Steady-State Policy Synthesis in Multichain Markov Decision Processes

George Atia[1] , **Andre Beckus**[1] , **Ismail Alkhouri**[1] and **Alvaro Velasquez**[2]
[1]Department of Electrical and Computer Engineering, University of Central Florida
[2]Information Directorate, Air Force Research Laboratory
george.atia@ucf.edu, {abeckus,ialkhouri}@knights.ucf.edu, alvaro.velasquez.1@us.af.mil

Abstract

The formal synthesis of automated or autonomous agents has elicited strong interest from the artificial intelligence community in recent years. This problem space broadly entails the derivation of decision-making policies for agents acting in an environment such that a formal specification of behavior is satisfied. Popular formalisms for such specifications include the quintessential Linear Temporal Logic (LTL) and Computation Tree Logic (CTL) which reason over infinite sequences and trees, respectively, of states. However, the related and relevant problem of reasoning over the frequency with which states are visited infinitely and enforcing behavioral specifications on the same has received little attention. That problem, known as Steady-State Policy Synthesis (SSPS) or steady-state control, is the focus of this paper. Prior related work has been mostly confined to unichain Markov Decision Processes (MDPs), while a tractable solution to the general multichain setting heretofore remains elusive. In this paper, we provide a solution to the latter within the context of multichain MDPs over a class of policies that account for all possible transitions in the given MDP. The solution policy is derived from a novel linear program (LP) that encodes constraints on the limiting distributions of the Markov chain induced by said policy. We establish a one-to-one correspondence between the feasible solutions of the LP and the stationary distributions of the induced Markov chains. The derived policy is shown to maximize the reward among the constrained class of stationary policies and to satisfy the specification constraints even when it does not exercise all possible transitions.

1 Introduction

There has been a focus in recent years on the verification of autonomous systems by leveraging techniques used for decades in the model checking of software [Fisher *et al.*, 2013]. While this verification step is crucial for the development of robust autonomous capabilities, a promising complementary approach is to design these capabilities in such a way that the search for a correct design is driven by the same specifications used for verification. This methodology is often called correct-by-design construction [Haesaert *et al.*, 2015] or formal/controller synthesis [Kress-Gazit *et al.*, 2018]. Our contribution is in the same vein and entails the search for policies which satisfy constraints on the steady-state distribution of the resulting agent as it interacts with its environment for an indefinite period of time following said policies. It is worth noting that progress in this area has interesting applications to problems where steady-state distributions are commonly used. This includes the derivation of maintenance plans for various systems such that asymptotic failure rate is minimized [Boussemart and Limnios, 2004] [Boussemart *et al.*, 2002] as well as to problems in constrained routing where average delay and packet loss metrics must be enforced [Lazar, 1983] [Skwirzynski, 1981].

Steady-State Policy Synthesis (SSPS) is framed in the context of constrained Markov Decision Processes (MDP) that model the agent-environment dynamics. This framework has long been studied in the stochastic dynamic control and operations research literature to handle multi-objective decision-making in the presence of uncertainty. The pioneering work of Derman [Derman, 1970] and Altman [Altman, 1999] developed a constrained optimization framework to dynamic control problems based on linear programming for both the discounted and total reward, as well as the expected average reward formulations. The vast majority of existing work, however, have focused on ergodic or unichain structures. This was pointed out recently by Altman in [Altman *et al.*, 2019], where it is stated that "*...the existing theory for solving such problems requires strong assumptions on the ergodic structure of the problem*". Under such assumptions, average-reward constrained MDPs have been shown to admit efficient solutions owing to an established one-to-one correspondence between the optimal solutions of a formulated linear program (LP) and the optimal policies of the MDP. The notable work of Kallenberg in [Kallenberg, 1983] has laid the groundwork for Markovian control problems and their characterizations in multichain settings and the construction of optimal policies based on linear programming under several optimality criteria. However, the algorithms developed to construct an optimal policy for general multichain structures were shown to be computationally prohibitive for the expected average reward formulation.

Summary of contributions. In this paper, we make three main contributions. First, we introduce the **M**ultich**a**in **St**eady-State Policy Synthesis over **E**dge **Pr**eserving Policies (MaStEr) problem as a generalization of the SSPS and steady-state control problems studied in [Velasquez, 2019] and [Akshay *et al.*, 2013], respectively. In particular, given a multichain MDP, we seek a policy in a predefined class of policies that maximizes an expected reward signal while enforcing steady-state specifications on the behavior of such a policy. In sharp contrast to [Akshay *et al.*, 2013], we dispense with the strong assumption about the ergodicity of the underlying MDP, which requires that every stationary deterministic policy induces an ergodic (i.e., recurrent and aperiodic) Markov chain. Further, unlike [Velasquez, 2019], we neither search for a policy that induces a recurrent Markov chain consisting of all the states within the given MDP, nor require the existence of such a chain. For example, it may very well be the case that such a recurrent chain does not exist if some states are inevitably transient. As our second contribution, we develop a scalable multi-step approach to synthesize a policy that provably meets the asymptotic steady-state specifications through the use of a novel linear programming formulation. Our third contribution lies in deriving important theoretical results, including sufficient conditions for the existence of a one-to-one correspondence between the feasible solutions of the proposed LP and the steady-state distributions of the Markov chains induced by the synthesized policies, and in turn establishing provable performance and behavior guarantees for the derived policy. To the best of our knowledge, this is the first work to allow synthesis of stationary steady-state policies with verifiable behavior in multichain MDPs.

2 Background

In this section, we introduce preliminary definitions and notation used throughout the paper. The vector e denotes the vector (of appropriate dimension) of all ones and ^{T} the transposition operator. Given a vector x and index set V, the vector $x_V := [x_v], v \in V$. By $|S|$, we denote the cardinality of a set S. For an integer $n > 0$, the set $[n] := \{1, \ldots, n\}$, and $A \setminus B$ denotes the set difference of sets A and B.

Definition 1 (Markov Chain)**.** *A Markov chain is a stochastic model given by a tuple $\mathcal{M} = (S, T, \beta)$, where S is the state space, T the transition function $T : S \times S \to [0, 1]$ with $T(s'|s)$ denoting the probability of transitioning from state s to state s', and β the initial state distribution. With slight abuse of notation, the transition function can also be thought of as a matrix $T \in [0, 1]^{|S|\times|S|}$, where the (s, s') entry $T(s, s') = T(s'|s)$. Its use will be clear from the context.*

Given a Markov chain $\mathcal{M} = (S, T, \beta)$, a state $s \in S$ is said to be *transient* if there is a non-zero probability of never returning to s given that we start in s. A set of transient states is termed a transient set. We define an *isolated* component I as a set of states in $\mathcal{M}$ that can never be visited, that is, $\beta_I = 0$ and I is not reachable from any state in $S \setminus I$, i.e., $\sum_{s' \in I} T(s'|s) = 0, \forall s \in S \setminus I$.

Definition 2 (Markov Decision Process (MDP))**.** *An MDP represented by the tuple $\mathcal{M} = (S, A, T, R, \beta)$ is a probabilistic automaton, in which S denotes the state space, A the set of actions, $T : S \times A \times S \to [0, 1]$ the transition function with $T(s'|s, a)$ denoting the probability of transitioning from state s to state s' under action a, $R : S \times A \times S \to \mathbb{R}$ a reward obtained when action a is taken in state s and we end up in state s', and β the initial distribution. We often use the alternative reward function $R : S \times A \to \mathbb{R}$, where $R(s, a) := \sum_{s' \in S} T(s'|s, a)R(s, a, s')$. By $A(s) \subseteq A$, we denote the set of actions available in state s.*

Definition 3 (Terminal Strongly Connected Component (TSCC))**.** *Consider the digraph formed by an arbitrary Markov chain or MDP $\mathcal{M}$ with state space S and initial distribution β. A Terminal Strongly Connected Component (TSCC) $S' \subseteq S$ is a strongly connected component reachable from some initial state $s, \beta(s) > 0$ and with no outgoing transitions to any state in $S \setminus S'$. We denote by $r(\mathcal{M}) = \bigcup_{k\in[m]} r_k(\mathcal{M})$ the set of states in the m TSCCs of $\mathcal{M}$, with $r_k(\mathcal{M}) \subseteq S$ denoting the k^{th} TSCC. The complement set is denoted by $\bar{r}(\mathcal{M}) := S \setminus r(\mathcal{M})$. In the case of Markov chains, this is the set of transient or isolated states.*

Definition 4 (Markov Chain Induced by a Policy)**.** *The tuple $\mathcal{M}_\pi = (S, T_\pi, \beta)$ is the Markov chain induced by a policy $\pi : S \times A \to [0, 1]$ in an underlying MDP $\mathcal{M} = (S, A, T, R, \beta)$, where $T_\pi(s'|s) = \sum_{s\in A(s)} T(s'|s, a)\pi(a|s)$ and $\pi(a|s)$ is the probability of taking action a in state s.*

Definition 5. *An MDP $\mathcal{M}$ is called unichain if the Markov chain $\mathcal{M}_\pi$ induced by any admissible deterministic policy π is unichain, that is, consists of exactly one closed recurrent set and possibly some transient states. An MDP is said to be multichain if it is not unichain.*

Definition 6. *Given an MDP $\mathcal{M}$ and policy π, the steady-state distribution $\Pr^\infty_\pi : S \times A \to [0, 1]$ over the state-action pairs (also known as the occupation measure) is the long-term proportion of time spent in state-action pair (s, a) as the number of transitions approaches ∞, i.e.,*

$$\Pr^\infty_\pi(s, a) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \Pr(S_t = s, A_t = a|\beta, \pi) \quad (1)$$

where S_t and A_t are the state and action at time t. Also, $\Pr^\infty_\pi(s) := \sum_{a\in A(s)} \Pr^\infty_\pi(s, a)$ is the steady-state probability of being in state $s \in S$.

Definition 7 (Steady-State Specification [Velasquez, 2019])**.** *Given an MDP $\mathcal{M} = (S, A, T, R, \beta)$ and a set of labels $L = \{L_1, \ldots, L_{n_L}\}$, where $L_i \subseteq S$, a set of steady-state specifications is given by $\Phi^\infty_L = \{(L_i, [l_i, u_i])\}^{n_L}_{i=1}$. Given a policy π, the specification $(L_i, [l, u]) \in \Phi^\infty_L$ is satisfied if and only if $\sum_{s\in L_i} \Pr^\infty_\pi(s) \in [l, u]$; that is, if the steady-state probability of being in a state $s \in L_i$ in the Markov chain $\mathcal{M}_\pi$ falls within the interval $[l, u]$. An MDP $\mathcal{M} = (S, A, T, R, \beta, L, \Phi^\infty_L)$ augmented with the label set L and specifications Φ^∞_L is termed a labeled MDP (LMDP).*

Lemma 1. *[Kallenberg, 1983] Given an MDP $\mathcal{M} = (S, A, T, R, \beta)$ and policy $\pi \in \Pi_S$, where Π_S is the set of stationary policies, the steady-state distribution $\Pr^\infty_\pi := \{\Pr^\infty_\pi(s, a)\}_{s,a}$ of the Markov chain $\mathcal{M}_\pi$ is given by (2), where $T^\infty_\pi := \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} T^t_\pi$ is the Cesàro limit.*

$$\Pr^\infty_\pi(s, a) = (\beta^T T^\infty_\pi)_s \pi(a|s),\ s \in S, a \in A(s) \quad (2)$$

3 Related Work

The problem of arbitrating control in discrete-time MDPs has been studied for many decades, with LP solutions based on occupation measures being proposed in [Manne, 1960; De Ghellinck, 1960] for unichain MDPs. While such LPs can easily handle steady-state constraints in unichain settings, serious issues arise when a similar approach is used for multichain MDPs, as described in the pioneering work [Kallenberg, 1983]. In particular, it was shown that there is not a one-to-one correspondence between the feasible solutions of the augmented LP and the stationary policies. Instead, the space of feasible solutions is partitioned into equivalence classes of various feasible solutions mapping to the same policy. The key deficiency is that the steady-state distribution of the Markov chain induced by the synthesized policy does not match the optimal solution to the LP in general, and so the derived policy does not always meet the steady-state specification constraints.

To work around these difficulties, existing works either impose severe restrictions on the input MDP, or allow the production of troublesome non-stationary policies. One common restriction excludes multichain MDPs altogether. The most restrictive requirement can be found in [Akshay *et al.*, 2013] and the reinforcement learning algorithms of [Bhatnagar and Lakshmanan, 2012], which assume that the input MDP is ergodic or irreducible, respectively. These MDPs guarantee that the Markov chain induced by any stationary and deterministic solution policy is itself ergodic or irreducible, thereby ensuring a valid solution to the steady-state equations. The authors proceed to find policies in such MDPs which satisfy a set of steady-state specifications. Though the steady-state control problem defined therein is described in the full generality of Markovian and history-dependent policies, an equivalence between the two is established and the solutions proposed focus on the latter. The works of [Ross, 1989; Altman, 1999; Feinberg, 2009] contain a slight relaxation in that unichain MDPs are allowed. Nonetheless, the strict underlying requirement of a single recurrent class remains.

The unichain MDP assumption is dissolved in [Velasquez, 2019], where the SSPS problem is introduced as a generalization of steady-state control. The solution proposed finds a policy which induces a recurrent Markov chain and maximizes the average reward objective while satisfying a set of steady-state specifications, if one exists. The work in [Velasquez, 2019], however, cannot reason about multichain MDPs, as it requires the *existence* of a policy which induces a strongly connected Markov chain containing all states in the MDP and optimizes over such policies. Therefore, this effort fails to produce a policy in the most general setting.

To eliminate restrictions on the input MDP, another line of work allows non-stationary policies as output. The first example of this appears in [Kallenberg, 1983], where the proposed algorithm produces a policy with a different decision rule for each time step. Not only is the policy impractical to apply, but the algorithm itself is computationally intractable.

In contrast to the aforementioned methods, our approach is computationally tractable, works with the most general MDPs, and always produces a stationary policy.

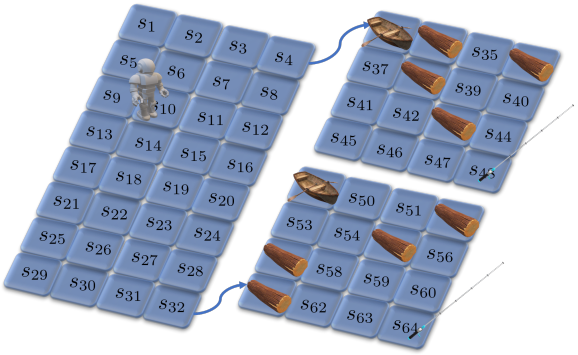

Figure 1: $L_{\text{log1}} = \{s_{34}, s_{36}, s_{38}, s_{43}\}$, $L_{\text{log2}} = \{s_{52}, s_{55}, s_{57}, s_{61}\}$, $L_{\text{canoe1}} = \{s_{33}\}$, $L_{\text{canoe2}} = \{s_{49}\}$ $L_{\text{fish1}} = \{s_{48}\}$, $L_{\text{fish2}} = \{s_{64}\}$.

4 Multichain Steady-State Policy Synthesis

We begin with a simple motivating example. Suppose some autonomous agent is stranded on three connected frozen islands as pictured in Figure 1. The agent's mission is to build a canoe to escape the islands while maximizing the amount of time it spends fishing for sustenance. The agent begins on the larger island of size $n \times n/2$ with an initial uniform distribution over those states, i.e., $\beta(s) = 2/n^2$ for every state belonging to the large island. Once the agent transitions into one of the two smaller islands, it can never go back. In these smaller islands, one quarter of the land consists of logs which can be used to build a canoe, and there is one fishing site as well. For each island, we have $(L_{\text{log1}}, [0.25, 1])$, $(L_{\text{log2}}, [0.25, 1])$, $(L_{\text{canoe1}}, [0.05, 1.0])$, $(L_{\text{canoe2}}, [0.05, 1.0])$, $(L_{\text{fish1}}, [0.1, 1.0])$, $(L_{\text{fish2}}, [0.1, 1.0])$, $R(\cdot, \cdot, L_{\text{fish1}}) = R(\cdot, \cdot, L_{\text{fish2}}) = 1$, $R(\cdot, \cdot, S \setminus (L_{\text{fish1}} \cup L_{\text{fish2}})) = 0$. Since these islands are frozen, the agent has a chance of slipping whenever it moves, causing a transition into one of three possible states. Namely, if the agent chooses to go right (left), there is a 90% chance that it will transition to the right (left), and the chance of transitioning to either of the states above or below it is 5%. Similarly, if the agent chooses to up (down), it will end up in the states above (below) it with 90% chance, and in the states to the right and left of it with chance 5% each. This Frozen Island scenario is similar to those found in OpenAI Gym's FrozenLake environment [Brockman *et al.*, 2016].

In this example, the feasible set of the LP of [Velasquez, 2019] will be empty. While the LP of [Kallenberg, 1983] returns a solution, the corresponding policy may not satisfy the specified constraints. We will demonstrate this deficiency by using a simple example.

Example 1. *Consider the MDP defined in Figure 2 with rewards $R(s_5, a_3) = R(s_8, a_1) = 1$ and zero otherwise. One optimal solution x^* to LP (4.7.6) in [Kallenberg, 1983] expressed in terms of state-action variables $x_{sa}, s \in S, a \in A(s)$ has $x^*_{s_5 a_3} = 0.5, x^*_{s_6 a_2} = 0.25, x^*_{s_8 a_1} = 0.25$, yielding a policy π with $\Pr^\infty_\pi(s_5, a_3) = 0.4173$, $\Pr^\infty_\pi(s_6, a_2) = 0.2724$ and $\Pr^\infty_\pi(s_8, a_1) = 0.2724$. For the remaining state-action pairs, x^*_{sa} and $\Pr^\infty_\pi(s, a)$ are both less than 10^{-10}. Clearly, $\Pr^\infty_\pi \neq x^*$, so any desired specifications for $\Pr^\infty_\pi$ (encoded as constraints on the variables x_{sa}) are generally not met.*

Definition 8 (Edge-Preserving Policies)**.** *Given an MDP $\mathcal{M}$, we define the 'Edge-Preserving' set of policies Π_{EP} as the set of stationary policies that exercise all existing transitions in the TSCCs $r(\mathcal{M})$ of $\mathcal{M}$ and such that $r(\mathcal{M}_\pi) = r(\mathcal{M})$, i.e.,*

$$\Pi_{EP} = \left\{ \pi \in \Pi_S : \begin{array}{c} r(\mathcal{M}_\pi) = r(\mathcal{M}) \wedge \\ \pi(a|s) > 0, \forall s \in r(\mathcal{M}), a \in A(s) \end{array} \right\} \quad (3)$$

Note that the condition $r(\mathcal{M}_\pi) = r(\mathcal{M})$ in Definition 8 implies that $\bar{r}(\mathcal{M})$ consists of transient or isolated states in $\mathcal{M}_\pi$ for any $\pi \in \Pi_{EP}$.

Example 2. *Figure 2 (left) shows a multichain MDP with two TSCCs designated by different colors. Figure 2 (right) shows the Markov chain $\mathcal{M}_\pi$ induced by an edge-preserving policy $\pi \in \Pi_{EP}$. As shown, π preserves all edges in the TSCCs of the MDP, however, the policy π does not necessarily exercise all transitions in the transient states. In $\mathcal{M}_\pi$, state s_1 is isolated and s_2 is transient.*

We can readily define MaStEr as the problem of finding a policy in Π_{EP} that maximizes the average expected reward and satisfies a given set of steady-state constraints.

Definition 9 (MaStEr)**.** *Given an LMDP $\mathcal{M} = \{S, A, T, R, \beta, L, \Phi_L^\infty\}$, we define MaStEr as the problem of finding a stochastic policy $\pi \in \Pi_{EP}$ that maximizes the objective in* (4) *and satisfies the steady-state specifications Φ_L^∞, i.e., solves*

$$\begin{aligned} &\max_{\pi \in \Pi_{EP}} \sum_{s \in S} \sum_{a \in A(s)} \Pr_\pi^\infty(s,a) R(s,a) \;\; \textit{subject to} \\ &\sum_{s \in L_i} \sum_{a \in A(s)} \Pr_\pi^\infty(s,a) \in [l,u], \;\; \forall (L_i, [l,u]) \in \Phi_L^\infty \, . \end{aligned} \quad (4)$$

In order to solve the MaStEr problem, we first determine the TSCCs $r(\mathcal{M})$ of $\mathcal{M}$ and the complement set $\bar{r}(\mathcal{M})$ using standard techniques from graph theory [Tarjan, 1971]. These are then used to define the LP (5) from which the solution policy is derived.

$$\begin{aligned} \max \quad & \sum_{s \in S} \sum_{a \in A(s)} x_{sa} R(s,a) \text{ subject to} \\ (i) \quad & \sum_{s \in S} \sum_{a \in A(s)} x_{sa} T(s' \mid s, a) = \sum_{a \in A(s')} x_{s'a}, \; \forall s' \in S \\ (ii) \quad & \sum_{s \in S} \sum_{a \in A(s)} y_{sa} T(s' \mid s, a) \\ & \qquad = \sum_{a \in A(s')} (x_{s'a} + y_{s'a}) - \beta_{s'}, \; \forall s' \in S \\ (iii) \quad & \sum_{s \in r_k(\mathcal{M})} \sum_{a \in A(s)} x_{sa} = \sum_{f \in \bar{r}(\mathcal{M})} \beta_f \, p_{fk} + \sum_{s \in r_k(\mathcal{M})} \beta_s, \; \forall k \in [m] \\ (iv) \quad & \sum_{k \in [m]} p_{fk} = 1, \; \forall f \in \bar{r}(\mathcal{M}) \\ (v) \quad & \sum_{f \in \bar{r}(\mathcal{M})} \sum_{a \in A(f)} x_{fa} = 0 \\ (vi) \quad & x_{sa} > 0, \; \forall s \in r_k(\mathcal{M}), k \in [m], a \in A(s) \end{aligned}$$

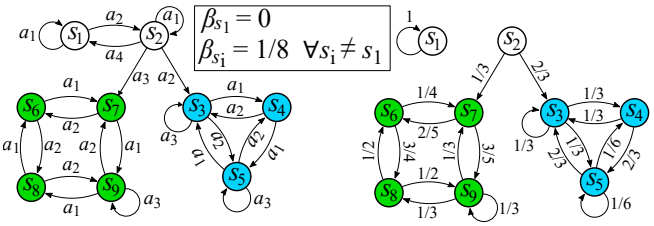

Figure 2: (*left*) Multichain MDP with two TSCCs. (*right*) Markov chain induced by some $\pi \in \Pi_{EP}$.

$$\begin{aligned} (vii) \quad & l_i \leq \sum_{s \in L_i} \sum_{a \in A(s)} x_{sa} \leq u_i, \; \forall (L_i, [l_i, u_i]) \in \Phi_L^\infty \quad (5) \\ & x_{sa} \in [0,1], y_{sa} \geq 0, \; \forall s \in S, a \in A(s) \\ & p_{fk} \in [0,1], \; \forall f \in \bar{r}(\mathcal{M}), k \in [m] \end{aligned}$$

Constraint (*i*) ensures that x is a stationary distribution [Altman, 1999] [Puterman, 1994]; constraint (*ii*), which is described in [Kallenberg, 1983] [Puterman, 1994], enforces consistency in the expected average number of visits y_{fa} for any transient state-action $f \in \bar{r}(\mathcal{M}), a \in A(f)$; constraints (*iii*), (*iv*) encode valid absorption probabilities p_{fk} ensuring that from any state $f \in \bar{r}(\mathcal{M})$, the process will be ultimately absorbed into the recurrent components $r_k(\mathcal{M}), k \in [m]$; constraint (*v*) preserves the non-recurrence of the states $f \in \bar{r}(\mathcal{M})$ by forcing zero steady-state occupancy; the strict positivity constraints (*vi*) preserve the transitions in the TSCCs for yielding edge-preserving policies; finally, constraints (*vii*) encode the steady-state specifications.

Theorem 1. *Given an LMDP $\mathcal{M}$, let $(x, y, p) \in Q$ and π be defined as in* (6)*, where Q is the feasible set of solutions to LP* (5)*, $x_s := \sum_{a \in A(s)} x_{sa}$, $y_s = \sum_{a \in A(s)} y_{sa}$, $E_x := \{s \in S : x_s > 0\}$ and $E_y := \{s \in S : y_s > 0\}$. Then, we have $\pi \in \Pi_{EP}$.*

$$\pi(a|s) = \begin{cases} \frac{x_{sa}}{x_s} & s \in E_x, a \in A(s) \\ \frac{y_{sa}}{y_s} & s \in E_y \setminus E_x, a \in A(s) \\ \textit{arbitrary} & \textit{o.w.} \end{cases} \quad (6)$$

Proof. Let $f \in \bar{r}(\mathcal{M})$. From constraint (*v*), we have $x_f = 0$. Two cases arise. If $f \notin E_y$, then $y_f = 0$. It follows from constraint (*ii*) and (6) that $\beta_f = 0$ and $T_\pi(f|f') = 0, \forall f' \in E_y$ as evidenced by

$$\begin{aligned} y_f &= \beta_f + \sum_{f' \in \bar{r}(\mathcal{M})} \sum_{a \in A(f')} y_{f'a} T(f|f', a) \quad (7) \\ &= \beta_f + \sum_{f'} y_{f'} \sum_{a} T(f|f', a) \pi(a|f') = \beta_f + \sum_{f'} y_{f'} T_\pi(f|f') \end{aligned}$$

Therefore, $f \in \bar{r}(\mathcal{M}_\pi)$. Now, consider the case where $f \in E_y$ and assume, for the sake of contradiction, that $f \in r(\mathcal{M}_\pi)$. Hence, $f \in F \subseteq r(\mathcal{M})$, for some TSCC F. Summing (7) over states $f' \in F$, we get that $\beta_{f'} = 0, \forall f' \in F$. From (7), we also have $T_\pi(f|f') = 0, \forall f' \in (\bar{r}(\mathcal{M}) \setminus F) \cap E_y$. Therefore, $F \subseteq \bar{r}(\mathcal{M}_\pi)$, yielding a contradiction. Hence, $f \in \bar{r}(\mathcal{M}_\pi)$. We conclude that $\bar{r}(\mathcal{M}) \subseteq \bar{r}(\mathcal{M}_\pi)$.

Consider $s \in r_k(\mathcal{M})$ for some $k \in [m]$ and assume, for the sake of contradiction, that $s \in \bar{r}(\mathcal{M}_\pi)$. From the

positivity constraint (*vi*), we have that $s \in r_k(\mathcal{M}) \cap E_x$. Since $s \in \bar{r}(\mathcal{M}_\pi)$, the column of the matrix T_π^∞ corresponding to state s is zero. Hence, from constraint (*i*), we have $x_s = 0$, i.e., $s \notin E_x$, yielding a contradiction. Hence, $r_k(\mathcal{M}) \subseteq r(\mathcal{M}_\pi), \forall k \in [m]$. Since we have already shown that $\bar{r}(\mathcal{M}) \subseteq \bar{r}(\mathcal{M}_\pi)$, we conclude that $r(\mathcal{M}_\pi) = r(\mathcal{M})$. The second requirement in (3) now follows from constraint (*vi*) and the definition of π in (6). Therefore, $\pi \in \Pi_{EP}$. □

We can readily state the following theorem establishing the correctness of the proposed LP.

Theorem 2. *Given an LMDP $\mathcal{M} = (S, A, T, R, \beta, L, \Phi_L^\infty)$, the linear program in* (5) *is feasible iff there exists a policy $\pi \in \Pi_{EP}$ such that the Markov chain $\mathcal{M}_\pi = (S, T_\pi)$ satisfies the specifications Φ_L^∞. Further, given an optimal solution x^*, y^* of* (5)*, the policy π^* as defined in* (6) *is optimal in the class of policies Π_{EP} and meets the specifications Φ_L^∞.*

Proof. ($\implies$) Let $(x, y, p) \in Q$ denote a feasible solution to (5) and let π be defined as in (6). We will show that $\Pr_\pi^\infty(s, a) = x_{sa}, s \in S, a \in A(s)$, which implies that $\mathcal{M}_\pi$ meets the specifications Φ_L^∞ per constraint (*vii*). By Theorem 1, $\pi \in \Pi_{EP}$. Therefore, $\bar{r}(\mathcal{M}_\pi) = \bar{r}(\mathcal{M})$, implying $\Pr_\pi^\infty(f, a) = 0 = x_{fa}, f \in \bar{r}(\mathcal{M}), a \in A(f)$, where the second equality follows from constraint (*v*).

First, note that from (7), we have $y_f = \beta_f + \sum_{f'} y_{f'} T_\pi(f|f')$ for $f \in \bar{r}(\mathcal{M})$. This can be written as

$$y_{\bar{r}(\mathcal{M})} = (I - Z_\pi^T)^{-1} \beta_{\bar{r}(\mathcal{M})} \tag{8}$$

where $Z_\pi = [T_\pi(f', f)] \in [0, 1]^{|\bar{r}(\mathcal{M})| \times |\bar{r}(\mathcal{M})|}$ defines the transitions between states in $\bar{r}(\mathcal{M})$ under policy π.

Second, recall that $\pi \in \Pi_{EP}$ and so the set $r_k(\mathcal{M})$ for any $k \in [m]$ is a TSCC of $\mathcal{M}_\pi$. For every $k \in [m]$, we have

$$\begin{aligned}\sum_{s \in r_k(\mathcal{M})} \beta_s &= \sum_{s \in r_k(\mathcal{M})} x_s + \sum_{s \in r_k(\mathcal{M})} y_s - \\ &\quad \sum_{s \in r_k(\mathcal{M})} \sum_{s' \in r_k(\mathcal{M}) \cup \bar{r}(\mathcal{M})} \sum_{a \in A(s')} T(s|s', a) y_{s'a} \\ &= \sum_{s \in r_k(\mathcal{M})} x_s - \sum_{s' \in \bar{r}(\mathcal{M})} y_{s'} \sum_{s \in r_k(\mathcal{M})} T_\pi(s|s') \end{aligned} \tag{9}$$

where the first equality follows from (*ii*) and the fact that $r_k(\mathcal{M})$ is only reachable from states in $r_k(\mathcal{M}) \cup \bar{r}(\mathcal{M})$. The second equality follows by breaking the summation over the union of the sets $r_k(\mathcal{M})$ and $\bar{r}(\mathcal{M})$ and the fact that $r_k(\mathcal{M})$ is closed. Combining (8) and (9), we get

$$\sum_{s \in r_k(\mathcal{M})} (x_s - \beta_s) = \beta_{\bar{r}(\mathcal{M})}^T (I - Z_\pi)^{-1} L_{\pi,k} e \tag{10}$$

where $L_{\pi,k}$ is the submatrix of T_π of transitions from $\bar{r}(\mathcal{M})$ to $r_k(\mathcal{M})$. Given the definition of $\Pr_\pi^\infty(s)$ in Lemma 1 and the known form of T_π^∞ [Puterman, 1994], we recognize that the right hand side of (10) is exactly equal to $\beta_{\bar{r}(\mathcal{M})}^T P_{\pi,k}$, where $P_{\pi,k} = [p_{fk}], f \in \bar{r}(\mathcal{M})$ are the absorption probabilities from $\bar{r}(\mathcal{M})$ to $r_k(\mathcal{M})$ under policy π [Feller, 1968; Kallenberg, 1983]. We conclude that $\sum_{s \in r_k(\mathcal{M})} x_s = \sum_{s \in r_k(\mathcal{M})} \Pr_\pi^\infty(s)$. Finally, $x_{r_k(\mathcal{M})}^T T_{\pi,k} = x_{r_k(\mathcal{M})}^T$ from constraint (*i*) and (6), where $T_{\pi,k}$ is the submatrix of T_π corresponding to transitions between states in $r_k(\mathcal{M})$. Hence, by the ergodic Theorem for unichain components [Altman, 1999], the solution is unique for each component $r_k(\mathcal{M}), k \in [m]$ so $x = \Pr_\pi^\infty$, the unique steady-state distribution.

($\impliedby$) Suppose there exists such a policy π as in the statement of Theorem 2. Then $\Pr_\pi^\infty$ is well-defined as in Lemma 1. Hence, we can set $x_{sa} = \Pr_\pi^\infty(s, a), s \in S, a \in A(s)$. Recall that $\bar{r}(\mathcal{M}_\pi) = \bar{r}(\mathcal{M})$ since $\pi \in \Pi_{EP}$, so we set $x_{sa} = \Pr_\pi^\infty(s, a) = 0, s \in \bar{r}(\mathcal{M})$. The variables $y_{fa}, f \in \bar{r}(\mathcal{M})$ and p_{fk} can be set as in (8) and (10). It can be easily verified that the variables $y_{sa}, s \in r_k(\mathcal{M})$ can now be defined in terms of $x_{sa}, y_{fa}, p_{fk}, T(s'|s, a)$ and β such that the corresponding constraints are satisfied. The optimality of π^* follows from the optimality of (x^*, y^*), Theorem 1 and the established equality $\Pr_{\pi^*}^\infty = x^*$. □

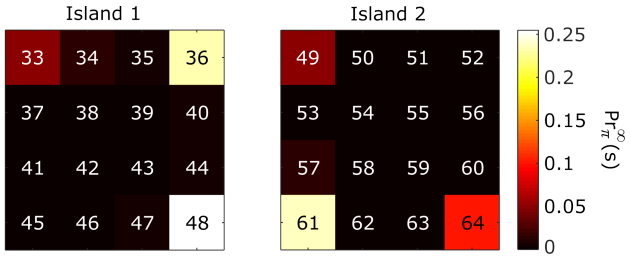

Figure 3: Heat maps showing the steady-state probabilities $\Pr^\infty(s)$ for states $s \in r(\mathcal{M})$ belonging to the two TSCCs of the Frozen Lakes example in Figure 1.

5 Numerical Results

We first run our proposed LP (5) to calculate the steady-state distribution $\Pr_\pi^\infty(s)$ for the Frozen Island example shown in Figure 1. The values for the two recurrent sets (the two islands) are shown in Figure 3. The heat map provides insight into the way in which the agent meets the specifications. Once entering the islands, the agent spends a large proportion of its time in states s_{33}, s_{36}, s_{48}, s_{49}, s_{61}, and s_{64}, in the sense of average expected number of visits. The agent satisfies constraints $(L_{\text{log1}}, [0.25, 1])$ and $(L_{\text{log2}}, [0.25, 1])$ largely by visiting states s_{36} and s_{61}, respectively. Likewise, the agent meets constraints $(L_{\text{canoe1}}, [0.05, 1.0])$, $(L_{\text{canoe2}}, [0.05, 1.0])$ $(L_{\text{fish1}}, [0.1, 1.0])$, $(L_{\text{fish2}}, [0.1, 1.0])$ by visiting states s_{33}, s_{49}, s_{48}, and s_{64}, respectively. While satisfying the constraints, the agent maximizes the accumulated reward by visiting state s_{48} over 25% of the time.

Figure 4 (*left*) shows the values of $\Pr_\pi^\infty(s)$ along with the optimal values x_s^* obtained from LP (5), demonstrating that the steady-state distribution matches that estimated by the LP for every state. This holds also for all state-action pairs, i.e., $\Pr^\infty = x^*$. This condition is critical to the proof of Theorem 2, and ensures that the policy is both optimal and meets the steady-state specifications. For comparison, we solve LP (4.7.6) from [Kallenberg, 1983] to obtain optimal values x^*, y^*, then form the corresponding policy π. Results are shown in Figure 4 (*right*). As with Example 1, the produced policy fails to yield a steady-state distribution equal to x^*.

Table 1 demonstrates the consequences when $\Pr_\pi^\infty \neq x^*$.

Method	Specifications												Rewards	
	Logs (≥ 0.25)				Canoe (≥ 0.05)				Fish Rod (≥ 0.1)					
	Island 1		Island 2		Island 1		Island 2		Island 1		Island 2		$\mathbf{R}^*$	$\mathbf{R}^\infty_\pi$
	x^*	$\Pr^\infty$	x^*	$\Pr^\infty$	x^*	$\Pr^\infty$	x^*	$\Pr^\infty$	x^*	$\Pr^\infty$	x^*	$\Pr^\infty$		
Proposed LP	0.25	0.25	0.25	0.25	0.05	0.05	0.05	0.05	0.25	0.25	0.10	0.10	0.3547	0.3547
Kallenberg	0.25	**0.17**	0.25	0.36	0.05	**0.04**	0.05	0.07	0.26	0.19	0.10	0.14	0.3621	**0.3278**

Table 1: Steady-state specification comparison. Bold red text indicates violated specifications. Constraints are specified in the header for each label type.

For each specification $(L_i, [l, u]) \in \Phi_L^\infty$, Table 1 shows $e^T x^*_{L_i}$ and $\Pr^\infty_\pi(L_i) := \sum_{s \in L_i} \Pr^\infty_\pi(s)$. For the proposed LP, all of the specifications are met. However, for Kallenberg's formulation, although $x^*_{L_{\text{log1}}}$ and $x^*_{L_{\text{canoe1}}}$ meet the specification, the policy yields steady-state distributions $\Pr^\infty_\pi(L_{\text{log1}})$ and $\Pr^\infty_\pi(L_{\text{canoe1}})$ which violate the specifications (the violations are highlighted with bold red text). Additionally, we show the optimal reward R^* output by the LP, as well as the average expected reward of the obtained policy $R^\infty_\pi := \sum_{s \in S} \sum_{a \in A(s)} \Pr^\infty_\pi(s, a) R(s, a)$. Although R^* obtained by Kallenberg's formulation is larger than that of the proposed LP, the proposed LP produces a policy which gives a larger R^∞.

Finally, we execute the policies to verify the validity of the formulations and to further demonstrate the failure of Kallenberg's formulation to meet specifications and yield optimal rewards. Define S_t and A_t as the state and action, respectively, of the Frozen Island example at time t assuming initial distribution β and policy π. The average number of visits $f_{\pi,n}$ and average reward $g_{\pi,n}$ up to time n are defined as

$$f_{\pi,n}(L) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_L(S_t), \quad \mathbb{1}_L(s) = \begin{cases} 1 & s \in L \\ 0 & s \notin L \end{cases} \tag{11}$$

$$g_{\pi,n} = \frac{1}{n} \sum_{t=1}^{n} R(S_t, A_t, S_{t+1}) \tag{12}$$

For $f_{\pi,n}(L)$ and $g_{\pi,n}$, we take an ensemble average over 5000 paths. In Figure 5 (*left*), the solid and dotted lines show the average number of visits to the states labeled as logs, the dashed and dash-dotted lines indicate the steady-state distributions, and the square markers show the specification lower bound for the logs. For the proposed LP, $f_{\pi,n}(L_{\text{log1}})$ converges to $\Pr^\infty_\pi(L_{\text{log1}})$ and meets the specification. For Kallenberg's policy, although $f_{\pi,n}(L_{\text{log1}})$ converges to $\Pr^\infty_\pi(L_{\text{log1}})$, it fails to meet the specifications for the reason described earlier. In Figure 5 (*right*), the solid and dotted lines show the average reward, and the dashed and dash-dotted lines indicate the reward which was output by the LP. While the proposed LP converges to the LP reward, as described earlier, Kallenberg's formulation converges to a different reward.

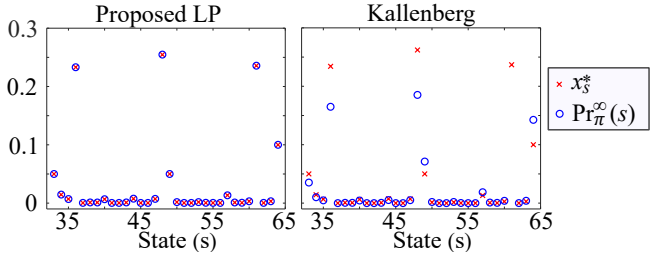

Figure 4: Example showing that $\Pr^\infty_\pi(s) = x^*_s, s \in r(\mathcal{M})$ for the Proposed LP, but not for Kallenberg's formulation.

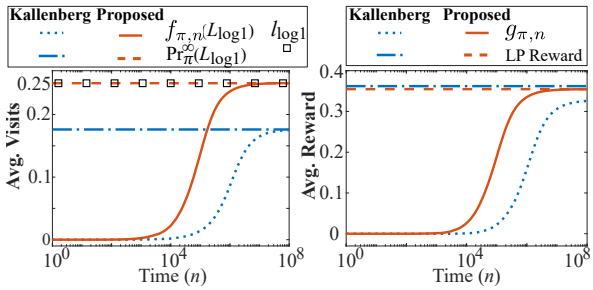

Figure 5: Execution of policy, showing (*left*) average visits and (*right*) average reward up to time n.

We also demonstrate the scalability of the proposed LP by running CPLEX 12.8 simulations on a standard desktop computer with 128GB of RAM for random instances of the Frozen Island problem. See Table 2 for runtime results.

Optimal LP	8×8	16×16	32×32	64×64	128×128
Runtime	0.0049	0.0837	0.6587	6.5122	410.75

Table 2: Average runtime (in seconds) of 20 instances per LP for the three-island problem described in Figure 1. These islands combined form an $n \times n$ grid. In each of the smaller islands, logs are randomly distributed over $1/4$ of the states and a canoe (fishing rod) is placed in the top-left (bottom-right) tile. For these experiments, we have the constraints $(L_{\text{log1}} \cup L_{\text{log2}}, [0.3, 1]), (L_{\text{canoe1}} \cup L_{\text{canoe2}}, [0.05, 1])$ and reward function $R(\cdot, \cdot, L_{\text{fish1}} \cup L_{\text{fish2}}) = 1, R(\cdot, \cdot, S \setminus L_{\text{fish1}} \cup L_{\text{fish2}}) = 0$.

6 Conclusion

The multichain SSPS problem was defined for deriving policies that satisfy constraints on the steady-state behavior of the agent. A linear programming solution was proposed and its correctness proved for the class of edge-preserving policies. Simulations of the resulting policies demonstrate that our approach overcomes limitations in the literature.

References

[Akshay *et al.*, 2013] Sundararaman Akshay, Nathalie Bertrand, Serge Haddad, and Loïc Hélouët. The steady-state control problem for Markov decision processes. In *Int. Conf. Quant. Eval.*, pages 290–304, Berlin Heidelberg, 2013. Springer.

[Altman *et al.*, 2019] Eitan Altman, Said Boularouk, and Didier Josselin. Constrained Markov decision processes with total expected cost criteria. In *Proceedings of the 12th EAI International Conference on Performance Evaluation Methodologies and Tools*, pages 191–192. ACM, 2019.

[Altman, 1999] Eitan Altman. *Constrained Markov decision processes*. CRC Press, Boca Raton, 1999.

[Bhatnagar and Lakshmanan, 2012] Shalabh Bhatnagar and K Lakshmanan. An online actor–critic algorithm with function approximation for constrained Markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

[Boussemart and Limnios, 2004] M Boussemart and N Limnios. Markov decision processes with asymptotic average failure rate constraint. *Communications in Statistics-Theory and Methods*, 33(7):1689–1714, 2004.

[Boussemart *et al.*, 2002] M Boussemart, N Limnios, and JC Fillion. Non-ergodic Markov decision processes with a constraint on the asymptotic failure rate: general class of policies. *Stochastic models*, 18(1):173–191, 2002.

[Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

[De Ghellinck, 1960] Guy De Ghellinck. Les problèmes de décisions séquentielles. *Cahiers du Centre d'Etudes de Recherche Opérationnelle*, 2(2):161–179, 1960.

[Derman, 1970] Cyrus Derman. *Finite State Markovian Decision Processes*. Academic Press, Inc., Orlando, FL, USA, 1970.

[Feinberg, 2009] E. A. Feinberg. Adaptive computation of optimal nonrandomized policies in constrained average-reward MDPs. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 96–100, March 2009.

[Feller, 1968] William Feller. *An Introduction to Probability Theory and its Applications*, volume 1. Wiley, 3 edition, 1968.

[Fisher *et al.*, 2013] Michael Fisher, Louise A Dennis, and Matthew P Webster. Verifying autonomous systems. *Commun. ACM*, 56(9):84–93, 2013.

[Haesaert *et al.*, 2015] Sofie Haesaert, Alessandro Abate, and Paul MJ Van den Hof. Correct-by-design output feedback of LTI systems. In *54th IEEE Conference on Decision and Control (CDC)*, pages 6159–6164, 2015.

[Kallenberg, 1983] L. C. M. Kallenberg. *Linear programming and finite Markovian control problems*. Mathematisch Centrum, Amsterdam, 1983.

[Kress-Gazit *et al.*, 2018] Hadas Kress-Gazit, Morteza Lahijanian, and Vasumathi Raman. Synthesis for robots: Guarantees and feedback for robot behavior. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:211–236, 2018.

[Lazar, 1983] Andreas Lazar. Optimal flow control of a class of queueing networks in equilibrium. *IEEE Transactions on Automatic Control*, 28(11):1001–1007, 1983.

[Manne, 1960] Alan S. Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

[Puterman, 1994] Martin Puterman. *Markov decision processes : discrete stochastic dynamic programming*. Wiley, New York, 1994.

[Ross, 1989] Keith W. Ross. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research*, 37(3):474–477, 1989.

[Skwirzynski, 1981] Joseph K Skwirzynski. *New concepts in multi-user communication*, volume 43. Springer Science & Business Media, 1981.

[Tarjan, 1971] R. Tarjan. Depth-first search and linear graph algorithms. In *12th Annual Symposium on Switching and Automata Theory*, pages 114–121, Oct 1971.

[Velasquez, 2019] Alvaro Velasquez. Steady-state policy synthesis for verifiable control. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5653–5661. AAAI Press, 2019.