2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Optimal Control Synthesis of Markov Decision Processes for Efficiency with Surveillance Tasks

Yu Chen, Xuanyuan Yin, Hao Ye, Shaoyuan Li and Xiang Yin

***Abstract*— We investigate the problem of optimal control synthesis for Markov Decision Processes (MDPs), addressing both qualitative and quantitative objectives. Specifically, we require the system to fulfill a qualitative surveillance task in the sense that a specific region of interest can be visited infinitely often with probability one. Furthermore, to quantify the performance of the system, we consider the concept of *efficiency*, which is defined as the *ratio* between rewards and costs. This measure is more general than the standard long-run average reward metric as it aims to maximize the reward obtained *per unit cost*. Our objective is to synthesize a control policy that ensures the surveillance task while maximizes the efficiency. We provide an effective approach to synthesize a stationary control policy achieving ϵ-optimality by integrating state classifications of MDPs and perturbation analysis in a novel manner. Our results generalize existing works on efficiency-optimal control synthesis for MDP by incorporating qualitative surveillance tasks.**

I. Introduction

Decision-making in dynamic environments is a fundamental challenge for autonomous systems. Markov Decision Processes (MDPs) offer a theoretical framework for sequential decision-making by abstracting uncertainties in both environments and system executions as transition probabilities. In the context of autonomous systems, MDPs have found extensive applications across various domains such as swarm robotics [1], autonomous driving [2], and underwater vehicles [3]; reader is referred to recent surveys for additional references [4], [5].

To assess the performance of infinite horizon behaviors, two widely recognized measures are the long-run average reward (or mean payoff) and the discounted reward [6]. The long-run average reward quantifies the average reward received per state as the system evolves infinitely towards a steady state. Recently, the notion of *efficiency* has emerged to capture the *reward-to-cost ratio* [7], [8]. Specifically, the efficiency of a system trajectory is defined as the ratio between accumulated reward and accumulated cost. The efficient controller synthesis problem thus aims to maximize the expected long-run efficiency [8].

In addition to maximizing quantitative performance measures, many applications also require achieving qualitative tasks. Recently, within the context of MDPs, there has been a growing interest in synthesizing control policies to maximize the probability of satisfying high-level logic tasks expressed in, for example, linear temporal logic. When the MDPs model is known precisely, offline algorithms have been proposed to synthesize optimal controller under LTL specifications; see, e.g., [9]–[11]. Recently, reinforcement learning for LTL tasks has also been investigated for MDPs with unknown transition probabilities [12], [13]. Motivated primarily by persistent surveillance needs in autonomous systems [14]–[17], one important qualitative task that has been extensively studied is the *surveillance task*, which is equivalent to the Büchi accepting condition requiring that certain desired target states can be visited infinitely often.

In this work, we investigate control policy synthesis for MDPs with both qualitative and quantitative requirements. Specifically, for the qualitative aspect, we require that the surveillance task is satisfied with probability one (w.p.1). Additionally, for the quantitative aspect, we adopt the efficiency measure. Our overarching objective is to maximize the expected long-run efficiency while ensuring the satisfaction of the surveillance task w.p.1. It is worth noting that existing works typically focus on either efficiency optimization (ratio objectives) without qualitative requirements [8], or they consider qualitative requirements but under the standard long-run average reward (mean payoff) measure [18]. In [9], the authors consider qualitative requirement expressed by LTL formulae, with a quantitative measure referred to as the *per cycle* average reward. However, the per cycle average reward is essentially a special instance of the ratio objective by setting unit cost for specific state on the denominator. To the best of our knowledge, the simultaneous maximization of efficiency while achieving the surveillance task has not been addressed in the existing literature.

To fill this gap in research, we present an effective approach to synthesize stationary policies achieving ϵ-optimality. Our approach integrates state classifications of MDPs [19] and perturbation analysis techniques [20]–[22] in a novel manner. Specifically, the key idea of our approach is as follows. Initially, we decompose the MDPs into accepting maximal end components (AMECs) using state classifications, where for each AMEC, we solve the standard efficiency optimization problem without considering the surveillance task [8]. Subsequently, we synthesize a basic policy that achieves optimal efficiency but may fail to fulfill the surveillance task. Finally, we perturb the basic policy "slightly" by a target seeking policy such that the quantitative performance is decreased to ϵ-optimal but the surveillance

This work was supported by the National Natural Science Foundation of China (62061136004, 62173226, 61833012).

Yu Chen, Shaoyuan Li and Xiang Yin are with Department of Automation and Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China. {yuchen26, syli, yinxiang}@sjtu.edu.cn. Xuanyuan Yin is with School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore. Hao Ye is with the Department of Automation, Tsinghua University, China.

task is fulfilled. Our approach suggests that perturbation analysis is a conceptually simple yet powerful technique for solving MDPs with both qualitative and quantitative tasks, which may offer new insights into addressing this class of problems.

II. Preliminary

A. Markov Decision Processes

A (finite) Markov decision process (MDP) is a 4-tuple $\mathcal{M} = (S, s_0, A, P)$, where S is a finite set of states, $s_0 \in S$ is the initial state, A is a finite set of actions, and $P : S \times A \times S \to [0,1]$ is a transition function such that $\forall s \in S, a \in A : \sum_{s' \in S} P(s' \mid s, a) \in \{0, 1\}$. We also write $P(s' \mid s, a)$ as $P_{s,a,s'}$. For each state $s \in S$, we define $A(s) = \{a \in A : \sum_{s' \in S} P(s' \mid s, a) = 1\}$ as the set of available actions at s. We assume that each state has at least one available action, i.e., $\forall s \in S : A(s) \neq \emptyset$. An MDP also induces an underlying directed graph (digraph), where each vertex is a state and an edge of form $\langle s, s' \rangle$ is defined if $P(s' \mid s, a) > 0$ for some $a \in A(s)$.

A Markov chain (MC) $\mathcal{C}$ is an MDP such that $|A(s)| = 1$ for all $s \in S$. The transition matrix of MC is denoted by a $|S| \times |S|$ matrix $\mathbb{P}$, i.e., $\mathbb{P}_{s,s'} = P(s' \mid s, a)$, where $a \in A(s)$ is the unique action at state s. Therefore, we can omit actions in MC and write it as $\mathcal{C} = (S, s_0, \mathbb{P})$. The *limit transition matrix* of MC is defined by $\mathbb{P}^\star = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n} \mathbb{P}^k$, which always exists for finite MC [6]. Let $\pi_0 \in \mathbb{R}^{|S|}$ be initial distribution with $\pi_0(s) = 1$ if s is initial state and $\pi_0(s) = 0$ otherwise. Then the *limit distribution* of MC is $\pi = \pi_0 \mathbb{P}^\star$. A state is said to be *transient* if its corresponding column in the limit transition matrix is a zero vector; otherwise, the state is *recurrent*.

A *policy* for an MDP $\mathcal{M}$ is a sequence $\mu = (\mu_0, \mu_1, ...)$, where $\mu_k : S \times A \to [0,1]$ is a function such that $\forall s \in S : \sum_{a \in A(s)} \mu_k(s, a) = 1$. A policy is said to be *stationary* if $\mu_i = \mu_j, \forall i, j$ and we write a stationary policy as $\mu = (\mu, \mu, \dots)$ for simplicity. Given an MDP $\mathcal{M}$, the sets of all policies and all stationary policies are denoted by $\Pi_\mathcal{M}$ and $\Pi^S_\mathcal{M}$, respectively. For policy $\mu \in \Pi_\mathcal{M}$, at each instant k, it induces a transition matrix $\mathbb{P}^{\mu_k}$, where $\mathbb{P}^{\mu_k}_{i,j} = \sum_{a \in A(i)} \mu_k(i, a) P_{i,a,j}$. A stationary policy $\mu \in \Pi^S_\mathcal{M}$ induces a time-homogeneous MC with transition matrix $\mathbb{P}^\mu$.

An infinite sequence $\rho = s_0 s_1 \cdots$ of states is said to be a *path* in MDP $\mathcal{M}$ under policy $\mu \in \Pi_\mathcal{M}$ if s_0 is initial state of $\mathcal{M}$ and $\forall k \geq 0 : \sum_{a \in A(s_k)} \mu_k(s_k, a) P(s_{k+1} \mid s_k, a) > 0$. We denote by $\mathsf{Path}^\mu(\mathcal{M}) \subseteq S^\omega$ the set of all paths in $\mathcal{M}$ under μ, where S^ω denotes the set of all infinite sequences of states. We use the standard cylinder-based probability measure $\mathsf{Pr}^\mu_\mathcal{M} : 2^{S^\omega} \to [0,1]$ for infinite paths; see, e.g., [19].

For MDP $\mathcal{M} = (S, s_0, A, P)$, a *sub-MDP* is a tuple $(\mathcal{S}, \mathcal{A})$, where $\mathcal{S} \subseteq S$ is a non-empty subset of states and $\mathcal{A} : \mathcal{S} \to 2^A \setminus \emptyset$ is a function such that (i) $\forall s \in \mathcal{S} : \mathcal{A}(s) \subseteq A(s)$; and (ii) $\forall s \in \mathcal{S}, a \in \mathcal{A}(s) : \sum_{s' \in \mathcal{S}} P_{s,a,s'} = 1$. Essentially, $(\mathcal{S}, \mathcal{A})$ induces a new MDP by restricting the state space to $\mathcal{S}$ and available actions to $\mathcal{A}(s)$ for each state $s \in \mathcal{S}$.

B. Ratio Objectives for Efficiency

In the context of MDPs, quantitative measures such as *average rewards* have been widely used for systems operating in infinite horizons. In [7], [8], a general quantitative measure called *ratio objective* is proposed to characterize the *efficiency* of policies. Specifically, two different functions are involved:

- a *reward function* $\mathtt{R} : S \times A \to \mathbb{R}_{\geq 0}$ assigning each state-action pair a non-negative reward; and
- a *cost function* $\mathtt{C} : S \times A \to \mathbb{R}_+$ assigning each state-action pair a positive cost.

Then the *efficiency value* from initial state s_0 under policy $\mu \in \Pi_\mathcal{M}$ w.r.t. reward-cost pair $(\mathtt{R}, \mathtt{C})$ is defined by

$$J^\mu(s_0, \mathtt{R}, \mathtt{C}) := \limsup_{N \to +\infty} E\left\{ \frac{\sum_{i=0}^{N} \mathtt{R}(s_i, a_i)}{\sum_{i=0}^{N} \mathtt{C}(s_i, a_i)} \right\},$$

where $E\{\cdot\}$ is the expectation of probability measure $\mathsf{Pr}^\mu_\mathcal{M}$. We omit the reward and cost functions if they are clear by context. Intuitively, $J^\mu(s_0)$ captures the average reward the system received *per cost*, i.e., the efficiency. Let $\Pi \subseteq \Pi_\mathcal{M}$ be a set of policies. Then optimal efficiency value among policy set Π is denoted by $J(s_0, \Pi) = \sup_{\mu \in \Pi} J^\mu(s_0)$.

Note that the standard long-run average reward is a special case of ratio objective by taking $\mathtt{C}(s, a) = 1, \forall s \in S, a \in A(s)$. For this case, we denote by $W^\mu(s_0, \mathtt{R}) := J^\mu(s_0, \mathtt{R}, \mathbf{1})$ the standard long-run average reward from initial state s_0 under policy μ, and denote by $W(s_0, \Pi) = \sup_{\mu \in \Pi} W^\mu(s_0)$ the optimal long-run average reward among policy set Π .

III. Problem Formulation

Note that efficiency does not take qualitative requirements into account, i.e., the system may maximize its efficiency by doing useless things. In this work, motivated by surveillance tasks in autonomous robots, in addition to the ratio objectives, we further consider the qualitative requirement by visiting target states *infinitely often*.

Formally, let $B \subseteq S$ be a set of *target states* that need to be visited infinitely. Then the probability of visiting B infinitely often under policy $\mu \in \Pi_\mathcal{M}$ is defined by

$$\mathsf{Pr}^\mu_\mathcal{M}(\square\lozenge B) = \mathsf{Pr}^\mu_\mathcal{M}(\{\tau \in \mathsf{Path}^\mu(\mathcal{M}) \mid \mathtt{inf}(\tau) \cap B \neq \emptyset\}),$$

where $\mathtt{inf}(\tau)$ denotes the set of states that occur infinite number of times in path $\tau \in \mathsf{Path}^\mu(\mathcal{M})$. We denote by $\Pi^B_\mathcal{M}$ the set of all policies under which B is visited infinitely often w.p.1, i.e.,

$$\Pi^B_\mathcal{M} := \{\mu \in \Pi_\mathcal{M} \mid \mathsf{Pr}^\mu_\mathcal{M}(\square\lozenge B) = 1\}.$$

For the sake of simplicity and without loss of generality, we assume that, starting from any state, there exists a policy such that the surveillance task can be satisfied.

Now we formulate the problem solved in this paper.

Problem 1: Given MDP $\mathcal{M} = (S, s_0, A, P)$, reward function $\mathtt{R}$, cost function $\mathtt{C}$ and a threshold value $\epsilon > 0$, find a stationary policy $\mu^\star \in \Pi^B_\mathcal{M} \cap \Pi^S_\mathcal{M}$ such that

$$J^{\mu^\star}(s_0) \geq J(s_0, \Pi^B_\mathcal{M}) - \epsilon. \tag{1}$$

Remark 1: Before proceeding further, we make several comments on the above problem formulation.

- First, here we seek to find an ϵ-optimal policy $\mu^\star$ among all policies satisfying surveillance tasks. The main motivation for this setting is that policies with finite memory are not sufficient to achieve the optimal efficiency value $J(s_0, \Pi^B_{\mathcal{M}})$. Furthermore, even if one employs an infinite memory policy to achieve the optimal efficiency value, the system will visit target states less and less frequently as time progresses. One is referred to [18] regarding this issue for the case of standard long-run average measure, which is a special case of our ratio objective.
- Second, we further restrict our attention to stationary policies in $\Pi^S_{\mathcal{M}}$ a priori. We will show in the following result that such a restriction is without loss of generality in the sense that a stationary solution always exists.

Proposition 1: Given MDP $\mathcal{M} = (S, s_0, A, P)$ and threshold value $\epsilon > 0$, there always exists a policy $\mu \in \Pi^B_{\mathcal{M}} \cap \Pi^S_{\mathcal{M}}$ such that $J^\mu(s_0) \geq J(s_0, \Pi^B_{\mathcal{M}}) - \epsilon$.

IV. Case of Communicating MDPs

Before tackling the general case, in this section, we consider a special scenario, where the MDP is communicating. Formally, an MDP $\mathcal{M}$ is said to be *communicating* if

$$\forall s, s' \in S, \exists \mu \in \Pi_{\mathcal{M}}, \exists n \geq 0 : (\mathbb{P}^\mu)^n_{s,s'} > 0. \tag{2}$$

In other words, for a communicating MDP, one state is always able to reach another state.

General Idea: We solve Problem 1 for the case of communicating MDP by the following three steps:

- First, we apply the standard algorithm in [8] to optimize the ratio objective without considering the surveillance task. The resulting policy is denoted by μ_{opt}.
- Second, we select an arbitrary policy μ_{sur} such that its induced MC is irreducible. Therefore, target states can be visited infinitely often under μ_{sur}.
- Finally, we *perturb* policy μ_{opt} "slightly" by μ_{sur} such that the efficiency value of the resulting policy is ϵ-close to that of μ_{opt}, and the surveillance task can still be achieved due to the presence of perturbation μ_{sur} .

Now, we proceed the above idea in more detail.

A. Efficiency Optimization for Communicating MDP

In this subsection, we review the existing solution for efficiency optimization. It has been shown in [8] that, for communicating MDP $\mathcal{M}$, there exists a stationary policy $\mu \in \Pi^S_{\mathcal{M}}$ such that $J^\mu(s_0) = J(s_0, \Pi_{\mathcal{M}})$ and the induced MC $\mathcal{M}^\mu$ is an *unichain* (MC with a single recurrent class and some transient states). Furthermore, we have

$$J^\mu(s_0) = \frac{\sum_{s\in S}\sum_{a\in A(s)} \pi(s)\mu(s,a)\mathtt{R}(s,a)}{\sum_{s\in S}\sum_{a\in A(s)} \pi(s)\mu(s,a)\mathtt{C}(s,a)}, \tag{3}$$

where $\pi^\top \in \mathbb{R}^{|S|}$ is the unique stationary distribution such that $\pi\mathbb{P}^\mu = \pi$. With this structural property for communicating MDP, [8] transforms the policy synthesis problem for efficiency optimization to to a steady-state parameter synthesis problem described by the nonlinear program (4)-(9) shown as follows:

$$\max_{\gamma(s,a)} \quad \frac{\sum_{s\in S}\sum_{a\in A(s)} \gamma(s,a)\mathtt{R}(s,a)}{\sum_{s\in S}\sum_{a\in A(s)} \gamma(s,a)\mathtt{C}(s,a)} \tag{4}$$

$$\text{s.t.} \quad q(s,t) = \sum_{a\in A(s)} \gamma(s,a)P(t \mid s,a), \forall s,t \in S \tag{5}$$

$$\lambda(s) = \sum_{a\in A(s)} \gamma(s,a), \forall s \in S \tag{6}$$

$$\lambda(t) = \sum_{s\in S} q(s,t), \forall t \in S \tag{7}$$

$$\sum_{s\in S} \lambda(s) = 1 \tag{8}$$

$$\gamma(s,a) \geq 0, \forall s \in S, \forall a \in A(s) \tag{9}$$

Since we will only leverage this existing result, the reader is referred to [8] for more details on the intuition of the above nonlinear program. The only point we would like to emphasize is that this nonlinear program is a linear-fractional programming, which can be solved efficiently by converting to a linear program by Charnes-Cooper transformation [23]. Now, let $\gamma^\star(s,a)$ be the solution to Equations (4)-(9). The *optimal policy*, denoted by μ_{opt}, can be decoded as follows. Let $Q = \{s \in S \mid \sum_{a\in A(s)} \gamma^\star(s,a) > 0\}$. Then for states in Q, we define

$$\mu_{opt}(s,a) = \frac{\gamma^\star(s,a)}{\sum_{a\in A(s)} \gamma^\star(s,a)}, \quad s \in Q. \tag{10}$$

For the remaining part, policy μ_{opt} only needs to ensure that states in $S \setminus Q$ are transient states in MC $\mathcal{M}^{\mu_{opt}}$; see, e.g., procedure in [6, Page 480]. Then such a policy μ_{opt} achieves $J^{\mu_{opt}}(s_0) = J(s_0, \Pi_{\mathcal{M}})$. Furthermore, it has been shown in [8] that μ_{opt} can be deterministic, i.e., $\forall s \in S, \exists a \in A(s) : \mu_{opt}(s,a) = 1$. Hereafter, we assume that the constructed policy μ_{opt} is deterministic.

B. Efficiency Optimization with Surveillance Tasks

Note that, under policy μ_{opt}, only states in Q are recurrent. Therefore, if $Q \cap B = \emptyset$, then the surveillance task fails. As we mentioned at the beginning of this section, our approach is to perturb μ_{opt} so that (i) its ratio value will not decrease more than ϵ; and (ii) the surveillance task can be achieved.

To this end, let us consider an arbitrary stationary policy $\mu_{sur} \in \Pi^S_{\mathcal{M}}$, which is referred to as the *surveillance policy*, such that $\mathcal{M}^{\mu_{sur}}$ is irreducible. For policy μ_{sur}, we have

- It is well-defined since we already assume that the MDP $\mathcal{M}$ is communicating. For example, one can simply use the uniform policy as μ_{sur}, i.e., each available action is enabled with the same probability at each state;
- The surveillance task can be achieved by μ_{sur} since all states can be visited infinitely often w.p.1.

Now, we perturb the optimal policy μ_{opt} by the surveillance policy μ_{sur} to obtain a new policy as follows

$$\mu_{pert} := (1-\delta)\mu_{opt} + \delta\mu_{sur}, \tag{11}$$

where $0 < \delta < 1$ is the perturbation degree and the above notation means that $\mu_{pert}(s,a) = (1-\delta)\mu_{opt}(s,a) + \delta\mu_{sur}(s,a), \forall s \in S, a \in A(s)$. Clearly, this perturbed policy μ_{pert} has the following two properties:

- First, we have $J^{\mu_{pert}}(s_0) \leq J^{\mu_{opt}}(s_0)$ as μ_{opt} is already the optimal one to achieve the ratio objective. Furthermore, $J^{\mu_{pert}}(s_0) \to J^{\mu_{opt}}(s_0)$ as $\delta \to 0$;
- Second, the surveillance task can still be achieved. This is because, under policy μ_{pert}, the system always has non-zero probability to execute surveillance policy μ_{sur}.

Now, it remains to quantify the relationship between perturbation degree δ and the performance decrease $J^{\mu_{opt}}(s_0) - J^{\mu_{pert}}(s_0)$. That is, how small δ should be in order to ensure ϵ-optimality.

To this end, we adopt the idea of perturbation analysis of MDP, which is originally developed to quantify the difference of long-run average rewards between two policies [20]. First, we introduce some related definitions.

Definition 1 (Utility Vectors & Potential Vectors): Let $\mu \in \Pi^S_{\mathcal{M}}$ be a stationary policy and $\mathtt{V} : S \times A \to \mathbb{R}$ be a generic utility function, which can be either the reward function $\mathtt{R}$ or the cost function $\mathtt{C}$. Then

- the *utility vector* of policy μ (w.r.t. utility function $\mathtt{V}$), denoted by $v^{\mu}_{\mathtt{V}} \in \mathbb{R}^{|S|}$, is defined by

$$v^{\mu}_{\mathtt{V}}(s) = \sum_{a \in A(s)} \mu(s,a)\mathtt{V}(s,a). \tag{12}$$

- the *potential vector* of policy μ (w.r.t. utility function $\mathtt{V}$), denoted by $g^{\mu}_{\mathtt{V}} \in \mathbb{R}^{|S|}$, is defined by

$$g^{\mu}_{\mathtt{V}} = (I - \mathbb{P}^{\mu} + (\mathbb{P}^{\mu})^{\star})^{-1} v^{\mu}_{\mathtt{V}}. \tag{13}$$

In the above definition, the potential vector is well-defined as matrix $I - \mathbb{P}^{\mu} + (\mathbb{P}^{\mu})^{\star}$ is always invertible [6], where $(\mathbb{P}^{\mu})^{\star}$ is the limit transition matrix of $\mathbb{P}^{\mu}$. Intuitively, the potential vector $g^{\mu}_{\mathtt{V}}$ contains the information regarding the long run average utility in MC $\mathcal{M}^{\mu}$. Specifically, let π_{μ} be the limit distribution of MC $\mathcal{M}^{\mu}$. Then we have

$$\pi^{\top}_{\mu} g^{\mu}_{\mathtt{V}} = \pi^{\top}_{\mu} v^{\mu}_{\mathtt{V}} = W^{\mu}(s_0, \mathtt{V}),$$

which computes the long run average utility under μ.

Next, we define notion of deviation vectors of two different policies.

Definition 2 (Deviation Vectors): Let $\mu, \mu' \in \Pi^S_{\mathcal{M}}$ be two stationary policies and $\mathtt{V} : S \times A \to \mathbb{R}$ be a utility function. Then the *deviation vector* from μ to μ' (w.r.t. utility function $\mathtt{V}$) is defined by

$$\mathbf{D}_{\mathtt{V}}(\mu, \mu') = (v^{\mu'}_{\mathtt{V}} - v^{\mu}_{\mathtt{V}}) + (\mathbb{P}^{\mu'} - \mathbb{P}^{\mu}) g^{\mu}_{\mathtt{V}}. \tag{14}$$

The deviation vector can be used to compute the difference between the long-run average utility of the original policy and the perturbed policy. Formally, let $\mu, \mu' \in \Pi^S_{\mathcal{M}}$ be two stationary policies, $\mathtt{V} : S \times A \to \mathbb{R}$ be a utility function and $\delta \in (0,1)$ be the perturbation degree. We define

$$\mu_{\delta} = (1-\delta)\mu + \delta\mu'$$

as the δ-perturbed policy of μ by μ'. It was shown in [20] that, when $\mathcal{M}^{\mu}$ is a unichain, the differences between the long run average utilities of the perturbed policy and the original policy can be calculated as follow:

$$W^{\mu_{\delta}}(s_0, \mathtt{V}) - W^{\mu}(s_0, \mathtt{V}) = \pi^{\top}_{\mu_{\delta}} v^{\mu_{\delta}}_{\mathtt{V}} - \pi^{\top}_{\mu} v^{\mu}_{\mathtt{V}} = \delta \pi^{\top}_{\mu_{\delta}} \mathbf{D}_{\mathtt{V}}(\mu, \mu'). \tag{15}$$

However, the above classical result can only be applied to the case of long-run average reward. The following proposition provides the key result of this subsection, which shows how to generalize Equation (15) from long-run average reward to the case of long-run efficiency under the ratio objective.

Proposition 2: Let $\mu, \mu' \in \Pi^S_{\mathcal{M}}$ be two stationary policies, $\mathtt{R} : S \times A \to \mathbb{R}_{\geq 0}$ be the reward function, $\mathtt{C} : S \times A \to \mathbb{R}_{+}$ be the cost function, and $\delta \in (0,1)$ be the perturbation degree. Let $\mu_{\delta} = (1-\delta)\mu + \delta\mu'$ be the perturbed policy. If $\mathcal{M}^{\mu}$ is unichain, then we have

$$\begin{aligned} &J^{\mu_{\delta}}(s_0, \mathtt{R}, \mathtt{C}) - J^{\mu}(s_0, \mathtt{R}, \mathtt{C}) \\ =&\frac{\delta}{\pi^{\top}_{\mu_{\delta}} v^{\mu_{\delta}}_{\mathtt{C}}} \pi^{\top}_{\mu_{\delta}} \left(\mathbf{D}_{\mathtt{R}}(\mu, \mu') - J^{\mu}(s_0, \mathtt{R}, \mathtt{C}) \mathbf{D}_{\mathtt{C}}(\mu, \mu')\right) \end{aligned} \tag{16}$$

Remark 2: Clearly, our new result in Equation (16) for ratio objective subsumes the classical result in Equation (15) for the case of long-run average reward. Specifically, when $\mathtt{C}(s,a) = 1, \forall s \in S, a \in A(s)$, $J^{\mu}(s_0, \mathtt{R}, \mathtt{C})$ reduces to $W^{\mu}(s_0, \mathtt{R})$. For this case, we know that $\pi^{\top}_{\mu_{\delta}} v^{\mu_{\delta}}_{\mathtt{C}} = 1$ as $v^{\mu}_{\mathtt{C}}(s) = 1, \forall s \in S$. Furthermore, we have $\mathbf{D}_{\mathtt{C}}(\mu, \mu') = 0$ as both policies achieve the same cost. Therefore, Equation (16) becomes to Equation (15) and our result provides a more general form of perturbation analysis.

Now let us discuss how to use Proposition 2 to determine the perturbation degree δ such that ϵ-optimality holds. Note that, in Equation (16), term $\mathbf{D}_{\mathtt{R}}(\mu, \mu') - J^{\mu}(s_0, \mathtt{R}, \mathtt{C}) \mathbf{D}_{\mathtt{C}}(\mu, \mu')$ can be computed explicitly based on μ and μ'. However, term $\frac{\pi^{\top}_{\mu_{\delta}}}{\pi^{\top}_{\mu_{\delta}} v^{\mu_{\delta}}_{\mathtt{C}}}$ cannot be directly computed. Our approach here is to estimate its bound as follows:

- Let $c_{min} = \min_{s \in S, a \in A(s)} \mathtt{C}(s,a)$ be minimum cost for all state-action pairs. Then we have $\pi^{\top}_{\mu_{\delta}} v^{\mu_{\delta}}_{\mathtt{C}} \geq c_{min}$.
- Let the infinity norm of the computable part be

$$\mathbf{D}^{\mu, \mu'}_{\infty} = \|\mathbf{D}_{\mathtt{R}}(\mu, \mu') - J^{\mu}(s_0) \mathbf{D}_{\mathtt{C}}(\mu, \mu')\|_{\infty}. \tag{17}$$

We have $|\pi^{\top}_{\mu_{\varsigma}} (\mathbf{D}_{\mathtt{R}}(\mu, \mu') - J^{\mu}(s_0) \mathbf{D}_{\mathtt{C}}(\mu, \mu'))| \leq \mathbf{D}^{\mu, \mu'}_{\infty}$.

These inequalities lead to the following result.

Proposition 3: Let $\mathcal{M} = (S, s_0, A, P)$ be a communicating MDP, $\mu_{opt} \in \Pi^S_{\mathcal{M}}$ be the optimal policy for ratio objective, $\mu_{sur} \in \Pi^S_{\mathcal{M}} \cap \Pi^B_{\mathcal{M}}$ be a surveillance policy, and μ_{pert} be defined in (11). If

$$\delta \leq \epsilon \frac{c_{min}}{\mathbf{D}^{\mu_{opt}, \mu_{sur}}_{\infty}}, \tag{18}$$

then we have $J^{\mu_{pert}}(s_0) \geq J^{\mu_{opt}}(s_0) - \epsilon$.

Finally, based on Proposition 3, we can establish the main theorem showing the correctness of our approach.

Theorem 1: Let $\mathcal{M} = (S, s_0, A, P)$ be a communicating MDP. If δ satisfies Equation (18), then policy μ_{pert} defined in (11) is a solution to Problem 1.

V. Solution to the General Case

A. Overview of Our Approach

The approach in the previous section assumes that MDP $\mathcal{M}$ is communicating. In general, however, the MDP may not be communicating and the optimal ratio objective policy may induce a multi-chain MC, i.e., an MC containing more than one recurrent classes. Our approach for handling the general case consists of the following steps:

1) First, we decompose the MDP into several communicating sub-MDPs containing target states, which are referred to as accepting maximal end components (AMEC). Eventually, the system needs to stay within these AMECs in order to achieve the surveillance task;
2) Next, for each AMEC, since it is communicating, we can compute the optimal efficiency value one can achieve within the AMEC by the nonlinear program (4)-(9) as discussed in Section IV-A;
3) Note that, since we consider long-run objectives, the efficiency value counts only when one decides to stay in some AMEC forever. Therefore, we construct a standard long-run average reward (per-stage) optimization problem, in which the reward for each state is determined by the optimal efficiency value of its associated AMEC (if any). This gives us a *basic policy* such that it attains the optimal efficiency value within all policies in $\Pi^B_{\mathcal{M}}$ (but may has not yet achieve the surveillance task);
4) Finally, for the basic policy, we perturb within each AMEC using the approach in Section IV-B such that the efficiency value decreases to ϵ-optimal but the surveillance task is achieved.

Before presenting our formal algorithm, we further introduce some necessary concepts.

Definition 3 (Accepting Maximal End Components): A sub-MDP $(\mathcal{S}, \mathcal{A})$ of $\mathcal{M} = (S, s_0, A, P)$ is said to be an *end component* if its underlying digraph is strongly connected. We say $(\mathcal{S}, \mathcal{A})$ is an *maximal end component* if it is an end component and there is no other end component $(\mathcal{S}', \mathcal{A}')$ such that (i) $\mathcal{S} \subseteq \mathcal{S}'$; and (ii) $\mathcal{A} \subseteq \mathcal{A}'$. We denote by $\texttt{MEC}(\mathcal{M})$ the set of all MECs in $\mathcal{M}$. An MEC is said to be an *accepting* MEC (AMEC) if $\mathcal{S} \cap B \neq \emptyset$; we denote by $\texttt{AMEC}(\mathcal{M}) \subseteq \texttt{MEC}(\mathcal{M})$ the set of AMECs.

Now suppose that $\mathcal{M}$ has n AMECs denoted by $\texttt{AMEC}(\mathcal{M}) = \{(\mathcal{S}_1, \mathcal{A}_1), \ldots, (\mathcal{S}_n, \mathcal{A}_n)\}$, which can be computed in polynomial time by Algorithm 47 in [19]. For each AMEC $(\mathcal{S}_i, \mathcal{A}_i)$, we denote by μ^i_{opt} and $J^{\mu^i_{opt}}$ the optimal policy for ratio objective $(\texttt{R}, \texttt{C})$ and its corresponding efficiency value computed by program (4)-(9), respectively[1].

Note that we already assume, without loss of generality that, μ^i_{opt} is deterministic. Let $K \in \mathbb{R}$ be a real number. Then based on K and $\mu^i_{opt}, i = 1, \ldots, n$, we define a new reward function $\texttt{R}_K : S \times A \to \mathbb{R}$ for the entire $\mathcal{M}$ by:

$$\texttt{R}_K(s,a) = \begin{cases} J^{\mu^i_{opt}} & \text{if } s \in \mathcal{S}_i \wedge \mu^i_{opt}(s,a) = 1 \\ K & \text{otherwise} \end{cases}. \quad (19)$$

[1]We omit initial state in $J^{\mu^i_{opt}}$ since for each communicating MDP, the efficiency value under the optimal policy is initial-state independent.

Algorithm 1: Policy Synthesis for the General Case

Input: MDP $\mathcal{M} = (S, s_0, A, P)$, target set $B \subseteq S$ and threshold value $\epsilon > 0$
Output: Policy $\mu^\star \in \Pi^S_{\mathcal{M}}$ which solve Problem 1

1 Compute all AMECs $\texttt{AMEC}(\mathcal{M})$ in $\mathcal{M}$;
2 For each AMEC $(\mathcal{S}_i, \mathcal{A}_i) \in \texttt{AMEC}(\mathcal{M})$, compute μ^i_{opt} and $J^{\mu^i_{opt}}$ by program (4)-(9) over $(\mathcal{S}_i, \mathcal{A}_i)$;
3 Define reward function $\texttt{R}_K$ according to Eq. (19), where K satisfies Eq. (20);
4 Compute policy $\mu^\star_K$ by solving the classical long-run average reward maximization problem w.r.t. $\texttt{R}_K$;
5 $\mu^\star \leftarrow \mu^\star_K$;
6 **for** $(\mathcal{S}_i, \mathcal{A}_i) \in \texttt{AMEC}(\mathcal{M})$ **do**
7 **if** $\mathcal{S}_i$ *contains a recurrent state in MC* $\mathcal{M}^{\mu^\star_K}$ **then**
8 Find a surveillance policy μ^i_{sur} for sub-MDP $(\mathcal{S}_i, \mathcal{A}_i)$
9 Pick $\delta > 0$ satisfying Equation (18)
10 Perturb the policy $\mu^\star$ by μ^i_{sur} with degree δ for the part of sub-MDP $(\mathcal{S}_i, \mathcal{A}_i)$, i.e.,
11 $$\mu^\star(s,a) \leftarrow \begin{cases} (1-\delta)\mu^\star(s,a) + \delta\mu^i_{sur}(s,a) & \text{if } s \in \mathcal{S}_i, a \in \mathcal{A}_i \\ \mu^\star(s,a) & \text{otherwise} \end{cases}$$
12 **Return** ϵ-optimal policy $\mu^\star$

Intuitively, for each optimal state-action pair in an AMEC, the above construction assigns exactly the same reward identical to the optimal efficiency value one can achieve within this AMEC. For the remaining state-action pairs that are either non-optimal or not in AMECs, we assign them value K. Clearly, for the purposes of being optimal or to fulfill the surveillance task, one needs to avoid executing such state-action pair with value K. Hence, one needs to select K to be sufficiently small and we will show later in Section V-C how small K can ensure so.

Later on, we also need to solve the classical long-run average reward maximization problem of $\mathcal{M}$ w.r.t. reward function $\texttt{R}_K$. We denote by $\mu^\star_K \in \Pi^S_{\mathcal{M}}$ the optimal long-run average reward policy, i.e.,

$$W^{\mu^\star_K}(s_0, \texttt{R}_K) = W(s_0, \texttt{R}_K, \Pi_{\mathcal{M}}).$$

Such optimal policy $\mu^\star_K$ can be obtained by the standard linear programming approach in [6].

B. Main Synthesis Algorithm

Based on the above informal discussions, our overall synthesis procedure for the entire MDP $\mathcal{M}$ is provided in Algorithm 1. Specifically, in line 1, we first compute all AMECs. Then we solve program (4)-(9) for each AMEC and record the constructed policy and optimal efficiency value in lines 2. These policies and values help us to define reward function $\texttt{R}_K$, for which the maximum average reward policy $\mu^\star_K$ is synthesized. These are done by lines 3-4. Note that K needs to be chosen sufficiently small so that MDP will not stay in those non-AMEC states. Then in line 5, we choose $\mu^\star_K$ as the initial policy to be perturbed.

Finally, in lines 7-11, based on the initial policy, we determine whether each AMEC $(\mathcal{S}_i, \mathcal{A}_i)$ contains some recurrent state in MC $\mathcal{M}^{\mu_K^\star}$. If so, it means that the MDP will achieve higher efficiency value when choosing to stay in this AMEC forever. Therefore, within this AMEC, we perturb the initial policy by μ_{sur} slightly to achieve ϵ-optimality and the surveillance task. Note that, since we perturbed each AMEC each containing recurrent states to each ϵ-optimality, the overall perturbed policy $\mu^\star$ is still ϵ-optimal.

Remark 3: In fact, for each recurrent class in MC $\mathcal{M}^{\mu_K^\star}$, we can first check if it already contains a target state in B. If so, then we can skip the perturbation procedure in lines 8-11, and the resulting policy within the associated AMEC will actually be optimal rather than ϵ-optimal.

C. Properties Analysis and Correctness

We conclude this section by formally analyzing the properties of the proposed algorithm.

Still, for $i = 1, \ldots, n$, we denote by $J^{\mu_{opt}^i}$ the optimal efficiency value one can achieve for AMEC $(\mathcal{S}_i, \mathcal{A}_i)$ and define

$$\begin{aligned} J_{max} &= \max\{J^{\mu_{opt}^1}, J^{\mu_{opt}^2}, \ldots, J^{\mu_{opt}^n}\} \\ J_{min} &= \min\{J^{\mu_{opt}^1}, J^{\mu_{opt}^2}, \ldots, J^{\mu_{opt}^n}\} \\ p_{min} &= \min\{P_{s,a,t} \mid s, t \in S, a \in A(s), P_{s,a,t} > 0\}. \end{aligned}$$

The following result shows that, by selecting K to be sufficiently small, the solution to the long-run average reward maximization problem w.r.t. reward function $\mathtt{R}_K$ indeed achieves the supremum efficiency value among all policies in $\Pi_{\mathcal{M}}^B$.

Proposition 4: If K is selected such that

$$K \leq -\frac{1}{p_{min}}(J_{max} - J_{min}), \tag{20}$$

then we have $W(s_0, \mathtt{R}_K, \Pi_{\mathcal{M}}) = J(s_0, \mathtt{R}, \mathtt{C}, \Pi_{\mathcal{M}}^B)$.

Based on the above criterion, we can finally establish the correctness result of the synthesis procedure for the general case of non-communicating MDPs.

Theorem 2: If K is selected such that Equation (20) holds, then Algorithm 1 correctly solves Problem 1.

VI. Conclusion

In this paper, we addressed the challenge of maximizing the long-run efficiency of control policies for Markov Decision Processes, which are characterized by the reward-to-cost ratio, while achieving the surveillance task by visiting target states infinitely often w.p.1. Our result showed that, by exploring stationary policies, one can achieve ϵ-optimality for any threshold value ϵ. Our approach was based on the perturbation analysis technique originally developed for the classical long-run average reward optimization problem. Here, we extended the perturbation analysis technique to the case of long-run efficiency optimization and derived a general formula. Our work not only extended the theory of perturbation analysis but also illustrated its conceptual simplicity and effectiveness in solving MDPs with both qualitative and quantitative tasks.

References

[1] R. N. Haksar and M. Schwager, "Constrained control of large graph-based MDPs under measurement uncertainty," *IEEE Transactions on Automatic Control*, vol. 11, pp. 6605–6620, 2023.

[2] N. Li, A. Girard, and I. Kolmanovsky, "Stochastic predictive control for partially observable Markov decision processes with time-joint chance constraints and application to autonomous vehicle control," *Journal of Dynamic Systems, Measurement, and Control*, vol. 141, no. 7, p. 071007, 2019.

[3] L. Paull, M. Seto, J. J. Leonard, and H. Li, "Probabilistic cooperative mobile robot area coverage and its application to autonomous seabed mapping," *The International Journal of Robotics Research*, vol. 37, no. 1, pp. 21–45, 2018.

[4] M. Luckcuck, M. Farrell, L. A. Dennis, C. Dixon, and M. Fisher, "Formal specification and verification of autonomous robotic systems: A survey," *ACM Computing Surveys*, vol. 52, no. 5, pp. 1–41, 2019.

[5] X. Yin, B. Gao, and X. Yu, "Formal Synthesis of Controllers for Safety-Critical Autonomous Systems: Developments and Challenges," *Annual Reviews in Control*, p. 100940, 2024.

[6] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA: John Wiley & Sons, Inc., 1st ed., 1994.

[7] R. Bloem, K. Chatterjee, K. Greimel, T. A. Henzinger, G. Hofferek, B. Jobstmann, B. Könighofer, and R. Könighofer, "Synthesizing robust systems," *Acta Informatica*, vol. 51, pp. 193–220, 2014.

[8] C. Von Essen, B. Jobstmann, D. Parker, and R. Varshneya, "Synthesizing efficient systems in probabilistic environments," *Acta Informatica*, vol. 53, pp. 425–457, 2016.

[9] X. Ding, S. L. Smith, C. Belta, and D. Rus, "Optimal control of Markov decision processes with linear temporal logic constraints," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1244–1257, 2014.

[10] M. Guo and M. M. Zavlanos, "Probabilistic motion planning under temporal tasks and soft constraints," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4051–4066, 2018.

[11] L. Niu and A. Clark, "Optimal secure control with linear temporal logic constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 6, pp. 2434–2449, 2019.

[12] M. Cai, M. Hasanbeig, S. Xiao, A. Abate, and Z. Kan, "Modular deep reinforcement learning for continuous motion planning with temporal logic," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7973–7980, 2021.

[13] C. Voloshin, H. Le, S. Chaudhuri, and Y. Yue, "Policy optimization with linear temporal logic constraints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17690–17702, 2022.

[14] S. L. Smith, J. Tůmová, C. Belta, and D. Rus, "Optimal path planning for surveillance with temporal-logic constraints," *The International Journal of Robotics Research*, vol. 30, no. 14, pp. 1695–1708, 2011.

[15] Y. Kantaros and M. M. Zavlanos, "STyLuS*: A Temporal Logic Optimal Control Synthesis Algorithm for Large-Scale Multi-Robot Systems," *The International Journal of Robotics Research*, vol. 39, no. 7, pp. 812–836, 2020.

[16] Y. Chen, S. Li, and X. Yin, "Entropy rate maximization of Markov decision processes for surveillance tasks," in *IFAC World Congress*, pp. 4601–4607, 2023.

[17] Y. Chen, S. Yang, R. Mangharam, and X. Yin, "You don't know when i will arrive: Unpredictable controller synthesis for temporal logic tasks," in *IFAC World Congress*, pp. 3967–3973, 2023.

[18] K. Chatterjee, T. A. Henzinger, B. Jobstmann, and R. Singh, "Measuring and synthesizing systems in probabilistic environments," *Journal of the ACM*, vol. 62, no. 1, pp. 1–34, 2015.

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

[20] X.-R. Cao, "The relations among potentials, perturbation analysis, and Markov decision processes," *Discrete Event Dynamic Systems*, vol. 8, pp. 71–87, 1998.

[21] X.-R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer, 2007.

[22] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*. Springer, 2008.

[23] S. Zionts, "Programming with linear fractional functionals," *Naval Research Logistics Quarterly*, vol. 15, no. 3, pp. 449–451, 1968.