Constrained Reinforcement Learning Beyond Rewards via ω-Regular Specifications

Dominik Wagner, Leon Witzman, Luke Ong
{dominik.wagner,luke.ong}@ntu.edu.sg, WITZ0001@e.ntu.edu.sg

NTU Singapore

Abstract

Constrained Markov decision processes (CMDPs) model sequential decision-making under scalar reward and cost signals, but cannot naturally express rich temporal requirements such as ordering, persistence, or liveness. Temporal logics and, more generally, ω-regular objectives provide such expressive specifications, but do not directly capture trade-offs between task performance and safety constraints. Following Forejt et al. (2011b), we study CMDPs with both ω-regular objectives and ω-regular constraints. We show that optimal policies in this setting form a fundamentally different class from those sufficient for either standard CMDPs or unconstrained ω-regular planning. Exploiting this characterisation, we develop a model-based reinforcement learning algorithm with finite-sample guarantees, and establish a reduction to discounted CMDPs by restricting optimisation to this sufficient policy class.

1 Introduction

Reinforcement learning (RL) (Sutton & Barto, 2018) studies how agents can learn to make sequential decisions in initially unknown environments through interaction alone. In standard RL, tasks are specified using scalar reward functions, and agents seek to maximise expected cumulative reward. While widely successful, reward-based formulations are often insufficient for safety-critical settings, where satisfying explicit constraints is as important as optimising performance. This has led to the study of constrained Markov decision processes (CMDPs), which augment rewards with cost signals and enforce safety requirements through constraints on expected cumulative cost (Altman, 2021).

Despite their utility, both rewards and costs in CMDPs remain scalar quantities, limiting their ability to express complex temporal and conditional requirements. Many practical specifications are inherently temporal ("eventually reach a target"), conditional ("if signal A occurs, ensure B"), or persistent ("visit regions A and B infinitely often"). Capturing such behaviours using scalar rewards and costs is often unnatural and can lead to misspecified objectives that fail to promote the intended behaviour.

To address these limitations, there has been growing interest (Wolff et al., 2012; Ding et al., 2014; Voloshin et al., 2022a; Cai et al., 2023; Fu & Topcu, 2014) in specifying tasks using temporal logics such as Linear Temporal Logic (LTL), or more generally ω-regular objectives (Baier & Katoen, 2008). These formalisms can naturally express rich combinations of *safety*, *liveness*, and *sequencing* requirements in a compositional manner. However, existing approaches based on ω-regular objectives typically optimise a single scalar quantity, namely the probability of satisfying a specification. As a consequence, they do not explicitly represent trade-offs between task performance and safety constraints. This motivates

Specification \ Policy	Determ.	Stationary Stoch.	Mix. determ.	Mix. stoch.
Unconstrained disc. reward	✓	✓	✓	✓
Unconstrained ω-regular	Q	Q	Q	Q
Discounted reward and cost	✗	✓	✓	✓
(C-Recurrence)	✗	✗	$\{0, 1\}$	✓
(C-ω-Regular)	✗	✗	$Q \times Q' \times \{0, 1\}$	$Q \times Q'$

Table 1: Memory for optimal policies. "✓" indicates stationary policies are sufficient. Q and Q' are the state spaces of suitable automata expressing the ω-regular languages. "✗" indicates non-existence of optimal such policies (Example 1). Example 2 shows that 1 bit of memory is necessary for Problem (C-Recurrence) with mixtures of deterministic policies.

combining the expressive power of ω-regular objectives with the constrained optimisation perspective of CMDPs.

In this work, we investigate a constrained RL framework integrating ω-regular objectives with ω-regular constraints. This allows task objectives and safety requirements to be specified separately within a unified temporal logic formalisation, motivating a study of the resulting class of optimal policies and its algorithmic implications.

Contributions. Our contributions are as follows:

- We identify policy classes sufficient for constrained ω-regular optimisation (see Table 1), which are neccesarily distinct from the standard sufficient classes for unconstrained ω-regular optimisation and discounted CMDPs.
- We develop a model-based reinforcement learning algorithm with finite-sample guarantees under a generative model.
- We establish a reduction to discounted constrained MDPs, showing that optimal policies can be approximated within a suitable policy class for sufficiently large discount factors, thereby connecting temporal-logic specifications with classical constrained reinforcement learning.

Related Work. Prior work has investigated multi-objective probabilistic model checking, including ω-regular objectives, when the MDP is assumed to be fully known (e.g. Etessami et al. (2008); Forejt et al. (2011a;b)).

Recent works have explored the reinforcement learning setting, where MDP transition dynamics are unknown. For unconstrained ω-regular objectives, Perez et al. (2024) propose a model-based approach that estimates a surrogate MDP from data and performs planning on the approximation, while Bozkurt et al. (2019) introduce translation to discounted rewards together with an optimality result for sufficiently large discount factors.

For multi-objective RL, Voloshin et al. (2022b); Shah et al. (2024) consider the problem of maximising discounted reward among policies which satisfy an LTL formula with maximal probability, while Hahn et al. (2023) study multi-objective problems with lexicographic preference ordering. In contrast, we study a CMDP-style threshold formulation in which both the optimization objective and the constraint are ω-regular specifications.

Finally, the field of Safe RL (see (Gu et al., 2024; Wachi et al., 2024) for surveys) typically studies constrained problems defined by scalar reward and cost signals under discounting, which thus have limited ability to specify natural temporal properties.

(a) MDP for Example 1 with $F := \{s_1\}$ and $F' := \{s_0, s_1\}$, i.e. the objective is to maximise the probability of reaching s_1, such that s_2 is avoided with probability at least 0.9.

(b) MDP for Example 2 with $F := \{s_1\}$ and $F' := \{s_2\}$, where no deterministic stationary policy visits both F and F' infinitely often.

Figure 1: Example MDPs for Problem (C-Recurrence).

2 Problem Statement

We work with MDPs $\mathcal{M} = (S, A, \mathrm{Act}, s_0, P)$, with S and A finite state and action sets, enabled actions $\mathrm{Act}(s) \subseteq A$, initial state s_0, and transition kernel P, and use $\mathcal{D}_\pi^{\mathcal{M}}$ for the distribution over runs generated by a policy π. We equip $\mathcal{M}$ with a labelling function $\ell : S \to 2^{\mathcal{AP}}$, with $\mathcal{AP}$ a finite set of atomic propositions (i.e. high-level symbolic features). A trajectory $\rho = s_0 a_0 s_1 a_1 \cdots$ therefore induces an infinite word $\ell(\rho) := \ell(s_0)\ell(s_1)\ell(s_2)\cdots \in (2^{\mathcal{AP}})^\omega$.

A temporal specification is a set $L \subseteq (2^{\mathcal{AP}})^\omega$ of acceptable (infinite) label sequences. ω-regular specifications, those recognised by finite state machines (see Definition 12), form a tractable subclass. ω-regular specifications subsume Linear Temporal Logic and hence capture safety, reachability, recurrence, sequencing, and conditional behaviour properties.

For example, if $u, t \in \mathcal{AP}$ label unsafe and target states, respectively, then the following are ω-regular: always avoid u, eventually reach t, and visit t infinitely often. For a more complex example, consider a logistics drone (see Fig. 3) that must repeatedly visit two pick-up locations before delivering a load. Such a drone may naturally be considered battery-safe if, whenever battery is critically low, a charging station is visited within two timesteps.

ω-Regular Objectives and Constraints. We distinguish between the specification to optimise and the specification to enforce as a constraint. Let $L, L' \subseteq (2^{\mathcal{AP}})^\omega$ be an ω-regular objective and constraint, resepctively, and $p \in [0, 1]$ be a satisfaction threshold. We study

$$\max_\pi \ \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L] \quad \text{s.t.} \quad \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L'] \geq p \tag{C-ω-Regular}$$

In the drone example, L may encode the package delivery objective, with L' encoding the battery-recovery requirement with high probability $p := 0.99$.

As is standard for reinforcement learning, we consider *ϵ-approximately feasible and optimal* solutions, which are feasible and optimal up to an ϵ-margin (see Definition 14).

Recurrence Specifications. Before extending to full ω-regular specifications, we first analyze the simpler special case of recurrence. Given target sets $F, F' \subseteq S$, the constrained recurrence problem is given by

$$\max_\pi \ \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] \quad \text{s.t.} \quad \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F' \neq \emptyset] \geq p, \tag{C-Recurrence}$$

where $\mathrm{Inf}(\rho)$ denotes the set of states visited infinitely often along ρ. Sections 3 to 5 develop results first for the recurrence setting, which are then lifted to the general ω-regular setting in Section 6.

2.1 Expressivity

While stochastic stationary policies $\pi : S \to \mathcal{D}(A)$ suffice for standard discounted CMDPs (Altman, 2021), ω-regular specifications require tracking past behaviour. Fortunately, one need not track the full state-action history as it suffices to recall only the current state of the finite specification automaton (Hahn et al., 2019). As such, we consider a general class of policies with explicit memory dependence (see Definition 13 in Section 8).

Furthermore, Problem (C-Recurrence) (hence Problem (C-ω-Regular) by extension) can encode trade-offs between objectives and constraints at various safety thresholds which cannot in general be captured by a singular ω-regular objective.

Example 1. The MDP in Fig. 1a is an instance of Problem (C-Recurrence) with $F = \{s_1\}$, $F' = \{s_0, s_1\}$, and $p = 0.9$. Clearly, $\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\exists i \in \mathbb{N}.\, \rho_i^S \in F] = 1 - \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\exists i \in \mathbb{N}.\, \rho_i^S \in F']$ for all policies π. Thus, any policy which (eventually) leaves s_0 with probability exactly 0.2 would be a solution. However, no deterministic nor stationary policy can achieve this, as any such policy either remains in s_0 almost surely or leaves s_0 almost surely.

Policy Mixtures. For policies π_0, π_1 and $p \in [0, 1]$, we use the notation $p \cdot \pi_0 + (1 - p) \cdot \pi_1$ for the policy which, with probability p, generates rollouts by following π_0, and otherwise produces rollouts by following π_1 (see Section 8 for the formal definition).

Example 1 is solved by the mixture $0.8 \cdot \pi_a + 0.2 \cdot \pi_b$ of deterministic stationary policies, where π_a is the policy always performing action a, likewise for b, which achieves objective satisfaction of 0.1. Hence, constraints *strictly increase expressivity* compared to unconstrained ω-regular objectives, for which deterministic optimal policies always exist, and standard discounted CMDPs, for which solutions can be assumed stationary (Altman, 2021).

Furthermore, mixtures of deterministic stationary policies alone are still insufficient to solve Problem (C-Recurrence) in general.

Example 2. Consider the MDP in Fig. 1b, where $F \coloneqq \{s_1\}$ and $F' \coloneqq \{s_2\}$. A deterministic stationary policy visits exactly one of F, F' infinitely often in all runs. Therefore, for $p > 0$, any feasible mixture of deterministic stationary policies satisfies the objective with probability less than 1, whereas alternating between a and b in s_0 satisfies both objective and constraint with probability 1.

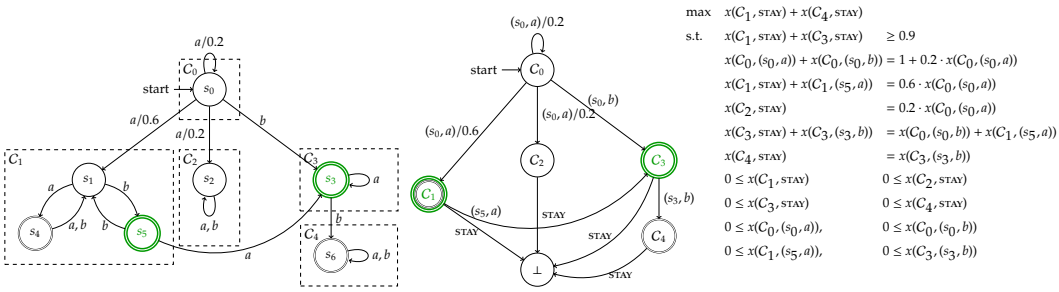
Augmented MDP. Intuitively, the need to oscillate between F and F' can preclude deterministic stationary policies. To address this limitation, we consider an *augmented MDP* with states $S \times \{0, 1\}$, where the additional bit is updated deterministically to 0 when visiting F, and to 1 for $F' \setminus F$ (see Definition 16 in Section 8.1).

3 Planning and Optimal Policies

Before addressing the RL setting, we first cover planning for Problem (C-Recurrence) when the MDP is fully known. Of particular interest are its end components (strongly connected sub-MDPs, see Section 9.1), and whether they intersect the relevant regions F and F'.

Almost surely, the state-action pairs taken infinitely often by a run $\rho \sim \mathcal{D}_\pi^{\mathcal{M}}$ constitute an end component $C(\rho)$ (Lemma 20). Hence $\mathrm{Inf}(\rho) \cap F \neq \emptyset$ can only occur if $C(\rho)$ contains a state in F. Conversely, once an end component containing F has been reached, one can ensure F is visited infinitely often by sampling from the actions within the end component. We therefore reduce recurrence optimisation to reachability-style optimisation over maximal end components.

Following Ciesinski et al. (2008); Forejt et al. (2011b), we construct a collapsed MDP $\mathcal{M}_{\mathrm{coll}}$ in which each maximal end component of $\mathcal{M}$ is represented by a single state (see Fig. 2b for an example and Section 9.4). From a collapsed state, the distinguished action STAY



(a) Decomposition of an MDP into MECs (dashed). Probabilities which are 1 are omitted.

(b) Collapsed MDP. Probabilities which are 1 are omitted.

(c) Corresponding LP (see Example 4).

Figure 2: Running example.

transitions to sink state $\perp$ and represents a commitment to remain indefinitely within the corresponding maximal end component. This commitment is rewarded via the binary reward functions R_{coll} and R'_{coll}, provided the component intersects F or F', respectively (Definition 30). Hence, expected reward is the probability of remaining in an appropriate end component. The remaining actions correspond to state-action pairs of $\mathcal{M}$ which leave the current maximal end component with positive probability.

As $\mathcal{M}_{\text{coll}}$ is absorbing (Section 9.2 and Lemma 29), the resulting constrained reward optimisation problem is amenable to standard occupancy-measure linear programming (see Section 9.3). In particular, the splitting procedure of (Feinberg & Rothblum, 2012, Alg. 2,Thm. 9.2) either declares such absorbing problems infeasible, or returns an optimal mixture of two deterministic stationary policies. Each deterministic stationary policy for $\mathcal{M}_{\text{coll}}$ lifts to a stochastic stationary policy for $\mathcal{M}$ (alternatively, deterministic stationary for $\mathcal{M}_{\text{aug}}$, Lemma 31), hence we obtain the following sufficient policy classes for solving Problem (C-Recurrence).

Theorem 3. *Feasible instances of Problem (C-Recurrence) have solutions which are a mixture of two stochastic stationary policies, likewise solutions which are a mixture of two deterministic stationary policies over the augmented MDP (which by Remark 17 can be viewed as mixtures of deterministic policies with 1 bit memory).*

Example 4. The MDP of Fig. 2a with $F := \{s_4, s_6\}$ and $F' := \{s_5, s_3\}$ produces the collapsed MDP in Fig. 2b with $R_{\text{coll}}(C_1, \text{STAY}) = 1 = R_{\text{coll}}(C_4, \text{STAY})$, $R'_{\text{coll}}(C_1, \text{STAY}) = 1 = R'_{\text{coll}}(C_3, \text{STAY})$ and 0 elsewhere. Subject to constraint threshold $p = 0.9$, its optimal objective value is 0.3, realised by the mixture of deterministic stationary policies $0.4 \cdot \pi_a + 0.6 \cdot \pi_b$ where π_a performs (s_0, a) in C_0, likewise for π_b, and both policies perform STAY at all other states.

We derive stochastic stationary policies $\overline{\pi}_a, \overline{\pi}_b$ for $\mathcal{M}$, where $\overline{\pi}_a$ always selects a in s_0 and $\overline{\pi}_b$ always selects b in s_0. They agree at all other states, sampling from actions guaranteed to stay in the current maximal end component uniformly at random. Finally, a solution for Fig. 2 is $0.4 \cdot \overline{\pi}_a + 0.6 \cdot \overline{\pi}_b$, with objective satisfaction 0.3 and constraint satisfaction 0.9.

4 Model-Based Reinforcement Learning Algorithm

Leveraging the LP-based planning framework developed in the previous section, we now consider the reinforcement learning setting in which the transition kernel is unknown. We assume access to a generative model (Agarwal et al., 2020; Vaswani et al., 2022; Agarwal et al., 2019) that, for any state–action pair (s, a), returns independent samples from the transition distribution $P(\cdot \mid s, a)$.

Algorithm 1 Reinforcement Learning for Problem (C-Recurrence)

Require: MDP without transition kernel $(S, A, \mathrm{Act}, s_0)$, sampling oracle for P, regions $F, F' \subseteq S$, threshold $p \in [0,1]$, sampling budget $N \in \mathbb{N}$

1: **for** $i \in \{1, \dots, N\}, s \in S, a \in \mathrm{Act}(s)$ **do**
2: $\quad s_{s,a}^{(i)} \sim P(\cdot \mid s, a)$
3: **end for**
4: $\widehat{\mathcal{M}} := (S, A, \mathrm{Act}, s_0, \widehat{P})$, where $\widehat{P}(s' \mid s, a) := \frac{1}{N} \cdot \left|\{1 \le i \le N \mid s_{s,a}^{(i)} = s'\}\right|$
5: Collapse $\widehat{\mathcal{M}}$ into $\widehat{\mathcal{M}}_{\mathrm{coll}}$
6: out $\leftarrow$ SplitAbsorbing$(\widehat{\mathcal{M}}_{\mathrm{coll}}, \widehat{R}_{\mathrm{coll}}, \widehat{R}'_{\mathrm{coll}}, p)$ {(Feinberg & Rothblum, 2012, Alg. 2)}
7: **if** out = Infeasible **then**
8: $\quad$ **return** Infeasible
9: **else**
10: $\quad$ let $\lambda \cdot \widehat{\pi}_1 + (1-\lambda) \cdot \widehat{\pi}_2 = \mathrm{out}$
11: $\quad$ convert $\widehat{\pi}_1, \widehat{\pi}_2$ to policies π_1, π_2 for $\mathcal{M}$ {Lemma 31}
12: $\quad$ **return** $\lambda \cdot \pi_1 + (1-\lambda) \cdot \pi_2$
13: **end if**

For each (s, a), Algorithm 1 collects N transition samples and constructs the empirical MDP $\widehat{\mathcal{M}}$ (line 4). It then solve the planning problem on $\widehat{\mathcal{M}}$ as described in the previous section (lines 5–13).

The following simulation lemma bounds the change in recurrence probabilities induced by perturbations of the transition kernel. For $\|A\|_\infty := \max_{s,a} \sum_{s'} |A(s' \mid s, a)|$, the discrepancy scales linearly with $\|P - P'\|_\infty$ under a matching-support assumption.

Lemma 5 (Simulation)**.** *Let $\mathcal{M} = (S, A, \mathrm{Act}, s_0, P)$ and $\mathcal{M}' = (S, A, \mathrm{Act}, s_0, P')$ be MDPs satisfying $P(s' \mid s, a) > 0$ iff $P'(s' \mid s, a) > 0$ for all $s, s' \in S$ and $a \in A$. Let $p_{\min} > 0$ be the smallest positive transition probability in P. Then for every $F \subseteq S$ and deterministic stationary policy π,*

$$\left|\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] - \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}'}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset]\right| \le \frac{2|S|}{p_{\min}^{|S|}} \|P - P'\|_\infty.$$

As unconstrained ω-regular objectives are not PAC-learnable in general (Alur et al., 2022; Yang et al., 2022), we follow Perez et al. (2024) in assuming a known lower bound $p_{\min} > 0$ on the positive transition probabilities of $\mathcal{M}$. This does not restrict the class of finite MDPs considered, since such a bound always exists, but is required to bound the sample complexity *a priori*.

We additionally assume the existence of a non-trivial feasibility margin to compensate for statistical estimation error in the empirical model:

Assumption 6. There exists a policy $\overline{\pi}$ satisfying $\overline{p} := \mathbb{P}_{\rho \sim \mathcal{D}_{\overline{\pi}}^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F' \neq \emptyset] > p$.

The following then follows from the simulation Lemma 5 and standard concentration bounds (Lemma 36, (McDiarmid et al., 1989)).

Theorem 7 (PAC Guarantee)**.** *Let $\delta, \epsilon > 0$ be arbitrary. Suppose $p_{\min} > 0$ is a lower bound on positive transition probabilities in $\mathcal{M}$. If*

$$N \ge \frac{|S| \log(8|S||A|/\delta)}{(\epsilon')^2} \qquad \text{where } \epsilon' := \min\left\{\frac{p_{\min}}{2}, \frac{\epsilon}{4\eta}, \frac{\epsilon}{2\eta}(\overline{p} - p)\right\}, \quad \eta := \frac{2|S|}{p_{\min}^{|S|}}$$

then with probability at least $1 - \delta$, Problem (C-Recurrence) for the empirical estimate $\widehat{\mathcal{M}}$ is feasible and any optimal solution $\widehat{\pi}$ is ϵ-approximately feasible and optimal.

Whilst knowledge of $p_{\min}$ enables an explicit finite-sample guarantee, Algorithm 1 remains well-defined without it, although no *a priori* stopping criterion is then available. Nevertheless, increasing the number of samples drives both the feasibility violation and the suboptimality gap arbitrarily close to zero with arbitrarily high probability.

5 Reduction to Discounted Constrained MDPs

Complementary to the direct approach of the previous section, we now study reductions from recurrence objectives to standard discounted CMDPs. For *unconstrained* recurrence and, more generally, ω-regular objectives, Bozkurt et al. (2019) show that recurrence probabilities can be approximated by discounted returns with state-dependent discounting. Specifically, for a target set F, they define

$$J^{\mathcal{M}}_{F,\gamma,\bar{\gamma}}(\pi) := (1-\gamma) \cdot \mathbb{E}_{\rho \sim \mathcal{D}^{\mathcal{M}}_{\pi}} \left[\sum_{i \in \mathbb{N}} R_F(s_i) \cdot \prod_{j=0}^{i-1} \Gamma_{F,\gamma,\bar{\gamma}}(s_j) \right]$$

where $R_F(s) := \mathbf{1}_{\{s \in F\}}$ and $\Gamma_{F,\gamma,\bar{\gamma}}(s) := \gamma$ if $s \in F$ and $\Gamma_{F,\gamma,\bar{\gamma}}(s) := \bar{\gamma}$ otherwise. For suitable sequences $(\gamma_k)_{k \in \mathbb{N}}$ and $(\bar{\gamma}_k)_{k \in \mathbb{N}}$ converging to 1, the discounted objective converges to the recurrence probability:

$$\lim_{k \to \infty} J^{\mathcal{M}}_{F,\gamma_k,\bar{\gamma}_k}(\pi) = \mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}}_{\pi}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset]$$

for all $\mathcal{M}$, $F \subseteq S$, and deterministic stationary policies π (Bozkurt et al., 2019, Thm. 1).

A natural approach is to apply this reduction simultaneously to the objective and constraint sets F and F'. However, this does not immediately yield a solution to Problem (C-Recurrence): optimal policies for discounted CMDPs can be chosen stochastic stationary (Altman, 2021), which are insufficient in general for constrained recurrence (Example 1).

To resolve this mismatch, we restrict optimisation to the established sufficient class Π_{mds} of mixtures of deterministic stationary policies (Theorem 3). This yields the following sequence of discounted constrained optimisation problems indexed by $k \in \mathbb{N}$:

$$\max_{\pi \in \Pi_{mds}} \quad J^{\mathcal{M}_{\mathrm{aug}}}_{F \times \{0,1\},\gamma_k,\bar{\gamma}_k}(\pi) \qquad \text{s. t.} \quad J^{\mathcal{M}_{\mathrm{aug}}}_{F' \times \{0,1\},\gamma_k,\bar{\gamma}_k}(\pi) \geq p \qquad \text{(C-Discounted-}k\text{)}$$

Crucially, the discounted objective and constraint converge uniformly over Π_{mds} to the corresponding recurrence probabilities as $k \to \infty$ (see Section 11.2). Together with Assumption 6, this implies Problem (C-Discounted-k) is feasible for all sufficiently large k.

We consequently obtain the following asymptotic correspondence between discounted and recurrence-constrained optimisation problems (proof in Section 11):

Theorem 8. *Let $\epsilon, \delta > 0$. There exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, Problem (C-Discounted-k) is feasible, and for any mixture π of deterministic stationary policies over $\mathcal{M}_{\mathrm{aug}}$, if π is δ-approximately feasible and optimal for Problem (C-Discounted-k), then the corresponding 1-bit memory policy for $\mathcal{M}$ (cf. Remark 17) is $\delta + \epsilon$-approximately feasible and optimal for Problem (C-Recurrence).*

Discussion. The reduction to discounted CMDPs enables the use of existing methods for discounted constrained RL, provided they return mixtures of deterministic stationary policies. For example, Feinberg & Rothblum (2012, Alg. 2) propose a planning algorithm for policies of the required form (interpreting discounted MDPs as absorbing MDPs, cf. Section 11.1).

The main limitation of the reduction is that the required discount factors are MDP-dependent and generally not known *a priori*. In contrast, the direct approach of the previous section remains executable without such prior knowledge, although its finite-sample guarantees likewise depend on problem-specific quantities.

6 Extensions to ω-Regular Objectives and Constraints

Thus far, we have focused on Problem (C-Recurrence). We now extend our results to arbitrary ω-regular objectives and constraints. As is standard in probabilistic model checking and ω-regular reinforcement learning (e.g. (Hasanbeig et al., 2023; Voloshin et al., 2023; Shah et al., 2024)), we reduce Problem (C-ω-Regular) to Problem (C-Recurrence) via a product construction synchronising the MDP with automata representing the objective and constraint specifications. In particular, we use nondeterministic Büchi automata $\mathcal{A} = (Q, 2^{\mathcal{AP}}, q_0, \Delta, F)$, which accept an infinite word over $2^{\mathcal{AP}}$ if there exists a run whose state sequence visits the accepting set $F \subseteq Q$ infinitely often, specifically *limit-deterministic Büchi automata* (*LDBA*, see (Sickert et al., 2016; Hasanbeig et al., 2023) and Definition 40), which combine an initial nondeterministic phase with a deterministic accepting phase.

For LDBAs $\mathcal{A}$ and $\mathcal{A}'$ (with state space Q and Q' respectively), we form the product MDP $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$ with state space $S \times Q \times Q'$, obtained by synchronising the MDP with the automata states under the labelling map ℓ of $\mathcal{M}$ (see Definition 42). In addition, we define the accepting recurrence sets $F^{\otimes} := S \times F \times Q'$ and $(F')^{\otimes} := S \times Q \times F'$, where F and F' are the accepting states of $\mathcal{A}$ and $\mathcal{A}'$, respectively.

Intuitively, each product state $s^{\otimes} \in S \times Q \times Q$ records the current MDP and LDBA states. In addition to actions of $\mathcal{M}$, the product MDP also has actions $(\bar{q}, \bar{q}') \in Q \times Q'$ which explicitly perform (simultaneous) ϵ-transitions in both automata. Limit-determinism ensures any such action is performed at most once in any run. Hence, any mixture of stationary policies $\pi^{\otimes}$ on $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$ induces a finite-memory policy π on $\mathcal{M}$ with no worse performance with memory $Q \times Q'$, where the memory tracks the current automaton states (see Lemma 43). Moreover, mixtures of deterministic policies $\pi^{\otimes}$ for $(\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{B}')_{\text{aug}}$ yield mixtures of deterministic policies π for $\mathcal{M}$ with memory $Q \times Q' \times \{0, 1\}$ (see Lemma 44).

As in the unconstrained setting (Hahn et al., 2020; 2015; Sickert et al., 2016; Schewe et al., 2023), regardless of the ω-regular objective and constraint languages, suitable LDBAs can always be constructed (see Section 12.1 for the proof):

Proposition 9. *Let MDP $\mathcal{M} = (S, A, \text{Act}, \mu, P)$ be an MDP with labelling function $\ell : S \to 2^{\mathcal{AP}}$, let $L, L' \subseteq (2^{\mathcal{AP}})^{\omega}$ be ω-regular languages and let $p \in [0, 1]$.*

There exist LDBAs $\mathcal{A} = (Q, 2^{\mathcal{AP}}, q_0, \Delta, F)$ and $\mathcal{A}' = (Q', 2^{\mathcal{AP}}, q_0', \Delta', F')$ such that for any policy π which is feasible for Problem (C-ω-Regular) there exists a policy $\pi^{\otimes}$ for $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$ satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}_{\pi^{\otimes}}^{\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'}} [\text{Inf}(\rho) \cap F^{\otimes} \neq \emptyset] \geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi}^{\mathcal{M}}} [\ell(\rho) \in L(\mathcal{A})] \qquad \mathbb{P}_{\rho \sim \mathcal{D}_{\pi^{\otimes}}^{\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'}} [\text{Inf}(\rho) \cap (F')^{\otimes} \neq \emptyset] \geq p$$

Consequently, all recurrence-based results developed in Sections 3 and 5 lift directly to ω-regular objectives after replacing $\mathcal{M}$ by $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{B}'$ and (F, F') by $(F^{\otimes}, (F')^{\otimes})$. In particular, Theorem 3 extends to the ω-regular setting, where policies have to additionally track the states of $\mathcal{A}$ and $\mathcal{A}'$, instead of just being stationary over $\mathcal{M}$:

Theorem 10. *Feasible instances of Problem (C-ω-Regular) have solutions which are a mixture of two policies with memory $Q \times Q'$, likewise solutions which are a mixture of two deterministic policies with memory $Q \times Q' \times \{0, 1\}$.*

Moreover, all learning and sample-complexity guarantees (Theorems 7 and 8) extend by replacing S with $S \times Q \times Q'$.

7 Concluding Remarks

We proposed a theoretical framework for reinforcement learning with ω-regular objectives and constraints, extending the expressivity of existing formulations. Our LP-based characterisation establishes the structure of optimal policies and guarantees their existence. The

corresponding model-based algorithm with sample complexity results converges almost surely to an optimal constrained policy, and our translation to constrained discounted reward problems yields approximate solutions. These results provide a formal foundation for learning under rich temporal and safety specifications.

Acknowledgments

This research is supported by the National Research Foundation, Singapore, under its RSS Scheme (NRFRSS2022-009).

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Rajeev Alur, Suguman Bansal, Osbert Bastani, and Kishor Jothimurugan. A framework for transforming specifications in reinforcement learning. In Jean-François Raskin, Krishnendu Chatterjee, Laurent Doyen, and Rupak Majumdar (eds.), *Principles of Systems Design: Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday*, pp. 604–624, Cham, 2022. Springer Nature Switzerland.

Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. The MIT Press, 2008.

Alper Kamil Bozkurt, Yu Wang, Michael M. Zavlanos, and Miroslav Pajic. Control synthesis from Linear Temporal Logic specifications using model-free reinforcement learning. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10349–10355, 2019. URL https://api.semanticscholar.org/CorpusID:202577521.

Mingyu Cai, Shaoping Xiao, Zhijun Li, and Zhen Kan. Optimal probabilistic motion planning with potential infeasible LTL constraints. *IEEE Transactions on Automatic Control*, 68(1):301–316, 2023. DOI: 10.1109/TAC.2021.3138704.

Frank Ciesinski, Christel Baier, Marcus Größer, and Joachim Klein. Reduction techniques for model checking markov decision processes. In *2008 Fifth International Conference on Quantitative Evaluation of Systems*, pp. 45–54, 2008. DOI: 10.1109/QEST.2008.45.

Luca de Alfaro. Computing minimum and maximum reachability times in probabilistic systems. In Jos C. M. Baeten and Sjouke Mauw (eds.), *CONCUR'99 Concurrency Theory*, pp. 66–81, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48320-5.

Xuchu Ding, Stephen L. Smith, Calin Belta, and Daniela Rus. Optimal control of Markov Decision Processes with Linear Temporal Logic constraints. *IEEE Transactions on Automatic Control*, 59(5):1244–1257, 2014. DOI: 10.1109/TAC.2014.2298143.

Kousha Etessami, Marta Kwiatkowska, Moshe Y Vardi, and Mihalis Yannakakis. Multi-objective model checking of markov decision processes. *Logical Methods in Computer Science*, 4, 2008.

Eugene A Feinberg and Uriel G Rothblum. Splitting randomized stationary policies in total-reward markov decision processes. *Mathematics of Operations Research*, 37(1):129–153, 2012.

Vojtěch Forejt, Marta Kwiatkowska, Gethin Norman, and David Parker. Automated verification techniques for probabilistic systems. In *International school on formal methods for the design of computer, communication and software systems*, pp. 53–113. Springer, 2011a.

Vojtěch Forejt, Marta Kwiatkowska, Gethin Norman, David Parker, and Hongyang Qu. Quantitative multi-objective verification for probabilistic systems. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 112–127. Springer, 2011b.

Jie Fu and Ufuk Topcu. Probably approximately correct MDP learning and control with Temporal Logic constraints. In Dieter Fox, Lydia E. Kavraki, and Hanna Kurniawati (eds.), *Robotics: Science and Systems X, University of California, Berkeley, USA, July 12-16, 2014*, 2014.

Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1997.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ernst Moritz Hahn, Guangyuan Li, Sven Schewe, Andrea Turrini, and Lijun Zhang. Lazy probabilistic model checking without determinisation. In Luca Aceto and David de Frutos-Escrig (eds.), *26th International Conference on Concurrency Theory, CONCUR 2015, Madrid, Spain, September 1.4, 2015*, volume 42 of *LIPIcs*, pp. 354–367. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2015.

Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. Omega-regular objectives in model-free reinforcement learning. In Tomáš Vojnar and Lijun Zhang (eds.), *Tools and Algorithms for the Construction and Analysis of Systems*, pp. 395–412, Cham, 2019. Springer International Publishing. ISBN 978-3-030-17462-0.

Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. Good-for-mdps automata for probabilistic analysis and reinforcement learning. In Armin Biere and David Parker (eds.), *Tools and Algorithms for the Construction and Analysis of Systems - 26th International Conference, TACAS 2020, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2020, Dublin, Ireland, April 25-30, 2020, Proceedings, Part I*, volume 12078 of *Lecture Notes in Computer Science*, pp. 306–323. Springer, 2020.

Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. Multi-objective ω-regular reinforcement learning. *Formal Aspects of Computing*, 35(2):1–24, 2023.

Hosein Hasanbeig, Daniel Kroening, and Alessandro Abate. Certified reinforcement learning with logic guidance. *Artificial Intelligence*, 322:103949, 2023. ISSN 0004-3702.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Mateo Perez, Fabio Somenzi, and Ashutosh Trivedi. A PAC Learning Algorithm for LTL and Omega-Regular Objectives in MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21510–21517, 2024.

Sven Schewe, Qiyi Tang, and Tansholpan Zhanabekova. Deciding what is good-for-mdps. In Guillermo A. Pérez and Jean-François Raskin (eds.), *34th International Conference on Concurrency Theory, CONCUR 2023, September 18-23, 2023, Antwerp, Belgium*, volume 279 of *LIPIcs*, pp. 35:1–35:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

Ameesh Shah, Cameron Voloshin, Chenxi Yang, Abhinav Verma, Swarat Chaudhuri, and Sanjit A. Seshia. Ltl-constrained policy optimization with cycle experience replay, 2024.

Salomon Sickert, Javier Esparza, Stefan Jaax, and Jan Křetínský. Limit-deterministic Büchi automata for Linear Temporal Logic. In Swarat Chaudhuri and Azadeh Farzan (eds.), *Computer Aided Verification*, pp. 312–332, Cham, 2016. Springer International Publishing. ISBN 978-3-319-41540-6.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Sharan Vaswani, Lin Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, 35:3110–3122, 2022.

Cameron Voloshin, Hoang Le, Swarat Chaudhuri, and Yisong Yue. Policy optimization with Linear Temporal Logic constraints. *Advances in Neural Information Processing Systems*, 35: 17690–17702, 2022a.

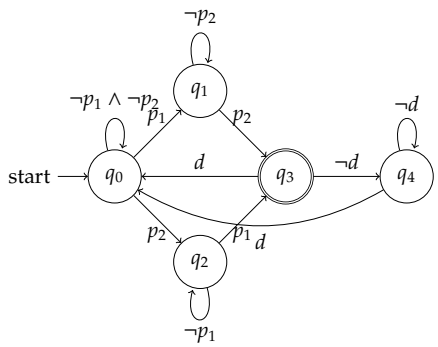
Cameron Voloshin, Hoang Minh Le, Swarat Chaudhuri, and Yisong Yue. Policy optimization with linear temporal logic constraints. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b.

Cameron Voloshin, Abhinav Verma, and Yisong Yue. Eventual discounting temporal logic counterfactual experience replay. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35137–35150. PMLR, 2023.

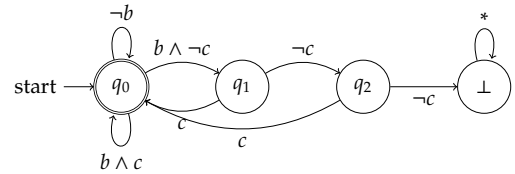
Akifumi Wachi, Xun Shen, and Yanan Sui. A survey of constraint formulations in safe reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 8262–8271. ijcai.org, 2024.

Eric M. Wolff, Ufuk Topcu, and Richard M. Murray. Robust control of uncertain Markov Decision Processes with Temporal Logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3372–3379, 2012. DOI: 10.1109/CDC.2012.6426174.

Cambridge Yang, Michael L. Littman, and Michael Carbin. On the (in)tractability of reinforcement learning for LTL objectives. In Luc De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 3650–3658. ijcai.org, 2022.



(a) Repeatedly visit pick-up locations p_1, p_2 and deliver the load to destination d.



(b) Visit a charging station (labelled c) within 2 time steps whenever the battery level gets critically low (labelled b).

Figure 3: Büchi automata for the logicstics drone example of Section 2.

Supplementary Materials

The following content was not necessarily subject to peer review.

8 Supplementary Materials for Section 2

For a countable sets X, let 2^X denote its power set, and $\mathcal{D}(X)$ denote the set of distributions over X.

We give a slightly more general definition of MDP than in Section 2, where we now allow initial state distributions.

Definition 11. An *MDP* $\mathcal{M}$ is a tuple $(S, A, \mathrm{Act}, \mu, P)$ with finite (non-empty) state set S, finite action set A, enabled action function $\mathrm{Act} : S \to (2^A \setminus \emptyset)$, initial state distribution $\mu \in \mathcal{D}(S)$, and transition kernel $P : S \times A \to \mathcal{D}(S)$.

For an MDP $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ and $\nu \in \mathcal{D}(S)$ let $\mathcal{M}[\nu] := (S, A, \mathrm{Act}, \nu, P)$ be the MDP with initial state distribution modified to ν.

Definition 12. A Büchi automaton is a tuple $\mathcal{A} = (Q, \Sigma, q_0, \Delta, F)$ with finite state set Q, finite alphabet Σ (we use $\Sigma := 2^{\mathcal{AP}}$), transition function $\Delta : Q \times (\Sigma \cup \{\epsilon\}) \to 2^Q$, and accepting states $F \subseteq Q$.

A word $w \in \Sigma^\omega$ is accepted by $\mathcal{A}$ (notation: $w \in L(\mathcal{A})$) if there exists $w_0, w_1, \ldots \in \Sigma \cup \{\epsilon\}$ such that $w = w_0 w_1 \cdots$ (removing ϵ symbols), $q_{i+1} \in \Delta(q_i, w_i)$ for all $i \in \mathbb{N}$, and $q_i \in F$ for infinitely many $i \in \mathbb{N}$.

Definition 13 (Policy)**.** A *policy* is a tuple $\pi = (\pi^{\mathrm{mem}}, \pi^{\mathrm{init}}, \pi^{\mathrm{act}})$, where π^{mem} is a potentially countably infinite set (representing memory), $\pi^{\mathrm{init}} \in \mathcal{D}(\pi^{\mathrm{mem}})$ is a distribution of initial memory states, and $\pi^{\mathrm{act}} : S \times \pi^{\mathrm{mem}} \to \mathcal{D}(A \times \pi^{\mathrm{mem}})$ is the kernel jointly selecting enabled actions and updating memory (i.e. $\pi^{\mathrm{act}}(a, m' \mid s, m) > 0$ implies $a \in \mathrm{Act}(s)$).

Moreover, π is *stationary* if $|\pi^{\mathrm{mem}}| = 1$, and it is *deterministic* if π^{init} and $\pi^{\mathrm{act}}(s, m)$ are Dirac distributions for all $s \in S$ and $m \in \pi^{\mathrm{mem}}$.

For stationary π we use standard notation $\pi(a \mid s) := \pi^{\mathrm{act}}(a, m \mid s, m)$ (for the unique $m \in \pi^{\mathrm{mem}}$), and for deterministic stationary π and $s \in S$ we use the notation $\pi(s) = a$ for the unique $a \in A$ and $m \in \pi^{\mathrm{mem}}$ satisfying $\pi^{\mathrm{act}}(a, m \mid s, m) = 1$.

Definition 14. For $\epsilon \geq 0$, a policy π is *ϵ-approximately feasible and optimal* if

1. $\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L'] \geq p - \epsilon$ and

2. $\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho)\in L] \geq \mathbb{P}_{\rho\sim\mathcal{D}_{\pi'}^{\mathcal{M}}}[\ell(\rho)\in L]$ for all policies π' satisfying $\mathbb{P}_{\rho\sim\mathcal{D}_{\pi'}^{\mathcal{M}}}[\ell(\rho)\in L'] \geq p$.

8.1 Supplementary Materials for Section 2.1

Definition 15 (Mixture of Policies)**.** Let $\pi_1 = (\pi_1^{\text{mem}}, \pi_1^{\text{init}}, \pi_1^{\text{act}})$ and $\pi_2 = (\pi_2^{\text{mem}}, \pi_2^{\text{init}}, \pi_2^{\text{act}})$ be policies and $\lambda \in [0,1]$. The *mixture* $p\cdot\pi_0 + (1-p)\cdot\pi_1$ of π_1 and π_2 (with mixing coefficient λ) is the policy $(\pi^{\text{mem}}, \pi^{\text{init}}, \pi^{\text{act}})$, where

$$\begin{aligned}
\pi^{\text{mem}} &:= (\pi_1^{\text{mem}} \times \{1\}) \cup (\pi_2^{\text{mem}} \times \{2\}) \\
\pi^{\text{init}}(m,i) &:= \begin{cases} \lambda\cdot\pi_1^{\text{init}}(m) & \text{if } i=1 \\ (1-\lambda)\cdot\pi_2^{\text{init}}(m) & \text{otherwise} \end{cases} \\
\pi^{\text{act}}(a,(m',i') \mid (m,i),s) &:= \mathbf{1}_{\{i=i'\}}\cdot\pi_i^{\text{act}}(a,m' \mid s,m)
\end{aligned}$$

Intuitively, this policy generates rollouts by following π_1 with probability λ, and otherwise produces rollouts by following π_2.

For a given $\mathcal{M}$, we use the following notation for the various policy classes: $\Pi(\mathcal{M})$ for all policies, $\Pi_s(\mathcal{M})$ for stationary policies, $\Pi_{ms}(\mathcal{M})$ for mixtures of stationary policies, $\Pi_{ds}(\mathcal{M})$ for deterministic stationary policies and $\Pi_{mds}(\mathcal{M})$ for mixtures of deterministic stationary policies. We omit $\mathcal{M}$ when clear from the context.

Definition 16. Let $\mathcal{M} = (S, A, \text{Act}, \mu, P)$ be an MDP and $F, F' \subseteq Q$. The *augmented MDP* $\mathcal{M}_{\text{aug}}$ is defined by $(S_{\text{aug}}, A_{\text{aug}}, \text{Act}_{\text{aug}}, \mu_{\text{aug}}, P_{\text{aug}})$ where

$$\begin{aligned}
S_{\text{aug}} &:= S \times \{0,1\} \\
A_{\text{aug}} &:= A \\
\text{Act}_{\text{aug}}(s,b) &:= \text{Act}(s) \\
\mu_{\text{aug}}(s,b) &:= \mu(s)\cdot\mathbf{1}_{\{b=0\}} \\
P_{\text{aug}}((s',b') \mid (s,b),a) &:= \begin{cases} P(s' \mid s,a) & \text{if } s\in F \text{ and } b'=0 \\ P(s' \mid s,a) & \text{if } s\in F'\setminus F \text{ and } b'=1 \\ P(s' \mid s,a) & \text{if } s\in S\setminus(F\cup F') \text{ and } b'=b \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Remark 17. Clearly, any policy π for $\mathcal{M}$ can be viewed as a policy π_{aug} for $\mathcal{M}_{\text{aug}}$ (by ignoring the additional bit of the state space) satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}_{\pi_{\text{aug}}}^{\mathcal{M}_{\text{aug}}}}[\text{Inf}(\rho)\cap(S'\times\{0,1\})\neq\emptyset] = \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\text{Inf}(\rho)\cap S'\neq\emptyset]$$

Conversely, a policy $\pi_{\text{aug}} = (\pi_{\text{aug}}^{\text{mem}}, \pi_{\text{aug}}^{\text{init}}, \pi_{\text{aug}}^{\text{act}})$ for $\mathcal{M}_{\text{aug}}$ naturally induces a policy π for $\mathcal{M}$ with one additional bit of memory tracking the bit of $\mathcal{M}_{\text{aug}}$:

$$\begin{aligned}
\pi^{\text{mem}} &:= \pi_{\text{aug}}^{\text{mem}} \times \{0,1\} \\
\pi^{\text{init}}(m,b) &:= \pi_{\text{aug}}^{\text{init}}(m)\cdot\mathbf{1}_{\{b=0\}} \\
\pi^{\text{act}}(a,(m',b') \mid s,(m,b)) &:= \begin{cases} \pi_{\text{aug}}^{\text{act}}(a,m' \mid (s,b),m) & \text{if } s\in F \text{ and } b'=0 \\ \pi_{\text{aug}}^{\text{act}}(a,m' \mid (s,b),m) & \text{if } s'\in F'\setminus F \text{ and } b'=1 \\ \pi_{\text{aug}}^{\text{act}}(a,m' \mid (s,b),m) & \text{if } s'\in S\setminus(F\cup F') \text{ and } b'=b \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Clearly, π is deterministic if π_{aug} is such, and for any $S' \subseteq S$,

$$\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\text{Inf}(\rho)\cap S'\neq\emptyset] = \mathbb{P}_{\rho\sim\mathcal{D}_{\pi_{\text{aug}}}^{\mathcal{M}_{\text{aug}}}}[\text{Inf}(\rho)\cap(S'\times\{0,1\})\neq\emptyset]$$

9 Supplementary Materials for Section 3

9.1 MDP End Components

Definition 18. An *end component* (EC) of an MDP $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ is a pair (S_C, Act_C) where $S_C \subseteq S$ and $\mathrm{Act}_C : S_C \to 2^A$ satisfy

1. $\sum_{s' \in S_C} P(s' \mid s, a) = 1$ for all $s \in S_C$, $a \in \mathrm{Act}_C(s)$,
2. the graph $(S_C, \to_{\mathrm{Act}_C})$ is strongly connected ($s \to_{\mathrm{Act}_C} s'$ iff $P(s' \mid s, a) > 0$ for some $a \in \mathrm{Act}_C(s)$)

We remark that this definition permits singleton ECs $(\{s\}, \mathrm{Act}_C)$, where $\mathrm{Act}_C(s) = \emptyset$, and call ECs other than these *proper*.

Definition 19. An EC C is *included* in an EC C' if $S_C \subseteq S_{C'}$ and $\mathrm{Act}_C(s) \subseteq \mathrm{Act}_{C'}(s)$ for all $s \in S_C$. A *maximal EC (MEC)* is a maximal element with respect to the inclusion relation.

It is readily verified that MECs have pairwise-disjoint state sets and hence partition S.

Lemma 20 ((de Alfaro, 1999, Thm. 3.2))**.** *Let $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an MDP and $\pi \in \Pi$. For a run ρ define*

$$S_{C(\rho)} := \mathrm{Inf}(\rho) \qquad \mathrm{Act}_{C(\rho)}(s) := \{a \in A \mid \textit{for infinitely many } i \in \mathbb{N}.\ \rho_i^S = s \textit{ and } \rho_i^A = a\}$$

and $C(\rho) := (S_{C(\rho)}, \mathrm{Act}_{C(\rho)})$. Then

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[C(\rho) \textit{ is an end component}] = 1$$

9.2 Absorbing MDPs

Definition 21. Let $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an MDP. A state $s \in S$ is *absorbing* if $P(s, a, s) = 1$ for all $a \in A$. The MDP $\mathcal{M}$ is *absorbing* if absorbing states are reachable from any state under any stationary policy . That is, for all $\nu \in \mathcal{D}(S)$ and $\pi \in \Pi_s$,

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}[\nu]}}\left[\exists i \in \mathbb{N}.\ \rho_i^S \in T\right] > 0$$

where $T \subseteq S$ is the set of absorbing states.

Let $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an absorbing MDP, where $T \subseteq S$ is the set of absorbing states, and let $R, R' : S \times A \to \mathbb{R}$ be functions satisfying $R(s, a) = 0 = R'(s, a)$ for $s \in T$. Since non-absorbing states can be visited only finitely often by any policy, the following constrained total reward problem is well-defined

$$\text{maximise } \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} R(\rho_i^S, \rho_i^A)\right] \text{ where } \pi \in \Pi \tag{C-Absorbing}$$

$$\text{subject to } \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} R'(\rho_i^S, \rho_i^A)\right] \geq p$$

Lemma 22. *Let $\mathcal{M}$ be an absorbing MDP. For $s \in S \setminus T$ and $a \in \mathrm{Act}(s)$, we have*

$$\sup_{\pi \in \Pi} \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i=0}^{\infty} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] < \infty$$

Proof. Fix $s \in S \setminus T$ and $a \in \mathrm{Act}(s)$. Define

$$r := \sup_{\pi \in \Pi} \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}[P(s,a)]}}[\exists i \in \mathbb{N}. \rho_i^S = s],$$

an upper bound on the probability of reaching s from initial state distribution $P(s,a) \in \mathcal{D}(S)$. If $r < 1$, we may conclude by observing

$$\sup_{\pi\in\Pi} \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i\in\mathbb{N}} \mathbf{1}_{\{\rho_i^S=s,\rho_i^A=a\}}\right] = \sup_{\pi\in\Pi}\sum_{i\in\mathbb{N}} \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{j\in\mathbb{N}} \mathbf{1}_{\{\rho_j^S=s,\rho_j^A=a\}} \geq i\right]$$
$$\leq \sup_{\pi\in\Pi}\sum_{i=1}^{\infty} r^{i-1} = \sum_{i\in\mathbb{N}} r^i = \frac{1}{1-r} < \infty$$

as needed. Now suppose towards a contradiction that $r = 1$. As this is an unconstrained reachability problem , there is a deterministic stationary policy π witnessing

$$\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}[P(s,a)]}}[\exists i \in \mathbb{N}.\rho_i^S = s] = r = 1.$$

Define a new deterministic stationary policy π^* by

$$\pi^*(a' \mid s') := \begin{cases} \mathbf{1}_{\{a'=a\}} & s' = s \\ \pi(a' \mid s') & \text{otherwise} \end{cases} \qquad s' \in S, a' \in \mathrm{Act}(s').$$

By construction, we have

$$\mathbb{P}_{\rho\sim\mathcal{D}_{\pi^*}^{\mathcal{M}[P(s,a)]}}\left[s \in \mathrm{Inf}(\rho)\right] = 1.$$

That is, π^* almost surely visits s infinitely often. However, as $\mathcal{M}$ is absorbing, we must have

$$\mathbb{P}_{\rho\sim\mathcal{D}_{\pi^*}^{\mathcal{M}[P(s,a)]}}\left[\exists i \in \mathbb{N}.\rho_i^S \in T\right] > 0.$$

Then we must have $s \in T$ since all states in T are absorbing, contradicting $s \in S \setminus T$ as needed. □

9.3 Linear Program for Absorbing MDPs

For absorbing $\mathcal{M}$, expected state-action occupation counts (aka occupancy measures) $x(s,a)$, where $s \in S \setminus T$ and $a \in \mathrm{Act}(s)$, can be axiomatised via flow equations, resulting in a linear programming formulation of Problem (C-Absorbing):

$$\begin{aligned}
\text{maximise} \quad & \sum_{s\in S\setminus T}\sum_{a\in\mathrm{Act}(s)} x(s,a)\cdot R(s,a) \\
\text{subject to} \quad & \sum_{s\in S\setminus T}\sum_{a\in\mathrm{Act}(s)} x(s,a)\cdot R'(s,a) \geq p \\
& \sum_{a\in\mathrm{Act}(s)} x(s,a) = \mu(s) + \sum_{s'\in S\setminus T}\sum_{a\in\mathrm{Act}(s')} P(s \mid s',a)\cdot x(s',a) && (s \in S\setminus T) \\
& x(s,a) \geq 0 && (s \in S\setminus T, a \in \mathrm{Act}(s))
\end{aligned} \tag{LP-Flow}$$

Note that arbitrary policies π give rise to assignments

$$x(s,a) = \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i=0}^{\infty} \mathbf{1}_{\{\rho_i^S=s,\rho_i^A=a\}}\right] \qquad (s \in S\setminus T, a \in \mathrm{Act}(s)) \tag{1}$$

satisfying Eq. (LP-Flow), see Lemma 25. Conversely, assignments satisfying Eq. (LP-Flow) give rise to stationary stochastic policies $\pi(a \mid s) \propto x(s,a)$ satisfying Eq. (1), see Lemma 23.

Lemma 23. *Let $\mathcal{M}$ be an absorbing MDP. If $x : (S \setminus T) \times A \to \mathbb{R}_{\geq 0}$ satisfies (LP-Flow), then there is a stochastic stationary policy π satisfying*

$$\mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] = x(s,a) \qquad s \in S \setminus T, a \in \mathrm{Act}(s)$$

and for $s \in S \setminus T$,

$$\sum_{a \in \mathrm{Act}(s)} \mathbf{1}_{\{\pi(a|s) > 0\}} = \max\left\{1, \sum_{a \in A} \mathbf{1}_{\{x(s,a) > 0\}}\right\}$$

Proof. Let x be a point in the LP feasible region. For each $s \in S$, fix an arbitrary $a_s \in \mathrm{Act}(s)$ and define

$$\pi(a \mid s) := \begin{cases} \frac{x(s,a)}{\sum_{a' \in \mathrm{Act}(s)} x(s,a')} & s \notin T \text{ and } \sum_{a' \in \mathrm{Act}(s)} x(s,a') > 0 \\ \mathbf{1}_{\{a = a_s\}} & \text{otherwise} \end{cases} \qquad s \in S, a \in \mathrm{Act}(s)$$

Let $s \in S \setminus T$ and $a \in \mathrm{Act}(s)$. First, suppose $\sum_{a' \in \mathrm{Act}(s)} x(s,a') > 0$. We observe

$$\begin{aligned}
&\pi(a \mid s) \cdot \left(\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')\right) \\
&= \frac{x(s,a)}{\sum_{a' \in \mathrm{Act}(s)} x(s,a')} \cdot \left(\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')\right) \\
&= x(s,a) \cdot \frac{\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')}{\sum_{a' \in \mathrm{Act}(s)} x(s,a')} \\
&= x(s,a) \cdot \frac{\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')}{\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')} \\
&= x(s,a)
\end{aligned}$$

by (LP-Flow). Now suppose otherwise that $\sum_{a' \in \mathrm{Act}(s)} x(s,a') = 0$. Then

$$x(s,a) = 0 = \sum_{a' \in \mathrm{Act}(s)} x(s,a') = \mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a'),$$

where we again use (LP-Flow). Thus, in either case, we have

$$x(s,a) = \pi(a \mid s) \cdot \left(\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')\right),$$

so Corollary 27 yields

$$x(s,a) = \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] \qquad s \in S \setminus T, a \in \mathrm{Act}(s).$$

□

Corollary 24. *The LP has bounded feasible region.*

Proof. Immediate consequence of Lemmas 22 and 23. □

Lemma 25. *Let $\pi \in \Pi$. Then the following assignment satisfies Eq. (LP-Flow):*

$$x(s,a) := \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i=0}^{\infty} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] \qquad s \in S \setminus T, a \in \mathrm{Act}(s)$$

Proof. Fix $s \in S \setminus T$. We verify

$$\begin{aligned}
\sum_{a\in\mathrm{Act}(s)} x(s,a) &= \sum_{a\in\mathrm{Act}(s)} \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i=0}^{\infty} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] = \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i=0}^{\infty} \mathbf{1}_{\{\rho_i^S = s\}}\right] \\
&= \mu(s) + \sum_{i\in\mathbb{N}} \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\rho_{i+1}^S = s\right] \\
&= \mu(s) + \sum_{i\in\mathbb{N}} \sum_{s'\in S\setminus T} \sum_{a\in\mathrm{Act}(s')} \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\rho_i^S = s', \rho_i^A = a, \rho_{i+1}^S = s\right] \\
&= \mu(s) + \sum_{i\in\mathbb{N}} \sum_{s'\in S\setminus T} \sum_{a\in\mathrm{Act}(s')} P(s \mid s', a) \cdot \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\rho_i^S = s', \rho_i^A = a\right] \\
&= \mu(s) + \sum_{s'\in S\setminus T} \sum_{a\in\mathrm{Act}(s')} P(s \mid s', a) \cdot \sum_{i\in\mathbb{N}} \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\rho_i^S = s', \rho_i^A = a\right] \\
&= \mu(s) + \sum_{s'\in S\setminus T} \sum_{a\in\mathrm{Act}(s')} P(s \mid s', a) \cdot \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i=0}^{\infty} \mathbf{1}_{\{\rho_i^S = s', \rho_i^A = a\}}\right] \\
&= \mu(s) + \sum_{s'\in S\setminus T} \sum_{a\in\mathrm{Act}(s')} P(s \mid s', a,) \cdot x(s', a)
\end{aligned}$$

as needed. □

Lemma 26 ((Grinstead & Snell, 1997, Thm. 11.4))**.** *Let $\mathcal{M} = (S, s_0, P)$ be an absorbing Markov chain (i.e. an absorbing MDP with $|A| = 1$). Then the map*

$$s \mapsto \mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}}}\left[\sum_{i\in\mathbb{N}} \mathbf{1}_{\{\rho_i^S = s\}}\right] \qquad s \in S \setminus T$$

is the unique $x : S \setminus T \to \mathbb{R}$ satisfying

$$x(s) = \mu(s) + \sum_{s'\in S\setminus T} P(s \mid s') \cdot x(s') \qquad s \in S \setminus T.$$

Corollary 27. *Let $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an absorbing MDP and π be stationary. Then*

$$x(s,a) := \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i\in\mathbb{N}} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] \qquad s \in S \setminus T, a \in \mathrm{Act}(s)$$

is the unique $x : \{(s,a) \in (S \setminus T) \times A \mid a \in \mathrm{Act}(s)\} \to \mathbb{R}$ satisfying

$$x(s,a) = \pi(a \mid s) \cdot \left(\mu(s) + \sum_{s'\in S\setminus T} \sum_{a'\in\mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a')\right) \qquad s \in S \setminus T, a \in \mathrm{Act}(s).$$

Proof. The "is a solution" direction follows from Lemma 25. Define the Markov chain $\mathcal{M}_\pi := (S, \mu, P_\pi)$ with P_π given by

$$P_\pi(s' \mid s) := \sum_{a\in\mathrm{Act}(s)} \pi(a \mid s) \cdot P(s' \mid s, a) \qquad s, s' \in S.$$

Let T_π denote the absorbing states of $\mathcal{M}_\pi$:

$$T_\pi := \{s \in S \mid P_\pi(s \mid s) = 1\} = \left\{ s \in S \mid \sum_{a \in \mathrm{Act}(s)} \pi(a \mid s) \cdot P(s \mid s, a) = 1 \right\}.$$

It is evident that $T \subseteq T_\pi$. Furthermore, if there was $s \in T_\pi \setminus T$, we would have

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}[s]}} \left[\exists i \in \mathbb{N}. \rho_i^S \in T \right] = 0,$$

which contradicts $\mathcal{M}$ being absorbing. Thus we must instead have $T_\pi = T$. By construction, we have

$$\mathcal{D}^{\mathcal{M}_\pi[\nu]} = \mathcal{D}_\pi^{\mathcal{M}[\nu]} \qquad \nu \in \mathcal{D}(S)$$

so

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}_\pi[\nu]}} \left[\exists i \in \mathbb{N}. \rho_i^S \in T \right] = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}[\nu]}} \left[\exists i \in \mathbb{N}. \rho_i^S \in T \right] > 0 \qquad \nu \in \mathcal{D}(S)$$

as $\mathcal{M}$ is absorbing. Thus $\mathcal{M}_\pi$ is also absorbing, and Lemma 26 yields

$$y(s) := \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}} \left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s\}} \right] \qquad s \in S \setminus T$$

is the unique $y : S \setminus T \to \mathbb{R}$ satisfying

$$y(s) = \mu(s) + \sum_{s' \in S \setminus T} P_\pi(s \mid s) \cdot y(s') \qquad s \in S \setminus T.$$

Now suppose $x : (S \setminus T) \times A \to \mathbb{R}$ satisfies

$$x(s, a) = \pi(a \mid s) \cdot \left(\mu(s) + \sum_{s' \in S \setminus T} \sum_{a' \in \mathrm{Act}(s')} P(s \mid s', a') \cdot x(s', a') \right) \qquad s \in S \setminus T, a \in \mathrm{Act}(s).$$

We define

$$\overline{x}(s) := \sum_{a \in \mathrm{Act}(s)} x(s, a) \qquad s \in S \setminus T.$$

Then for $s \in S \setminus T$, we have

$$\begin{aligned} \overline{x}(s) = \sum_{a \in \mathrm{Act}(s)} x(s, a) &= \sum_{a \in \mathrm{Act}(s)} \pi(a \mid s) \cdot \left(\mu(s) + \sum_{s' \in S \setminus T} \sum_{a \in \mathrm{Act}(s')} P(s \mid s', a) \cdot x(s', a) \right) \\ &= \mu(s) + \sum_{s' \in S \setminus T} \sum_{a \in \mathrm{Act}(s')} P(s \mid s', a) \cdot x(s', a), \end{aligned}$$

so

$$x(s, a) = \pi(a \mid s) \cdot \overline{x}(s) \qquad s \in S \setminus T, a \in \mathrm{Act}(s)$$

and

$$\begin{aligned} \overline{x}(s) &= \mu(s) + \sum_{s' \in S \setminus T} \sum_{a \in \mathrm{Act}(s')} P(s \mid s', a) \cdot \pi(a \mid s') \cdot \overline{x}(s') \\ &= \mu(s) + \sum_{s' \in S \setminus T} \left(\sum_{a \in \mathrm{Act}(s')} \pi(a \mid s') \cdot P(s \mid s', a) \right) \cdot \overline{x}(s') \\ &= \mu(s) + \sum_{s' \in S \setminus T} P_\pi(s \mid s') \cdot \overline{x}(s') \end{aligned}$$

for $s \in S \setminus T$. By uniqueness, we must have

$$\overline{x}(s) = y(s) = \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s\}}\right] \qquad s \in S \setminus T.$$

Thus, we conclude

$$\begin{aligned} x(s,a) &= \pi(a \mid s) \cdot \overline{x}(s) \\ &= \pi(a \mid s) \cdot \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s\}}\right] \\ &= \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] \end{aligned}$$

for $s \in S \setminus T$ and $a \in \mathrm{Act}(s)$ as needed. □

9.4 Collapsed MDP Construction

We first introduce the notation $\mathfrak{M}$ and $\mathfrak{M}_p$ for the set of MECs and proper MECs, respectively. Besides, for an EC $C = (S_C, \mathrm{Act}_C)$ we define $\mathrm{out}(C)$ to be the state-action pairs which may leave C:

$$\mathrm{out}(C) \coloneqq \{(s,a) \in S_C \times A \mid a \in \mathrm{Act}(s) \setminus \mathrm{Act}_C(s)\}$$

Definition 28. Let $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an MDP. The *collapsed MDP* $\mathcal{M}_{\mathrm{coll}}$ is the MDP $(S_{\mathrm{coll}}, A_{\mathrm{coll}}, \mathrm{Act}_{\mathrm{coll}}, \mu_{\mathrm{coll}}, P_{\mathrm{coll}})$, where

$$\begin{aligned} S_{\mathrm{coll}} &\coloneqq \mathfrak{M} \cup \{\bot\} \\ A_{\mathrm{coll}} &\coloneqq \bigcup_{C \in \mathfrak{M}} \mathrm{out}(C) \cup \{\textsc{stay}\} \\ \mathrm{Act}_{\mathrm{coll}}(s_{\mathrm{coll}}) &\coloneqq \begin{cases} \mathrm{out}(s_{\mathrm{coll}}) \cup \{\textsc{stay}\} & \text{if } s_{\mathrm{coll}} \in \mathfrak{M}_p \\ \mathrm{out}(s_{\mathrm{coll}}) & \text{if } s_{\mathrm{coll}} \in \mathfrak{M} \setminus \mathfrak{M}_p \\ \{\textsc{stay}\} & \text{otherwise } (s_{\mathrm{coll}} = \bot) \end{cases} \\ \mu_{\mathrm{coll}}(s_{\mathrm{coll}}) &\coloneqq \begin{cases} \sum_{s \in S_C} \mu(s) & \text{if } s_{\mathrm{coll}} = (S_C, \mathrm{Act}_C) \in \mathfrak{M} \\ 0 & \text{otherwise} \end{cases} \\ P_{\mathrm{coll}}(s'_{\mathrm{coll}} \mid s_{\mathrm{coll}}, a_{\mathrm{coll}}) &\coloneqq \begin{cases} \sum_{s' \in S_{C'}} P(s' \mid s, a) & \text{if } C \coloneqq s_{\mathrm{coll}} \in \mathfrak{M}, (s,a) \coloneqq a_{\mathrm{coll}} \in \mathrm{out}(C), C' \coloneqq s'_{\mathrm{coll}} \in \mathfrak{M} \\ 1 & \text{if } a_{\mathrm{coll}} = \textsc{stay}, s'_{\mathrm{coll}} = \bot \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

That is, the probability to transition to C' (in one step) when performing a in s, whilst STAY (if enabled) transitions deterministically to $\bot$, which is the only absorbing state.

Lemma 29. *$\mathcal{M}_{\mathrm{coll}}$ is absorbing.*

Proof. Observe that $\bot$ is the only absorbing state in $\mathcal{M}_{\mathrm{coll}}$ and assume towards contradiction that there exists $C \in S_{\mathrm{coll}}$ and a deterministic stationary policy for $\mathcal{M}_{\mathrm{coll}}$ satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}_{\mathrm{coll}}[C]}}\left[\exists i \in \mathbb{N}.\, \rho_i^S = \bot\right] = 0$$

Then there exists an EC C_{coll} for $\mathcal{M}_{\mathrm{coll}}$ satisfying $\bot \notin S_{C_{\mathrm{coll}}}$. Note that by definition of $\mathcal{M}_{\mathrm{coll}}$ and end components, $S_{C_{\mathrm{coll}}}$ cannot be a singleton. Now consider,

$$S_{C'} \coloneqq \bigcup_{C'' \in S_{C_{\mathrm{coll}}}} S_{C''} \subseteq S \qquad \mathrm{Act}_{C'}(s) \coloneqq \mathrm{Act}_{C_s}(s) \cup \{a \in A \mid (s,a) \in \mathrm{Act}_{C_{\mathrm{coll}}}(C_s)\} \subseteq \mathrm{Act}(s)$$

where $s \in S_{C'} \subseteq S$ and $C_s \in S_{C_{\text{coll}}}$ is the unique MEC in $\mathcal{M}$ containing.

By construction $C' := (S_{C'}, \text{Act}_{C'})$ is an EC, and it strictly extends any EC in $S_{C_{\text{coll}}}$, which is a contradiction to their supposed maximality. □

Definition 30. Give a recurrence specification $F, F' \subseteq S$, we define natural reward and constraint functions $R_{\text{coll}}, R'_{\text{coll}} : S_{\text{coll}} \times A_{\text{coll}} \to \mathbb{R}$ for $\mathcal{M}_{\text{coll}}$ as follows:

$$R_{\text{coll}}(C, \textsc{stay}) := \mathbf{1}_{\{S_C \cap F \neq \emptyset\}} \qquad R'_{\text{coll}}(C, \textsc{stay}) := \mathbf{1}_{\{S_C \cap F' \neq \emptyset\}}$$

where $C \in \mathfrak{M}_p$, with all other transitions assigned 0 reward/constraint.

Intuitively, R_{coll} and R'_{coll} rewards to committing to proper MECs intersecting F and F', respectively, because once a proper MEC has been reached, visiting all its states infinitely often is readily achievable (e.g. by performing all MEC-enabled actions uniformly at random).

Lemma 31. *Let π_{coll} be a deterministic stationary policy for the absorbing $\mathcal{M}_{\text{coll}}$. There exists a stochastic stationary policy $\pi_s \in \Pi_s$ over the original $\mathcal{M}$ and a deterministic stationary $\pi_{ds} \in \Pi_{ds}$ over the augmented original $\mathcal{M}_{\text{aug}}$ satisfying*

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}}_{\pi_s}}[\text{Inf}(\rho) \cap F \neq \emptyset] \geq \mathbb{E}_{\rho \sim \mathcal{D}^{\mathcal{M}_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i \in \mathbb{N}} R_{\text{coll}}(\rho^S_i, \rho^A_i)\right]$$

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}}_{\pi_s}}[\text{Inf}(\rho) \cap F' \neq \emptyset] \geq \mathbb{E}_{\rho \sim \mathcal{D}^{\mathcal{M}_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i \in \mathbb{N}} R'_{\text{coll}}(\rho^S_i, \rho^A_i)\right]$$

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}_{\text{aug}}}_{\pi_{ds}}}[\text{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] \geq \mathbb{E}_{\rho \sim \mathcal{D}^{\mathcal{M}_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i \in \mathbb{N}} R_{\text{coll}}(\rho^S_i, \rho^A_i)\right]$$

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}_{\text{aug}}}_{\pi_{ds}}}[\text{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] \geq \mathbb{E}_{\rho \sim \mathcal{D}^{\mathcal{M}_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i \in \mathbb{N}} R'_{\text{coll}}(\rho^S_i, \rho^A_i)\right]$$

Proof. First, let π_s be the stochastic stationary policy selecting actions $\text{Act}_C(s)$ uniformly at random for $s \in S_C$ unless $\pi_{\text{coll}}(C) = (s, a)$ in which case π_s selects a deterministically in state s. Note that any run can only get a reward 1 if it performs action STAY in a non-degenerate $C \in \mathfrak{M}$ satisfying $S_C \cap F \neq \emptyset$. The corresponding run in $\mathcal{M}$ under π_s almost-surely visits all states in S_C, hence some state in F, infinitely often. For analogous reasons the second inequality for holds.

Next, we define a deterministic policy π_{ds} over the augmented MDP which achieves the same. For $s \in S_C$ and $\emptyset \neq S' \subseteq S_C$ let $a \in \text{Act}_C$ be a fixed action on the shortest (in terms of the number of edges) path in the graph $(S_C, \to_{\text{Act}_C})$ from s to S'. We define $\pi_{ds}(s, b)$, where $b \in \{0, 1\}$ is the additional augmentation bit, by case distinction on the unique $C = (S_C, \text{Act}_C)$ with $s \in S_C$:

- If $\pi_{\text{coll}}(C) = (s', a)$ then $\pi_{ds}((s, b)) := a_{s,\{s'\}}$ for $s \neq s'$ and $\pi_{ds}(s') := a$.
- If $\pi_{\text{coll}}(C) = \textsc{stay}$ then
  - If $S_C \cap F \neq \emptyset$ but $S_C \cap F' = \emptyset$ then $\pi_{ds}(s, b) := a_{s,S_C \cap F}$.
  - If $S_C \cap F' \neq \emptyset$ but $S_C \cap F = \emptyset$ then $\pi_{ds}(s, b) := a_{s,S_C \cap F'}$.
  - If $S_C \cap F \cap F' \neq \emptyset$ then $\pi_{ds}(s, b) := a_{s,S_C \cap F}$.
  - If $S_C \cap F \neq \emptyset$ and $S_C \cap F' \neq \emptyset$ and $S_C \cap F \cap F' = \emptyset$ then $\pi_{ds}(s, 0) := a_{s,S_C \cap F'}$ and $\pi_{ds}(s, 1) := a_{s,S_C \cap F}$.

In the last case, we can almost surely visit states in both F and F' infinitely often. Our definition alternates between visiting states in $F \cap S_C$ and $F' \cap S_C$ by again taking actions

on the shortest path to the respective set in C. The additional bit of the augmented MDP (which tracks whether F or $F' \setminus F$ have last been visited) allows us to alternately pick the right target amongst $F \cap S_C$ and $F' \cap S_C$. □

Lemma 32. *Let π be a policy for $\mathcal{M}$. Define*

$$x(C,(s,a)) \coloneqq \mathbb{E}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i\in\mathbb{N}} \mathbf{1}_{\{\rho_i^S=s,\rho_i^A=a\}}\right] \qquad (C\in\mathfrak{M},(s,a)\in\mathrm{out}(C))$$

$$x(C,\textsc{stay}) \coloneqq \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\exists i\in\mathbb{N}.\,\forall j\geq i.\,\rho_j^S\in S_C\right] \qquad (C\in\mathfrak{M}_p)$$

Then x satisfies (LP-Flow) *and*

$$\sum_{s_{\mathrm{coll}}\in S_{\mathrm{coll}}\setminus T_{\mathrm{coll}}}\ \sum_{a_{\mathrm{coll}}\in\mathrm{Act}_{\mathrm{coll}}(s_{\mathrm{coll}})} x(s_{\mathrm{coll}},a_{\mathrm{coll}})\cdot R(s_{\mathrm{coll}},a_{\mathrm{coll}}) \geq \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\mathrm{Inf}(\rho)\cap F\neq\emptyset\right]$$

$$\sum_{s_{\mathrm{coll}}\in S_{\mathrm{coll}}\setminus T_{\mathrm{coll}}}\ \sum_{a_{\mathrm{coll}}\in\mathrm{Act}_{\mathrm{coll}}(s_{\mathrm{coll}})} x(s_{\mathrm{coll}},a_{\mathrm{coll}})\cdot R'(s_{\mathrm{coll}},a_{\mathrm{coll}}) \geq \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\mathrm{Inf}(\rho)\cap F'\neq\emptyset\right]$$

Proof. By definition all variables are non-negative. Note that

$$\begin{aligned}\sum_{s_{\mathrm{coll}}\in S_{\mathrm{coll}}\setminus T_{\mathrm{coll}}}\ \sum_{a_{\mathrm{coll}}\in\mathrm{Act}_{\mathrm{coll}}(s_{\mathrm{coll}})} x(s_{\mathrm{coll}},a_{\mathrm{coll}})\cdot R(s_{\mathrm{coll}},a_{\mathrm{coll}}) &= \sum_{C\in\mathfrak{M}_p}\mathbf{1}_{\{F\cap S_C\neq\emptyset\}}\cdot x(C,\textsc{stay})\\ &\geq \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}\left[\mathrm{Inf}(\rho)\cap F\neq\emptyset\right]\end{aligned}$$

since $R(C,a_{\mathrm{coll}})=1$ only if $a_{\mathrm{coll}}=\textsc{stay}$ and $S_C\cap F\neq\emptyset$, and $\textsc{stay}$ is only enabled in $C\in\mathfrak{M}_p$, and almost surely $\mathrm{Inf}(\rho)\cap F\neq\emptyset$ only if the (proper) maximal end component ρ enters and never leaves intersects with F. (The second inequality is verified analogously.)

Finally, we turn towards the flow equation. Because policies almost surely eventually remain within one proper EC, hence one proper MEC (equivalently, $\sum_{C\in\mathfrak{M}_p} x(C,\textsc{stay})=1$), we have for $C\in\mathfrak{M}$,

$$\lim_{i\to\infty}\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\rho_i^S\in S_C] = \begin{cases} x(C,\textsc{stay}) & \text{if } C\in\mathfrak{M}_p\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

Next, observe that for all runs ρ, $C\in\mathfrak{M}$, and $i\in\mathbb{N}$,

$$\mathbf{1}_{\{\rho_i^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{\rho_j^S\in S_C,\rho_{j+1}^S\notin S_C\}} = \mathbf{1}_{\{\rho_0^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{\rho_j^S\notin S_C,\rho_{j+1}^S\in S_C\}}.$$

That is, up to time i, there is a correspondence between the number of times each C was entered and exited in run ρ. Hence we have

$$\begin{aligned}\mathbf{1}_{\{\rho_i^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{(\rho_j^S,\rho_j^A)\in\mathrm{out}(C)\}} &= \mathbf{1}_{\{\rho_i^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{(\rho_j^S,\rho_j^A)\in\mathrm{out}(C),\rho_{j+1}^S\notin S_C\}} + \sum_{j<i}\mathbf{1}_{\{(\rho_j^S,\rho_j^A)\in\mathrm{out}(C),\rho_{j+1}^S\in S_C\}}\\ &= \left(\mathbf{1}_{\{\rho_i^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{\rho_j^S\in S_C,\rho_{j+1}^S\notin S_C\}}\right) + \sum_{j<i}\mathbf{1}_{\{(\rho_j^S,\rho_j^A)\in\mathrm{out}(C),\rho_{j+1}^S\in S_C\}}\\ &= \left(\mathbf{1}_{\{\rho_0^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{\rho_j^S\notin S_C,\rho_{j+1}^S\in S_C\}}\right) + \sum_{j<i}\mathbf{1}_{\{(\rho_j^S,\rho_j^A)\in\mathrm{out}(C),\rho_{j+1}^S\in S_C\}}\\ &= \mathbf{1}_{\{\rho_0^S\in S_C\}} + \sum_{j<i}\mathbf{1}_{\{(\rho_j^S,\rho_j^A)\in\mathrm{out},\rho_{j+1}^S\in S_C\}}\end{aligned}$$

where $\text{out} := \bigcup_{C \in \mathfrak{M}} \text{out}(C)$. Taking expectation over ρ yields

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\rho_i^S \in S_C\right] + \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{j<i} \mathbf{1}_{\{(\rho_j^S, \rho_j^A) \in \text{out}(C)\}}\right] = \mu_{\text{coll}}(C) + \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{j<i} \mathbf{1}_{\{(\rho_j^S, \rho_j^A) \in \text{out}, \rho_{j+1}^S \in S_C\}}\right]$$

We now inspect the limiting behavior of both sides of this equation as $i \to \infty$. By Eq. (2), for proper C (and similarly for improper C),

$$\begin{aligned}
&\lim_{i \to \infty}\left(\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\rho_i^S \in S_C\right] + \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{j<i} \mathbf{1}_{\{(\rho_j^S, \rho_j^A) \in \text{out}(C)\}}\right]\right) \\
&= x(C, \textsc{stay}) + \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{(\rho_i^S, \rho_i^A) \in \text{out}(C)\}}\right] \\
&= x(C, \textsc{stay}) + \sum_{(s,a) \in \text{out}(C)} \underbrace{\mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{(\rho_i^S, \rho_i^A) = (s,a)\}}\right]}_{x(C,(s,a))} \\
&= \sum_{a_{\text{coll}} \in \text{Act}_{\text{coll}}(C)} x(C, a_{\text{coll}})
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\lim_{i \to \infty}\left(\mu_{\text{coll}}(C) + \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{j<i} \mathbf{1}_{\{(\rho_j^S, \rho_j^A) \in \text{out}, \rho_{j+1}^S \in S_C\}}\right]\right) \\
&= \mu_{\text{coll}}(C) + \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{(\rho_i^S, \rho_i^A) \in \text{out}, \rho_{i+1}^S \in S_C\}}\right] \\
&= \mu_{\text{coll}}(C) + \sum_{C' \in \mathfrak{M}} \sum_{(s,a) \in \text{out}(C')} \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{(\rho_i^S, \rho_i^A) = (s,a), \rho_{i+1}^S \in S_C\}}\right] \\
&= \mu_{\text{coll}}(C) + \sum_{C' \in \mathfrak{M}} \sum_{(s,a) \in \text{out}(C')} \underbrace{\mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{(\rho_i^S, \rho_i^A) = (s,a)\}}\right]}_{x(C',(s,a))} \cdot P_{\text{coll}}(C \mid C', (s,a)) \\
&= \mu_{\text{coll}}(C) + \sum_{C' \in S_{\text{coll}} \setminus T_{\text{coll}}} \sum_{a_{\text{coll}} \in \text{Act}_{\text{coll}}(C')} P_{\text{coll}}(C \mid C', a_{\text{coll}}) \cdot x(C', a_{\text{coll}})
\end{aligned}$$

since $P(C \mid C', a_{\text{coll}}) > 0$ only if $a_{\text{coll}} \neq \textsc{stay}$. Thus we conclude

$$\sum_{a_{\text{coll}} \in \text{Act}_{\text{coll}}(C)} x(C, a_{\text{coll}}) = \mu_{\text{coll}}(C) + \sum_{s'_{\text{coll}} \in S_{\text{coll}} \setminus T_{\text{coll}}} \sum_{a_{\text{coll}} \in \text{Act}_{\text{coll}}(s'_{\text{coll}})} P_{\text{coll}}(C \mid s'_{\text{coll}}, a_{\text{coll}}) \cdot x(s'_{\text{coll}}, a_{\text{coll}})$$

for all $C \in \mathfrak{M} = S_{\text{coll}} \setminus T_{\text{coll}}$ as needed. □

9.5 Proof of Theorem 3

Theorem 3. *Feasible instances of Problem (C-Recurrence) have solutions which are a mixture of two stochastic stationary policies, likewise solutions which are a mixture of two deterministic stationary policies over the augmented MDP (which by Remark 17 can be viewed as mixtures of deterministic policies with 1 bit memory).*

Proof. Let π' be a policy for $\mathcal{M}$ satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\mathrm{Inf}(\rho)\cap F'\neq\emptyset\right]\geq p$$

By Lemmas 23 and 32 there exists a policy π'_{coll} for $\mathcal{M}_{\mathrm{coll}}$ satisfying

$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi'_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]\geq\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi'}}\left[\mathrm{Inf}(\rho)\cap F\neq\emptyset\right]$$
$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi'_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R'(\rho^S_i,\rho^A_i)\right]\geq\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi'}}\left[\mathrm{Inf}(\rho)\cap F'\neq\emptyset\right]\geq p$$

In particular, Problem (C-Absorbing) is feasible. By (Feinberg & Rothblum, 2012, Alg. 2,Thm. 9.2) there exists a mixture of deterministic policies π_{coll} for $\mathcal{M}_{\mathrm{coll}}$ satisfying

$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi'_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]$$
$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R'(\rho^S_i,\rho^A_i)\right]\geq p$$

By Lemma 31 there exists a mixture π_s of stochastic stationary policies over $\mathcal{M}$ satisfying

$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi_s}}\left[\mathrm{Inf}(\rho)\cap F\neq\emptyset\right]\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]$$
$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi_s}}\left[\mathrm{Inf}(\rho)\cap F'\neq\emptyset\right]\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R'(\rho^S_i,\rho^A_i)\right]$$

Hence, π_s is a feasible solution for Problem (C-Recurrence) and it is optimal since π' was an arbitrary feasible solution for which we have shown

$$\begin{aligned}\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi_s}}\left[\mathrm{Inf}(\rho)\cap F\neq\emptyset\right]&\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]\\&\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi'_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]\geq\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi'}}\left[\mathrm{Inf}(\rho)\cap F\neq\emptyset\right]\end{aligned}$$

Likewise, by Lemma 31 there exists a mixture π_{ds} of deterministic stationary policies over $\mathcal{M}_{\mathrm{aug}}$

$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{aug}}}_{\pi_{ds}}}\left[\mathrm{Inf}(\rho)\cap(F\times\{0,1\})\neq\emptyset\right]\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R(\rho^S_i,\rho^A_i)\right]$$
$$\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{aug}}}_{\pi_{ds}}}\left[\mathrm{Inf}(\rho)\cap(F'\times\{0,1\})\neq\emptyset\right]\geq\mathbb{E}_{\rho\sim\mathcal{D}^{\mathcal{M}_{\mathrm{coll}}}_{\pi_{\mathrm{coll}}}}\left[\sum_{i\in\mathbb{N}}R'(\rho^S_i,\rho^A_i)\right]$$

which by Remark 17 can be viewed as a mixture of deterministic policies with 1-bit memory. □

10 Supplementary Materials for Section 4

Let $\mathcal{M}=(S,s_0,P)$ and $\mathcal{M}'=(S,s_0,P')$ be Markov chains with shared finite state space S and $T\subseteq S$. Let $p_{\min}>0$ the smallest positive entry of P.

For a matrix $A \in \mathbb{R}^{n\times m}$ recall $\|A\|_\infty := \max_{1\le i\le n} \sum_{1\le j\le m} |A_{ij}|$.

For $s \in S$ and $t \in \mathbb{N}$ let

$$h(s) := \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}[s]}}\left[\exists i \in \mathbb{N}.\, \rho_i^S \in T\right] \qquad h_t(s) := \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}[s]}}\left[\{\rho_0^S, \dots, \rho_t^S\} \cap T \neq \emptyset\right]$$

be the hitting probability of T and the at most t-step hitting probability of T, respectively, from $s \in S$. Moreover, for $s, s' \in S$ and $t \in \mathbb{N}$ let

$$q_t(s, s') := \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}[s]}}\left[\rho_t^S = s' \text{ and } \{\rho_0^S, \dots, \rho_t^S\} \subseteq S \setminus T\right]$$

be the t-step sub-probability kernel restricted to $S \setminus T$. Analogously, we can define h', h'_t and q'_t for $\mathcal{M}'$.

Lemma 33. *For every $t \in \mathbb{N}$,*

$$\max_{s\in S} |h_t(s) - h'_t(s)| \le t \cdot \|P - P'\|_\infty$$
$$\max_{s\in S} \sum_{s'\in S} |q_t(s, s') - q'_t(s, s')| \le t \cdot \|P - P'\|_\infty$$

Proof. We prove the claim by induction on $t \in \mathbb{N}$. The $t = 0$ case is immediate.

Hence, suppose $t \ge 0$. Using the Markov property,

$$h_{t+1}(s) = \begin{cases} 1 & \text{if } s \in T \\ \sum_{s'\in S\setminus T} P(s' \mid s) \cdot h_t(x) & \text{otherwise} \end{cases}$$

and analogously for h'. Thus, for any $s \in S \setminus T$

$$|h_{t+1}(s) - h'_{t+1}(s)| \le \left|\sum_{s'\in S} P(s, s') \cdot (h_t(s') - h'_t(s'))\right| + \left|\sum_{s'\in S} (P(s' \mid s) - P'(s' \mid s)) \cdot h'_t(s')\right| \tag{3}$$

Since P is stochastic, by the inductive hypothesis,

$$\left|\sum_{s'\in S} P(s' \mid s) \cdot (h_t(s') - h'_t(s'))\right| \le \sum_{s'\in S} \underbrace{P(s' \mid s)}_{\le 1} \cdot \left|h_t(s') - h'_t(s')\right| \le t \cdot \|P - P'\|_\infty \tag{4}$$

and by definition of $\|P - P'\|_\infty$,

$$\left|\sum_{s'\in S} (P(s' \mid s) - P'(s' \mid s)) \cdot h'_t(s')\right| \le \sum_{s'\in S\setminus T} \underbrace{h'_t(s')}_{\le 1} \cdot \underbrace{|P(s' \mid s) - P'(s' \mid s)|}_{\le \|P-P'\|_\infty} \le \|P - P'\|_\infty \tag{5}$$

The first part of the claim follows immediately by Eqs. (3) to (5) since $s \in S \setminus T$ was arbitrary (and for $s \in T$, $h_{t+1}(s) = 1 = h'_{t+1}$).

For the second part, analogously note that

$$q_{t+1}(s, s') = \begin{cases} 0 & \text{if } s \in T \text{ or } s' \in T \\ \sum_{s''\in S\setminus T} P(s'' \mid s) \cdot q_t(s'', s') & \text{otherwise} \end{cases}$$

and hence for any $s \in S \setminus T$,

$$\sum_{s'\in S} |q_{t+1}(s, s') - q'_{t+1}(s, s')| \le \sum_{s'\in S\setminus T} \left|\sum_{s''\in S\setminus T} P(s'' \mid s) \cdot (q_t(s'', s') - q'_t(s'', s'))\right| + \sum_{s'\in S\setminus T} \left|\sum_{s''\in S\setminus T} (P(s'' \mid s) - P'(s'' \mid s)) \cdot q'_t(s'', s')\right| \tag{6}$$

Since P is stochastic, by the inductive hypothesis,

$$\sum_{s' \in S \setminus T} \left| \sum_{s'' \in S \setminus T} P(s'' \mid s) \cdot (q_t(s'', s') - q'_t(s'', s')) \right| \leq \sum_{s'' \in S \setminus T} P(s'' \mid s) \cdot \underbrace{\sum_{s' \in S \setminus T} \left| q_t(s'', s') - q'_t(s'', s') \right|}_{\leq t \cdot \|P - P'\|_\infty}$$

$$\leq t \cdot \|P - P'\|_\infty \tag{7}$$

and by definition of $\|P - P'\|_\infty$,

$$\sum_{s' \in S \setminus T} \left| \sum_{s'' \in S \setminus T} (P(s'' \mid s) - P'(s'' \mid s)) \cdot q'_t(s'', s') \right| \leq \sum_{s'' \in S \setminus T} |P(s'' \mid s) - P'(s'' \mid s)| \cdot \underbrace{\sum_{s' \in S \setminus T} q'_t(s'', s')}_{\leq 1}$$

$$\leq \|P - P'\|_\infty \tag{8}$$

The second part of the claim follows immediately by Eqs. (6) to (8) since $s \in S \setminus T$ was arbitrary (and for $s \in T$ this is obvious). □

Lemma 34. *Let $\mathcal{M} = (S, s_0, P)$ and $\mathcal{M}' = (S, s_0, P')$ be Markov chains satisfying $P(s' \mid s) > 0$ iff $P'(s' \mid s) > 0$ (for all $s, s' \in S$). Furthermore, let $T \subseteq S$ and $p_{\min} > 0$ the smallest positive entry of P. Then for all $s \in S$*

$$\left| \mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}[s]}}[\exists i \in \mathbb{N}.\, \rho_i^S \in T] - \mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}'[s]}}[\exists i \in \mathbb{N}.\, \rho_i^S \in T] \right| \leq \frac{2 \cdot |S|}{p_{\min}^{|S|}} \cdot \|P - P'\|_\infty$$

Proof. Let $S' \subseteq S$ be the set of states which can reach T with positive probability in $\mathcal{M}$ (and hence $\mathcal{M}'$). Then there exists $t \in \mathbb{N}$ such that

$$\alpha \coloneqq \min_{s \in S'} \mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}[s]}}[\{\rho_0^S, \ldots, \rho_t^S\} \cap T \neq \emptyset] > 0$$

Note that we can choose $\alpha \coloneqq p_{\min}^{|S|}$ and $t \coloneqq |S|$. Furthermore, for $s \in S'$,

$$\sum_{s' \in S' \setminus T} q_t(s, s') \leq 1 - \alpha \qquad\qquad h(s) = h_t(s) + \sum_{s' \in S' \setminus T} q_t(s, s') \cdot h(s')$$

and similarly for h'. Therefore, by Lemma 33, for $s \in S'$,

$$\begin{aligned}
|h(s) - h'(s)| &= \left| h_t(s) + \sum_{s' \in S' \setminus T} q_t(s, s') \cdot h(s') - h'_t(s) - \sum_{s' \in S' \setminus T} q'_t(s, s') \cdot h'(s') \right| \\
&\leq \underbrace{\left| h_t(s) - h'_t(s) \right|}_{\leq t \cdot \|P - P'\|_\infty} + \sum_{s' \in S' \setminus T} q_t(s, s') \cdot \underbrace{|h(s') - h'(s')|}_{\leq \max_{s' \in S' \setminus T} |h(s') - h'(s')|} + \sum_{s' \in S' \setminus T} \left| q_t(s, s') - q'_t(s, s') \right| \cdot \underbrace{h'(s')}_{\leq 1} \\
&\leq 2 \cdot t \cdot \|P - P'\|_\infty + (1 - \alpha) \cdot \max_{s' \in S' \setminus T} |h(s') - h'(s')|
\end{aligned}$$

Hence,

$$\max_{s \in S'} |h(s) - h'(s)| \leq \frac{2 \cdot t}{\alpha} \cdot \|P - P'\|_\infty \leq \frac{2 \cdot |S|}{p_{\min}^{|S|}} \cdot \|P - P'\|_\infty$$

and the claim follows because for $s \in S \setminus S'$, $h(s) = 0 = h'(s)$ by assumption. □

Lemma 5 (Simulation). *Let $\mathcal{M} = (S, A, \mathrm{Act}, s_0, P)$ and $\mathcal{M}' = (S, A, \mathrm{Act}, s_0, P')$ be MDPs satisfying $P(s' \mid s, a) > 0$ iff $P'(s' \mid s, a) > 0$ for all $s, s' \in S$ and $a \in A$. Let $p_{\min} > 0$ be the smallest positive transition probability in P. Then for every $F \subseteq S$ and deterministic stationary policy π,*

$$\left|\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] - \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}'}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset]\right| \leq \frac{2|S|}{p_{\min}^{|S|}} \|P - P'\|_\infty.$$

Corollary 35 (Simulation). *Let $\mathcal{M} = (S, A, \mathrm{Act}, s_0, P)$ and $\mathcal{M}' = (S, A, \mathrm{Act}, s_0, P')$ be MDPs satisfying $P(s' \mid s, a) > 0$ iff $P'(s' \mid s, a) > 0$ for all $s, s' \in S$ and $a \in A$. Let $p_{\min} > 0$ be the smallest positive transition probability in P. Then for every $F \subseteq S$ and $\pi \in \Pi_{mds}$,*

$$\left|\mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] - \mathbb{P}_{\rho\sim\mathcal{D}_\pi^{\mathcal{M}'}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset]\right| \leq \frac{2|S|}{p_{\min}^{|S|}} \|P - P'\|_\infty.$$

Prof of Lemma 5. Suppose $\pi \in \Pi_{ds}$ is deterministic. We consider the Markov chains $\mathcal{M}_\pi$ and $\mathcal{M}'_\pi$ induced by the MDPs and the stationary deterministic policy π. Note that

$$\|P_\pi - P'_\pi\|_\infty = \max_{s\in S} \sum_{s'\in S} |P(s' \mid s, \pi(s)) - P'(s' \mid s, \pi(s))| \leq \|P - P'\|_\infty$$

Furthermore, since π is deterministic $p_{\min} > 0$ still bounds the smallest possible probability in $\mathcal{M}_\pi$.

Let T be the set of states in ECs for $\mathcal{M}_\pi$ (and hence $\mathcal{M}'_\pi$) which intersect with F. The claim follows from Lemma 34 since acceptance probabilities correspond to the probabilities of reaching (and hence staying in) ECs intersecting with F. □

Lemma 36 (Concentration Bound, (McDiarmid et al., 1989; Agarwal et al., 2019)). *Let Z be a discrete random variable supported on $\{1, \ldots, d\}$ with distribution $q \in \mathcal{D}([d])$, and let $\widehat{q}$ be the empirical distribution from N i.i.d. samples. Then with probability at least $1 - \delta$,*

$$\left\|\widehat{q} - q\right\|_2 \leq \frac{1}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}$$

Theorem 7 (PAC Guarantee). *Let $\delta, \epsilon > 0$ be arbitrary. Suppose $p_{\min} > 0$ is a lower bound on positive transition probabilities in $\mathcal{M}$. If*

$$N \geq \frac{|S| \log(8|S||A|/\delta)}{(\epsilon')^2} \qquad \text{where } \epsilon' := \min\left\{\frac{p_{\min}}{2}, \frac{\epsilon}{4\eta}, \frac{\epsilon}{2\eta}(\overline{p} - p)\right\}, \quad \eta := \frac{2|S|}{p_{\min}^{|S|}}$$

then with probability at least $1 - \delta$, Problem (C-Recurrence) for the empirical estimate $\widehat{\mathcal{M}}$ is feasible and any optimal solution $\widehat{\pi}$ is ϵ-approximately feasible and optimal.

Proof. By Lemma 36, for all $s \in S$ and $a \in A$ with probability $1 - \delta'$,

$$\left\|\widehat{P}(s, a) - P(s, a)\right\|_2 \leq \frac{1}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta')}{2N}} \leq \sqrt{\frac{\log(8/\delta')}{N}}$$

Therefore, by union bound with probability $1 - \delta$ (using $\delta' := \delta/(|S||A|)$)

$$\|P - P'\|_\infty \leq \max_{s,a} \left\|\widehat{P}(s, a, -) - P(s, a, -)\right\|_1 \leq \sqrt{\frac{|S| \cdot \log(8/\delta')}{N}} \leq \epsilon'$$

using $N \geq |S| \cdot \log(8|S||A|/\delta)/(\epsilon')^2$. In particular,

$$\max_{s,a,s'} \left|P(s, a, s') - \widehat{P}(s, a, s')\right| \leq \|P - \widehat{P}\|_\infty \leq \epsilon'$$

and since $\epsilon' < p_{\min}$, for every $s, s' \in S$ and $a \in A$, $P(s,a,s') > 0$ iff $\widehat{P}(s,a,s') > 0$.

Henceforth, assume that high-probability event.

Let $\pi \in \Pi(\mathcal{M})$ be a policy satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F' \neq \emptyset] \geq p$$

By Remark 17 π can be viewed as $\pi_{\mathrm{aug}} \in \Pi_{mds}(\mathcal{M}_{\mathrm{aug}})$ satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap F \times \{0,1\} \neq \emptyset] = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F' \neq \emptyset]$$
$$\mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap F' \times \{0,1\} \neq \emptyset] = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F' \neq \emptyset] \geq p$$

and similarly $\overline{\pi}_{\mathrm{aug}}$ for $\overline{\pi}$. Define $\pi'_{\mathrm{aug}} := (1-\epsilon/2) \cdot \pi_{\mathrm{aug}} + \epsilon/2 \cdot \overline{\pi}_{\mathrm{aug}}$. By Corollary 35 applied to $\mathcal{M}_{\mathrm{aug}}$, $\widehat{\mathcal{M}}_{\mathrm{aug}}$, π_{aug} and $\overline{\pi}_{\mathrm{aug}}$ (clearly $\|P_{\mathrm{aug}} - \widehat{P}_{\mathrm{aug}}\|_\infty = \|P - \widehat{P}\|_\infty$),

$$\begin{aligned}
&\mathbb{P}_{\rho \sim \mathcal{D}_{\pi'_{\mathrm{aug}}}^{\widehat{\mathcal{M}_{\mathrm{aug}}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] \\
&= (1 - \frac{\epsilon}{2}) \cdot \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\widehat{\mathcal{M}_{\mathrm{aug}}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] + \frac{\epsilon}{2} \cdot \mathbb{P}_{\rho \sim \mathcal{D}_{\overline{\pi}_{\mathrm{aug}}}^{\widehat{\mathcal{M}_{\mathrm{aug}}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] \\
&\geq (1 - \frac{\epsilon}{2}) \cdot \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] + \frac{\epsilon}{2} \cdot \mathbb{P}_{\rho \sim \mathcal{D}_{\overline{\pi}_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] - \eta \cdot \left\|P - \widehat{P}\right\|_\infty \\
&\geq (1 - \frac{\epsilon}{2}) \cdot p + \frac{\epsilon}{2} \cdot \overline{p} - \frac{\epsilon}{2} \cdot (\overline{p} - p) = p
\end{aligned}$$

Hence, π'_{aug} is feasible for the estimated $\widehat{\mathcal{M}}_{\mathrm{aug}}$. Let $\widehat{\pi}$ be an optimal policy for $\widehat{\mathcal{M}}$ and $\widehat{\pi}_{\mathrm{aug}}$ be the corresponding policy for $\widehat{\mathcal{M}}_{\mathrm{aug}}$, for which

$$\begin{aligned}
\mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}_{\mathrm{aug}}}^{\widehat{\mathcal{M}}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] &= \mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}}^{\widehat{\mathcal{M}}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] \\
&\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi'}^{\widehat{\mathcal{M}}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] = \mathbb{P}_{\rho \sim \mathcal{D}_{\pi'_{\mathrm{aug}}}^{\widehat{\mathcal{M}}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset]
\end{aligned}$$

By Corollary 35,

$$\mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] \geq \mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}_{\mathrm{aug}}}^{\widehat{\mathcal{M}}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F' \times \{0,1\}) \neq \emptyset] - \eta \cdot \left\|P - \widehat{P}\right\|_\infty \geq p - \epsilon$$

and

$$\begin{aligned}
\mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}}^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] &= \mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] \\
&\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\widehat{\pi}_{\mathrm{aug}}}^{\widehat{\mathcal{M}}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] - \eta \cdot \left\|P - \widehat{P}\right\|_\infty \\
&\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi'_{\mathrm{aug}}}^{\widehat{\mathcal{M}}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] - \eta \cdot \left\|P - \widehat{P}\right\|_\infty \\
&\geq (1 - \frac{\epsilon}{2}) \cdot \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\widehat{\mathcal{M}}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] - \eta \cdot \left\|P - \widehat{P}\right\|_\infty \\
&\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] - \frac{\epsilon}{2} - 2\eta \cdot \left\|P - \widehat{P}\right\|_\infty \\
&\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}}^{\mathcal{M}_{\mathrm{aug}}}}[\mathrm{Inf}(\rho) \cap (F \times \{0,1\}) \neq \emptyset] - \epsilon \\
&= \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\mathrm{Inf}(\rho) \cap F \neq \emptyset] - \epsilon
\end{aligned}$$

□

11 Supplementary Materials for Section 5

11.1 Discounted MDPs

To leverage the results of Sections 9.2 and 9.3, we identify discounted MDPs with a subclass of absorbing MDPs. See (Altman, 2021, Sec. 10.1) for a complete account of equivalence to standard discounting formulations.

Definition 37. Let $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an MDP and $\Gamma : S \to [0, 1)$ be a state-dependent discount factor. We define the MDP $\mathcal{M}_\Gamma := (S \cup \{\perp\}, A, \mathrm{Act}_\Gamma, \mu, P_\Gamma)$ by

$$\mathrm{Act}_\Gamma(s) := \begin{cases} \mathrm{Act}(s) & s \in S \\ A & s = \perp \end{cases} \qquad s \in S \cup \{\perp\}$$

$$P_\Gamma(s' \mid s, a) := \begin{cases} \Gamma(s) \cdot P(s' \mid s, a) & s, s' \in S \\ 1 - \Gamma(s) & s \in S, s' = \perp \\ 1 & s, s' = \perp \end{cases} \qquad s, s' \in S \cup \{\perp\}, a \in \mathrm{Act}_\Gamma(s)$$

Note in particular that $\mathcal{M}_\Gamma$ is absorbing as $\max_{s \in S} \Gamma(s) < 1$.

11.2 Asymptotic Analysis

We first develop basic abstract theory, which is useful for showing asymptotic feasibility and optimality.

Let X be a finite-dimensional VS with basis B and let $\overline{B}$ be the convex hull of B. Let f_k and g_k be sequences of linear (hence continuous) non-negative functions $X \to \mathbb{R}_{\geq 0}$ converging pointwise to (necessarily linear) f, g respectively on B. Furthermore, let $p \in (0, 1)$ and $\epsilon \geq 0$.

Lemma 38. *$f_k \xrightarrow{unif} f$ on $\overline{B}$ and $g_k \xrightarrow{unif} g$ on $\overline{B}$.*

Proof. Let $\epsilon > 0$. By pointwise convergence on B, which is finite, there exists k_0 such that for all $k \geq k_0$ and $b \in B$, $|f_k(b) - f(b)| \leq \frac{\epsilon}{|B|}$. Now let $x \in \overline{B}$ and $k \geq k_0$ be arbitrary. There exist $\alpha_b \in [0, 1]$ such that $x = \sum_{b \in B} \alpha_b \cdot b$. Hence, by linearity

$$\left|f_k(x) - f(x)\right| \leq \sum_{b \in B} \alpha_b \cdot \left|f_k(b) - f(b)\right| \leq |B| \cdot \frac{\epsilon}{|B|} = \epsilon$$

for g_k. □

Lemma 39. *Let $\epsilon, \delta \geq 0$, $Y := \{x \in \overline{B} : g(x) \geq p\}$ and $Y_k := \{x \in \overline{B} \mid g_k(x) \geq p - \delta\}$. Assume there exists $x' \in B$ such that $g(x') > p$. Then there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$*

1. *$Y_k \neq \emptyset$*
2. *If $x_k \in Y_k$ is such that $f_k(x_k) \geq f_k(x) - \delta$ for all $x \in Y_k$ then for all $x \in Y$,*

$$g(x_k) \geq p - \delta - \epsilon \qquad\qquad f(x_k) \geq f(x) - \delta - \epsilon$$

NB by assumption $Y \neq \emptyset$ and Y_k is a closed non-empty (by the first part) subset of the compact space $\overline{B}$, so Y_k is itself compact. Hence f_k attains a maximum value on Y_k.

Proof. By assumption and convergence (Lemma 38), $\lim_k g_k(x') = g(x') > p$. Hence, there exists $k_1 \in \mathbb{N}$ such that $g_k(x') > p$ (thus $x' \in Y_k$) for all $k \geq k_1$.

Next, for $k \geq k_1$ suppose $x_k \in Y_k$ is such that $f_k(x_k) \geq f_k(x) - \delta$ for all $x \in Y_k$. By definition of Y_k and uniform convergence (Lemma 38),

$$\liminf_k g(x_k) \geq \liminf_k g_k(x_k) - \limsup_k |g_k(x_k) - g(x_k)| \geq p - \delta \tag{9}$$

Besides, let $x \in Y$ be arbitrary. For $k \in \mathbb{N}$ define

$$x'_k := \alpha_k \cdot x + (1 - \alpha_k) \cdot x' \qquad \alpha_k := \begin{cases} 1 & \text{if } g_k(x) \geq p \\ \frac{p - g_k(x')}{g_k(x) - g_k(x')} & \text{otherwise} \end{cases}$$

Note that for all $k \in \mathbb{N}$, $\alpha_k \in [0,1]$ and $g_k(x'_k) \geq p$. Hence, $f_k(x_k) \geq f_k(x'_k) - \delta$. Moreover, by uniform convergence (Lemma 38), $\lim_k \alpha_k = 1$. Therefore,

$$\begin{aligned} \liminf_k f(x_k) &= \liminf_k f_k(x_k) - \underbrace{\limsup_k \left| f_k(x_k) - f(x_k) \right|}_{0} \\ &\geq \liminf_k f_k(x'_k) - \delta \\ &\geq \liminf_k \alpha_k \cdot f_k(x) - \delta = f(x) - \delta \end{aligned}$$

The existence of a suitable $k_0 \geq k_1$ follows immediately from this and Eq. (9). □

Theorem 8. *Let $\epsilon, \delta > 0$. There exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, Problem (C-Discounted-k) is feasible, and for any mixture π of deterministic stationary policies over $\mathcal{M}_{\text{aug}}$, if π is δ-approximately feasible and optimal for Problem (C-Discounted-k), then the corresponding 1-bit memory policy for $\mathcal{M}$ (cf. Remark 17) is $\delta + \epsilon$-approximately feasible and optimal for Problem (C-Recurrence).*

Proof. We wish to apply Lemma 39 with $B := \Pi_{ds}(\mathcal{M}_{\text{aug}})$, for which $\overline{B} = \Pi_{mds}(\mathcal{M}_{\text{aug}})$. We also instantiate $f(\pi) := \mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}}_{\pi}}[\text{Inf}(\rho) \cap F \neq \emptyset]$, $f'(\pi) := \mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}}_{\pi}}[\text{Inf}(\rho) \cap F' \neq \emptyset]$ and use the discounted approximations $f_k(\pi) := J_{F,\gamma_k,\bar{\gamma}_k}(\pi)$ and $g_k(\pi) := J_{F',\gamma_k,\bar{\gamma}_k}(\pi)$, which are linear and satisfy $f_k(\pi) \to f(\pi)$ and $g_k(\pi) \to g(\pi)$ for $\pi \in \Pi_{ds}(\mathcal{M}_{\text{aug}})$ as $k \to \infty$ by (Bozkurt et al., 2019). The theorem immediately follows from Assumption 6 and Lemma 39. □

12 Supplementary Materials for Section 6

Definition 40. An NBA $(Q, \Sigma, q_0, \Delta, F)$ is *limit-deterministic* if Q can be partitioned into $Q_i \cup Q_d$ such that (i) $F \subseteq Q_d$, (ii) $|\Delta(q, \sigma)| = 1$ for all $q \in Q$ and $\sigma \in \Sigma$, (iii) $\Delta(q, \epsilon) = \emptyset$ for $q \in Q_d$, (iv) $\Delta(q, \epsilon) \subseteq Q_d$ for $q \in Q_i$, and (v) $\Delta(q, \sigma) \subseteq Q_d$ for $q \in Q_d$ and $\sigma \in \Sigma$.

Intuitively, accepting states are in the purely deterministic part of the automaton, which is closed, and ϵ-transitions move between the initial and the deterministic part.

Definition 41. A *non-deterministic finite semi-automaton* is a quadruple $\mathcal{A} = (Q, \Sigma, q_0, \Delta)$, where Q is a finite set of states, Σ is a finite alphabet, $q_0 \in Q$ is the initial state and $\Delta : S \times (\Sigma \cup \{\epsilon\}) \to 2^Q$ is the transition function satisfying $|\Delta(q, \sigma)| = 1$ for all $q \in Q$ and $\sigma \in \Sigma$.

That is, the only non-determinism is caused by ϵ-transitions. Clearly, limit-deterministic Büchi automata according to Definition 40 and DRA can be viewed as non-deterministic finite semi-automaton by dropping their acceptance criterion.

Henceforth, we fix an MDP

Definition 42 (Product MDP)**.** Let $\mathcal{M} = (S, A, \text{Act}, \mu, P)$ be an MDP, $\ell : S \to 2^{\mathcal{AP}}$ be a labelling function and $\mathcal{A} = (Q, 2^{\mathcal{AP}}, q_0, \Delta)$ and $\mathcal{A}' = (Q', 2^{\mathcal{AP}}, q'_0, \Delta')$ be non-deterministic finite semi-automaton . The *product MDP* $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$ is an MDP $(S^{\otimes}, A^{\otimes}, \text{Act}^{\otimes}, \mu^{\otimes}, P^{\otimes})$,

where

$$
\begin{aligned}
S^{\otimes} &:= S \times Q \times Q' \\
A^{\otimes} &:= A \cup (Q \times Q') \\
\mathrm{Act}^{\otimes}(s, q, q') &:= \mathrm{Act}(s) \cup \{(\bar{q}, \bar{q}') \mid \bar{q} \in \Delta(q, \epsilon), \bar{q}' \in \Delta'(q', \epsilon)\} \\
\mu^{\otimes}(s, q, q') &:= \mu(s) \cdot \mathbf{1}_{\{q = q_0, q' = q'_0\}} \\
P^{\otimes}((\bar{s}, \bar{q}, \bar{q}') \mid (s, q, q'), a) &:= \begin{cases} P(\bar{s} \mid s, a) & \text{if } a \in A, \bar{q} \in \Delta(q, \ell(s)), \bar{q}' \in \Delta'(q', \ell(s)) \\ 0 & \text{otherwise} \end{cases} \\
P^{\otimes}((s, \bar{q}, \bar{q}') \mid (s, q, q'), (\bar{q}, \bar{q}')) &:= \begin{cases} 1 & \text{if } \bar{q} \in \Delta(q, \epsilon), \bar{q}' \in \Delta'(q', \epsilon) \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Define:

$$
F^{\otimes} := S \times F \times Q' \qquad\qquad (F')^{\otimes} := S \times Q \times F'
$$

Lemma 43. *Let MDP $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an MDP with labelling function $\ell : S \to 2^{\mathcal{AP}}$, let $\mathcal{A} = (Q, 2^{\mathcal{AP}}, q_0, \Delta, F)$ and $\mathcal{A}' = (Q', 2^{\mathcal{AP}}, q'_0, \Delta', F')$ be limit-deterministic NBAs and let $\pi^{\otimes}$ be a mixture of stationary policies for $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$. Let π be the finite-memory policy which employs memory to track the automata states and acts according to $\pi^{\otimes}$. Then*

$$
\begin{aligned}
\mathbb{P}_{\rho \sim \mathcal{D}_{\pi}^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{A})\right] &\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi^{\otimes}}^{\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'}}\left[\inf(\rho) \cap F^{\otimes} \neq \emptyset\right] \\
\mathbb{P}_{\rho \sim \mathcal{D}_{\pi}^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{A}')\right] &\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi^{\otimes}}^{\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'}}\left[\inf(\rho) \cap (F')^{\otimes} \neq \emptyset\right]
\end{aligned}
$$

Proof. Suppose $\pi^{\otimes} = \lambda \cdot \pi_1^{\otimes} + (1 - \lambda) \cdot \pi_1^{\otimes}$. For $i \in \{1, 2\}$, define

$$
\begin{aligned}
\pi_i^{\mathrm{mem}} &:= Q \times Q' \\
\pi_i^{\mathrm{init}}(q, q') &:= \mathbf{1}_{\{q = q_0, q' = q'_0\}} \\
\pi_i^{\mathrm{act}}(a, (\bar{q}, \bar{q}') \mid s, (q, q')) &:= \pi_i^{\otimes}(a \mid (s, q, q')) \cdot \mathbf{1}_{\{\bar{q} \in \Delta(q, \ell(s)), \bar{q}' \in \Delta(q', \ell(s))\}} \\
&\quad + \sum_{\tilde{q} \in Q, \tilde{q}' \in Q'} \pi_i^{\otimes}((\tilde{q}, \tilde{q}') \mid (s, q, q')) \cdot \pi_i^{\otimes}(a \mid (s, \tilde{q}, \tilde{q}')) \cdot \mathbf{1}_{\{\bar{q} \in \Delta(\tilde{q}, \ell(s)), \bar{q}' \in \Delta(\tilde{q}', \ell(s))\}}
\end{aligned}
$$

and $\pi := \lambda \cdot (\pi_1^{\mathrm{mem}}, \pi_1^{\mathrm{init}}, \pi_1^{\mathrm{act}}) + (1 - \lambda) \cdot (\pi_1^{\mathrm{mem}}, \pi_1^{\mathrm{init}}, \pi_1^{\mathrm{act}})$.

Since ϵ-actions can be performed at most once in a limit-deterministic NBA (in the sense of Definition 40), actions $(\bar{q}, \bar{q}') \in Q \times Q'$ can be performed at most once in any run of the product MDP. Hence, there is a 1-to-1 correspondence between runs of π in $\mathcal{M}$ and those under $\pi^{\otimes}$ in $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$.

□

Lemma 44. *Let MDP $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ be an MDP with labelling function $\ell : S \to 2^{\mathcal{AP}}$, let $\mathcal{A} = (Q, 2^{\mathcal{AP}}, q_0, \Delta, F)$ and $\mathcal{A}' = (Q', 2^{\mathcal{AP}}, q'_0, \Delta', F')$ be limit-deterministic NBAs and let $\pi_{\mathrm{aug}}^{\otimes}$ be a mixture of deterministic stationary policies for $(\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}')_{\mathrm{aug}}$. Let π be the mixture of deterministic finite-memory policy which employs memory to track the automata states and the augmentation bit and acts according to $\pi^{\otimes}$. Then*

$$
\begin{aligned}
\mathbb{P}_{\rho \sim \mathcal{D}_{\pi}^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{A})\right] &\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}^{\otimes}}^{(\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}')_{\mathrm{aug}}}}\left[\inf(\rho) \cap (F^{\otimes} \times \{0, 1\}) \neq \emptyset\right] \\
\mathbb{P}_{\rho \sim \mathcal{D}_{\pi}^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{A}')\right] &\geq \mathbb{P}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{aug}}^{\otimes}}^{(\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}')_{\mathrm{aug}}}}\left[\inf(\rho) \cap ((F')^{\otimes} \times \{0, 1\}) \neq \emptyset\right]
\end{aligned}
$$

Proof. Let $\pi^{\otimes}_{\text{aug}}$ be a mixture of deterministic stationary policies for $(\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}')_{\text{aug}}$. By definition and Remark 17, there exist $\lambda \in [0,1]$ and deterministic policies π_1 and π_2 for $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$ with $\pi_1^{\text{mem}} = \{0,1\} = \pi_2^{\text{mem}}$ satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\inf(\rho) \cap F^{\otimes} \neq \emptyset\right] = \mathbb{P}_{\rho \sim \mathcal{D}^{(\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}')_{\text{aug}}}_{\pi^{\otimes}_{\text{aug}}}}\left[\inf(\rho) \cap (F^{\otimes} \times \{0,1\}) \neq \emptyset\right]$$

$$\mathbb{P}_{\rho \sim \mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\inf(\rho) \cap (F')^{\otimes} \neq \emptyset\right] = \mathbb{P}_{\rho \sim \mathcal{D}^{(\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}')_{\text{aug}}}_{\pi^{\otimes}_{\text{aug}}}}\left[\inf(\rho) \cap ((F')^{\otimes} \times \{0,1\}) \neq \emptyset\right]$$

where $\pi^{\otimes} := \lambda \cdot \pi_1^{\otimes} + (1-\lambda) \cdot \pi_2^{\otimes}$.

Now, for $i \in \{1,2\}$, define

$$\pi_i^{\text{mem}} := Q \times Q' \times \{0,1\}$$
$$\pi_i^{\text{init}}(q, q', j) := \mathbf{1}_{\{q=q_0, q'=q'_0, j=0\}}$$

If $(a, j') := \pi_i^{\otimes}((s, q, q'), j) \in A \times \{0,1\}$ we define

$$\pi_i^{\text{act}}(s, (q, q', j)) := (a, (\delta(q, \ell(s)), \delta'(q', \ell(s)), j'))$$

where $\bar{q} := \delta(q, \ell(s))$ and $\bar{q}' := \delta(q', \ell(s))$. Otherwise, $((\tilde{q}, \tilde{q}'), j') := \pi^{\otimes}((s, q, q'), j) \in (Q \times Q') \times \{0,1\}$ and we define

$$\pi_i^{\text{act}}(s, (q, q', j)) := (a, (\delta(\tilde{q}, \ell(s)), \delta'(\tilde{q}', \ell(s)), j''))$$

where $(a, j'') := \pi^{\otimes}((s, \tilde{q}, \tilde{q}'), j') \in A \times \{0,1\}$. Finally, $\pi := \lambda \cdot (\pi_1^{\text{mem}}, \pi_1^{\text{init}}, \pi_1^{\text{act}}) + (1-\lambda) \cdot (\pi_1^{\text{mem}}, \pi_1^{\text{init}}, \pi_1^{\text{act}})$. □

12.1 Proof of Proposition 9

In addition to NBAs, we consider a *deterministic* class of automaton, which accepts all ω-regular languages:

Definition 45 (DRA)**.** A *deterministic Rabin automaton (DRA)* is a tuple $\mathcal{A} = (Q, \Sigma, q_0, \delta, \mathcal{F})$ with finite state set Q, finite alphabet Σ, transition function $\delta : Q \times \Sigma \to Q$, and acceptance criterion $\mathcal{F} = \{(F_1, U_1), \ldots, (F_n, U_n)\}$, where $F_1, U_1, \ldots, F_n, U_n \subseteq Q$. A word $w = w_0 w_1 w_2 \ldots \in \Sigma^{\omega}$ is accepted by $\mathcal{A}$ (notation: $w \in L(\mathcal{A})$) if there exists $1 \leq i \leq n$ such that for the unique $q_0 q_1 q_2 \ldots$ of $\mathcal{A}$, i.e. $q_{i+1} = \delta(q_i, w_i)$, a state in F_i occurs infinitely often and all states in U_i only occur finitely often.

For a DRA $\mathcal{A} = (Q, \Sigma, q_0, \delta, \{(F_1, U_1), \ldots, (F_n, U_n)\})$ we define $\text{LDBA}(\mathcal{A}) := (Q', \Sigma, q'_0, \Delta', F')$, where

$$\begin{aligned}
Q' &:= Q \times \{\top, \bot, 1, \ldots, n\} \\
q'_0 &:= (q_0, \top) \\
\Delta'((q, f), \sigma) &:= \{(\delta(q, \sigma), f)\} && q \in Q, f \in \{\top, \bot\}, \sigma \in \Sigma \\
\Delta'((q, i), \sigma) &:= \begin{cases} \{(\delta(q, \sigma), \bot)\} & \text{if } q \in U_i \\ \{(\delta(q, \sigma), i)\} & \text{otherwise} \end{cases} && q \in Q, i \in \{1, \ldots, n\}, \sigma \in \Sigma \\
\Delta'((q, \top), \epsilon) &:= \{(q, f) \mid f \in \{\bot, 1, \ldots, n\}\} && q \in Q \\
F' &:= \{(q, i) \in Q \times \{1, \ldots, n\} \mid q \in F_i\}
\end{aligned}$$

This automaton spontaneously commits to a Rabin pair (via ϵ-transitions) and then checks adherence to it (violations result in a non-recoverable $\bot$ flag).

Lemma 46. *Let $\mathcal{A}$ be a DRA.* $\text{LDBA}(\mathcal{A})$ *is a limit-deterministic NBA and it satisfies* $L(\mathcal{A}) = L(\text{LDBA}(\mathcal{A}))$.

Proof. □

Henceforth, we fix an MDP $\mathcal{M} = (S, A, \mathrm{Act}, \mu, P)$ and a labelling function $\ell : S \to 2^{\mathcal{AP}}$ and DRAs $\mathcal{B} = (Q, 2^{\mathcal{AP}}, q_0, \delta, \{(F_1, U_1), \ldots, (F_m, U_m)\})$ and $\mathcal{B}' = (Q', 2^{\mathcal{AP}}, q'_0, \delta, \{(F'_1, U'_1), \ldots, (F'_n, U'_n)\})$.

Definition 47 (Accepting EC)**.** An EC $C = (S_C, \mathrm{Act}_C)$ of $\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'$ is *accepting for* $\mathcal{B}$ if $S_C \cap (S \times F_i \times Q') \neq \emptyset$ and $S_C \cap (S \times U_i \times Q') = \emptyset$ for some $1 \leq i \leq m$. Likewise, it is *accepting for* $\mathcal{B}'$ if $S_C \cap (S \times Q \times F'_i) \neq \emptyset$ and $S_C \cap (S \times Q \times U'_i) = \emptyset$ for some $1 \leq i \leq n$.

Note that a policy π for $\mathcal{M}$ can be viewed as a policy for $\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'$ (by ignoring the additional DRA components of the state). As a consequence of Lemma 20:

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L(\mathcal{B})] = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'}}[C(\rho) \text{ accepting for } \mathcal{B}] \tag{10}$$

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L(\mathcal{B}')] = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'}}[C(\rho) \text{ accepting for } \mathcal{B}'] \tag{11}$$

We consider a modified collapsed MDP construction $(\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}')_{\mathrm{coll}}$ for $\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'$, where for each proper EC C (not just maximal ones) we use stay actions $\textsc{stay}_C$, which are enabled in MECs containing C and transition deterministically to $\perp$. We also define functions $R_{\mathrm{coll}}, R'_{\mathrm{coll}} : S_{\mathrm{coll}} \times A_{\mathrm{coll}} \to \mathbb{R}$ by

$$R_{\mathrm{coll}}(C, \textsc{stay}_{C'}) = \mathbf{1}_{\{C' \text{ accepting for } \mathcal{B}\}} \qquad R'_{\mathrm{coll}}(C, \textsc{stay}_{C'}) = \mathbf{1}_{\{C' \text{ accepting for } \mathcal{B}'\}}$$

for $C \in \mathfrak{M}$ and proper C' contained in C, and defined to be 0 elsewhere.

Lemma 48. *Let π be a policy for $\mathcal{M}$ (and hence for $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$). Define*

$$x(C, (s, a)) := \mathbb{E}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'}}\left[\sum_{i \in \mathbb{N}} \mathbf{1}_{\{\rho_i^S = s, \rho_i^A = a\}}\right] \qquad (C \in \mathfrak{M}, (s, a) \in \mathrm{out}(C))$$

$$x(C, \textsc{stay}_{C'}) := \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'}}[C(\rho) = C'] \qquad (C \in \mathfrak{M}_p, C' \textit{ included in } C)$$

Then x satisfies (LP-Flow) *and*

$$\sum_{s_{\mathrm{coll}} \in S_{\mathrm{coll}}} \sum_{a_{\mathrm{coll}} \in \mathrm{Act}_{\mathrm{coll}}(s_{\mathrm{coll}})} x(s_{\mathrm{coll}}, a_{\mathrm{coll}}) \cdot R_{\mathrm{coll}}(s_{\mathrm{coll}}, a_{\mathrm{coll}}) = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L(\mathcal{A})]$$

$$\sum_{s_{\mathrm{coll}} \in S_{\mathrm{coll}}} \sum_{a_{\mathrm{coll}} \in \mathrm{Act}_{\mathrm{coll}}(s_{\mathrm{coll}})} x(s_{\mathrm{coll}}, a_{\mathrm{coll}}) \cdot R'_{\mathrm{coll}}(s_{\mathrm{coll}}, a_{\mathrm{coll}}) = \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}[\ell(\rho) \in L(\mathcal{A}')]$$

Proof. (LP-Flow) is verified as in Lemma 32 and the additional equalities follow immediately from Eqs. (10) and (11) and the definition of $R_{\mathrm{coll}}, R'_{\mathrm{coll}} : S_{\mathrm{coll}} \times A_{\mathrm{coll}} \to \mathbb{R}$. □

Analogous to Lemma 31:

Lemma 49. *Let π_{coll} be a deterministic stationary policy for the modified collapsed product $(\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}')_{\mathrm{coll}}$. There exists a stationary policy π over $\mathcal{M}^\otimes := \mathcal{M} \otimes \mathrm{LDBA}(\mathcal{B}) \otimes \mathrm{LDBA}(\mathcal{B}')$ satisfying*

$$\mathbb{P}_{\rho \sim \mathcal{D}_{\pi_s}^{\mathcal{M}^\otimes}}[\mathrm{Inf}(\rho) \cap F^\otimes \neq \emptyset] = \mathbb{E}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{coll}}}^{(\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}')_{\mathrm{coll}}}}\left[\sum_{i \in \mathbb{N}} R_{\mathrm{coll}}(\rho_i^S, \rho_i^A)\right]$$

$$\mathbb{P}_{\rho \sim \mathcal{D}_{\pi_s}^{\mathcal{M}^\otimes}}[\mathrm{Inf}(\rho) \cap (F')^\otimes \neq \emptyset] = \mathbb{E}_{\rho \sim \mathcal{D}_{\pi_{\mathrm{coll}}}^{(\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}')_{\mathrm{coll}}}}\left[\sum_{i \in \mathbb{N}} R'_{\mathrm{coll}}(\rho_i^S, \rho_i^A)\right]$$

Proof. For $(s,q,q') \in S \times Q \times Q'$ let $C_{(s,q,q')}$ be the unique MEC of $\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'$ containing (s,q,q'). We define

$$\pi\left(a \mid (s,(q,\top),(q',\top))\right) := \begin{cases} \mathbf{1}_{\{a=((q,f_C),(q',f'_C))\}} & \text{if } \pi_{\text{coll}}(C_{(s,q,q')}) = \textsc{stay}_C, (s,q,q') \in S_C \\ \mathbf{1}_{\{a=a'\}} & \text{if } \pi_{\text{coll}}(C_{(s,q,q')}) = (s,a) \\ \frac{\mathbf{1}_{\{a \in \text{Act}_{C_{(s,q,q')}}(s,q,q')\}}}{|\text{Act}_{C_{(s,q,q')}}(s,q,q')|} & \text{otherwise} \end{cases}$$

where $f_C \in \{1,\dots,m\}$ witnesses acceptance of C for $\mathcal{B}$[1] or $f_C := \bot$ if C is not accpeting for $\mathcal{B}$. (f'_C is defined analogously). Intuitively, prior to commiting to two Rabin conditions, we select actions uniformly at random within the current MEC unless π_{coll} selects a state-action pair which may leave the current MEC or it commits to a particular EC C' within the current MEC and we have reached a state in it (via the aforementioned randomisation). In the latter case, we simultaneously commit to suitable Rabin pairs (or $\bot$) for this EC C' for both DRAs.

After having committed to C' we follow its enabled actions uniformly at random, which guarantees almost surely visiting all its states infinitely often:

$$\pi\left(a \mid (s,(q,f),(q',f'))\right) := \frac{\mathbf{1}_{\{a \in \text{Act}_{C'}(s,q,q')\}}}{|\text{Act}_{C'}(s,q,q')|}$$

for $s \in S$, $q \in Q$, $q' \in Q'$ and $f, f' \in \{\bot, 1, \dots, n\}$ such that $\pi_{\text{coll}}(C_{(s,q,q')}) = \textsc{stay}_C$ and $(s,q,q') \in S_C$. Elsewhere, π can be defined arbitrarily.

□

Proposition 9. *Let MDP $\mathcal{M} = (S, A, \text{Act}, \mu, P)$ be an MDP with labelling function $\ell : S \to 2^{\mathcal{AP}}$, let $L, L' \subseteq (2^{\mathcal{AP}})^\omega$ be ω-regular languages and let $p \in [0,1]$.*

There exist LDBAs $\mathcal{A} = (Q, 2^{\mathcal{AP}}, q_0, \Delta, F)$ and $\mathcal{A}' = (Q', 2^{\mathcal{AP}}, q'_0, \Delta', F')$ such that for any policy π which is feasible for Problem (C-ω-Regular) there exists a policy $\pi^\otimes$ for $\mathcal{M} \otimes \mathcal{A} \otimes \mathcal{A}'$ satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}_{\pi^\otimes}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}}\left[\text{Inf}(\rho) \cap F^\otimes \neq \emptyset\right] \geq \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{A})\right] \qquad \mathbb{P}_{\rho \sim \mathcal{D}_{\pi^\otimes}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}}\left[\text{Inf}(\rho) \cap (F')^\otimes \neq \emptyset\right] \geq p$$

Proof. Let $\mathcal{B}$ and $\mathcal{B}'$ be DRAs satisfying $L = L(\mathcal{B})$ and $L' = L(\mathcal{B}')$. Define $\mathcal{A} := \text{LDBA}(\mathcal{B})$ and $\mathcal{A}' := \text{LDBA}(\mathcal{B}')$.

Let π be a policy for $\mathcal{M}$ satisfying

$$\mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\ell(\rho) \in L'\right] \geq p$$

Since DRAs are deterministic, π can be viewed as a policy for $\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}'$ (by ignoring the additional components of the state corresponding to $\mathcal{B}$ and $\mathcal{B}'$). By Lemma 48 there exists a solution x to the LP for the modified $(\mathcal{M} \otimes \mathcal{B} \otimes \mathcal{B}')_{\text{coll}}$ satisfying

$$\sum_{s_{\text{coll}} \in S_{\text{coll}}} \sum_{a_{\text{coll}} \in \text{Act}_{\text{coll}}(s_{\text{coll}})} x(s_{\text{coll}}, a_{\text{coll}}) \cdot R_{\text{coll}}(s_{\text{coll}}, a_{\text{coll}}) \geq \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{B})\right]$$
$$\sum_{s_{\text{coll}} \in S_{\text{coll}}} \sum_{a_{\text{coll}} \in \text{Act}_{\text{coll}}(s_{\text{coll}})} x(s_{\text{coll}}, a_{\text{coll}}) \cdot R'_{\text{coll}}(s_{\text{coll}}, a_{\text{coll}}) \geq \mathbb{P}_{\rho \sim \mathcal{D}_\pi^{\mathcal{M}}}\left[\ell(\rho) \in L(\mathcal{B}')\right] \geq p$$

Hence, by Lemma 23, Problem (C-Absorbing) is feasible and by (Feinberg & Rothblum, 2012, Alg. 2,Thm. 9.2) there exists a mixture π_{coll} of deterministic stationary policies for

[1] i.e. $f_C \in \{1,\dots,m\}$ is such that $S_C \cap F_{f_C} \neq \emptyset$ and $S_C \cap U_{f_C} = \emptyset$ and $f_C = \bot$

Problem (C-Absorbing) satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{(\mathcal{M}\otimes\mathcal{B}\otimes\mathcal{B}')_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i\in\mathbb{N}} R_{\text{coll}}(\rho_i^S,\rho_i^A)\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{B})\right]$$
$$\mathbb{P}_{\rho\sim\mathcal{D}^{(\mathcal{M}\otimes\mathcal{B}\otimes\mathcal{B}')_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i\in\mathbb{N}} R'_{\text{coll}}(\rho_i^S,\rho_i^A)\right] \geq p$$

By Lemma 49 there exists a mixture $\pi^{\otimes}$ of stationary policies over $\mathcal{M}^{\otimes} := \mathcal{M}\otimes \text{LDBA}(\mathcal{A})\otimes \text{LDBA}(\mathcal{A}')$ satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}^{\otimes}}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap F^{\otimes}\neq\emptyset\right] = \mathbb{E}_{\rho\sim\mathcal{D}^{(\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}')_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i\in\mathbb{N}} R_{\text{coll}}(\rho_i^S,\rho_i^A)\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{B})\right]$$
$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}^{\otimes}}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap (F')^{\otimes}\neq\emptyset\right] = \mathbb{E}_{\rho\sim\mathcal{D}^{(\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}')_{\text{coll}}}_{\pi_{\text{coll}}}}\left[\sum_{i\in\mathbb{N}} R'_{\text{coll}}(\rho_i^S,\rho_i^A)\right] \geq p$$

The claim follows with Lemma 46. □

12.2 Proof of Theorem 10

Theorem 10. *Feasible instances of Problem (C-ω-Regular) have solutions which are a mixture of two policies with memory $Q\times Q'$, likewise solutions which are a mixture of two deterministic policies with memory $Q\times Q'\times\{0,1\}$.*

Proof. Let π be a policy for $\mathcal{M}$ satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}[\ell(\rho)\in L'] \geq p$$

By Proposition 9 there exists limit-deterministic Büchi automata $\mathcal{A}$ and $\mathcal{A}'$ satisfying $L(\mathcal{A}) = L$, $L(\mathcal{A}') = L'$ and a policy $\pi^{\otimes}$ satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap F^{\otimes}\neq\emptyset\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{A})\right]$$
$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap (F')^{\otimes}\neq\emptyset\right] \geq p$$

By Theorem 3 and Lemma 43 there exists a mixture π of two finite-memory policies for $\mathcal{M}$ satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{A})\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap F^{\otimes}\neq\emptyset\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{A})\right]$$
$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{A}')\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap (F')^{\otimes}\neq\emptyset\right] \geq p$$

Likewise, by By Theorem 3 and Lemma 44 there exists a mixture π_d of two deterministic finite-memory policies for $\mathcal{M}$ satisfying

$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi_d}}\left[\ell(\rho)\in L(\mathcal{A})\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap F^{\otimes}\neq\emptyset\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi}}\left[\ell(\rho)\in L(\mathcal{A})\right]$$
$$\mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}}_{\pi_d}}\left[\ell(\rho)\in L(\mathcal{A}')\right] \geq \mathbb{P}_{\rho\sim\mathcal{D}^{\mathcal{M}\otimes\mathcal{A}\otimes\mathcal{A}'}_{\pi^{\otimes}}}\left[\text{Inf}(\rho)\cap (F')^{\otimes}\neq\emptyset\right] \geq p$$

□