d) MDP for San Francisco

Figure 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 satisfie $\phi^{\mathcal{A}}$ with high probability. 'KL', 'L_1', and 'L_2' refer to the candidate deceptive policies.

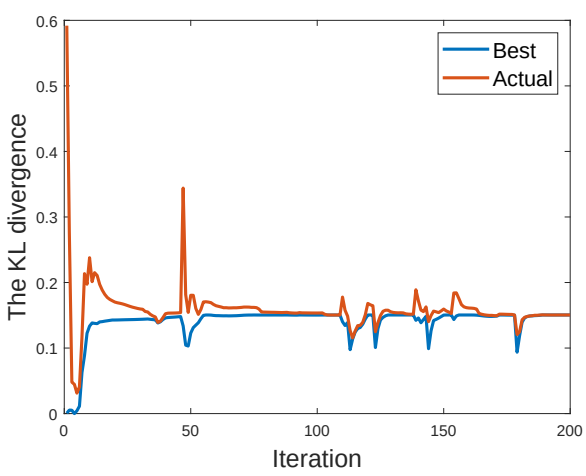
C. Optimal Reference Policies

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 have labels 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. At the green state there is an extra action that allows self transition. The initial state is the top-left state.

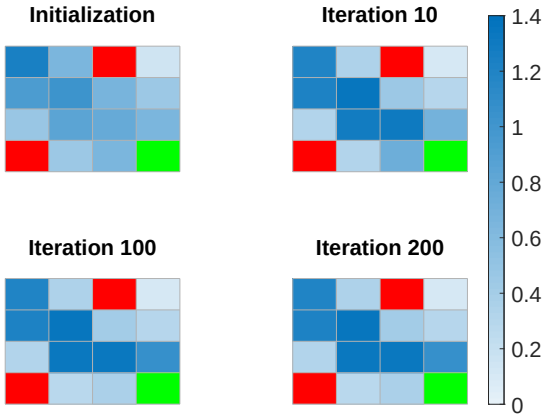
The specification of the supervisor is to reach the green state. The specification is encoded with the co-safe LTL formula $\phi^{\mathcal{S}} = \Diamond g$. Note that the specification of the supervisor is satisfied with probability 1 under any policy. The specification of the agent is to reach one of the red states. The specification is encoded with the co-safe LTL formula $\phi^{\mathcal{A}} = \Diamond r$. The probability threshold for the agent's task is 0.3.

We synthesize the reference policy via Algorithm 1 given in Section V-B. 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 expected residence times 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.



(a)



(b)

Figure 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 expected residence times 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 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-C 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 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 proposed a method based on the ADMM to compute a locally optimal solution.

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.

ACKNOWLEDGMENTS

This work was supported in part by ARO W911NF-15-1-0592, DARPA D19AP00004, and DARPA W911NF-16-1-0001.

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/~rdoody/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] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "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.
[14] M. Bakshi and V. M. Prabhakaran, "Plausible deniability over broadcast channels," *IEEE Transactions on Information Theory*, vol. 64, no. 12, pp. 7883–7902, 2018.
[15] 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.
[16] 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.
[17] 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.
[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] A. Pnueli, "The temporal logic of programs," in *18th Annual Symposium on Foundations of Computer Science*, 1977, pp. 46–57.
[20] C. Baier and J.-P. Katoen, *Principles of Model Checking*. MIT Press, 2008.
[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[22] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, pp. 1–35, Jun 2015.
[23] 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.
[24] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.
[25] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, 2014.
[26] MOSEK ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 8.1.*, 2017. [Online]. Available: http://docs.mosek.com/8.1/toolbox/index.html
[27] 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.
[28] Google, "Map of San Francisco," https://www.google.com/maps/@37.789463,-122.4068681,16.98z, accessed: Jan. 25, 2019.
[29] 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.
[30] K. Conrad, "Probability distributions and maximum entropy," *Entropy*, vol. 6, no. 452, pp. 1–10, 2004.
[31] F. H. Clarke, "Generalized gradients and applications," *Transactions of the American Mathematical Society*, vol. 205, pp. 247–262, 1975.

APPENDIX

We use the following definition and lemmas in the proof of Proposition 1.

Definition 4. 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 [30])**.** *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 $H(X^*) = c\left(-\frac{1}{c}\log\left(\frac{1}{c}\right) - \left(1-\frac{1}{c}\right)\log\left(1-\frac{1}{c}\right)\right)$.*

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^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. Let d^* be the optimal value of Problem 1. For a state $s \in S_d \backslash C_{cl}$, consider first the case $\Pr_{\mathcal{M}_p}^{\pi^{\mathcal{S}}}(s_{0_p} \models \Diamond s) = 0$. In this case, the agent's policy $\pi^{\mathcal{A}}$ must satisfy $\Pr_{\mathcal{M}_p}^{\pi^{\mathcal{A}}}(s_{0_p} \models \Diamond s) = 0$, i.e., s must be unreachable, otherwise the KL divergence is infinite. Hence the expected residence time is zero in this case.

Consider now $\Pr_{\mathcal{M}_p}^{\pi^{\mathcal{S}}}(s_{0_p} \models \Diamond s) > 0$. For this case, we will show that if the expected residence time 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 $\Pr_{\mathcal{M}_p}^{\pi^{\mathcal{S}}}(s_{0_p} \models \Diamond s) > 0$, $\Pr_{\mathcal{M}_p}^{\pi^{\mathcal{S}}}(s \models \bigcirc\Diamond s) \in [0,1)$, 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 $KL(\Gamma_{\mathcal{M}_p}^{\pi^{\mathcal{A}}} || \Gamma_{\mathcal{M}_p}^{\pi^{\mathcal{S}}}) > d^*$.

Consider a partitioning of paths that partitions according to the number of times s appears in a path. By the data processing inequality given in (1), we know that $KL(\Gamma_{\mathcal{M}_p}^{\pi^{\mathcal{A}}} || \Gamma_{\mathcal{M}_p}^{\pi^{\mathcal{S}}}) \geq KL(N_s^{\pi^{\mathcal{A}}} || N_s^{\pi^{\mathcal{S}}})$. Therefore it suffices to prove the following claim: Given $\Pr_{\mathcal{M}_p}^{\mathcal{S}}(s_0 \models \Diamond s) > 0$, $\Pr_{\mathcal{M}_p}^{\mathcal{S}}(s \models \bigcirc\Diamond s) \in [0,1)$, 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 $r^{\mathcal{S}} = 1 - \Pr_{\mathcal{M}_p}^{\mathcal{S}}(s_0 \models \Diamond s)$, $l^{\mathcal{S}} = \Pr_{\mathcal{M}_p}^{\mathcal{S}}(s \models \bigcirc\Diamond s)$, $p_i = \Pr(N_s^{\pi^{\mathcal{A}}} = i)$ and $\hat{p}_i = \Pr(\hat{N}_s^{\pi^{\mathcal{A}}} = i)$. Also let $\mathbb{E}[N_s^{\pi^{\mathcal{A}}}] = M^{\mathcal{A}}$, $\mathbb{E}[\hat{N}_s^{\pi^{\mathcal{A}}}] = \frac{M^{\mathcal{A}}}{1-p_0} = \hat{M}^{\mathcal{A}}$, and $\mathbb{E}[N_s^{\pi^{\mathcal{S}}}] = M^{\mathcal{S}}$.

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

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

where the equality (17a) follows from Lemma 2. The inequality in (17d) holds since the removed terms correspond to $KL(Ber(r^{\mathcal{A}})||Ber(r^{\mathcal{S}}))$ which is nonnegative.

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

$$KL(N_s^{\pi^{\mathcal{A}}}||N_s^{\pi^{\mathcal{S}}}) = -(1-p_0)H(\hat{N}_s^{\pi^{\mathcal{A}}}) - (1-p_0)\sum_{i=1}^{\infty}\hat{p}_i \log\left((l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})\right) \tag{18a}$$

$$\geq -(1-p_0)\hat{M}^{\mathcal{A}}\left(-\frac{1}{\hat{M}^{\mathcal{A}}}\log\left(\frac{1}{\hat{M}^{\mathcal{A}}}\right) - \left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\log\left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\right)$$
$$- (1-p_0)\sum_{i=1}^{\infty}\hat{p}_i \log\left((l^{\mathcal{S}})^{i-1}(1-l^{\mathcal{S}})\right) \tag{18b}$$

$$= (1-p_0)\hat{M}^{\mathcal{A}}\left(\frac{1}{\hat{M}^{\mathcal{A}}}\log\left(\frac{1}{\hat{M}^{\mathcal{A}}}\right) + \left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\log\left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\right)$$
$$- (1-p_0)\sum_{i=1}^{\infty}\hat{p}_i \left((i-1)\log(l^{\mathcal{S}}) + \log(1-l^{\mathcal{S}})\right) \tag{18c}$$

$$= (1-p_0)\hat{M}^{\mathcal{A}}\left(\frac{1}{\hat{M}^{\mathcal{A}}}\log\left(\frac{1}{\hat{M}^{\mathcal{A}}}\right) + \left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\log\left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\right)$$
$$- (1-p_0)\left(\log(1-l^{\mathcal{S}}) + \left(\hat{M}^{\mathcal{A}}-1\right)\log(l^{\mathcal{S}})\right) \tag{18d}$$

$$= (1-p_0)\hat{M}^{\mathcal{A}}\left(\frac{1}{\hat{M}^{\mathcal{A}}}\log\left(\frac{1}{\hat{M}^{\mathcal{A}}}\right) + \left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\log\left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\right)$$
$$- (1-p_0)\hat{M}^{\mathcal{A}}\left(\frac{1}{\hat{M}^{\mathcal{A}}}\log(1-l^{\mathcal{S}}) + \left(1-\frac{1}{\hat{M}^{\mathcal{A}}}\right)\log(l^{\mathcal{S}})\right) \tag{18e}$$

$$= M^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1}{\hat{M}^{\mathcal{A}}}\right)||Ber(1-l^{\mathcal{S}})\right)\right) \tag{18f}$$

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

$$KL(N_s^{\pi^{\mathcal{A}}}||N_s^{\pi^{\mathcal{S}}}) \geq M^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1}{\hat{M}^{\mathcal{A}}}\right)||Ber(1-l^{\mathcal{S}})\right)\right) \tag{19a}$$

$$\geq M^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1}{M^{\mathcal{A}}}\right)||Ber\left(1-l^{\mathcal{S}}\right)\right)\right) \tag{19b}$$

$$\geq M^{\mathcal{A}}\left(KL\left(Ber\left(\frac{1-l^{\mathcal{S}}}{c}\right)||Ber\left(1-l^{\mathcal{S}}\right)\right)\right) \tag{19c}$$

since $\hat{M}^{\mathcal{A}} > M^{\mathcal{A}}$ 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 $M^{\mathcal{A}} > M_s$.

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

Sketch of Proof for Proposition 2. Assume that the KL divergence between the path distributions is finite. Note that the expected residence times of $\pi^{\mathcal{A}}$ are finite for all $s \in S_d$.

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

Let $X_s^{\mathcal{S}}$ 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 $KL(X_{s,a}||X_s^{\mathcal{S}})$. 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 $C^{\mathcal{A}}$ 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 expected state-action residence times. An optimal policy can be characterized by the expected state-action residence times.

The expected residence times must be finite for all $s \in S_d$ as we showed in Proposition 1. Since every finite expected residence time vector of S_d can also be achieved by a stationary policy, there exists a stationary policy which shares the same expected residence times with an optimal policy. 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 $X_s^{\pi^*}$ be the successor state distribution at state s under π^*. Note that at state s there exists an action a^* that has $P(s, a^*, q) = X_s^{\pi}(q)$ since the action space is convex for the semi-infinite MDP. Also due to the convexity of KL divergence we have

$$\int KL(X_{s,a}||X_s^{\mathcal{S}})d\pi^*_s(a) \geq KL(X_s^{\pi^*}||X_s^{\mathcal{S}}).$$

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 $\mathcal{M}_p$. Hence the proposition holds.

■

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 [17].

Definition 5. 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 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 $\pi^{\mathcal{S}}$, i.e,*

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

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

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

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_{\xi_k\in\Xi_k} \Pr^{\pi^{\mathcal{A}}}(\xi_k)\sum_{t=0}^{k-1}\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)\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_{s\in S_d}\mathbb{1}_s(s_t)\sum_{q\in Succ(s)}\mathbb{1}_q(s_{t+1}|s_t=s)\sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(\xi_k)\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_{t=0}^{k-1}\sum_{s\in S_d}\mathbb{1}_s(s_t)\sum_{q\in Succ(s)}\sum_{a\in A(s_t)} P_{s,a,q}\pi^{\mathcal{A}}_{s,q}\sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(\xi_k)\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_{t=0}^{k-1}\sum_{s\in S_d}\sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(\xi_k)\mathbb{1}_s(s_t)\sum_{q\in Succ(s)}\sum_{a\in A(s_t)} 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_{t=0}^{k-1}\sum_{s\in S_d}\sum_{\xi_k\in\Xi_k}\Pr^{\pi^{\mathcal{A}}}(s_t=s)\sum_{q\in Succ(s)}\sum_{a\in A(s_t)} 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_{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)} 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)
\end{aligned}
$$

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

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

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_p. Finally, we have

$$
KL(\Gamma^{\pi^{\mathcal{A}}}_{\mathcal{M}}||\Gamma^{\pi^{\mathcal{S}}}_{\mathcal{M}}) = \sum_{s\in S_d}\sum_{q\in S_p}\sum_{a\in A(s)} P_{s,a,q}x^{\mathcal{A}}_{s,a}\log\left(\frac{\sum_{a'\in A(s)} P_{s,a',q}x^{\mathcal{A}}_{s,a'}}{\pi^{\mathcal{S}}_{s,q}\sum_{a'\in A(s)} x^{\mathcal{A}}_{s,a'}}\right).
$$

■

Proof of 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 (7) 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 (6b)-(6c) define the stationary policies that make the states in S_d have finite expected residence time and the constraint (6d) encodes the reachability constraint.

Note that

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

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 (6) is a sum of convex functions and the constraints are affine, (6) is a convex optimization problem.

We now show that there exists a stationary policy that achieves the optimal value of (1). By Proposition 1, we have that for all $s \in S_d$, the expected residence times 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 (6). 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 optimal residence time values, and consequently a stationary policy that achieves the optimal value of (6). ■

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 2, the expected residence times are bounded for all $s \in S_d$ under $\pi^{\mathcal{S}}$. Hence, the objective function of (8) 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 (8) 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 expected residence times 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 $x^{\mathcal{S}}$ [31]. 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 (8). ■