LOGO

On Minimizing Total Discounted Cost in MDPs Subject to Reachability Constraints

Yagiz Savas, Christos K. Verginis, Michael Hibbard, and Ufuk Topcu

Abstract— We study the synthesis of a policy in a Markov decision process (MDP) following which an agent reaches a target state in the MDP while minimizing its total discounted cost. The problem combines a reachability criterion with a discounted cost criterion and naturally expresses the completion of a task with probabilistic guarantees and optimal transient performance. We first establish that an optimal policy for the considered formulation may not exist but that there always exists a near-optimal stationary policy. We additionally provide a necessary and sufficient condition for the existence of an optimal policy. We then restrict our attention to stationary deterministic policies and show that the decision problem associated with the synthesis of an optimal stationary deterministic policy is NP-complete. Finally, we provide an exact algorithm based on mixed-integer linear programming and propose an efficient approximation algorithm based on linear programming for the synthesis of an optimal stationary deterministic policy.

Index Terms— Markov decision processes, discounting, reachability, optimization.

I. INTRODUCTION

Markov decision processes (MDPs) provide a framework to model the behavior of an agent, e.g., humans or autonomous robots, operating under uncertainty [1]. A typical objective in an MDP is to synthesize a policy under which the agent reaches a set of target states with *maximum probability* [2]. Such a reachability objective may express, e.g., the completion of a surveillance mission in robotics applications [3], the achievement of a drug concentration level in blood in healthcare applications [4], and the alignment of a portfolio with investors' preferences in finance applications [5].

In many planning problems, achieving a desired transient behavior is as important as the task completion. A widely used performance criterion in MDPs is the total *discounted* cost accumulated by the agent along its trajectories [1], [6]. By associating the agent's actions with *nonnegative* costs and discounting the costs incurred in the future, such a criterion naturally expresses the agent's short- and long-term considerations. Discounting may represent, e.g., the importance of early detection of an intruder in robotic applications [7] or the agent's opportunity cost in healthcare [4] applications.

In this paper, we present a comprehensive analysis for the problem of synthesizing a policy under which an agent reaches a desired set of target states with maximum probability while minimizing its total discounted cost. Such a policy ensures the completion of a task with probabilistic guarantees. Moreover, when the task can be completed at a minimum total cost by following multiple trajectories, the synthesized policy allows the agent to decide on the ordering of the events happening until completion thanks to discounting.

In the literature, several problem formulations have been proposed to synthesize policies satisfying multiple criteria. Extensively studied formulations include the so-called constrained MDP problems [6], stochastic shortest path (SSP) problems [8], and multi-objective model checking problems [9]. The constrained MDP problem associates the agent's actions with multiple costs that are discounted

with either the same [6], [10] or different [11]–[14] discount factors. In general, probabilistic reachability objectives cannot be expressed as total discounted criteria without making restrictive assumptions on the MDP structure [15]. In the SSP [8], [16], [17] and multi-objective model checking problems [9], [18], one aims to synthesize a policy that satisfies reachability constraints while minimizing the agent's total *undiscounted* cost. Due to lack of discounting, however, the agent cannot adjust the importance of its short- and long-term considerations. Overall, existing methods for multi-objective planning fail when one needs to synthesize policies that minimize the total discounted cost while satisfying a task with probabilistic guarantees.

Similar problems have also been studied in the reinforcement learning (RL) literature in the context of multi-objective RL [19]–[21]. These works concern the learning of a policy in systems with unknown dynamics, e.g., transition functions, or cost functions. In this paper, we consider a planning problem in systems with known dynamics and costs; however, the analysis contained within this paper can serve as a foundation for future multi-objective RL algorithms.

It is known that an optimal policy always exists for the previously described problem formulations [6], [8], [9], [22]. However, to the best of our knowledge, the formulation considered in this paper has never been studied in the literature; hence, the existence of optimal solutions is an open problem. Our first contribution is establishing that, in an MDP, a policy that minimizes the total discounted cost among the ones that maximize the probability of reaching a target state may not exist. This result illustrates a fundamental difference of the problem considered in this paper from the formulations considered in the literature.

Since optimal policies do not exist for the general case, it is critical to verify whether there exists an optimal policy for a given problem instance. As the second contribution, we present an efficiently verifiable necessary and sufficient condition for the existence of optimal policies in a given problem instance. When there are no optimal policies, one typically searches for near-optimal policies. As the third contribution, we show that, for any positive constant ϵ, there exists an ϵ-optimal *stationary* policy that can be synthesized efficiently.

In many applications, it is desirable to generate stationary deterministic policies due to their low computational requirements. It is known [12] that, in general, the synthesis of such policies is NP-hard for constrained MDPs in which multiple costs are discounted with the same discount factor. Since we consider a problem involving a probabilistic constraint, however, the existing complexity results on constrained MDPs do not apply to the problem considered in this paper. As the fourth contribution, we establish that it is NP-complete to decide whether there exists a *stationary deterministic* policy that maximizes the probability of reaching a set of target states while attaining a total discounted cost below a desired threshold. Motivated by this complexity result, we synthesize an optimal stationary deterministic policy by formulating a mixed-integer linear program (MILP) which is an extension of the algorithms developed in [11], [23]. For small problem instances, one can compute optimal solutions to MILPs using off-the-shelf solvers, e.g., [24]. However, for large problem instances, MILP formulations become intractable. To remedy this limitation, our fifth contribution is the development of an approximation algorithm based on linear programming that efficiently synthesizes stationary

deterministic policies with theoretical suboptimality guarantees.

In numerical simulations, we present an application of the studied formulation to motion planning. Specifically, we consider an agent that aims to deliver a package to a certain location while minimizing the risk of being attacked by adversaries in the environment. We illustrate that the proposed approximation algorithm generates agent trajectories that guarantee reachability to the desired location while visiting the minimum number of risky regions in the environment.

II. PRELIMINARIES

Notation: The sets of natural and real numbers are denoted by $\mathbb{N}$ and $\mathbb{R}$, respectively, while the set of nonnegative reals is denoted by $\mathbb{R}_{\geq 0}$. Finally, $|S|$ denotes the cardinality of a set S.

Definition 1: A *Markov decision process* (MDP) is a tuple $\mathcal{M}:=(S, \alpha, \mathcal{A}, \mathcal{P})$ where S is a finite set of states, $\alpha:S\to[0,1]$ is the initial distribution such that $\sum_{s\in S} \alpha(s)=1$, $\mathcal{A}$ is a finite set of actions, and $\mathcal{P}:S\times\mathcal{A}\times S\to[0,1]$ is a transition function such that $\sum_{s'\in S} \mathcal{P}(s,a,s')=1$ for all $s\in S$ and $a\in\mathcal{A}(s)$, where $\mathcal{A}(s)\subseteq\mathcal{A}$ denotes the set of available actions in $s\in S$.

We denote the transition probability $\mathcal{P}(s,a,s')$ by $\mathcal{P}_{s,a,s'}$. A state $s\in S$ is *absorbing* if $\mathcal{P}_{s,a,s}=1$ for all $a\in\mathcal{A}(s)$.

Definition 2: For an MDP $\mathcal{M}$, a *policy* $\pi:=(d_1, d_2, d_3, \dots)$ is a sequence where, for each $t\in\mathbb{N}$, $d_t:S\times\mathcal{A}\to[0,1]$ is a mapping such that $\sum_{a\in\mathcal{A}(s)} d_t(s,a)=1$ for all $s\in S$. A *stationary* policy is a policy of the form $\pi=(d_1, d_1, d_1, \dots)$. A stationary *deterministic* policy is a stationary policy where, for each $s\in S$, $d_1(s,a)=1$ for some $a\in\mathcal{A}(s)$. We denote the set of all policies, all stationary policies, and all stationary deterministic policies by $\Pi(\mathcal{M})$, $\Pi^S(\mathcal{M})$, and $\Pi^{SD}(\mathcal{M})$, respectively.

A policy $\pi\in\Pi(\mathcal{M})$ is traditionally referred to as a Markovian policy [1]. Although it is possible to consider more general policy classes, the consideration of the set $\Pi(\mathcal{M})$ is without loss of generality for the purposes of this paper due to Theorem 5.5.1 in [1]. For notational simplicity, we denote the probability of taking an action $a\in\mathcal{A}$ in a state $s\in S$ under a *stationary* policy π by $\pi(s,a)$.

A *path* is a sequence $\varrho^\pi=s_1 a_1 s_2 a_2 s_3 \dots$ of states and actions generated under π which satisfies $\mathcal{P}_{s_t,a_t,s_{t+1}}>0$ for all $t\in\mathbb{N}$. We define the set of all paths in $\mathcal{M}$ with initial distribution α under π by $Paths_{\mathcal{M}}^{\pi,\alpha}$ and use the standard probability measure over the set $Paths_{\mathcal{M}}^{\pi,\alpha}$ [25]. Let $\varrho^\pi[t]:=s_t$ denote the state visited at the t-th step along ϱ^π. We define

$$\Pr_{\mathcal{M}}^{\pi,\alpha}(Reach[B]) := \Pr\{\varrho^\pi \in Paths_{\mathcal{M}}^{\pi,\alpha} : \exists t \in \mathbb{N}, \varrho^\pi[t] \in B\}$$

as the probability with which the paths generated in $\mathcal{M}$ with initial distribution α under π reach the set $B\subseteq S$.

III. PROBLEM STATEMENT

We consider an agent that aims to reach a set of target states with maximum probability while minimizing the expected total discounted cost it accumulates along its path. Formally, let $c:S\times\mathcal{A}\to\mathbb{R}_{\geq 0}$ be a *non-negative* cost function, $\beta\in(0,1)$ be a discount factor, $B\subseteq S$ be a set of *absorbing* target states, and $J:\Pi(\mathcal{M})\to\mathbb{R}_{\geq 0}$ be a cost-to-go function such that

$$J(\pi) := \mathbb{E}_\alpha^\pi\left[\sum_{t=1}^\infty \beta^{t-1} c(s_t, a_t)\right], \tag{1}$$

where the expectation is taken over the paths generated in $\mathcal{M}$ with initial distribution α under π. Moreover, let $\Xi(\mathcal{M}, B)\subseteq\Pi(\mathcal{M})$ denote the set of policies under which the paths in $\mathcal{M}$ reach the target set B with maximum probability, i.e., $\overline{\pi}\in\Xi(\mathcal{M}, B)$ if and only if

$$\overline{\pi} \in \arg\max_{\pi\in\Pi(\mathcal{M})} \Pr_{\mathcal{M}}^{\pi,\alpha}(Reach[B]). \tag{2}$$
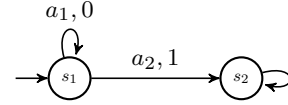

Fig. 1: An MDP example to illustrate the non-existence of optimal policies. The initial distribution is α such that $\alpha(s_1)=1$ and $\alpha(s_2)=0$, and $B=\{s_2\}$. The tuples (a,c) indicate the action a and the cost c.

Existence of the maximum in (2) is due to Lemma 10.102 in [25].

In this paper, we study the synthesis of a policy $\pi^\star$ such that

$$\pi^\star \in \arg\inf_{\pi\in\Xi(\mathcal{M},B)} J(\pi). \tag{3}$$

In the following sections, we analyze the problem in (3) and present efficient algorithms to synthesize policies that either exactly or approximately satisfy the condition in (3).

IV. NON-EXISTENCE OF OPTIMAL POLICIES

We now present a numerical example to illustrate that an optimal policy satisfying the condition in (3) may not exist. This example demonstrates the significant difference of the problem considered in this work from the traditional constrained MDP problems, e.g., [6], [11], [14], for which an optimal policy always exists.

Consider the example shown in Fig. 1. The agent starts from the state s_1 and aims to reach the state s_2 with probability one, i.e., $B=\{s_2\}$, while minimizing its total discounted cost. In this example, we can focus on stationary policies without loss of generality. Suppose that the agent follows the stationary policy $\pi\in\Pi^S(\mathcal{M})$ such that $\pi(s_1,a_1)=1-\delta$ and $\pi(s_1,a_2)=\delta$ where $\delta\in[0,1]$. For $\delta\in(0,1]$, the agent reaches the state s_2 with probability one under π. However, if $\delta=0$, we have $\pi\notin\Xi(\mathcal{M}, B)$. Hence, the set $\Xi(\mathcal{M}, B)\cap\Pi^S(\mathcal{M})$ of feasible stationary policies is

$$\{\pi \in \Pi^S(\mathcal{M}) \mid \pi(s_1, a_1) = 1-\delta, \ \pi(s_1, a_2) = \delta, \ \delta \in (0,1]\}.$$

For a given stationary policy $\pi\in\Xi(\mathcal{M}, B)\cap\Pi^S(\mathcal{M})$, we have $J(\pi)=\frac{\delta}{1-\beta(1-\delta)}$. Note that $\lim_{\delta\to 0}\frac{\delta}{1-\beta(1-\delta)}=0$, which implies that $\inf_{\pi\in\Xi(\mathcal{M},B)} J(\pi) = 0$.

The *only* policy that attains the infimum is the *stationary* policy $\overline{\pi}$ such that $\overline{\pi}(s_1,a_1)=1$ and $\overline{\pi}(s_1,a_2)=0$ since any other policy incurs a non-zero cost by taking the action a_2 with non-zero probability. Note that the policy $\overline{\pi}$ is not in the feasible policy space as it reaches the set B with probability zero. Since the infimum is not attainable by any feasible policy, we conclude that there exists no optimal policy.

One may be tempted to think that the existence of a zero cost action a_1 is the reason for not having an optimal policy in this example. It can be shown that, even if we assign a positive cost for the action a_1, e.g., $c(s_1,a_1)=0.1$, an optimal policy still does not exist so long as we choose a small discount factor, e.g., $\beta<0.1$. An optimal policy does not exist in this example because the agent exploits the discounting in the costs and stays in the initial state for as long as possible before reaching the target state. By doing so, the agent ensures that its cost-to-go approaches zero while still satisfying the reachability constraint asymptotically. Such a behavior is specific to the problem in (3), which involves a discounted criterion in the objective and a probabilistic criterion in the constraint, and does not arise in existing MDP formulations.

V. EXISTENCE OF NEAR-OPTIMAL POLICIES

In the previous section, we showed that an optimal policy solving the problem in (3) may not exist. Here, we show that there always exists a near-optimal policy which can be synthesized efficiently.

Definition 3: For a given constant $\epsilon>0$, a policy $\overline{\pi}\in\Xi(\mathcal{M},B)$ is said to be an ϵ-optimal policy for the problem in (3) if $J(\overline{\pi}) \leq \inf_{\pi\in\Xi(\mathcal{M},B)} J(\pi) + \epsilon$.

We establish the existence of ϵ-optimal policies for the problem in (3) in three steps. First, we *clean up* the MDP $\mathcal{M}$ by removing the actions $a\in\mathcal{A}(s)$, from each $s\in S$, that are guaranteed to yield infeasible policies $\pi\in\Pi(\mathcal{M})\backslash\Xi(\mathcal{M},B)$. Second, we synthesize a policy on the resulting MDP under which the total discounted cost is minimized in the absence of the reachability constraint. Finally, we perturb the synthesized policy to obtain a policy that satisfies the reachability constraint and is ϵ-optimal for the problem in (3).

A. Cleaning up the MDP

For a given MDP $\mathcal{M}$, we first partition the set S of states into three disjoint sets. Let $B\subseteq S$ be the set of target states, and S_0 be the set of states that have zero probability of reaching the target set, i.e., $S_0=\{s \in S \mid \Pr_{\mathcal{M}}^{\pi,\alpha(s)=1}(Reach[B]) = 0, \ \forall \pi \in \Pi(\mathcal{M})\}$. Finally, we let $S_r:=S\backslash(B\cup S_0)$. These sets can be computed efficiently using graph search algorithms [25].

Let $\mathbf{x}:=(x_s)_{s\in S}\in\mathbb{R}^{|S|}$ be a vector such that

$$x_s := \begin{cases} 1 & \text{if } s \in B \\ 0 & \text{if } s \in S_0 \\ \max_{a\in\mathcal{A}(s)} \left\{ \sum_{s'\in S} \mathcal{P}_{s,a,s'} x_{s'} \right\} & \text{otherwise.} \end{cases} \quad (4)$$

It is known [25, Chapter 10] that each element x_s of $\mathbf{x}$ corresponds to the maximum probability of reaching B from $s\in S$. Moreover, $\mathbf{x}$ can be efficiently computed via linear programming [25].

Now, let $S_\to^\pi\subseteq S$ be the set of states that are reachable under the policy $\pi\in\Pi(\mathcal{M})$, i.e.,

$$S_\to^\pi := \{s \in S \mid \Pr_{\mathcal{M}}^{\pi,\alpha}(Reach[s]) > 0\}.$$

The following result, which is due to Theorem 10.100 in [25], characterizes a necessary condition for a policy $\pi\in\Pi(\mathcal{M})$ to maximize the probability of reaching the target set B.

Proposition 1: [25] If $\pi\in\Xi(\mathcal{M},B)$, then, for all $s\in S_\to^\pi\cap S_r$,

$$x_s = \sum_{s'\in S} \mathcal{P}_{s,a,s'} x_{s'} \quad (5)$$

for all $a\in\mathcal{A}(s)$ satisfying $d_t(s,a)>0$ for some $t\in\mathbb{N}$.

Equation (5) constitutes a necessary (but not sufficient) condition for the feasibility of a policy $\pi\in\Pi(\mathcal{M})$ for the problem in (3). Therefore, without loss of generality, we can remove from the MDP all actions that violate the equality in (5). Specifically, given an MDP $\mathcal{M}$, we obtain the *cleaned-up MDP* $\mathcal{M}'$ by removing all actions $a\in\mathcal{A}(s)\backslash\mathcal{A}_{\max}(s)$ from each state $s\in S_r$, where

$$\mathcal{A}_{\max}(s) := \left\{ a \in \mathcal{A}(s) \ \middle| \ x_s = \sum_{s'\in S} \mathcal{P}_{s,a,s'} x_{s'} \right\}. \quad (6)$$

B. Minimizing the total cost on the cleaned-up MDP

On the cleaned-up MDP $\mathcal{M}'$, we synthesize a stationary deterministic policy $\widetilde{\pi}\in\Pi^{SD}(\mathcal{M}')$ such that

$$\widetilde{\pi} \in \arg\min_{\pi\in\Pi(\mathcal{M}')} J(\pi). \quad (7)$$

The existence of a stationary deterministic policy $\widetilde{\pi}$ satisfying the condition in (7) follows from the fact that the problem in (7) is an *unconstrained* discounted MDP problem [1]. The policy $\widetilde{\pi}$ satisfies

$$J(\widetilde{\pi}) \leq \inf_{\pi\in\Xi(\mathcal{M},B)} J(\pi) \quad (8)$$

since $\Xi(\mathcal{M},B)\subseteq\Pi(\mathcal{M}')$. Therefore, if $\widetilde{\pi}\in\Xi(\mathcal{M},B)$, then $\widetilde{\pi}$ is an optimal policy for the problem in (3). However, in general, we have $\widetilde{\pi}\notin\Xi(\mathcal{M},B)$, in which case we need to perturb the policy $\widetilde{\pi}$ in a certain way to obtain an ϵ-optimal solution to the problem in (3).

C. Perturbations to maximize reachability

Given the stationary policy $\widetilde{\pi}$, for each state $s\in S$, let $\mathcal{A}_{act}(s)$ be the set of actions that are taken by a non-zero probability, i.e.,

$$\mathcal{A}_{act}(s) := \{a \in \mathcal{A}_{\max}(s) \mid \widetilde{\pi}(s,a) > 0\}.$$

Similarly, let $\mathcal{A}_{pass}(s):=\mathcal{A}_{\max}(s)\backslash\mathcal{A}_{act}(s)$. We define the perturbed stationary policy $\widetilde{\pi}'\in\Pi^S(\mathcal{M}')$ as

$$\widetilde{\pi}'(s,a) := \begin{cases} \epsilon' & \text{if } a \in \mathcal{A}_{pass}(s) \\ \widetilde{\pi}(s,a) - \epsilon'\frac{|\mathcal{A}_{pass}(s)|}{|\mathcal{A}_{act}(s)|} & \text{otherwise,} \end{cases} \quad (9)$$

where $\epsilon'>0$ is a sufficiently small constant such that $\widetilde{\pi}'(s,a)\geq0$ for all $s\in S$ and $a\in\mathcal{A}$. Note that the policy $\widetilde{\pi}'$ is well-defined since $\sum_{a\in\mathcal{A}_{\max}(s)} \widetilde{\pi}'(s,a)=1$ for all $s\in S$.

Definition 4: For an MDP $\mathcal{M}$ and a stationary policy $\pi\in\Pi^S(\mathcal{M})$, an *induced Markov chain* (MC) $\mathcal{M}_\pi=(S,\mathcal{P}^\pi)$ is a tuple where the transition function $\mathcal{P}^\pi:S\times S\to[0,1]$ is such that, for all $s,s'\in S$,

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

We now prove that $\widetilde{\pi}'$ maximizes the probability of reaching B.

Lemma 1: It holds that $\widetilde{\pi}'\in\Xi(\mathcal{M},B)$.

Proof: The policy $\widetilde{\pi}'$ is such that, for all $s\in S_\to^{\widetilde{\pi}'}$, all actions $a\in\mathcal{A}_{\max}(s)$ are taken with a non-zero probability. Therefore, in the induced MC $\mathcal{M}'_{\widetilde{\pi}'}$, there exists a path to B from each state $s\in S_\to^{\widetilde{\pi}'}\cap S_r$. Then, it follows from Lemma 1 in [17] that there exists a constant $N\in\mathbb{N}$ such that, for all $M\geq N$, $\varrho^{\widetilde{\pi}'}[M]\notin S_\to^{\widetilde{\pi}'}\cap S_r$. In other words, the agent eventually leaves the states $s\in S_\to^{\widetilde{\pi}'}\cap S_r$ with probability 1. By the construction of the cleaned-up MDP $\mathcal{M}'$, the agent reaches the set B with maximum probability under any policy leaving the set $S_\to^{\widetilde{\pi}'}\cap S_r$ with probability 1. Hence, we have $\widetilde{\pi}'\in\Xi(\mathcal{M}',B)$. Then, the result follows as $\Xi(\mathcal{M}',B)\subseteq\Xi(\mathcal{M},B)$. $\square$

We now show that, for an appropriately chosen constant ϵ', the stationary policy $\widetilde{\pi}'$ constitutes an ϵ-optimal policy to the problem in (3). With an abuse of notation, for a given $\pi\in\Pi^S(\mathcal{M}')$, let $\mathcal{P}^\pi\in\mathbb{R}^{|S|\times|S|}$ be the transition matrix of the induced MC $\mathcal{M}_\pi$. Let the vector $\boldsymbol{\alpha}=(\alpha(s))_{s\in S}\in\mathbb{R}^{|S|}$ be the initial state distribution. Finally, let $\boldsymbol{c}^\pi:=(c_s^\pi)_{s\in S}\in\mathbb{R}^{|S|}$ be a vector such that

$$c_s^\pi := \sum_{a\in\mathcal{A}_{\max}(s)} \pi(s,a)c(s,a).$$

It is known [1], [6] that, for any $\pi\in\Pi^S(\mathcal{M}')$, we have

$$J(\pi) = \boldsymbol{\alpha}^T(\mathbb{I} - \beta\mathcal{P}^\pi)^{-1}\boldsymbol{c}^\pi \quad (10)$$

where $(\cdot)^T$ denotes the transpose operation, and $\mathbb{I}\in\mathbb{R}^{|S|\times|S|}$ is the identity matrix. Moreover, it follows from (9) that

$$\mathcal{P}^{\widetilde{\pi}'} = \mathcal{P}^{\widetilde{\pi}} + \epsilon' M \quad \text{and} \quad \boldsymbol{c}^{\widetilde{\pi}'} = \boldsymbol{c}^{\widetilde{\pi}} + \epsilon'\boldsymbol{v}$$

where $M\in\mathbb{R}^{|S|\times|S|}$ and $\boldsymbol{v}:=(v_s)_{s\in S}\in\mathbb{R}^{|S|}$ are, respectively, the perturbation matrix and the perturbation vector satisfying

$$M(s,s') := \sum_{a\in\mathcal{A}_{pass}(s)} \mathcal{P}_{s,a,s'} - \frac{|\mathcal{A}_{pass}(s)|}{|\mathcal{A}_{act}(s)|} \sum_{a\in\mathcal{A}_{act}(s)} \mathcal{P}_{s,a,s'},$$

$$v_s := \sum_{a\in\mathcal{A}_{pass}(s)} c(s,a) - \frac{|\mathcal{A}_{pass}(s)|}{|\mathcal{A}_{act}(s)|} \sum_{a\in\mathcal{A}_{act}(s)} c(s,a).$$

Theorem 1: For any given $\epsilon>0$, the policy $\widetilde{\pi}'$ defined in (9) is an ϵ-optimal policy for the problem in (3) if $\epsilon'>0$ is chosen such that

$$\epsilon' \leq \frac{\epsilon}{\gamma_1 + \gamma_2} \tag{11}$$

where $\gamma_1 := \beta \boldsymbol{\alpha}^T (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}})^{-1} M (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1} \boldsymbol{c}^{\widetilde{\pi}}$ and $\gamma_2 := \boldsymbol{\alpha}^T (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1} \boldsymbol{v}$.

Proof: We show in Lemma 1 that the policy $\widetilde{\pi}'$ is feasible for the problem in (3). Here, we show that the cost-to-go $J(\widetilde{\pi}')$ is at most ϵ larger than the minimum achievable one.

Using (10) and the definition of γ_2, we have

$$J(\widetilde{\pi}') = \boldsymbol{\alpha}^T (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1} \boldsymbol{c}^{\widetilde{\pi}'} \tag{12a}$$

$$= \boldsymbol{\alpha}^T (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1} (\boldsymbol{c}^{\widetilde{\pi}} + \epsilon' \boldsymbol{v}) \tag{12b}$$

$$= \boldsymbol{\alpha}^T (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1} \boldsymbol{c}^{\widetilde{\pi}} + \epsilon' \gamma_2. \tag{12c}$$

It follows from equation (26) in [26] that

$$(\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1} = (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}} - \epsilon' \beta M)^{-1} \tag{13a}$$

$$= (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}})^{-1}$$
$$+ \epsilon' \beta (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}})^{-1} M (\mathbb{I} - \beta \mathcal{P}^{\widetilde{\pi}'})^{-1}. \tag{13b}$$

Plugging (13b) into (12c), we obtain $J(\widetilde{\pi}')=J(\widetilde{\pi})+\epsilon'(\gamma_1+\gamma_2)$. Using (11) and the definition of $\widetilde{\pi}$ given in (7), we conclude that

$$J(\widetilde{\pi}') \leq \min_{\pi \in \Pi(\mathcal{M}')} J(\pi) + \epsilon. \tag{14}$$

Then, the result follows as $\Xi(\mathcal{M}, B) \subseteq \Pi(\mathcal{M}')$. $\square$

We conclude this section by summarizing the three main steps of the efficient synthesis of an ϵ-optimal stationary policy for the problem in (3). First, obtain the cleaned-up MDP $\mathcal{M}'$ by removing all actions $a \in \mathcal{A}(s) \backslash \mathcal{A}_{\max}(s)$ where $\mathcal{A}_{\max}(s)$ is as defined in (6). Second, synthesize a stationary deterministic policy $\widetilde{\pi} \in \Pi^{SD}(\mathcal{M}')$ that satisfies (7) via linear programming. Finally, obtain an ϵ-optimal policy $\widetilde{\pi}'$ given in (9) by choosing ϵ' as shown in (11).

VI. A Necessary and Sufficient Condition for the Existence of Optimal Policies

In the previous sections, we showed that an optimal policy solving the problem in (3) may not exist, but an ϵ-optimal stationary policy can be synthesized efficiently. To complete the analysis, we now provide an efficiently verifiable necessary and sufficient condition for the existence of an optimal policy that solves the problem in (3).

The following result is an immediate consequence of the inequality in (14), which shows that the optimal value of the problem in (3) coincides with the optimal value of the problem in (7).

Corollary 1: The following equality holds:

$$\inf_{\pi \in \Xi(\mathcal{M}, B)} J(\pi) = \min_{\pi \in \Pi(\mathcal{M}')} J(\pi). \tag{15}$$

Theorem 2: There exists an optimal policy $\pi^\star \in \Pi(\mathcal{M})$ that satisfies the condition in (3) if and only if there exists a policy $\widetilde{\pi} \in \Xi(\mathcal{M}', B)$ that satisfies the condition in (7).

Proof: ($\Leftarrow$) Suppose that there exists a policy $\widetilde{\pi} \in \Xi(\mathcal{M}', B)$ that satisfies the condition in (7). Then, $\widetilde{\pi}$ is an optimal policy satisfying the condition in (3) due to (8).

($\Rightarrow$) Suppose that there exists an optimal policy $\pi^\star \in \Pi(\mathcal{M})$ that satisfies the condition in (3). Since $\pi^\star$ is a feasible solution to the problem in (3), it satisfies $\pi^\star \in \Xi(\mathcal{M}', B)$. Additionally, it follows from (15) that $\pi^\star$ satisfies the condition in (7). $\square$

Theorem 2 establishes that, for a given MDP, the existence of an optimal policy solving the problem in (3) can be verified in three steps as follows. First, construct the cleaned-up MDP $\mathcal{M}'$ as described in Section V-A. Second, find all policies that minimize the total discounted cost on the cleaned-up MDP $\mathcal{M}'$, i.e., policies that satisfy the condition in (7). Finally, check whether any of these policies belong to the set $\Xi(\mathcal{M}', B)$. In Section V-A, we show that the first step can be performed efficiently. In what follows, we show that the second and the third steps of the above procedure can also be performed efficiently.

Let $\mathbf{y}:=(y_s)_{s \in S} \in \mathbb{R}^{|S|}$ be a vector such that

$$y_s := \min_{a \in \mathcal{A}_{\max}(s)} \left\{ c(s, a) + \beta \sum_{s' \in S} \mathcal{P}_{s,a,s'} y_{s'} \right\}.$$

It is known [1] that each element y_s of $\mathbf{y}$ corresponds to the minimum total discounted cost accumulated along the paths that start from the state $s \in S$. The vector $\mathbf{y}$ can be efficiently computed via value iteration or linear programming [1]. The following result provides a necessary and sufficient condition for a policy to minimize the total discounted cost in an MDP in the absence of constraints.

Proposition 2: [1, Chapter 6] A policy $\widetilde{\pi}=(\widetilde{d}_1, \widetilde{d}_2, \widetilde{d}_3, \ldots) \in \Pi(\mathcal{M}')$ satisfies the condition in (7) if and only if, for all $s \in S^{\widetilde{\pi}}_{\rightarrow}$,

$$y_s = c(s, a) + \beta \sum_{s' \in S} \mathcal{P}_{s,a,s'} y_{s'} \tag{16}$$

for all $a \in \mathcal{A}_{\max}(s)$ satisfying $\widetilde{d}_t(s, a)>0$ for some $t \in \mathbb{N}$.

Given a cleaned-up MDP $\mathcal{M}'$, we obtain the *modified MDP* $\overline{\mathcal{M}}'$ by removing all actions $a \in \mathcal{A}_{\max}(s) \backslash \mathcal{A}_{opt}(s)$ from each $s \in S$, where

$$\mathcal{A}_{opt}(s) := \left\{ a \in \mathcal{A}_{\max}(s) \;\middle|\; y_s = c(s, a) + \beta \sum_{s' \in S} \mathcal{P}_{s,a,s'} y_{s'} \right\}.$$

Then, it follows from Proposition 2 that the policies on the modified MDP $\overline{\mathcal{M}}'$, and only them, minimize the total discounted cost on $\mathcal{M}'$.

Finally, we verify whether there exists a policy $\pi \in \Pi(\overline{\mathcal{M}}')$ on the modified MDP such that $\pi \in \Xi(\mathcal{M}', B)$. Let $\overline{\mathbf{x}}:=(\overline{x}_s)_{s \in S} \in \mathbb{R}^{|S|}$ be a vector such that the maximum in (4) is taken over the set $\mathcal{A}_{opt}(s)$ instead of the set $\mathcal{A}(s)$. Note that $\overline{x}_s$ corresponds to the probability of reaching the target set B from the state $s \in S$ in $\overline{\mathcal{M}}'$, which can be computed efficiently via linear programming.

Proposition 3: There exists $\pi \in \Pi(\overline{\mathcal{M}}')$ such that $\pi \in \Xi(\mathcal{M}', B)$ if and only if $\overline{x}_s=x_s$ for all $s \in S$ such that $\alpha(s)>0$.

Proof: ($\Rightarrow$) If $\pi \in \Pi(\overline{\mathcal{M}}')$ satisfies the condition $\pi \in \Xi(\mathcal{M}', B)$, then the result follows from Proposition 1 and the fact that $\overline{x}_s \leq x_s$.

($\Leftarrow$) If $\overline{x}_s=x_s$ for all $s \in S$ such that $\alpha(s)>0$, then a policy $\pi \in \Xi(\mathcal{M}', B)$ can be constructed as explained in Lemma 10.102 of [25]. On the modified MDP, we first remove the actions from each state $s \in S_r$ that do not satisfy the condition $\overline{x}_s=\sum_{s' \in S} \mathcal{P}_{s,a,s'} \overline{x}_{s'}$. On the graph corresponding to the resulting MDP, let $T_{\min}(s)$ be the length of the shortest path from the state $s \in S_r$ to the target set B (see Section VII-C for details). We construct $\pi \in \Xi(\mathcal{M}', B)$ by selecting an action in each $s \in S_r$ under which the agent transitions with a nonzero probability to s' such that $T_{\min}(s')<T_{\min}(s)$. $\square$

We conclude this section by noting that, once the existence of an optimal policy solving the problem in (3) is verified using the procedure explained above, one can synthesize such a policy through the procedure described in the proof of Proposition 3.

VII. An Analysis Over Deterministic Policies

In many applications, it is desirable to optimize performance using deterministic policies due to their low computational requirements and ease of implementation in distributed systems [27]. Accordingly,

in this section, we focus on stationary deterministic policies and consider the problem of synthesizing a policy $\pi_D^\star$ such that

$$\pi_D^\star \in \arg \min_{\pi \in \Xi^{SD}(\mathcal{M}, B)} J(\pi) \qquad (17)$$

where the set $\Xi^{SD}(\mathcal{M}, B) \subseteq \Pi^{SD}(\mathcal{M})$ denotes the set of stationary deterministic policies that maximize the probability of reaching the target set B. Specifically, it holds $\pi \in \Xi^{SD}(\mathcal{M}, B)$ if and only if $\pi \in \Pi^{SD}(\mathcal{M})$ satisfies the condition in (2). Since the set $\Xi^{SD}(\mathcal{M}, B)$ is finite, an optimal policy for the problem in (17) always exists.

A. A complexity result

We now show that the synthesis of a policy that satisfies the condition in (17) is, in general, intractable. We remark that the synthesis of a stationary deterministic policy that minimizes the total discounted cost in MDPs subject to total discounted reward constraints is known to be NP-hard [13]. The problem in (17) involves a probabilistic constraint which, in general, cannot be represented as a discounted criterion without changing the feasible policy space. Therefore, none of the existing complexity results on constrained MDPs apply to the problem in (17).

We prove the result by a reduction from the Hamiltonian path problem (HAMPATH) which is known to be NP-complete [28]. Let $G=(V, E)$ be a directed graph (digraph) where V is a finite set of vertices and $E \subseteq V \times V$ is a finite set of edges. For a given digraph G, a *finite path* $v_1 v_2 \ldots v_n$ of length $n \in \mathbb{N}$ from vertex v_1 to v_n is a sequence of vertices such that $(v_k, v_{k+1}) \in E$ for all $1 \le k < n$.

Definition 5: (HAMPATH) Given a digraph $G=(V, E)$ and an origin-destination pair $(o, d) \in V \times V$, decide whether there exists a Hamiltonian path on G, i.e., a finite path from the vertex o to the vertex d that visits each vertex $v \in V$ exactly once.

Theorem 3: For $K \in \mathbb{N}$, deciding whether there exists a policy $\pi \in \Xi^{SD}(\mathcal{M}, B)$ such that $J(\pi) \le K$ is NP-complete.

Proof: The decision problem is in NP since for any given policy $\pi \in \Xi^{SD}(\mathcal{M}, B)$, we can verify whether $J(\pi) \le K$ in polynomial-time using the formula in (10).

To show the NP-hardness, we reduce an arbitrary HAMPATH instance to an instance of the problem in (17).

Given a graph G and a pair (o, d), we construct an instance of the problem in (17) such that the agent starts from o and aims to reach d. We define the MDP $\mathcal{M}$ as follows. The set of states are $S=V$, the initial distribution is $\alpha(o)=1$ and $\alpha(s)=0$ otherwise, and the target set is $B=\{d\}$. For a state $v \in V$, we associate an action $a_{v'}$ for each edge $(v, v') \in E$, i.e., $\mathcal{A}(v)=\{a_{v'} \in \mathcal{A} : (v, v') \in E\}$. The transition function $\mathcal{P}$ is such that $\mathcal{P}_{v, a_{v'}, v'}=1$ if $v \in V \backslash B$, and $\mathcal{P}_{v, a_{v'}, v}=1$ if $v \in B$. Finally, we define the cost function c such that $c(v, a_{v'})=K/\beta^{|V|-1}$ if $v'=d$, and $c(v, a_{v'})=0$ otherwise.

We now show that, there exists a policy $\pi \in \Xi^{SD}(\mathcal{M}, B)$ such that $J(\pi) \le K$ if and only if there exists a Hamiltonian path on the graph G with the origin-destination pair (o, d).

Suppose that there is a Hamiltonian path $v_1 v_2 \ldots v_{|V|}$ where $v_1=o$ and $v_{|V|}=d$. Consider the policy $\pi \in \Pi^{SD}(\mathcal{M})$ such that $\pi(v_k, a_{v_{k+1}})=1$ for all $k \in \mathbb{N}$ such that $k < |V|$. Under the policy π, the agent reaches the target set B with probability one; hence, it holds $\pi \in \Xi^{SD}(\mathcal{M}, B)$. Moreover, the agent reaches the set B in exactly $|V|-1$ steps; hence, it holds that $J(\pi)=K$.

Suppose that there exists a policy $\pi \in \Xi^{SD}(\mathcal{M}, B)$ such that $J(\pi) \le K$. Then, under the policy π, the agent reaches the target set B with probability one in at least $|V|-1$ steps. Moreover, since π is a deterministic policy, the agent visits each state at most once because, otherwise, the probability of reaching the target set B must be zero. The combination of the above arguments imply that the agent reaches

the target set B in $|V|-1$ steps by visiting each state $v \in V \backslash B$ exactly once. By definition, such a path constitutes a Hamiltonian path. $\square$

B. An exact algorithm

In this section, we solve the problem in (17) by formulating it as a mixed-integer linear program (MILP). The MILP formulation is an extension of the results in [11], [23]. In [11], the authors synthesize stationary deterministic policies for MDPs involving multiple discount factors each of which is *strictly* less than one. In [23], an MILP formulation is presented for the synthesis of stationary deterministic policies under which an agent satisfies a temporal logic specification *with probability one* while minimizing its total discounted cost. Here, we formulate an MILP for the synthesis of a policy that reaches a desired set of target states with maximum probability (which is potentially less than one) while minimizing the total discounted cost.

Consider the following MILP:

$$\underset{\lambda_{s,a}^1, \lambda_{s,a}^2, \Delta_{s,a}}{\text{minimize}} \sum_{s \in S} \sum_{a \in \mathcal{A}} c(s, a) \lambda_{s,a}^1 \qquad (18a)$$

subject to:

$$\sum_{a \in \mathcal{A}} \lambda_{s,a}^1 - \beta \sum_{s' \in S} \sum_{a \in \mathcal{A}} \mathcal{P}_{s', a, s} \lambda_{s', a}^1 = \alpha(s), \ \forall s \in S \qquad (18b)$$

$$\sum_{a \in \mathcal{A}} \lambda_{s,a}^2 - \sum_{s' \in S} \sum_{a \in \mathcal{A}} \mathcal{P}_{s', a, s} \lambda_{s', a}^2 = \alpha(s), \quad \forall s \in S_r \qquad (18c)$$

$$\sum_{s \in S_r} \sum_{a \in \mathcal{A}} \lambda_{s,a}^2 r(s, a) = x_\alpha \qquad (18d)$$

$$\lambda_{s,a}^k / M \le \Delta_{s,a}, \quad \forall k \in \{1, 2\}, \ \forall s \in S, \ \forall a \in \mathcal{A} \qquad (18e)$$

$$\sum_{a \in \mathcal{A}} \Delta_{s,a} \le 1, \qquad\qquad \forall s \in S \qquad (18f)$$

$$\lambda_{s,a}^1 \ge 0, \lambda_{s,a}^2 \ge 0, \Delta_{s,a} \in \{0, 1\} \qquad (18g)$$

where M is a large constant whose precise value will be discussed shortly. x_α denotes the maximum probability of reaching the target set B in an MDP with initial distribution α and can be computed via linear programming as discussed in Section V-A. Finally, $r: S \times \mathcal{A} \to \mathbb{R}_{\ge 0}$ is a function such that

$$r(s, a) := \begin{cases} \sum_{s' \in B} \mathcal{P}_{s, a, s'} & \text{if } s \in S_r \\ 0 & \text{otherwise.} \end{cases}$$

There are $(|S||\mathcal{A}|)^2$ continuous and $|S||\mathcal{A}|$ integer variables in (18a)-(18g). The sets $\{\lambda_{s,a}^1 \ge 0 : s \in S, a \in \mathcal{A}\}$ and $\{\lambda_{s,a}^2 \ge 0 : s \in S, a \in \mathcal{A}\}$ of continuous variables correspond, respectively, to the *discounted* and *undiscounted* occupation measures [6], [9]. The set $\{\Delta_{s,a} \in \{0, 1\} : s \in S, a \in \mathcal{A}\}$ of integer variables correspond to the deterministic actions taken by the agent.

The constraints in (18b) and (18c) represent the balance equations for the discounted and undiscounted occupation measures, respectively [6]. The constraint in (18d) ensures that the agent reaches the target set B with maximum probability x_{s_1}. Finally, the constraints in (18e)-(18f) ensure that the agent follows a deterministic policy.

Let $f: S \times \mathcal{A} \to \mathbb{R}_{\ge 0}$ be a function such that $f(s, a):=1$ for all $s \in S_r$ and $f(s, a):=0$ otherwise. Moreover, let $M^\star$ be a constant such that

$$M^\star := \max_{\pi \in \Xi^{SD}(\mathcal{M}, B)} \mathbb{E}_\alpha^\pi \left[\sum_{t=1}^\infty f(s_t, a_t) \right].$$

$M^\star$ corresponds to the maximum expected number of steps taken under any stationary deterministic policy before reaching the set B with maximum probability. $M^\star$ is finite since all policies $\pi \in \Xi^{SD}(\mathcal{M}, B)$ induce MCs in which the set $S_r \cap S_{\to}^\pi$ consists only of transient states.

Theorem 4: Let the constant M in (18e) be chosen such that $M \geq M^\star$. Then, for $K \in \mathbb{R}$, there exists a policy $\pi \in \Xi^{SD}(\mathcal{M}, B)$ such that $J(\pi) \leq K$ if and only if the optimal value of the problem in (18a)-(18g) is less than or equal to K. Furthermore, an optimal policy $\pi_D^\star \in \Xi^{SD}(\mathcal{M}, B)$ such that $J(\pi_D^\star) \leq K$ can be obtained from the optimal variables $\{\Delta_{s,a}^\star : s \in S, a \in \mathcal{A}\}$ of the problem in (18a)-(18g) by the following rule:

$$\pi_D^\star(s,a) = \begin{cases} \Delta_{s,a}^\star, & \text{if } \sum_{a \in \mathcal{A}} \Delta_{s,a}^\star = 1 \\ \text{arbitrary}, & \text{otherwise}. \end{cases} \quad (19)$$

Proof: ($\Rightarrow$) Suppose that there exists a stationary deterministic policy $\pi \in \Xi^{SD}(\mathcal{M}, B)$ such that $J(\pi) \leq K$. We set $\Delta_{s,a} = \pi(s,a)$,

$$\lambda_{s,a}^1 = \sum_{t=1}^\infty \beta^{t-1} \Pr^\pi\left(S_t = s, A_t = a\right), \text{ and}$$

$$\lambda_{s,a}^2 = \begin{cases} \sum_{t=1}^\infty \Pr^\pi\left(S_t = s, A_t = a\right), & \text{if } s \in S_r \\ 0, & \text{otherwise}. \end{cases}$$

$\Pr^\pi(S_t = s, A_t = a)$ denotes the probability with which the state-action pair (s,a) is occupied in the induced MC $\mathcal{M}^\pi$. The variables $\lambda_{s,a}^1, \lambda_{s,a}^2$ are finite for all $s \in S$ and $a \in \mathcal{A}$ since the states $s \in S_r \cap S_\rightarrow^\pi$ are transient in the induced MC $\mathcal{M}^\pi$.

It can be shown that the above choices of the variables satisfy the balance equations in (18b)-(18c) (see, e.g., Theorem 3.1 in [6] for a similar derivation). The variables $\lambda_{s,a}^2$ also satisfy the constraint in (18d), which follows from the fact (see Theorem 10.15 in [25]) that, for any $\pi \in \Pi(\mathcal{M})$, we have

$$\Pr_\mathcal{M}^{\pi,\alpha}(Reach[B]) = \mathbb{E}_\alpha^\pi\left[\sum_{t=1}^\infty r(s_t, a_t)\right]. \quad (20)$$

The satisfaction of the constraint in (18e) follows from the definition of $M^\star$. The constraints in (18f) and (18g) are satisfied by construction. Finally, it follows from classical MDP theory, (see, e.g., [1, Chapter 6]), that these variables attain an objective value that is equal to $J(\pi)$. Since we constructed a feasible solution with the objective value $J(\pi)$, we conclude that the optimal value of the problem in (18a)-(18g) is less than or equal to K.

($\Leftarrow$) Suppose that the optimal value of the problem in (18a)-(18g) is less than or equal to K. From the optimal variables, we construct a policy $\pi_D^\star$ through the formula in (19). It follows from Proposition 2 in [11] that $\pi_D^\star \in \Pi^{SD}(\mathcal{M})$ and $J(\pi_D^\star) \leq K$. Due to the equality in (20), we also have $\pi^\star \in \Xi^{SD}(\mathcal{M}, D)$, which concludes the proof. $\square$

Theorem 4 establishes that, if we choose $M \geq M^\star$, we can obtain a solution to the problem in (17) by solving the MILP in (18a)-(18g) and constructing a policy through the formula in (19). It can be shown by a reduction from the Hamiltonian path problem that the computation of the constant $M^\star$ is NP-hard in general. For MDPs with deterministic transitions, we have $M^\star \leq |S|$; hence we set $M = |S|$. Although a finite upper bound on $M^\star$ for MDPs with stochastic transitions can be computed efficiently, we omit the details as the procedure is more involved and requires one to analyze the structure of the MDP. For such MDPs, we simply set $M = k|S|$ for some large $k \in \mathbb{N}$.

C. An approximation algorithm

We now present an algorithm to obtain an *approximate* solution to the NP-hard policy synthesis problem given in (17). We make the following assumptions throughout this section.

Assumption: The cost function satisfies $c(s,a) = 0$ for all $s \in B \cup S_0$.

Assumption: MDP $\mathcal{M}$ has a unique initial state $s_1 \in S$, i.e., $\alpha(s_1) = 1$.

The first assumption states that the agent incurs no cost at the target states and at the states from which there is no path to the target

states. This assumption typically holds in practice since costs are incurred only until a task is completed or failed to be completed. The second assumption is introduced because the approximation algorithm described below utilizes the minimum number of steps to reach a state from a specific initial state.

The main idea behind the approximation algorithm is that the discounted immediate cost $\beta^{t-1} c(s,a)$ incurred at a state-action pair (s,a) at step $t \in \mathbb{N}$ is upper bounded by $\beta^{T_{\min}(s)-1} c(s,a)$ where $T_{\min}(s)$ is the minimum number of steps taken to reach the state s from s_1. Using the derived upper bound, we define a surrogate objective function and formulate a constrained MDP problem for which we synthesize an optimal policy via linear programming.

For an MDP $\mathcal{M}$, let $G_\mathcal{M} = (S, E_\mathcal{M})$ be a digraph where S is the set of vertices and $E_\mathcal{M}$ is the set of edges such that

$$E_\mathcal{M} := \left\{(s, s') \in S \times S : \sum_{a \in \mathcal{A}(s)} \mathcal{P}_{s,a,s'} > 0\right\}. \quad (21)$$

On the digraph $G_\mathcal{M}$, let $\mathbb{PATH}_n(s)$ be the set of finite paths $v_1 v_2 \ldots v_n$ of length $n \in \mathbb{N}$ such that $v_1 = s_1$ and $v_n = s$, i.e., the set of finite paths that reach the state s starting from the initial state s_1. Then, on the MDP $\mathcal{M}$, the agent can reach the state s in minimum $T_{\min}(s)$ steps where

$$T_{\min}(s) := \min\{n \in \mathbb{N} : \mathbb{PATH}_n(s) \neq \emptyset\}. \quad (22)$$

Note that $T_{\min}(s)$ can be efficiently computed using standard shortest path algorithms, e.g., Dijkstra's algorithm [29].

Let $\widetilde{c}: S \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ be a modified cost function such that

$$\widetilde{c}(s, a) := \beta^{T_{\min}(s)-1} c(s, a), \quad (23)$$

and $\widetilde{J}: \Pi(\mathcal{M}) \to \mathbb{R}_{\geq 0}$ be a surrogate function such that

$$\widetilde{J}(\pi) := \mathbb{E}_\alpha^\pi\left[\sum_{t=1}^\infty \widetilde{c}(s_t, a_t)\right].$$

As the approximation algorithm, we propose to synthesize a policy $\pi_{D,app}^\star \in \Pi^{SD}(\mathcal{M})$ such that

$$\pi_{D,app}^\star \in \arg\min_{\pi \in \Xi^{SD}(\mathcal{M}, B)} \widetilde{J}(\pi). \quad (24)$$

In what follows, we first present an efficient method to synthesize a stationary deterministic policy $\pi_{D,app}^\star$ that satisfies the condition in (24). We then derive an upper bound on the suboptimality of the synthesized policy for the original problem given in (17).

1) Policy synthesis: The problem in (24) is a total *undiscounted* cost minimization problem subject to a constraint on the total *undiscounted* reward given in (20). We recently established the existence of optimal stationary deterministic policies for such problems in [30], where we also present an efficient method to synthesize optimal policies. For completeness, we provide the developed method below.

We solve two linear programs (LPs) to synthesize the policy $\pi_{D,app}^\star$. First, we solve the following LP:

$$\underset{\lambda_{s,a} \geq 0}{\text{minimize}} \sum_{s \in S_r} \sum_{a \in \mathcal{A}} \widetilde{c}(s, a) \lambda_{s,a} \quad (25a)$$

subject to:

$$\sum_{a \in \mathcal{A}} \lambda_{s,a} - \sum_{s' \in S} \sum_{a \in \mathcal{A}} \mathcal{P}_{s',a,s} \lambda_{s',a} = \alpha(s), \ \forall s \in S_r \quad (25b)$$

$$\sum_{s \in S_r} \sum_{a \in \mathcal{A}} \lambda_{s,a} r(s, a) = x_{s_1}. \quad (25c)$$

The above LP computes a set $\{\lambda_{s,a} \geq 0 : s \in S, a \in \mathcal{A}\}$ of occupancy measures from which one can synthesize a (randomized) policy π that minimizes the surrogate function $\widetilde{J}(\pi)$ over $\Pi(\mathcal{M})$ while satisfying

the maximum reachability constraint. Let $v^\star$ be the optimal value of the LP in (25a)-(25c). Next, we solve the following LP.

$$\underset{\lambda_{s,a} \geq 0}{\text{minimize}} \quad \sum_{s \in S} \sum_{a \in \mathcal{A}} \lambda_{s,a} \tag{26a}$$

$$\text{subject to:} \quad \sum_{s \in S_r} \sum_{a \in \mathcal{A}} \lambda_{s,a} \widetilde{c}(s,a) = v^\star \tag{26b}$$

$$(25b) - (25c). \tag{26c}$$

Let $\{\lambda_{s,a}^\star \geq 0 \; : \; s \in S, a \in \mathcal{A}\}$ be the set of optimal variables for the LP in (26a)-(26c). Moreover, for a given state $s \in S$, let $\mathcal{A}^\star(s) := \{a \in \mathcal{A} \; : \; \lambda_{s,a}^\star > 0\}$ be the set of optimal actions. Then, a stationary deterministic policy $\pi_{D,app}^\star \in \Pi^{SD}(\mathcal{M})$ satisfying the condition in (24) can be synthesized from this set by choosing

$$\pi_{D,app}^\star(s,a) = 1 \text{ for an arbitrary } a \in \mathcal{A}^\star(s). \tag{27}$$

Proposition 4: [30] A policy $\pi_{D,app}^\star \in \Pi^{SD}(\mathcal{M})$ generated by the rule in (27) satisfies the condition in (24).

The proposed approximation algorithm can be summarized as follows. First, construct the modified cost function given in (23). Then, sequentially solve the LPs in (25a)-(25c) and (26a)-(26c). Finally, synthesize the policy $\pi_{D,app}^\star \in \Pi^{SD}(\mathcal{M})$ via the rule in (27). Note that the computational complexity of the proposed approximation algorithm is that of solving a linear program, which can be achieved in matrix-multiplication time [31].

2) Suboptimality analysis: Let $c_{\min} := \min\{c(s,a) \; : \; s \in S_r, a \in \mathcal{A}\}$ and $\widetilde{c}_{\max} := \max\{\widetilde{c}(s,a) \; : \; s \in S_r, a \in \mathcal{A}\}$ be the minimum immediate cost and maximum modified immediate cost, respectively. Moreover, let $\underline{M}$ be a constant such that

$$\underline{M} := \min_{\pi \in \Xi^{SD}(\mathcal{M},B)} \mathbb{E}_\alpha^\pi \left[\sum_{t=1}^\infty \beta^{t-1} f(s_t, a_t) \right].$$

$\underline{M}$ is the minimum number of steps (in expectation) taken under any deterministic policy before reaching the target set B with maximum probability and can be computed efficiently via linear programming.

Proposition 5: For any $\pi \in \Xi^{SD}(\mathcal{M}, B)$, we have

$$\underline{M} c_{\min} \overset{(1)}{\leq} J(\pi) \overset{(2)}{\leq} \widetilde{J}(\pi) \overset{(3)}{\leq} M^\star \widetilde{c}_{\max}.$$

Proof: The inequality (1) holds since, under the assumption that $C(s,a) = 0$ for all $s \in B \cup S_0$, for any $\pi \in \Pi(\mathcal{M})$,

$$J(\pi) \geq c_{\min} \mathbb{E}_\alpha^\pi \left[\sum_{t=1}^\infty \beta^{t-1} f(s_t, a_t) \right] = \underline{M} c_{\min}.$$

To show that the inequality (2) holds, note that

$$J(\pi) = \sum_{t=1}^\infty \beta^{t-1} \text{Pr}^\pi \left(S_t = s_t, A_t = a | S_1 = s_1 \right) c(s_t, a_t).$$

It follows from the definition of $T_{\min}(s)$ that we have $\text{Pr}^\pi(S_t = s_t, A_t = a | S_1 = s_1) > 0$ if and only if $t \geq T_{\min}(s)$. Then, for each nonzero term in the right hand side of the above equation, we have $\beta^{t-1} \leq \beta^{T_{\min}(s)-1}$ since $\beta < 1$. Consequently, it follows from the definition of $\widetilde{c}$ that we have $J(\pi) \leq \widetilde{J}(\pi)$.

The inequality (3) holds for all $\pi \in \Xi^{SD}(\mathcal{M}, B)$ since, under the assumption that $c(s,a) = 0$ for all $s \in B \cup S_0$, we have

$$\widetilde{J}(\pi) \leq \widetilde{c}_{\max} \mathbb{E}_\alpha^\pi \left[\sum_{t=1}^\infty f(s_t, a_t) \right] \leq M^\star \widetilde{c}_{\max}. \quad \square$$

Corollary 2: Let $\pi_{D,app}^\star \in \Pi^{SD}(\mathcal{M})$ be a policy satisfying the condition in (24), and $\pi_D^\star \in \Pi^{SD}(\mathcal{M})$ be a policy satisfying the condition in (17). Then, we have

$$J(\pi_{D,app}^\star) - J(\pi_D^\star) \leq M^\star \widetilde{c}_{\max} - \underline{M} c_{\min}. \tag{28}$$

Furthermore, if the MDP $\mathcal{M}$ has only deterministic transitions, i.e., $\mathcal{P}_{s,a,s'} \in \{0,1\}$ for all $s \in S$ and for all $a \in \mathcal{A}$, then

$$J(\pi_{D,app}^\star) - J(\pi_D^\star) \leq |S| \widetilde{c}_{\max}. \tag{29}$$

Proof: It follows from Proposition 5 that $J(\pi) \leq \widetilde{J}(\pi) \leq M^\star \widetilde{c}_{\max}$ for all $\pi \in \Xi^{SD}(\mathcal{M}, B)$. Then, we have $J(\pi_{D,app}^\star) \leq M^\star \widetilde{c}_{\max}$. Similarly, since $\underline{M} c_{\min} \leq J(\pi)$ for all $\pi \in \Xi^{SD}(\mathcal{M}, B)$, we have $\underline{M} c_{\min} \leq J(\pi_D^\star)$. Combining the above inequalities, we obtain the inequality in (28). The result in (29) follows from the fact that $M^\star \leq |S|$ if the the MDP has only deterministic transitions. $\square$

The bound in (29) shows that, for MDPs with deterministic transitions, the suboptimality of the proposed approximation algorithm grows *at most* linearly in the size of the set S of states. Note also that, for deterministic systems, the suboptimality gap $|S| \widetilde{c}_{\max}$ decreases as the discount factor β decreases.

VIII. SIMULATION RESULTS

We illustrate the performance of the proposed algorithms on a motion planning example. We run the computations on a 3.2 GHz desktop with 8 GB RAM and use GUROBI [24] for optimization.

Consider an autonomous vehicle that aims to deliver a package to a target region in an adversarial environment. Specifically, in each region in the environment, the vehicle faces the risk of being attacked by an adversary whose objective is to prevent the vehicle from reaching its target. Being aware of the threat, the vehicle's objective is to reach its target by following minimum-risk trajectories.

We model the environment as a grid world illustrated in Figure 2. The agent starts from the initial state (brown) and aims to reach the target state (green) while avoiding the states that are occupied by obstacles (red). The set of actions available to the agent are $\{up, down, left, right, stay\}$. Under an action $a \in \{up, down, left, right\}$, the agent moves to the successor state in the desired direction with probability 0.9 and stays in its current state with probability 0.1. Under the $stay$ action, the agent stays in its current state with probability 1.

Each state in the environment has an associated cost that represents the risk of being attacked by an adversary. The states along the minimum length trajectory have a high risk, i.e., $c(s,a) = 4$. The states that are between the target and the obstacles have a moderate risk, i.e., $c(s,a) = 2$. Finally, all other states have a low risk, i.e., $c(s,a) = 1$.

We synthesize three stationary deterministic policies for the agent using two existing methods in the literature and the proposed approximation algorithm. With 309 binary variables in the MILP formulation, the computation exceeds the memory limit after 1325 seconds.

We first synthesize a policy based on the classical constrained MDP formulation in which both the incurred costs and the collected rewards are discounted [6]. In particular, we express the reachability criterion as a total reward criterion as shown in (20). We then discount both the costs and the rewards with $\beta < 1$ and synthesize a policy that minimizes the total discounted cost among the ones that maximizes the total discounted reward. The agent's trajectory under the synthesized policy is shown in Figure 2 (as "discounted cost + discounted reachability"). Due to the discounting in the rewards, the agent assigns an artificial importance to the reachability objective and follows a trajectory visiting the high-risk regions in the environment.

The second policy is based on a multi-objective MDP formulation [10] in which neither the costs nor the rewards are discounted. Specifically, we express the reachability criterion as a total reward criterion and consider undiscounted costs by setting $\beta = 1$. The agent's trajectory under the synthesized policy is shown in Figure 2 (as "undiscounted cost + reachability"). The agent minimizes the total risk along its trajectory. However, due to lack of discounting in the
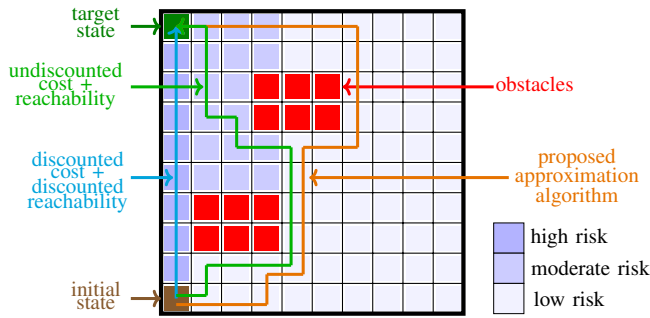
Fig. 2: An illustration of the environment in the motion planning example. The agent aims to reach the target state while avoiding risky regions for as long as possible. By combining a discounted cost criterion with a reachability criterion, the proposed approximation algorithm generates a trajectory that postpones the entrance to risky regions while ensuring the satisfaction of the task.

costs, the synthesized policy generates a trajectory that visits more states with moderate risk than the minimum achievable one.

We synthesize the third policy through the approximation algorithm presented in Section VII-C by choosing $\beta=0.9$. The agent's trajectory under the synthesized policy is shown in Figure 2 (as "proposed approximation algorithm"). The synthesized policy attains the same total risk with the second policy explained above. However, thanks to the discounting in the costs, the agent establishes a hierarchy between the policies with the same total risk and follows a trajectory that visits the minimum number of states with moderate risk.

The above example illustrates the two main benefits of the formulation that combines a discounted cost criterion with a probabilistic reachability criterion. First, by not discounting the reachability constraint, the agent is able to follow long trajectories if it is optimal to do so. Second, by discounting the costs, the agent is able to create an ordering between the policies that incur the same total cost.

IX. CONCLUSIONS AND DISCUSSIONS

We studied the problem of synthesizing a policy in a Markov decision process (MDP) which, when followed, causes an agent to reach a target state in the MDP while minimizing its total discounted cost. We showed that, in general, an optimal policy for this problem might not exist, but there always exists a near-optimal policy which can be synthesized efficiently. We also considered the synthesis of an optimal stationary deterministic policy and established that the synthesis of such a policy is NP-hard. Finally, we proposed a linear programming-based algorithm to synthesize stationary deterministic policies with a theoretical suboptimality guarantee.

We presented an analysis over *stationary* deterministic policies. In general, *Markovian* deterministic policies achieve lower total discounted costs than stationary deterministic policies when subject to a reachability constraint. For such instances, one can synthesize Markovian policies by taking a Cartesian product of the MDP with a deterministic finite automaton representing the "time-dependency" of the policy and applying the algorithms developed in this paper.

REFERENCES

[1] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
[2] L. De Alfaro, "Computing minimum and maximum reachability times in probabilistic systems," in *International Conference on Concurrency Theory*, 1999, pp. 66–81.
[3] 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.

[4] A. J. Schaefer, M. D. Bailey, S. M. Shechter, and M. S. Roberts, "Modeling medical treatment using Markov decision processes," in *Operations Research and Health Care*, 2005, pp. 593–612.
[5] G. Pola and G. Pola, "A stochastic reachability approach to portfolio construction in finance industry," *IEEE Transactions on Control Systems technology*, vol. 20, no. 1, pp. 189–195, 2011.
[6] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999.
[7] C. Kiennert, Z. Ismail, H. Debar, and J. Leneutre, "A survey on game-theoretic approaches for intrusion detection and response optimization," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–31, 2018.
[8] D. P. Bertsekas and J. N. Tsitsiklis, "An analysis of stochastic shortest path problems," *Mathematics of Operations Research*, vol. 16, no. 3, pp. 580–595, 1991.
[9] K. Etessami, M. Kwiatkowska, M. Y. Vardi, and M. Yannakakis, "Multi-objective model checking of Markov decision processes," in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2007, pp. 50–65.
[10] K. Chatterjee, R. Majumdar, and T. A. Henzinger, "Markov decision processes with multiple objectives," in *Annual Symposium on Theoretical Aspects of Computer Science*, 2006, pp. 325–336.
[11] D. Dolgov and E. Durfee, "Stationary deterministic policies for constrained MDPs with multiple rewards, costs, and discount factors," in *International Joint Conference on Artificial Intelligence*, 2005.
[12] E. A. Feinberg and A. Shwartz, "Constrained dynamic programming with two discount factors: applications and an algorithm," *IEEE Transactions on Automatic Control*, vol. 44, no. 3, pp. 628–631, 1999.
[13] E. A. Feinberg, "Constrained discounted Markov decision processes and Hamiltonian cycles," *Mathematics of Operations Research*, vol. 25, no. 1, pp. 130–140, 2000.
[14] R. C. Chen and G. L. Blankenship, "Dynamic programming equations for discounted constrained stochastic control," *IEEE Transactions on Automatic Control*, vol. 49, no. 5, pp. 699–709, 2004.
[15] E. A. Feinberg and J. Huang, "On the reduction of total-cost and average-cost MDPs to discounted MDPs," *Naval Research Logistics*, vol. 66, no. 1, pp. 38–56, 2019.
[16] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.
[17] F. Teichteil-Königsbuch, "Stochastic safest and shortest path problems," in *AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012.
[18] F. Delgrange, J.-P. Katoen, T. Quatmann, and M. Randour, "Simple strategies in multi-objective MDPs," in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2020, pp. 346–364.
[19] Z. Gábor, Z. Kalmár, and C. Szepesvári, "Multi-criteria reinforcement learning," in *International Conference on Machine learning*, vol. 98, 1998, pp. 197–205.
[20] R. Yang, X. Sun, and K. Narasimhan, "A generalized algorithm for multi-objective reinforcement learning and policy adaptation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[21] L. Barrett and S. Narayanan, "Learning all optimal policies with multiple criteria," in *International Conference on Machine learning*, 2008.
[22] E. A. Feinberg and A. Shwartz, "Constrained Markov decision models with weighted discounted rewards," *Mathematics of Operations Research*, vol. 20, no. 2, pp. 302–320, 1995.
[23] K. C. Kalagarla, R. Jain, and P. Nuzzo, "Synthesis of discounted-reward optimal policies for Markov decision processes under linear temporal logic specifications," *arXiv preprint arXiv:2011.00632*, 2020.
[24] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2021. [Online]. Available: http://www.gurobi.com
[25] C. Baier and J.-P. Katoen, *Principles of Model Checking*. MIT Press, 2008.
[26] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.
[27] P. Paruchuri, M. Tambe, F. Ordonez, and S. Kraus, "Towards a formalization of teamwork with resource constraints," in *International Joint Conference on Autonomous Agents and Multiagent Systems*, 2004.
[28] M. Sipser, "Introduction to the theory of computation," *ACM Sigact News*, vol. 27, no. 1, pp. 27–29, 1996.
[29] E. W. Dijkstra *et al.*, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
[30] Y. Savas, V. Gupta, M. Ornik, L. J. Ratliff, and U. Topcu, "Incentive design for temporal logic objectives," in *IEEE Conference on Decision and Control*, 2019, pp. 2251–2258.
[31] M. B. Cohen, Y. T. Lee, and Z. Song, "Solving linear programs in the current matrix multiplication time," *Journal of the ACM*, vol. 68, no. 1, pp. 1–39, 2021.