On Reachability of Markov Chains: A Long-Run Average Approach

Daniel Ávila and Mauricio Junca

***Abstract*—We consider a Markov control model in discrete time with countable both state space and action space. Using the value function of a suitable long-run average reward problem, we study various reachability/controllability problems. First, we characterize the domain of attraction and escape set of the system, and a generalization called p-domain of attraction, using the aforementioned value function. Next, we solve the problem of maximizing the probability of reaching a set A while avoiding a set B. Finally, we consider a constrained version of previous problem, where we ask for the probability of reaching the set B to be bounded. In the finite case, we use linear programming formulations to solve these problems.**

***Index Terms*—Long-run average, Markov decision processes (MDP), probabilistic constraints, reach avoid.**

Notation

Sets

$\mathbb{X}$:	State space set.
$\mathbb{U}$:	Actions space.
$\mathbb{U}(x)$:	Admissible actions for state $x \in \mathbb{X}$.
$\mathbb{K}$:	The set of feasible states and actions $\mathbb{K} = \{(x,u) \vert x \in \mathbb{X}, u \in \mathbb{U}(x)\}$.
Π:	The set of admissible policies.
Π_M:	The set of Markovian policies.

Measures

$Q(\cdot\vert\cdot)$:	Stochastic kernel on $\mathbb{X}$ given $\mathbb{K}$.
P_ν^π:	Probability under initial distribution ν and $\pi \in \Pi$.
E_ν^π:	Expected value with respect to P_ν^π.
P_x^π / E_x^π:	Notation when $\nu = \delta_x$ for $x \in \mathbb{X}$.

Random variables

$\{X_t\}_{t\geq 0}$:	Stochastic $\mathbb{X}$-valued process.
$\{U_t\}_{t\geq 0}$:	Stochastic $\mathbb{U}$-valued process.
τ_A:	Hitting time of $A \subset \mathbb{X}$.

I. Introduction

Markov decision processes (MDP) provide a mathematical framework for modeling decision making problems, where outcomes are uncertain. Formally, these are discrete time Markov control stochastic processes. The most common and studied setting are stochastic systems over a discrete state space with discrete action space, see for example [1], [2], but general spaces are also part of the literature, see [3], [4]. Applications of MDPs range from inventory control and investment planning to economics and behavioral ecology.

In this article, we will be interested in using the MDP framework to solve some problems concerning the controllability/reachability of a controlled Markov chain. As we will see, such problems do not only evaluate the state of the system at each time, but the whole evolution of the process. The first problem we aim to solve is to characterize the domain of attraction and escape set of a set A. The first one refers to the initial states for which there exist a control that takes the system to A with positive probability, while the escape set are the initial states such that no control can take the system to A. A somehow related problem was studied in [5] and [6], where the idea was to use entropy methods to maximize the number of recurrent states.

Definition I.1: Given a set $A \subset \mathbb{X}$, let

$$\Lambda_A = \left\{ x \in \mathbb{X} \middle| \liminf_{t\to\infty} P_x^\pi(X_t \in A) > 0 \textit{ for some policy } \pi \right\}$$

$$\Gamma_A = \left\{ x \in \mathbb{X} \middle| \liminf_{t\to\infty} P_x^\pi(X_t \in A) = 0 \textit{ for all policies } \pi \right\}.$$

The domain of attraction appears in the context of deterministic differential equations when describing the initial states under which the system will approach to a stable point. Such technique is commonly referred in the literature as Zubov's method, see [7]. It allows to characterize the domain of attraction and the escape set in terms of an appropriate value function, which is the solution of a differential equation. In [8], Zubov's method is generalized for deterministic controlled systems, and in [9], it is further generalized for stochastic differential equations. For the present article, we took as guide the constructions made in this last work. Inspired by the literature of stochastic target problem, see [10], we study the p-domain of attraction for any $p \in (0,1]$, defined as follows.

Definition I.2: Given a set $A \subset \mathbb{X}$ and $p \in (0,1]$, let

$$\Lambda_A^p = \left\{ x \in \mathbb{X} \middle| \liminf_{t\to\infty} P_x^\pi(X_t \in A) \geq p \textit{ for some policy } \pi \right\}.$$

A similar problem in the context of stochastic hybrid systems and finite horizon is studied in [11].

The next problem consists on finding a control policy that maximize the probability of reaching some set A while avoiding some set B. Given an initial distribution ν over the state space, our main objective will be to find a control policy π that solves the problem

$$\max_\pi P_\nu^\pi(\tau_A < \tau_B, \tau_A < \infty). \tag{P1}$$

Note that the set B acts as a cemetery set since the evolution of the controlled Markov process is meaningless after this set is reached. This problem is studied in [12], where the authors consider general state and actions spaces but assume that $\tau_A \wedge \tau_B < \infty$ a.s for every policy π, hence, they use a total reward approach. Finite horizon

Manuscript received January 31, 2020; revised July 17, 2020 and February 1, 2021; accepted March 28, 2021. Date of publication April 6, 2021; date of current version March 29, 2022. This work was supported by the Vice Presidency for Research & Creation publication fund by the Universidad de los Andes. Recommended by Associate Editor Q.-S. Jia. *(Corresponding author: Mauricio Junca.)*

Daniel Ávila is with the Center for Operations Research and Econometrics, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium (e-mail: daniel.avila@uclouvain.be).

Mauricio Junca is with the Department of Mathematics, Universidad de los Andes, Bogotá 111121, Colombia (e-mail: mj.junca20@uniandes.edu.co).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2021.3071334.

Digital Object Identifier 10.1109/TAC.2021.3071334

related problems for hybrid systems and a continuous time version for controlled diffusion can be found in [13] and [14], respectively.

Finally, we consider a constrained version of the previous problem as follows:

$$\max_{\pi} P_\nu^\pi(\tau_A < \infty)$$
$$\text{s.t. } P_\nu^\pi(\tau_B < \infty) \leq \epsilon. \tag{P2}$$

In this case the set B is no longer a cemetery set, making the evolution of the controlled process different, depending on whether the set B has been reached or not. This fact suggests that this cannot be a Markovian problem. To the best of our knowledge, such problem, or any similar, has not been studied in the literature.

The contributions of this article are the following. 1) We consider reachability problems in infinite horizon and relate them with long-run average reward problems in the context of MDPs. The main result in this direction is Theorem II.3 that calculates the probability of reaching a closed set in finite time in terms of such reward problems. Then, we use this result to characterize p-domains of attraction (see Theorem III.4) and solutions of reach-avoid problems (see Theorem IV.2). An important feature of our results is that it includes multichain models. To the best of our knowledge, this is the first time that long-run average reward problems are used to characterize reachability problems. 2) We consider a reachability problem with a constraint in the probability of hitting a given set of states. In general, stochastic control problems with probabilistic constrains are hard to solve since in this case dynamic programming principle is unclear. Using a state-space augmentation technique, we are able to formulate the problem in terms of long-run average reward problems (see Theorem V.4). In the case of finite state and action spaces we use linear programming duality to solve the problem (see Theorem VI.2).

This article is organized as follows. Section II states the framework and known results for MDP and long-run average reward problems. It also presents the result that relates the probability of reaching closed set in finite time with the MDP reward problem. Sections III–V present the results for p-domains of attraction, Problem (P1) and Problem (P2), respectively. In Section VI, we consider finite state and action spaces and formulate Problem (P2), as a linear program. Finally, in Section VII, we present numerical examples to illustrate our findings, and conclude afterwards.

II. Markov Control Model

In this section, we establish the framework of MDP and some results about closed sets that allow to formulate properly the problems described previously. A Markov control model (see [1], [4] for details and the nomenclature section for definitions) is a tuple $(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x)|x \in \mathbb{X}\}, Q, r)$, where we assume that $\mathbb{X}$ and $\mathbb{U}$ are countable. $r : \mathbb{K} \to \mathbb{R}$ is the reward function.

A control policy is a sequence $\pi = \{\mu_t\}_{t\geq 0}$ of stochastic kernels on $\mathbb{U}$ given the set of admissible histories up to time t, $H_t := \mathbb{K}^t \times \mathbb{X}$, such that for all $t \geq 0$ and $h_t = (\cdot, x_t) \in H_t$, $\mu_t(\cdot|h_t)$ is supported on $\mathbb{U}(x_t)$.

Consider the measurable space $(\Omega, \mathcal{F})$, where $\Omega := (\mathbb{X} \times \mathbb{U})^{\mathbb{N}}$ and $\mathcal{F}$ is the product sigma algebra. Given a policy $\pi \in \Pi$ and a distribution ν over $\mathbb{X}$ there exist a unique probability measure P_ν^π on $(\Omega, \mathcal{F})$ and a $\mathbb{X} \times \mathbb{U}$-valued stochastic process $(\{(X_t, U_t)\}_{t\geq 0}, P_\nu^\pi)$ such that for each $B \subset \mathbb{X}$ and $V \subset \mathbb{U}$, (see [4]):

1) $P_\nu^\pi(\mathbb{K}^{\mathbb{N}}) = 1$;
2) $P_\nu^\pi(X_0 \in B) = \nu(B)$;
3) $P_\nu^\pi(U_t \in V|h_t) = \mu_t(V|h_t)$;
4) $P_\nu^\pi(X_{t+1} \in B|h_t, U_t = u) = Q(B|x_t, u)$.

In general the process $\{X_t\}_{t\geq 0}$ is non-Markovian, so, in order to make the process Markovian, we also consider Markovian control policies, Π_M. In this case, we have a sequence $\pi = \{\mu_t\}_{t\geq 0}$, where μ_t is a stochastic kernel on $\mathbb{U}$ given $\mathbb{X}$ only. In this case, the process $\{X_t\}_{t\geq 0}$ is Markovian. If μ_t is the same for all $t \geq 0$ the Markovian policy is called stationary and in this case $\{X_t\}_{t\geq 0}$ is a time-homogeneous Markov process. Finally, if π is stationary and $\mu(u|x) \in \{0, 1\}$ for all $u \in \mathbb{U}(x)$ and $x \in \mathbb{X}$, then we say that π is a stationary deterministic policy.

We will denote by $\overline{x}^t$ the history up to time t of the process $\{X_t\}$, and similarly for the process $\{U_t\}$. Hence, admissible histories can be written as $h_t = (\overline{x}^{t-1}, \overline{u}^{t-1}, x_t)$. We also say that $\overline{x}^t \notin B$ if none of its components belong to B.

1) Long-Run Average Reward Problems: Given a Markov control model and an initial state $x \in \mathbb{X}$, we want to find a policy $\pi \in \Pi$ that maximizes the reward function

$$v^\pi(x) := \limsup_{N\to\infty} \frac{1}{N} \mathbb{E}_x^\pi \left[\sum_{t=0}^{N-1} r(X_t, U_t) \right].$$

The first important result about these problems is that for any $\pi \in \Pi$ there exists $\pi' \in \Pi_M$ such that

$$P_\nu^\pi(X_t = x,' U_t = u) = P_\nu^{\pi'}(X_t = x,' U_t = u) \tag{1}$$

for any $t \geq 0$, $x' \in \mathbb{X}$, and $u \in \mathbb{U}(x')$, which implies that the reward functions are the same, see [1]. Therefore, optimal policies, when they exist, can always be found among Markovian policies. Given $x \in \mathbb{X}$, let $V(x) = \sup_{\pi\in\Pi} v^\pi(x)$. An important tool to calculate an optimal policy is the so-called optimality equations for the multichain model. For all $x \in \mathbb{X}$

$$\max_{u\in\mathbb{U}(x)} \left\{ \sum_{x'\in\mathbb{X}} P_{x,x'}^u v(x') - v(x) \right\} = 0 \tag{MC1}$$

$$\max_{u\in\mathbb{U}(x)} \left\{ r(x, u) - v(x) + \sum_{x'\in\mathbb{X}} P_{x,x'}^u h(x') - h(x) \right\} = 0 \tag{MC2}$$

where $P_{x,x'}^u := Q(x'|x, u)$. Multichain models are those where there exists a stationary Markovian policy for which the induced Markov chain has at least two recurrent classes. Uni-chain models are easier since there is just one recurrent class and the function V is constant, so there is no need to introduce the first equation. Here, we consider the multichain case since we will deal with closed sets. It is known that if equations (MC1)-(MC2) have a solution (v, h) with h bounded, then, $v = V$. When the state and actions spaces are finite the equations always have a unique solution (see [1, Sec. 9.1]). For countable spaces, conditions for existence and uniqueness are given in [15] and [16].

1) Closed Sets: Closed subsets of the state space are essential for the correct formulation of the problems described in the introduction. Note that the notion of closed set is relative to a policy, hence, given a policy $\pi \in \Pi$ and a set $A \subset \mathbb{X}$, we say A is closed under π if given that $x_s \in A$ for some $s \leq t$, then

$$P_\nu^\pi(X_{t+1} \notin A|h_t) = \sum_{u_t\in\mathbb{U}(x_t)} Q(A^c|x_t, u_t)\mu_t(u_t|h_t) = 0.$$

Remark II.1: Note that if $Q(x'|x, u) = 0$ for all $x \in A$, $x' \notin A$, and $u \in \mathbb{U}(x)$, then A is a closed set under any policy $\pi \in \Pi$.

Now, given a set $A \subset \mathbb{X}$, we have the following result about closed sets and its proof is included in Appendix IX.

Proposition II.2: Assume A is a closed set under $\pi \in \Pi$. Then, for any $x \in \mathbb{X}$

$$\lim_{t\to\infty} P_x^\pi(X_t \in A) = P_x^\pi(\tau_A < \infty).$$

The importance of the previous proposition is that it allows to express the probability of an event that depends on the joint distribution of the process, in terms of probabilities of events that only depend on the marginal distributions. Therefore, we have the following theorem, which is the main result of this section. Consider a Markov control model with long-run average reward given by the function $r = \mathbf{1}_A$, the indicator function of set A. Hence, for $x \in \mathbb{X}$ and $\pi \in \Pi$ the associated long-run average reward is given by

$$v^\pi(x) := \limsup_{N\to\infty} \frac{1}{N}\mathbb{E}_x^\pi\left[\sum_{t=0}^{N-1} \mathbf{1}_A(X_t)\right]. \tag{2}$$

Theorem II.3: Let $x \in \mathbb{X}$ and assume that A is a closed set under $\pi \in \Pi$. Then, $v^\pi(x) = P_x^\pi(\tau_A < \infty)$.

Proof: For any $N \in \mathbb{N}$, we have that

$$\frac{1}{N}\mathbb{E}_x^\pi\left[\sum_{t=0}^{N-1} \mathbf{1}_A(X_t)\right] = \frac{1}{N}\sum_{t=0}^{N-1} P_x^\pi(X_t \in A).$$

Taking lim sup on both sides we get

$$v^\pi(x) = \limsup_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} P_x^\pi(X_t \in A).$$

Now, recall that given a sequence $\{s_t\}_{t\geq 0}$ such that the limit L exists, then the Cesàro limit also exists and it is equal to L, i.e.,

$$\lim_{N\to\infty} \frac{s_0 + \ldots + s_{N-1}}{N} = L.$$

Since A is closed under π, Proposition II.2 implies that

$$\lim_{t\to\infty} P_x^\pi(X_t \in A) = P_x^\pi(\tau_A < \infty).$$

Therefore,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} P_x^\pi(X_t \in A) = P_x^\pi(\tau_A < \infty).$$

III. Domain of Attraction

Given a Markov control model $(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x)|x \in \mathbb{X}\}, Q)$, the first problem that we consider is the characterization of the domain of attraction and the escape set of a set $A \subset \mathbb{X}$, which are defined in Definition I.1. To ensure some stability we assume the following.

Assumption III.1: Set A is closed under some policy $\pi \in \Pi$.

The previous Assumption is in fact easy to satisfy by modifying the stochastic kernel as in (3) below. The idea is to characterize both sets in terms of value functions. The first question we ask is whether we can restrict our attention to policies that make A a closed set.

Definition III.2: Let $\Pi_A \subset \Pi$ be the set of control polices that make A a closed set.

The following proposition, proved in Appendix XI, is fundamental to answer the question mentioned above.

Proposition III.3: Given $\pi \in \Pi$ there exists a policy $\pi' \in \Pi_A$ such that for any $x \in \mathbb{X}$ and $t \geq 0$ it holds that $P_x^\pi(X_t \in A) \leq P_x^{\pi'}(\tau_A \leq t)$. Furthermore, for any $x \in \mathbb{X}$

$$\liminf_{t\to\infty} P_x^\pi(X_t \in A) \leq P_x^{\pi'}(\tau_A < \infty).$$

Then, we can focus on policies that make A a closed set. Let us consider the average reward function v^π defined in (2) and define the value function $V^*(x) = \sup_{\pi\in\Pi} v^\pi(x)$. Given any $\pi \in \Pi$, let $\pi' \in \Pi_A$ be the policy defined in Lemma XI.1. If $x \in A$ it follows that $v^\pi(x) \leq v^{\pi'}(x)$ and if $x \notin A$, by Proposition III.3 and Theorem II.3, we obtain that $v^\pi(x) \leq v^{\pi'}(x)$. Hence, $V^*(x) = \sup_{\pi\in\Pi_A} v^\pi(x)$. We obtain the following description of our sets of interest.

Theorem III.4: The following characterization of the sets hold:

$$\Lambda_A = \left\{x \in \mathbb{X} \,\middle|\, V^*(x) > 0\right\}, \quad \Gamma_A = \left\{x \in \mathbb{X} \,\middle|\, V^*(x) = 0\right\}$$

and

$$\Lambda_A^p = \left\{x \in \mathbb{X} \,\middle|\, V^*(x) \geq p\right\}.$$

Proof: Let $x \in \Lambda_A$ so there exists a policy $\pi \in \Pi$ such that $\liminf_{t\to\infty} P_x^\pi(X_t \in A) > 0$. Let $\pi' \in \Pi_A$ be the policy given by the proposition. Therefore, $0 < \liminf_{t\to\infty} P_x^\pi(X_t \in A) \leq P_x^{\pi'}(\tau_A < \infty)$ and $V^*(x) > 0$. The other inclusion is clear by Proposition II.2. Similarly, the result holds for the $p-$domain of attraction.

IV. Reach-Avoid Problem

The problem described in the previous section can be seen as a feasibility problem. In this section, we will solve a related maximization problem, namely Problem (P1). Consider a Markov control model $(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x)|x \in \mathbb{X}\}, Q)$. Let $A, B \subset \mathbb{X}$ be disjoint sets, with their respective hitting times τ_A, τ_B, and an initial distribution ν over the state space. We will show that this problem is equivalent to a long-run average reward problem with a particular reward function. Our first step will be to rewrite the objective function of (P1). To achieve this, let us define a modified stochastic kernel that make sets A and B closed under any policy π. Given a pair $(x, u) \in \mathbb{K}$ construct the stochastic kernel $\widetilde{Q}$ as follows:

$$\begin{cases} \widetilde{Q}(A|x,u) = 1, & \text{if } x \in A \\ \widetilde{Q}(B|x,u) = 1, & \text{if } x \in B \\ \widetilde{Q}(\cdot|x,u) = Q(\cdot|x,u) & \text{otherwise.} \end{cases} \tag{3}$$

The measure induced by $\widetilde{Q}$ will be denoted as $\widetilde{P}$ and $\widetilde{\mathbb{E}}$ will denote the expectation with respect to $\widetilde{P}$. By Remark II.1, the sets A and B are closed for any control policy in Π. Then, we have the following result proved in Appendix XIII.

Proposition IV.1: Given a policy $\pi \in \Pi$, we have that

$$P_\nu^\pi(\tau_A < \tau_B, \tau_A < \infty) = \widetilde{P}_\nu^\pi(\tau_A < \infty). \tag{4}$$

Now, we consider the Markov control model $(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x)|x \in \mathbb{X}\}, \widetilde{Q}, \mathbf{1}_A)$, i.e., the given Markov model with the modified kernel and with the characteristic function of set A as reward function. Let $\widetilde{v}^\pi(x)$ be as in (2) with the modified stochastic kernel, hence, the following theorem is a direct consequence of (4) and Theorem II.3.

Theorem IV.2:

$$\begin{aligned}\sup_{\pi\in\Pi} P_\nu^\pi(\tau_A < \tau_B, \tau_A < \infty) &= \sup_{\pi\in\Pi_M} \sum_{x\in\mathbb{X}} \widetilde{v}^\pi(x)\nu(x) \\ &= \sum_{x\in\mathbb{X}} \widetilde{V}(x)\nu(x)\end{aligned}$$

where $\widetilde{V}(x) = \sup_{\pi\in\Pi} \widetilde{v}^\pi(x)$.

The importance of the result abovementioned is that the problem of maximizing the probability of reaching some set A while avoiding a set B can be cast as a long-run average reward problem over stationary Markovian policies [see (1)].

V. Reach With Hitting Constraint

In this section, we will solve a constrained version of the previous problem, that is Problem (P2). So, again consider a given Markov control model $(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x)|x \in \mathbb{X}\}, Q)$, along with $A, B \subset \mathbb{X}$ disjoint sets and an initial distribution ν. Our objective will be to find a control policy that maximizes the probability of reaching A in such a way

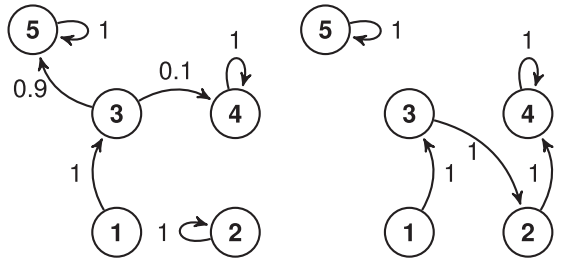

Fig. 1. Control Matrices u_1, u_2.

that the probability of reaching B is less than some $\epsilon > 0$. In this case, however, a Markovian policy might not be the best to solve this problem (recall Theorem IV.2). The following example shows this fact. To avoid cumbersome notation, in the sequel, we will use π to denote the stochastic kernels associated with such policy.

Example. Assume $\mathbb{X} = \{1, 2, 3, 4, 5\}$, $\mathbb{U} = \{u_1, u_2\}$ and $\mathbb{U}(x) = \mathbb{U}$ for all x. The corresponding control matrices are described in Fig. 1. Consider the sets $A = \{4\}, B = \{1, 2\}$. Assuming the uniform initial distribution let us consider the problem

$$\sup_{\pi\in\Pi} P_\nu^\pi(\tau_A < \infty)$$
$$\text{s.t. } P_\nu^\pi(\tau_B < \infty) \leq 0.5.$$

First of all, regardless of the kind of policy (Markovian or not), we must have that for any policy π, $P_1^\pi(\tau_B < \infty) = P_2^\pi(\tau_B < \infty) = 1, P_4^\pi(\tau_B < \infty) = P_5^\pi(\tau_B < \infty) = 0$. Thus, to satisfy the restriction, we must have that $P_3^\pi(\tau_B < \infty) \leq 0.5$. Moreover, since $\pi(u_2|3) = P^\pi(X_1 = 2|X_0 = 3) = P_3^\pi(\tau_B < \infty)$ then we need that $\pi(u_2|3) \leq 0.5$. We also have that for any policy

$$P_4^\pi(\tau_A < \infty) = 1, \quad P_5^\pi(\tau_A < \infty) = 0.$$

Thus, a first obvious choice to maximize $P_\nu^\pi(\tau_A < \infty)$ is to select $\pi(u_2|2) = 1$, so that $P_2^\pi(\tau_A < \infty) = 1$. Let π be a Markovian policy, so that $P_1^\pi(\tau_A < \infty) = P_3^\pi(\tau_A < \infty)$. If we select $\pi(u_2|3) = 1$, we would obtain $P_1^\pi(\tau_A < \infty) = P_3^\pi(\tau_A < \infty) = 1$. However, since $\pi(u_2|3) > 0.5$ we can not select such a policy. Moreover, note that for any policy for which $\pi(u_2|3) < 1$, we would have $P_1^\pi(\tau_A < \infty) = P_3^\pi(\tau_A < \infty) < 1$.

On the other hand, if we allow non-Markovian policies we can define a policy for which the value of $P_3^\pi(\tau_A < \infty)$ is the same as if we have used a Markovian policy, but $P_1^\pi(\tau_A < \infty) = 1$. Indeed, note that

$$\begin{aligned} &P_1^\pi(\tau_A < \infty) \\ &\quad = P_1^\pi(X_1 = 3, X_2 = 4) + P_1^\pi(X_1 = 3, X_2 = 2, X_3 = 4) \\ &\quad = \sum_{a_0,a_1\in\mathbb{U}} P_1^\pi(a_0, 3, a_1, 4) + \sum_{a_0,a_1,a_2\in\mathbb{U}} P_1^\pi(a_0, 3, a_1, 2, a_2, 4). \end{aligned}$$

By conditioning, the terms in the first sum can be written as

$$Q(4|3, a_1)\pi(a_1|h_1)Q(3|1, a_0)\pi(a_0|1).$$

Similarly, the terms in the second sum can be written as

$$Q(4|2, a_2)\pi(a_2|h_2)Q(2|3, a_1)\pi(a_1|h_1)Q(3|1, a_0)\pi(a_0|1).$$

If we define $\pi(u_2|1) = \pi(u_2|1, u_2, 3) = \pi(u_2|1, u_2, 3, u_2, 2) = 1$, we obtain that $P_1^\pi(\tau_A < \infty) = 1$. Now, we select $\pi(u_2|3) \leq 0.5$ so that the restriction is satisfied. Repeating the same procedure as before we obtain

$$\begin{aligned} P_3^\pi(\tau_A < \infty) &= P_3^\pi(X_1 = 4) + P_3^\pi(X_1 = 2, X_2 = 4) \\ &= \sum_{a_0\in\mathbb{U}} Q(4|3, a_0)\pi(a_0|3) \\ &+ \sum_{a_0,a_1\in\mathbb{U}} Q(4|2, a_1)\pi(a_1|h_1)Q(2|3, a_0)\pi(a_0|3). \end{aligned}$$

Thus, by defining $\pi(a_1|h_1)$ in such a way that only depends of the previous state, we can obtain the same value for $P_3^\pi(\tau_A < \infty)$ as if we have worked with Markovian policies. Therefore, $P_\nu^\pi(\tau_A < \infty)$ for such policy is bigger than for any stationary Markovian policy.

The key point is to realize that if the process $\{X_t\}$ has already gone through the set B, then it does not matter if the process hits the set again. But instead, if the chain has not gone through B it is better not to reach the set in order to satisfy the constraint. As a consequence, we will set the problem using control policies that remember whether the process has reached B or not.

Optimization problems with constraints can be rewritten using Lagrange multiplier. Hence, the problem above is equivalent to the problem $\sup_{\pi\in\Pi}\inf_{\lambda\geq 0}\mathcal{L}(\pi, \lambda)$, where

$$\mathcal{L}(\pi, \lambda) = P_\nu^\pi(\tau_A < \infty) + \lambda(\epsilon - P_\nu^\pi(\tau_B < \infty)).$$

The idea will be to write the Lagrangian function $\mathcal{L}$ as an average reward function with similar ideas as in the previous section. In order to do this, we consider the set $\widehat{\mathbb{X}} = \mathbb{X} \times \{0, 1\}$. Intuitively a state $(x, 0)$ indicates that the process has not reached B, while a state $(x, 1)$ indicates that the process has already reached B. Let $\mathbb{U}(x, i) = \mathbb{U}(x)$ for $i = 0, 1$ and $\widehat{\mathbb{K}}$ the corresponding set of feasible states and actions. We also define the stochastic kernel $\widehat{Q}$ on $\widehat{\mathbb{X}}$ given $\widehat{\mathbb{K}}$ as follows:

$$\begin{cases} \widehat{Q}((y,0)|(x,0),u) = Q(y|x,u), & \text{if } y \notin B \\ \widehat{Q}((y,0)|(x,0),u) = 0, & \text{if } y \in B \\ \widehat{Q}((y,1)|(x,0),u) = 0, & \text{if } y \notin B \\ \widehat{Q}((y,1)|(x,0),u) = Q(y|x,u), & \text{if } y \in B \\ \widehat{Q}((y,0)|(x,1),u) = 0 & \\ \widehat{Q}((y,1)|(x,1),u) = Q(y|x,u). & \end{cases} \tag{5}$$

Therefore, we have the augmented Markov control model $(\widehat{\mathbb{X}}, \mathbb{U}, \{\mathbb{U}(x, i)|(x, i) \in \widehat{\mathbb{X}}\}, \widehat{\mathbb{K}}, \widehat{Q})$. Let $\widehat{\Pi}$ be the set of policies over the augmented model, $\widehat{\Pi}_M$ the set of Markovian policies, and for any policy $\widehat{\pi} \in \widehat{\Pi}$ we denote by $\widehat{P^{\widehat{\pi}}}$ the measure induced by the policy. The corresponding $\widehat{\mathbb{X}} \times \mathbb{U}$-valued stochastic process is denoted by $\{(X_t, I_t, U_t)\}_{t\geq 0}$. Histories in this model are denoted by $\widehat{h}_t = (h_t, \overline{i}^t)$, with $\overline{i}^t$ the history up to time t of the process $\{I_t\}$.

Definition V.1: Given a $\pi \in \Pi$ we define a policy $\widehat{\pi} \in \widehat{\Pi}$ as follows: for any $t \geq 0$

$$\widehat{\pi}(u|\widehat{h}_t) = \pi(u|h_t)$$

whenever $\widehat{h}_t$ satisfies that $\overline{x}^t \notin B$ and $\overline{i}^t \notin \{1\}$, or there is some $0 \leq s \leq t$ such that $\overline{x}^{s-1} \notin B, x_s \in B, \overline{i}^{s-1} \notin \{1\}$ and $i_r = 1$ for $s \leq r \leq t$. Otherwise, we define $\widehat{\pi}(u|\widehat{h}_t) = \delta_{u_0}$ for any $u_0 \in \mathbb{U}(x_t)$. Note that $s = 0$ implies that $\overline{i}^t \notin \{0\}$.

The next result allows to express joint distributions of the original model in terms of joint distributions of the augmented model. Its proof can be found in Appendix XV.

Lemma V.2: Let $\pi \in \Pi$ and consider the policy $\widehat{\pi} \in \widehat{\Pi}$ of the previous definition. Given $h_t \in H_t$ and $u_t \in \mathbb{U}(x_t)$, let $\widehat{h}_t = ((x, 0), u_0, \ldots, (x_{t-1}, 0), u_{t-1}, (x_t, 0))$ if $\overline{x}^t \notin B$, and $\widehat{h}_t = ((x, 0), \ldots, (x_{s-1}, 0), u_{s-1}, (x_s, 1), \ldots, (x_t, 1))$ if $\overline{x}^{s-1} \notin B, x_s \in$

B. Then

$$P_x^{\pi}(h_t,u_t)=\begin{cases}\widehat{P}^{\widehat{\pi}}_{(x,0)}(\widehat{h}_t,u_t),\ \textit{if } x\notin B\\ \widehat{P}^{\widehat{\pi}}_{(x,1)}(\widehat{h}_t,u_t),\ \textit{if } x\in B.\end{cases}$$

Note that by the definition of $\widehat{Q}$, the set $\mathbb{X}\times\{1\}$ is closed under every policy in the augmented Markov model. Also, as in Section IV, we can redefine the kernel $\widehat{Q}$ to make the set $A\times\{0,1\}$ closed under any policy, i.e., $\widehat{Q}(A\times\{i\}|(x,i),u)=1$ if $x\in A$ and $i=0,1$. With this redefined kernel, we consider the Markov model along with the reward function

$$r(x,i)=(\mathbf{1}_{A\times\{0,1\}}-\lambda\mathbf{1}_{\mathbb{X}\times\{1\}})(x,i).$$

Given a policy $\widehat{\pi}\in\widehat{\Pi}$ and $(x,i)\in\widehat{\mathbb{X}}$, we consider the average reward function given by

$$v^{\widehat{\pi}}(x,i):=\limsup_{N\to\infty}\frac{1}{N}\widehat{\mathbb{E}}^{\widehat{\pi}}_{(x,i)}\left[\sum_{t=0}^{N-1}[\mathbf{1}_{A\times\{0,1\}}-\lambda\mathbf{1}_{\mathbb{X}\times\{1\}}](X_t,I_t)\right]. \tag{6}$$

The following theorem proved in Appendix XVII shows that the Lagrangean $\mathcal{L}$ can be written in terms of this function.

Theorem V.3: Let $\pi\in\Pi$ and $\widehat{\pi}\in\widehat{\Pi}$ the policy defined in Definition V.1. Let $x\in\mathbb{X}$, then

$$P_x^{\pi}(\tau_A<\infty)+\lambda\left(\epsilon-P_x^{\pi}(\tau_B<\infty)\right)=\begin{cases}v^{\widehat{\pi}}(x,0)+\lambda\epsilon,\ \textit{if } x\notin B\\ v^{\widehat{\pi}}(x,1)+\lambda\epsilon,\ \textit{if } x\in B.\end{cases}$$

Now, since Definition V.1 does not necessarily recover all policies in $\widehat{\Pi}$, the previous theorem shows that the following is an upper bound for (P2):

$$\sup_{\widehat{\pi}\in\widehat{\Pi}}\inf_{\lambda\geq 0}\lambda\epsilon+\sum_{x\notin B}v^{\widehat{\pi}}(x,0)\nu(x)+\sum_{x\in B}v^{\widehat{\pi}}(x,1)\nu(x). \tag{7}$$

Also, note that in the problem abovementioned, we can consider only Markovian policies by (1). Furthermore, given a policy $\widehat{\pi}\in\widehat{\Pi}_M$ and stationary, we can define a policy $\pi\in\Pi$ by

$$\pi(u|h_t)=\begin{cases}\widehat{\pi}(u|x_t,0), & \text{if } \overline{x}^t\notin B\\ \widehat{\pi}(u|x_t,1), & \text{otherwise.}\end{cases} \tag{8}$$

Let the set of policies obtained by (8) be denoted as $\Pi_B\subset\Pi$, i.e., policies that do not depend on the whole history but only whether the process has gone through set B or not. Hence, Definition V.1 and (8) define a one-to-one relation between Π_B and $\widehat{\Pi}_M$.

Theorem V.4: Problem (P2) is equivalent to

$$\sup_{\pi\in\Pi_B}P_\nu^{\pi}(\tau_A<\infty)\quad s.t.\quad P_\nu^{\pi}(\tau_B<\infty)\leq\epsilon \tag{9}$$

its optimal value is equal to

$$\sup_{\widehat{\pi}\in\widehat{\Pi}_M}\inf_{\lambda\geq 0}\lambda\epsilon+\sum_{x\notin B}v^{\widehat{\pi}}(x,0)\nu(x)+\sum_{x\in B}v^{\widehat{\pi}}(x,1)\nu(x) \tag{10}$$

and any optimal policy of (10) produces an optimal policy of (P2) through (8).

Proof: Let P^* be the optimal value of Problem (P2). Then, P^* is bounded previously by (7), which has the same value as (10), which by (8) has the same value as the optimal value of (9), which is bounded above by P^*. Hence, all these expressions are equivalent.

When state and action spaces are finite we can interchange sup and inf in (10) as we will see in the following section.

VI. Linear Programming Formulations: Finite Case

When the state space $\mathbb{X}$ is finite and $\mathbb{U}(x)$ is finite for all $x\in\mathbb{X}$, solutions of optimality equations (MC1)-(MC2) and optimal stationary policies can be found via linear programs, see [1, Sec. 9.3] for further details. Hence, using Theorem III.4, to be able to find the p-domain of attraction of a given set A, we need to solve the following linear program:

$$\begin{aligned}&\min\sum_{x\in\mathbb{X}}v(x)\\ &\text{s.t.}\\ &\quad v(x)\geq\sum_{j\in\mathbb{X}}P^u_{x,x'}v(x')\\ &\quad v(x)\geq\mathbf{1}_A(x)+\sum_{x'\in\mathbb{X}}P^u_{x,x'}h(x')-h(x)\\ &\quad\forall x\in\mathbb{X},u\in\mathbb{U}(x)\end{aligned}\tag{ReaP}$$

where $P^u_{x,j}=Q(j|x,u)$. Similarly, in order to solve Problem (P1), from Theorem IV.2, we need to solve the same linear program as above with $\widetilde{P}^u_{x,j}=\widetilde{Q}(j|x,u)$ instead of $P_{x,j}$, where $\widetilde{Q}$ is defined in (3). In both cases, an optimal stationary Markovian policy can be found by solving the corresponding dual problem

$$\begin{aligned}&\max\sum_{x\in A}\sum_{u\in\mathbb{U}(x)}\alpha(x,u)\\ &\text{s.t.}\\ &\sum_{u\in\mathbb{U}(x)}\alpha(x,u)-\sum_{\substack{x'\in\mathbb{X}\\ u\in\mathbb{U}(x')}}P^u_{x',x}\alpha(x,'u)=0\\ &\sum_{u\in\mathbb{U}(x)}\alpha(x,u)+\beta(x,u)-\sum_{\substack{x'\in\mathbb{X}\\ u\in\mathbb{U}(x')}}P^u_{x',x}\beta(x,'u)=1\\ &\alpha(x,u)\geq 0,\beta(x,u)\geq 0,\quad\forall x\in\mathbb{X},u\in\mathbb{U}(x).\end{aligned}\tag{ReaD}$$

Given (α^*,β^*) optimal solution of (ReaD), an optimal stationary policy $\pi=\{\mu\}$ is

$$\mu(u^+|x)=\begin{cases}\dfrac{\alpha^*(x,u^+)}{\sum_{u\in\mathbb{U}(x)}\alpha^*(x,u)},\ \text{if}\ \sum_{u\in\mathbb{U}(x)}\alpha^*(x,u)>0\\ \dfrac{\beta^*(x,u^+)}{\sum_{u\in\mathbb{U}(x)}\beta^*(x,u)},\ \text{otherwise.}\end{cases}$$

Remark VI.1: Note that in both cases the initial distribution ν does not play any role in order to find either the value functions V^* and $\widetilde{V}$ and the optimal policies.

For Problem (P2), the linear programming formulation is not as straightforward, as in the previous problems. In particular, we would like to switch the inf with the sup in (10), i.e., we would like to show that it is equivalent to its dual problem. Recall that (P2) can be written as

$$P^*=\sup_{\pi\in\Pi_B}\inf_{\lambda\geq 0}P_\nu^{\pi}(\tau_A<\infty)+\lambda(\epsilon-P_\nu^{\pi}(\tau_B<\infty))$$

which is bounded previously by the optimal value of its dual problem

$$D^*=\inf_{\lambda\geq 0}\sup_{\pi\in\Pi_B}P_\nu^{\pi}(\tau_A<\infty)+\lambda(\epsilon-P_\nu^{\pi}(\tau_B<\infty)).$$

By Theorem V.4, we can write the above problem as

$$\inf_{\lambda\geq 0} \lambda\epsilon + \max \sum_{\substack{j\in A\times\{0,1\},\\ u\in\mathbb{U}(j)}} \alpha(j,u) - \lambda \sum_{\substack{j\in\mathbb{X}\times\{1\}\\ u\in\mathbb{U}(j)}} \alpha(j,u)$$

s.t.

$$\sum_{u\in\mathbb{U}(j)} \alpha(j,u) - \sum_{\substack{j'\in\widehat{\mathbb{X}}\\ u\in\mathbb{U}(j')}} \widehat{P}^u_{j,'j}\alpha(j,u) = 0$$

$$\sum_{u\in\mathbb{U}(j)} \alpha(j,u) + \beta(j,u) - \sum_{\substack{j'\in\widehat{\mathbb{X}}\\ u\in\mathbb{U}(j')}} \widehat{P}^u_{j,'j}\beta(j,'u) = \widehat{\nu}(j)$$

$$\alpha(j,u)\geq 0, \beta(j,u)\geq 0, \quad \forall j\in\widehat{\mathbb{X}}, u\in\mathbb{U}(j) \tag{D1}$$

with $\widehat{P}^u_{(x,i),(x,'i')} = \widehat{Q}((x,'i')|(x,i),u)$, where $\widehat{Q}$ is defined in (5), and

$$\widehat{\nu}(x,i) = \begin{cases} \nu(x), & \text{if } x\notin B, i=0\\ \nu(x), & \text{if } x\in B, i=1\\ 0, & \text{otherwise.}\end{cases}$$

Now, by strong duality of linear programming we further obtain that

$$D^* = \max \sum_{\substack{j\in A\times\{0,1\},\\ u\in\mathbb{U}(j)}} \alpha(j,u)$$

s.t.

$$\sum_{u\in\mathbb{U}(j)} \alpha(j,u) - \sum_{\substack{j'\in\widehat{\mathbb{X}}\\ u\in\mathbb{U}(j')}} \widehat{P}^u_{j,'j}\alpha(j,u) = 0$$

$$\sum_{u\in\mathbb{U}(j)} \alpha(j,u) + \beta(j,u) - \sum_{\substack{j'\in\widehat{\mathbb{X}}\\ u\in\mathbb{U}(j')}} \widehat{P}^u_{j,'j}\beta(j,'u) = \widehat{\nu}(j)$$

$$\sum_{\substack{j\in\mathbb{X}\times\{1\}\\ u\in\mathbb{U}(j)}} \alpha(j,u) \leq \epsilon$$

$$\alpha(j,u)\geq 0, \beta(j,u)\geq 0, \quad \forall j\in\widehat{\mathbb{X}}, u\in\mathbb{U}(j). \tag{D2}$$

If Problem (D2) is infeasible, then $D^* = -\infty$ and, therefore, $P^* = -\infty$, i.e., Problem (P2) is infeasible. On the other hand, if (D2) is finite (note that it cannot be unbounded) with optimal solution (α^*, β^*), the inf over $\lambda\geq 0$ in (D1) is attained at some λ^* such that

$$\lambda^*\left(\epsilon - \sum_{x\in\mathbb{X}, u\in\mathbb{U}(x)} \alpha^*((x,1),u)\right) = 0$$

by complementary slackness condition. Let $\widehat{\pi}^*\in\widehat{\Pi}_M$ the stationary optimal policy induced by (α^*,β^*) and $\pi^*\in\Pi_B$ its corresponding over $\mathbb{X}$ given by (8). Therefore

$$P^* = \sup_{\pi\in\Pi_B}\inf_{\lambda\geq 0} P^\pi_\nu(\tau_A<\infty) + \lambda(\epsilon - P^\pi_\nu(\tau_B<\infty))$$

$$\geq \inf_{\lambda\geq 0} P^{\pi^*}_\nu(\tau_A<\infty) + \lambda(\epsilon - P^{\pi^*}_\nu(\tau_B<\infty))$$

$$= \inf_{\lambda\geq 0} \lambda\epsilon + \sum_{j\in\widehat{\mathbb{X}}} v^{\widehat{\pi^*}}(j)\widehat{\nu}(j)$$

$$= \inf_{\lambda\geq 0} \lambda\epsilon + \sum_{\substack{j\in A\times\{0,1\},\\ u\in\mathbb{U}(j)}} \alpha^*(j,u) - \lambda \sum_{\substack{j\in\mathbb{X}\times\{1\}\\ u\in\mathbb{U}(j)}} \alpha^*(j,u)$$

$$= \lambda^*\epsilon + \sum_{\substack{j\in A\times\{0,1\},\\ u\in\mathbb{U}(j)}} \alpha^*(j,u) - \lambda \sum_{\substack{j\in\mathbb{X}\times\{1\}\\ u\in\mathbb{U}(j)}} \alpha^*(j,u)$$

$$= \sum_{\substack{j\in A\times\{0,1\},\\ u\in\mathbb{U}(j)}} \alpha^*(j,u) = D^* \geq P^*.$$

Hence, we just proved the following result.

Theorem VI.2: Suppose $\mathbb{X}, \mathbb{U}$ are finite. Then, (P2) satisfies strong duality, i.e., it is equivalent to $\min_{\lambda\geq 0}\max_{\pi\in\Pi}\mathcal{L}(\pi,\lambda)$. Furthermore, it can be solved by the linear program (D2) in the augmented model and recover an optimal policy in Π_B through (8).

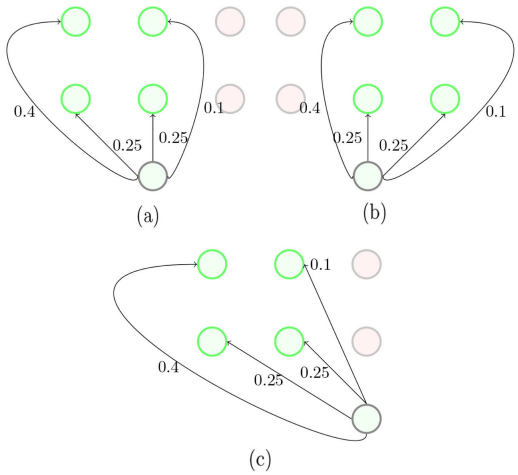

Fig. 2. Controls. (a) Up control. (b) Right control. (c) Left control.

TABLE I
CPU TIME (S)

Dom. Att.	P1		P2		
				(1, 10)	(17, 60)
35.6	(50, 1)	(1, 30)	$\epsilon = 0.01$	13.8	8.9
	10.79	11.04	$\epsilon = 0.2$	12.9	5.5
			$\epsilon = 0.8$	10.25	5.8

VII. NUMERICAL EXAMPLES

To illustrate our results, we consider an object that navigates over a grid under the influence of a north-west wind. The object has three controls available, as shown in Fig. 2. We assume that the states at the upper boundary of the grid are absorbing states. The controls are modified in the left and right boundaries to ensure the object does not leave the grid. This example is similar to the Zermelo navigation problem presented in [14] in the context of continuous time problems. The CPU time needed to solve each problem can be found in Table I.

A. Domain of Attraction

In order to show the findings of Section III, we consider a 100 by 100 grid with a closed set A in the central region of the grid marked with black squares. Fig. 3 shows the surface and level sets of the function $V^*(x)$, which defines the p-domains Λ_p. The escape set Γ corresponds to the states with value function equal to zero. This function was found by solving the linear program (ReaP). We note that no state outside of A belongs to Λ_1.

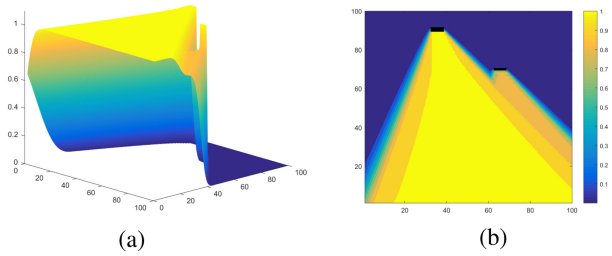

Fig. 3. p-domains Λ_p and escape set Γ. (a) Surface of $V^*(x)$. (b) Level sets of $V^*(x)$.

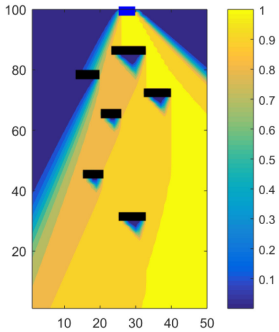

Fig. 4. Level sets of $\widetilde{V}(x)$.

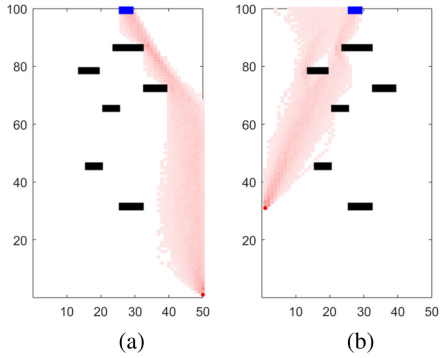

Fig. 5. Trajectories. (a) Initial state (50,1). (b) Initial state (1,30).

B. Reach and Avoid

For Problem (P1), we consider a 100 by 50 grid with the set A a portion of the upper boundary of the grid, marked with dark blue squares, and set B a number of obstacles spread over the grid, marked with black squares. Fig. 4 shows the level sets of the function $\widetilde{V}(x)$ computed by solving the linear program (ReaP). In Fig. 5, we show the paths of 500 simulated trajectories of the object under the optimal policy obtained from the linear program (ReaD). We choose two different starting states from different level sets according to Fig. 4. Note that in the first case all trajectories hit the target set A, while in the second case most of them drift away from the set. This situation agrees with the Fig. 4.

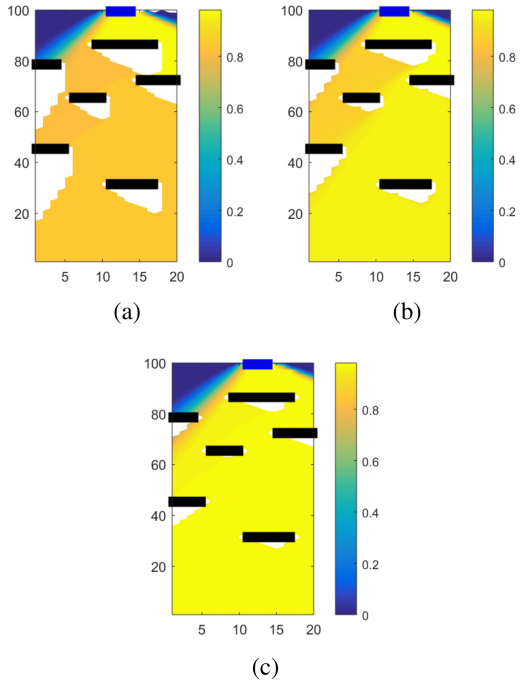

Fig. 6. Optimal value of (P2). (a) $\epsilon = 0.01$. (b) $\epsilon = 0.2$. (c) $\epsilon = 0.8$.

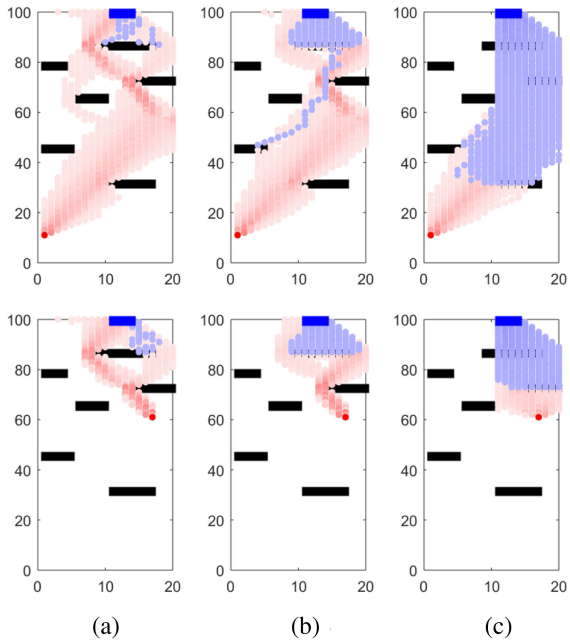

Fig. 7. Trajectories. (a) $\epsilon = 0.01$. (b) $\epsilon = 0.2$. (c) $\epsilon = 0.8$.

C. Reach With Constraint

For Problem (P2), we consider a 100 by 20 grid and sets A and B marked with dark blue and black squares, respectively. It is important to note that the initial distribution ν plays a key role in the feasibility of the problem. Fig. 6 shows the level sets of the function

$$\widehat{V}(x) := \max_{\pi \in \Pi_B} P_x^\pi(\tau_A < \infty) \quad \text{s.t.} \quad P_x^\pi(\tau_B < \infty) \leq \epsilon$$

for different values of ϵ. White regions represents the states for which the problem abovementioned is infeasible. As expected, the number of infeasible states decreases with bigger values of ϵ. In Fig. 7, we show the paths of 500 trajectories under the optimal policies with different initial distributions and values of ϵ. Red trajectories correspond to optimal trajectories before hitting set B and blue trajetories after hitting this set. Note also that blue trajectories do not avoid the obstacles since they were already hit.

VIII. Conclusion

We formulated reach/avoid problems in infinite horizon as long-run average reward problems in the context of MDPs and showed optimality of Markovian policies. We also considered a reachability problem with constrained probability of hitting a given set of states, where optimal polices are not in general Markovian, and show optimality of Markovian policies in an augmented state space. Finally, when state and actions spaces are finite, we use linear programs to find optimal policies.

Acknowledgment

All authors acknowledge financial support provided by the Vice Presidency for Research & Creation publication fund at the Universidad de los Andes.

References

[1] M. L. Puterman, *Markov Decision Processes: Siscrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.
[2] D. P. Bertsekas, "Dynamic programming and optimal control," *Athena Sci.*, vol. 1, p. 576, 1995.
[3] D. P. Bertsekas and S. E. Shreve, "Stochastic optimal control: The discrete time case," *Athena Sci.*, p. 330, 1996.
[4] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov Control Processes: Basic Optimality Criteria*. Berlin, Germany: Springer, 1996.
[5] E. Arvelo, E. Kim, and N. C. Martins, "Maximal persistent surveillance under safety constraints," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 4048–4053.
[6] E. Arvelo and N. Martins, "Maximizing the set of recurrent states of an MDP, subject to convex constraints," *Automatica*, vol. 50, no. 3, pp. 994–998, 2014.
[7] V. I. Zubov, *Methods of A.M. Lyapunov and their application*. Gronigen, The Netherlands: P. Noordhoff, 1964.
[8] F. Camilli, L. Grüne, and F. Wirth, "Control Lyapunov functions and Zubov's method," *SIAM J. Control Optim.*, vol. 47, no. 1, pp. 301–326, 2008.
[9] F. Camilli and P. Loreti, "A Zubov's method for stochastic differential equations," *Nonlinear Differ. Equ. Appl. NoDEA*, vol. 13, no. 2, pp. 205–222, 2006.
[10] H. M. Soner and N. Touzi, "Stochastic target problems, dynamic programming, and viscosity solutions," *SIAM J. Control Optim.*, vol. 41, no. 2, pp. 404–424, 2002.
[11] A. Abate, M. Prandini, J. Lygeros, and S. Sastry, "Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems," *Automatica*, vol. 44, no. 11, pp. 2724–2734, 2008.
[12] D. Chatterjee, E. Cinquemani, and J. Lygeros, "Maximizing the probability of attaining a target prior to extinction," *Nonlinear Anal., Hybrid Syst.*, vol. 5, no. 2, pp. 367–381, 2011.
[13] S. Summers and J. Lygeros, "Verification of discrete time stochastic hybrid systems: A stochastic reach-avoid decision problem," *Automatica*, vol. 46, no. 12, pp. 1951–1961, 2010.
[14] P. M. Esfahani, D. Chatterjee, and J. Lygeros, "The stochastic reach-avoid problem and set characterization for diffusions," *Automatica*, vol. 70, pp. 43–56, 2016.
[15] H. Zijm, "The optimality equations in multichain denumerable state Markov decision processes with the average cost criterion: The bounded cost case," *Statist. Decis.*, vol. 3, pp. 143–165, 1985.
[16] E. Mann, "Optimality equations and sensitive optimality in bounded Markov decision processes," *Optim., A J. Math. Program. Operations Res.*, vol. 16, no. 5, pp. 767–781, 1985.