

# Approximate Solutions To Constrained Risk-Sensitive Markov Decision Processes

<sup>1</sup>Uday Kumar M, <sup>1</sup>Sanjay P Bhat,  
<sup>2</sup>Veeraruna Kavitha, <sup>2</sup>Nandyala Hemachandra  
<sup>1</sup>TCS Research, Hyderabad, India and  
<sup>2</sup>IEOR, IIT Bombay

September 30, 2022

## Abstract

This paper considers the problem of finding near-optimal Markovian randomized (MR) policies for finite-state-action, infinite-horizon, constrained risk-sensitive Markov decision processes (CRSMDPs). Constraints are in the form of standard expected discounted cost functions as well as expected risk-sensitive discounted cost functions over finite and infinite horizons. The main contribution is to show that the problem possesses a solution if it is feasible, and to provide two methods for finding an approximate solution in the form of an ultimately stationary (US) MR policy. The latter is achieved through two approximating finite-horizon CRSMDPs which are constructed from the original CRSMDP by time-truncating the original objective and constraint cost functions, and suitably perturbing the constraint upper bounds. The first approximation gives a US policy which is  $\epsilon$ -optimal and feasible for the original problem, while the second approximation gives a near-optimal US policy whose violation of the original constraints is bounded above by a specified  $\epsilon$ . A key step in the proofs is an appropriate choice of a metric that makes the set of infinite-horizon MR policies and the feasible regions of the three CRSMDPs compact, and the objective and constraint functions continuous. A linear-programming-based formulation for solving the approximating finite-horizon CRSMDPs is also given.

**Keywords:** Discrete Optimization, Markov decision processes (MDP), risk-sensitive, constrained MDP,  $\epsilon$ -feasible policy, linear programming, policy space exponential utility.

# 1 Introduction

A Markov decision process (MDP) evolving over a finite set of states at discrete decision epochs under the influence of a finite number of actions is specified in terms of immediate rewards (or costs) and controlled state transition probabilities satisfying the Markov property.

A policy is a sequence of decisions rules, one for each decision epoch. A policy is **stationary** if it applies the same decisions rule at all the epochs, and **ultimately stationary** (US) if it applies the same decision rule beyond a certain epoch. The decision problem involving an MDP is to choose the policy that optimizes the cost obtained by appropriately aggregating immediate costs incurred over several decision epochs. Examples of aggregate costs include total, average or discounted costs over a horizon that may be finite or infinite. In standard MDPs, one seeks to optimize the expected aggregate cost. However, this results in a risk-neutral approach which does not take into account the risk preferences of possibly risk-sensitive decision makers.

The literature on MDPs considers various means for incorporating risk sensitivity into the decision problem. These include penalizing the long run variance of the costs discussed in [20], [13] and [29] attempting mean-variance tradeoff by simultaneously imposing thresholds on the expectation and variance of the aggregate cost discussed in [24], maximizing the probability of maintaining the aggregate cost within a budget is considered by authors in [6], using the conditional value-at-risk of the aggregate cost either as a constraint is discussed in [2] or as the objective is discussed in [30], and enforcing a threshold on the probability that the system enters a high-risk state is discussed in [11]. The most widely followed approach for taking risk sensitivity into account is to consider optimization of the expected exponential utility of the aggregate cost or, equivalently, its certainty equivalent dealt in [3, 16]. We follow the same approach in this paper, and use the term risk-sensitive MDP (RSMDP) to refer to an MDP with expected exponential utility of the total discounted cost as the risk-sensitive (RS) cost to be optimized. For the sake of clarity, we will use the terminology “standard discounted cost” to refer to the expectation of the discounted sum of immediate costs over the horizon of interest.

Work in [18, 19] contain two of the earliest treatments of infinite-horizon RSMDPs. Authors in [18] examined the relationship between optimality in terms of expected exponential utility and optimality in terms of higher moments of the discounted cost for infinite-horizon RSMDPs. The reference [19] showed that while optimal policies in infinite-horizon RSMDPS are not always stationary, US optimal policies do exist under certain assumptions. Work in [22] gave a solution to finite-horizon unconstrained RSMDPs with

finite state and action spaces using dynamic and linear programming (LP).

Emphasizing ease of computation and practical implementation over exact optimality in [23] focused on  $\epsilon$ -optimal US policies for infinite horizon RSMDPS. The results of [23] exploited two key ideas for obtaining  $\epsilon$ -optimal policies. The first idea, which was also considered earlier in [4] and [5], that involves truncating the tail cost beyond a finite horizon and using an arbitrarily chosen stationary policy thereafter. Discounting ensures that the resulting truncation error is small. Consequently, optimizing the truncated RS cost results in an  $\epsilon$ -optimal policy referred in [23] as the ultimately stationary tail off (USTO) policy. The second idea is based on the fact that the certainty equivalent of the RS cost approaches the standard discounted cost as the risk factor  $\gamma$  approaches 0 as discussed by authors in [18, 19] and [23, Thm. 3]. Authors in [23] makes subtle use of this fact by observing that the effective risk factor for the immediate costs sufficiently far out into the tail can be taken to be arbitrarily small due to the effect of discounting. The RS tail cost may therefore be approximated by the exponential of the standard discounted cost of the tail of the policy. These ideas lead to an  $\epsilon$ -optimal US policy called ultimately stationary linear discounted (USLD), which is obtained by appending the stationary optimal policy for the infinite-horizon discounted cost to the optimal policy for a finite-horizon terminal-cost RSMDP whose terminal cost is determined by the optimal infinite-horizon standard discounted cost.

Real-life optimization problems invariably involve constraints, and MDPs are no exceptions. Previous applications of constrained MDPs in risk-neutral as well as risk-sensitive settings include pavement management systems in [12], multi-arm bandits in [7] and delay torrent networks in [22]. The objective of this paper, therefore, is to extend the ideas of [23] to constrained RSMDPS (CRSMDPs) that involve constraints based on standard discounted and RS costs over finite and infinite horizons. In particular, this paper seeks to extend the USTO idea in [23] to infinite-horizon CRSMDPs.

While the literature on constrained MDPs (CMDPs) is not as extensive as that on unconstrained MDPs, constrained MDPs have been fairly well studied in the risk-neutral context. In one of the earliest developments, [8] gave a LP-based method to compute the optimal policy for a finite-horizon, total cost CMDP. [21] gave a more detailed exposition of the same. The influential paper [10] showed that a finite-state infinite-horizon CMDP having both objective and constraints of the standard discounted type possesses stationary randomized optimal policies as well as ultimately deterministic US optimal policies, and gave a LP-based algorithm and an iterative procedure to compute the latter. [17] analyzed the sensitivity of the optimal cost for the CMDP given in [10], and also gave a lower bound on the horizon beyond

which the optimal policy is stationary and deterministic. The reader is referred to book [1] for additional background and references on CMDPs in the risk-neutral case.

In contrast to CMDPs in the risk-neutral setting, CRSMDPs seem to have received very little attention in the literature. [22] provide a LP-based solution to finite-horizon CRSMDPs involving only finite-horizon standard discounted costs as constraints. [14] consider infinite-horizon risk constrained MDPs involving the minimization of a general law-invariant risk measure applied to infinite-horizon discounted cost. They also consider minimization of the expected utility of infinite-horizon discounted cost subject to a single constraint which is either a stochastic dominance constraint or a chance constraint. However, to the best of our knowledge, infinite-horizon CRSMDPs with multiple constraints of risk-neutral and risk-sensitive type over finite and infinite horizons have not been considered before. This motivates us to consider the problem of minimizing the expected exponential utility of total discounted cost over an infinite horizon scaled by a risk-factor under a fairly general set of finite- and infinite-horizon RS and standard discounted cost constraints. Our work thus seeks to extend that of [19] and [23] by including constraints. It also extends the work reported in [22] by allowing the horizon to be infinite and including RS cost constraints in addition to standard discounted cost constraints.

The main contribution of the current paper is to show that the problem possesses a solution if it is feasible, and provide two approximation techniques for finding a near-optimal US policy for the problem. The approximation is through two finite-horizon CRSMDPs which are constructed from the original CRSMDP by time-truncating the original objective and infinite-horizon constraint cost functions, and suitably perturbing the bounding constants defining the constraints. The construction of the approximating CRSMDPs is such that the feasible region of the first (second) approximating CRSMDPs provides an inner (outer) approximation to the feasible region of the original CRSMDP, while the optimal values of both converge to the optimal value of the original CRSMDP as the truncation horizon is extended. Consequently, a solution of the first approximating CRSMDP provides an  $\epsilon$ -optimal US solution to the original CRSMDP under an additional assumption. On the other hand, a solution to the second approximating CRSMDP provides a near-optimal US policy which is only guaranteed to be  $\epsilon$ -feasible for the original problem in the sense that constraint violations, if any, do not exceed  $\epsilon$ . A key step in the development is the introduction of a metric in which the set of MR policies is compact, and the objective and constraint functions are continuous. For the sake of completeness, we also provide a LP formulation for computing the solutions to the two approximating CRSMDPs.

The paper is organized as follows. The mathematical framework and the formal problem statement are given in Section 2. The Section 3 gives the main results on the existence of solutions to the infinite-horizon CRSMDP, and the  $\epsilon$ -optimality and  $\epsilon$ -feasibility of the solutions obtained from the two finite-horizon approximating CRSMDPs described above. The proofs of the main results along with related preliminaries are given in Section 4, while the LP formulation is described in Section 5. The proofs of all subsidiary results are given in the appendices.

## 2 Model Framework

As in the case of an MDP with standard discounted cost, the description of a RSMDP involves a state space, action space, immediate costs or rewards, and controlled transition probabilities. However, unlike the former, an RSMDP involves optimizing the expected exponential utility of the aggregated cost built up from costs collected over several decision epochs. In this paper, the aggregated cost is taken as the discounted sum of costs.

Let  $\mathcal{S} = \{s_1, s_2, \dots, s_m\}$  and  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$  denote the sets of all possible states and actions, respectively, with cardinalities  $|\mathcal{S}| = m$  and  $|\mathcal{A}| = n$ . The controlled transition probability, denoted by  $p(s_j | s_i, a_k)$ , represents the conditional probability that the system transitions to state  $s_j$  given that action  $a_k$  is taken in the current state  $s_i$ . Observe that the controlled transition probabilities are time homogeneous. Let  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  denote the immediate cost function.

Decisions are made at discrete epochs. In general, one can consider history-dependent decisions, in which the decision at any epoch requires the history of the process up to that epoch. However, in this paper, we only consider **Markovian** decisions, in which the decision at any epoch is based only on the current state at that epoch. More precisely, a **Markovian randomized (MR)** decision rule is a mapping  $d : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$  from the state space  $\mathcal{S}$  to the space  $\mathcal{P}(\mathcal{A})$  of probability distributions on the action space. Here,  $d(s)$  is the conditional probability mass function of the action chosen given that the state at the decision epoch is  $s$ . Note that decision rules which deterministically assign a unique action to each state are special cases of MR decision rules. An MR policy is an infinite sequence of MR decision rules, one for each decision epoch. We will denote the set of MR policies by  $\Pi_{\text{MR}}$ , and use  $\pi = \{d_t\}_{t=0}^{\infty}$  to denote an arbitrary policy  $\pi \in \Pi_{\text{MR}}$ , with  $d_t$  representing the decision rule applied at epoch  $t$ . An MR policy  $\pi = \{d_t\}_{t=0}^{\infty}$  is **ultimately stationary (US)** if there exists  $T \geq 0$  such that  $d_i = d_j$  for all  $i, j \geq T$ , and **stationary** if  $d_i = d_j$  for all  $i, j \geq 0$ .

For every initial state  $x \in \mathcal{S}$ , each policy  $\pi$  induces a probability measure on the set of sequences of state-action pairs  $\Omega = (\mathcal{S} \times \mathcal{A}) \times (\mathcal{S} \times \mathcal{A}) \times \dots$  (see, for example, Chapter 2 of [27]). We denote the state and action at a time  $t \in \{0, 1, \dots\}$  by random variables  $X_t$  and  $A_t$ , respectively, on  $\Omega$ . Let  $E_x^\pi[\cdot]$  represent the expectation conditioned on the initial state  $X_0 = x$  under the probability measure induced by  $\pi$ . Throughout this paper, we fix a discount factor  $\beta \in (0, 1)$  and a risk factor  $\gamma \in \mathbb{R} \setminus \{0\}$ . Unlike [18, 19], we allow  $\gamma$  to take both negative and positive values.

**Definition 2.1 [Standard discounted and RS cost functions:]** Given a policy  $\pi \in \Pi_{\text{MR}}$ , an immediate cost function  $R$  and an initial condition  $x$ , the RS discounted cost and the standard discounted cost over an infinite horizon as well as over a finite horizon  $T \geq 1$  are given by

$$J_{\gamma,R}^\pi(x) := \mathbb{E}_x^\pi \left[ e^{\gamma \sum_{t=0}^{\infty} \beta^t R(X_t, A_t)} \right], \quad J_{\gamma,R,T}^\pi(x) := \mathbb{E}_x^\pi \left[ e^{\gamma \sum_{t=0}^{T-1} \beta^t R(X_t, A_t)} \right], \quad (2.1)$$

$$v_R^\pi(x) := \mathbb{E}_x^\pi \left[ \sum_{t=0}^{\infty} \beta^t R(X_t, A_t) \right], \quad v_{R,T}^\pi(x) := \mathbb{E}_x^\pi \left[ \sum_{t=0}^{T-1} \beta^t R(X_t, A_t) \right] \quad (2.2)$$

Let  $J_{\gamma,R}^\pi$ ,  $J_{\gamma,R,T}^\pi$ ,  $v_R^\pi$  and  $v_{R,T}^\pi$ , represent the respective  $m$ -dimensional cost vectors whose elements indexed by the state  $x \in \mathcal{S}$  are given by (2.1) and (2.2).

## 2.1 Problem Statement

In this paper, we consider CRSMDPs involving a fairly general set of constraints. Specifically, we consider constraints involving RS and standard discounted costs over both finite and infinite horizons. To this end, given nonnegative integers  $M$ ,  $\hat{M}$ ,  $\bar{M}$  and  $\check{M}$  at least one of which is nonzero, we introduce a set of immediate cost functions  $C_i$ ,  $\hat{C}_j$ ,  $\bar{C}_k$ ,  $\check{C}_l$ , constants  $b_i$ ,  $\hat{b}_j$ ,  $\bar{b}_k$ ,  $\check{b}_l$ , and finite times  $\bar{T}_k$ ,  $\check{T}_l$  for  $i \in \{1, 2, \dots, M\}$ ,  $j \in \{1, 2, \dots, \hat{M}\}$ ,  $k \in \{1, 2, \dots, \bar{M}\}$ , and  $l \in \{1, 2, \dots, \check{M}\}$ . We define our constraints in terms of subsets of  $\Pi_{\text{MR}}$  given by

$$\mathcal{C} := \{\pi \in \Pi_{\text{MR}} : v_{C_i}^\pi(x) \leq b_i \text{ for all } i = 1, 2, \dots, M\}, \quad (2.3)$$

$$\hat{\mathcal{C}} := \{\pi \in \Pi_{\text{MR}} : J_{\gamma, \hat{C}_j}^\pi(x) \leq \hat{b}_j \text{ for all } j = 1, 2, \dots, \hat{M}\}, \quad (2.4)$$

$$\bar{\mathcal{C}} := \{\pi \in \Pi_{\text{MR}} : v_{\bar{C}_k, \bar{T}_k}^\pi(x) \leq \bar{b}_k \text{ for all } k = 1, 2, \dots, \bar{M}\}, \quad (2.5)$$

$$\check{\mathcal{C}} := \{\pi \in \Pi_{\text{MR}} : J_{\gamma, \check{C}_l, \check{T}_l}^\pi(x) \leq \check{b}_l \text{ for all } l = 1, 2, \dots, \check{M}\}, \quad (2.6)$$

where the notation follows the definitions in (2.1)-(2.2). Our assumption that  $\mathcal{S}$  and  $\mathcal{A}$  are finite implies that all the immediate cost functions appearing in (2.3)-(2.6) as well as the immediate cost function  $R$  are uniformly bounded above in absolute value by a common bound, which we denote by  $C > 0$ .

We are interested in the infinite-horizon CRSMDP problem given by

$$\min_{\pi \in \mathcal{F}} J_{\gamma, R}^{\pi}(x), \quad (P)$$

where  $\mathcal{F} := \mathcal{C} \cap \hat{\mathcal{C}} \cap \bar{\mathcal{C}} \cap \check{\mathcal{C}}$ . Note that the CRSMDP problem (P) involves minimizing an infinite-horizon RS cost function subject to  $M + \bar{M} + \hat{M} + \check{M}$  number of constraints involving finite- and infinite-horizon RS and standard discounted constraints.

Our approach is to approximate the infinite-horizon problem (P) by a finite-horizon problem where the objective function as well as the constraints are appropriately time truncated. For this purpose, denote  $K = C/(1 - \beta)$ . For each  $T \in \{1, 2, \dots\}$ , let  $K_T = e^{|\gamma|K\beta^T}$ , and define

$$\mathcal{C}_T^- := \{\pi \in \Pi_{\text{MR}} : v_{\mathcal{C}_i, T}^{\pi}(x) \leq b_i - K\beta^T \text{ for all } i = 1, 2, \dots, M\}, \quad (2.7)$$

$$\hat{\mathcal{C}}_T^- := \left\{ \pi \in \Pi_{\text{MR}} : J_{\gamma, \hat{\mathcal{C}}_j, T}^{\pi}(x) \leq \frac{\hat{b}_j}{K_T} \text{ for all } j = 1, 2, \dots, \hat{M} \right\}, \quad (2.8)$$

$$\mathcal{C}_T^+ := \{\pi \in \Pi_{\text{MR}} : v_{\mathcal{C}_i, T}^{\pi}(x) \leq b_i + K\beta^T, \text{ for all } i = 1, 2, \dots, M\}, \quad (2.9)$$

$$\hat{\mathcal{C}}_T^+ := \{\pi \in \Pi_{\text{MR}} : J_{\gamma, \hat{\mathcal{C}}_j, T}^{\pi}(x) \leq \hat{b}_j K_T, \text{ for all } j = 1, 2, \dots, \hat{M}\}. \quad (2.10)$$

The subsets of  $\Pi_{\text{MR}}$  defined in (2.7)-(2.10) represent constraints defined in terms of the time-truncated versions of the cost functions appearing in (2.3)-(2.4). Taken together with the time-truncated version of the infinite-horizon RS cost of problem (P), the constraints (2.7)-(2.10) allow us to define, for each  $T \geq 1$ , the finite-horizon problems  $(P_T^-)$  and  $(P_T^+)$  given by

$$(P_T^-) : \min_{\pi \in \mathcal{F}_T^-} J_{\gamma, R, T}^{\pi}(x) \quad \text{and} \quad (P_T^+) : \min_{\pi \in \mathcal{F}_T^+} J_{\gamma, R, T}^{\pi}(x), \quad (2.11)$$

where  $\mathcal{F}_T^- := \mathcal{C}_T^- \cap \hat{\mathcal{C}}_T^- \cap \bar{\mathcal{C}} \cap \check{\mathcal{C}}$  and  $\mathcal{F}_T^+ := \mathcal{C}_T^+ \cap \hat{\mathcal{C}}_T^+ \cap \bar{\mathcal{C}} \cap \check{\mathcal{C}}$ . We will show that the finite-horizon problems  $(P_T^-)$  and  $(P_T^+)$  serve as approximations to the infinite-horizon problem (P). More precisely, we will show that the optimal values of the problems  $(P_T^-)$  and  $(P_T^+)$  converge to the value of (P) as  $T \rightarrow \infty$  under appropriate conditions. Moreover, the feasible sets of the problems are related in that, the feasible sets of  $(P_T^-)$  for different values of  $T$  form an increasing nested sequence contained in the feasible set of (P), while the feasible sets of  $(P_T^+)$  form an decreasing nested sequence containing the feasible set of (P). As we will show in the next section, these properties

allow us to conclude that solutions of  $(P_T^-)$  and  $(P_T^+)$  for sufficiently large  $T$  serve as approximate solutions to  $(P)$ .

As mentioned in the introduction, reference [23] followed an identical approach to approximate the unconstrained version of the problem  $(P)$  by the unconstrained version of the problem  $(P_T^-)$  or, equivalently, the unconstrained version of  $(P_T^+)$ . This paper extends the approach of [23] to constrained RSMDPs.

We state the main results in the next section.

### 3 Main results

Our first result below asserts that the CRSMDPs  $(P)$  and  $(P_T^+)$ , for every  $T$ , possess solutions if  $(P)$  is feasible, while the CRSMDP  $(P_T^-)$  has a solution if it is feasible. The proof is given in section 4.

**Theorem 3.1 [Existence of optimal solutions:]** *If the feasible region  $\mathcal{F} \subseteq \Pi_{\text{MR}}$  of the original problem  $(P)$  is non-empty, then  $(P)$  has a solution and, for each  $T > 0$ ,  $(P_T^+)$  has a solution. On the other hand, if  $\mathcal{F}_{T^*}^- \neq \emptyset$  for some  $T^* > 0$ , then  $\mathcal{F} \neq \emptyset$ , and  $(P_T^-)$  has a solution for each  $T \geq T^*$ .*

While Theorem 3.1 guarantees the existence of a solution to  $(P)$ , computing a solution is a challenge. One reason is that the solution is generally non-stationary (see, for instance, [19]). To the best of our knowledge, there is no numerical method to compute such a solution. One expects that if the optimal policies are stationary or of the US type, then it may be relatively easier to compute the solution. Hence, it is natural to ask: i) can US policies approximate the optimal value of the problem  $(P)$ ; and ii) can a numerical method be devised to compute these approximate solutions? In this paper we provide affirmative answers to both these questions. Specifically, we show below that any policy that solves either  $(P_T^-)$  or  $(P_T^+)$  for sufficiently large  $T$  can be extended to a US policy that approximately solves the original problem  $(P)$ . Furthermore, we discuss the solution of  $(P_T^-)$  and  $(P_T^+)$  using a LP-based approach in section 5. Our next two results show how  $(P_T^-)$  and  $(P_T^+)$  approximate  $(P)$ .

Theorem 3.2 given below is our second main result. It gives sufficient conditions under which a solution of  $(P_T^-)$  yields an approximate solution to  $(P)$  for  $T$  sufficiently large. The sufficient condition involves checking if 0 is a local minimum value for the *maximum constraint violation* map  $h : \Pi_{\text{MR}} \mapsto \mathbb{R}$  which maps a policy  $\pi \in \Pi_{\text{MR}}$  to the largest constraint violation achieved by  $\pi$  across all the  $M + \bar{M} + \hat{M} + \check{M}$  constraints that form



part of the problem  $(P)$ . The qualifier “local” above refers to the topology on  $\Pi_{\text{MR}}$  that we introduce in section 4.1.2.

**Theorem 3.2** [*Finite-horizon approximation with feasibility:*] *Define the maps  $\theta, \hat{\theta}, \bar{\theta}, \check{\theta}, h : \Pi_{\text{MR}} \rightarrow \mathbb{R}$  by*

$$\begin{aligned} \theta(\pi) &:= \max_{i \in \{1, \dots, M\}} \{v_{\hat{C}_i}^\pi(x) - b_i\}; \quad \hat{\theta}(\pi) := \max_{j \in \{1, \dots, \hat{M}\}} \{J_{\gamma, \hat{C}_j}^\pi(x) - \hat{b}_j\}, \\ \bar{\theta}(\pi) &:= \max_{k \in \{1, \dots, \bar{M}\}} \{v_{\bar{C}_k, \bar{T}_k}^\pi(x) - \bar{b}_k\}; \quad \check{\theta}(\pi) := \max_{l \in \{1, \dots, \check{M}\}} \{J_{\gamma, \check{C}_l, \check{T}_l}^\pi(x) - \check{b}_l\}, \\ \text{and} \quad h(\pi) &:= \max\{\theta(\pi), \hat{\theta}(\pi), \bar{\theta}(\pi), \check{\theta}(\pi)\}. \end{aligned} \quad (3.1)$$

Suppose  $\mathcal{F} \neq \emptyset$  and 0 is not a local minimum value of the map  $h$ . Then, for every  $\epsilon > 0$ , there exists  $T^* > 0$  such that, for every  $T \geq T^*$  and every policy  $\eta \in \mathcal{F}_T^-$  that solves the problem  $(P_T^-)$ ,  $\eta$  is  $\epsilon$ -optimal for the problem  $(P)$ , that is,  $\eta \in \mathcal{F}$  and  $\eta$  satisfies

$$0 \leq J_{\gamma, R}^\eta(x) - \inf_{\pi \in \mathcal{F}} J_{\gamma, R}^\pi(x) < \epsilon. \quad (3.2)$$

Theorem 3.2 guarantees a feasible and  $\epsilon$ -optimal solution to the original infinite-horizon problem  $(P)$ . However, the theorem requires checking a local minimum condition on the function  $h$  given by (3.1), which could be difficult to check in practice. At the same time, we show through a counterexample in D that the conclusion of the theorem need not hold if the condition is violated. Our next result provides an alternative means of obtaining an approximate solution to problem  $(P)$  in situations where the local minimum condition on  $h$  either does not hold or is difficult to verify. The theorem shows that, for sufficiently large  $T$ , a solution of  $(P_T^+)$  provides a policy which is approximately optimal and approximately feasible for the problem  $(P)$ , where approximate feasibility, made precise below, means that the maximum constraint violation can be bounded by an arbitrarily specified small quantity.

**Definition 3.3** [ *$\epsilon$ -feasibility:*] *Given  $\epsilon > 0$ , a policy  $\pi \in \Pi_{\text{MR}}$  is  $\epsilon$ -feasible for the problem  $(P)$  if it satisfies*

$$v_{\hat{C}_i}^\pi(x) \leq b_i + \epsilon, \quad J_{\gamma, \hat{C}_j}^\pi(x) \leq \hat{b}_j + \epsilon, \quad v_{\bar{C}_k, \bar{T}_k}^\pi(x) \leq \bar{b}_k + \epsilon, \quad \text{and} \quad J_{\gamma, \check{C}_l, \check{T}_l}^\pi(x) \leq \check{b}_l + \epsilon$$

for all  $i = 1, \dots, M$ ,  $j = 1, \dots, \hat{M}$ ,  $k = 1, \dots, \bar{M}$  and  $l = 1, \dots, \check{M}$ .

**Theorem 3.4** [*Finite-horizon approximation with  $\epsilon$ -feasibility:*] *Suppose  $\mathcal{F}$  is nonempty. Then, for every  $\epsilon > 0$ , there exists  $T^* > 0$  such that, for every  $T \geq T^*$  and every policy  $\eta \in \mathcal{F}_T^+$  such that  $\eta$  is a solution to the problem  $(P_T^+)$ ,  $\eta$  is  $\epsilon$ -feasible for the problem  $(P)$ , and satisfies*

$$\left| J_{\gamma, R}^\eta(x) - \inf_{\pi \in \mathcal{F}} J_{\gamma, R}^\pi(x) \right| < \epsilon. \quad (3.3)$$

It is worthwhile to compare theorems 3.2 and 3.4. Both theorems show that an approximate solution to the infinite-horizon CRSMDP  $(P)$  can be obtained by solving either of the finite-horizon CRSMDPs  $(P_T^-)$  or  $(P_T^+)$  for a sufficiently large finite horizon  $T$ . While Theorem 3.2 requires an additional condition to hold, the approximate solution it provides is feasible for the original problem  $(P)$ . In contrast, Theorem 3.4 does not involve any additional conditions, but yields an approximate solution that is only approximately feasible in the sense of Definition 3.3.

**Remark 3.5** *Observe from (2.7)-(2.10) that, given any policy  $\pi \in \Pi_{\text{MR}}$ , the feasibility as well as optimality of  $\pi$  for the problems  $(P_T^-)$  and  $(P_T^+)$  are determined only by the first  $T$  decision rules of  $\pi$ . Consequently, appending arbitrary decision rules after the horizon  $T$  to a solution of  $(P_T^-)$  or  $(P_T^+)$  does not affect its feasibility or optimality with respect to  $(P_T^-)$  or  $(P_T^+)$ . In particular, one may extend a solution of either problem to a US policy by choosing all decision rules beyond the horizon  $T$  to be the same. Hence, the approximate solutions to the problem  $(P)$  guaranteed by theorems 3.2-3.4 may be chosen to be US policies.*

**Remark 3.6** *Theorems 3.1, 3.2 and 3.4 hold as stated even in the case where the discount factors and risk factors used in (2.1) and (2.3)-(2.6) are different. While the modifications needed to accommodate the added generality is only minor, we have sacrificed the slight increase in generality in favour of simplicity. The case of multiple discount factors is considered in [26] for CMDPs in the standard-discounted-cost setting.*

The next section provides proofs of the three main results stated in this section.

## 4 Proofs of the Main Results

In this section, we provide proofs of theorems 3.1, 3.2 and 3.4. In particular, we motivate the main steps in the proofs, and provide the auxiliary results which are needed to complete the proofs.

### 4.1 Proof Outline of Theorem 3.1

The proof of Theorem 3.1 uses the well known fact that a continuous function achieves its infimum on a nonempty compact set. Thus the major steps in proving Theorem 3.1 are: (i) constructing a topology on the set of policies  $\Pi_{\text{MR}}$ , (ii) showing that the objective functions in  $(P)$ ,  $(P_T^-)$  and  $(P_T^+)$  are

continuous in this topology, and (iii) showing that the feasible regions are compact in the same topology. Steps (i), (ii) and (iii) above are achieved in sub-sections 4.1.2 - 4.1.5 below. Before proceeding, we introduce the required notations.

#### 4.1.1 Notations

The set  $\mathbb{R}^{m \times n}$  of all real matrices of order  $m \times n$  is a real vector space under element-wise addition and scalar multiplication. The matrix  $\mathbb{1}_{m \times n}$  is the  $m \times n$  matrix with all elements equal to 1. Any vector in  $\mathbb{R}^n$  is represented by a column vector.

For any  $B = [b_{i,j}] \in \mathbb{R}^{m \times n}$  and  $y = [y_i] \in \mathbb{R}^n$ ,  $y'$  and  $B'$ , respectively represent their transposes. The  $l_\infty$  norm (or max norm) and the  $l_1$  norm of  $y \in \mathbb{R}^n$  are defined by  $\|y\|_\infty := \max_{i=1,\dots,n} |y_i|$  and  $\|y\|_1 = \sum_{i=1}^n |y_i|$ , respectively.

We denote by  $\|B\|_\infty$  the matrix norm of  $B \in \mathbb{R}^{m \times n}$  induced by  $l_\infty$  which equals  $\max_{i=1,\dots,m} \sum_{j=1}^n |b_{i,j}|$  (Refer [15, Ex. 5.6.5, p. 345]). We also define  $\|B\|_{\max} = \max_{i,j} |b_{i,j}|$ . For any two matrices  $B_1, B_2 \in \mathbb{R}^{m \times n}$ , the Schur product, denoted by  $B_1 \odot B_2$ , is the matrix obtained by element-wise multiplication of matrices  $B_1$  and  $B_2$ . Similarly, the Schur exponential, denoted by  $e^{\odot B}$ , is the element-wise exponential of the matrix  $B$ .

#### 4.1.2 Construction of a Topology on Policy Space

The topological structure that we define on  $\Pi_{\text{MR}}$  is based on the observation that an MR policy is a sequence of MR decision rules, one for each decision epoch, while each MR decision rule can be represented by a row-stochastic matrix of dimension  $m \times n$ , (recall that  $m = |\mathcal{S}|$  and  $n = |\mathcal{A}|$ ). This allows us to construct a metric, and hence a topology, on  $\Pi_{\text{MR}}$  by using the matrix norm  $\|\cdot\|_\infty$ . To this end, let  $\mathcal{R} \subset \mathbb{R}^{m \times n}$  denote the set of row-stochastic matrices, that is,

$$\mathcal{R} := \left\{ B = [b_{i,j}] \in \mathbb{R}^{m \times n} : b_{i,j} \geq 0 \forall i, j \text{ and } \sum_{j=1}^n b_{i,j} = 1 \forall i \right\}. \quad (4.1)$$

As a subset of  $\mathbb{R}^{m \times n}$ ,  $\mathcal{R}$  is closed and bounded, and hence compact, in the topology induced by the norm  $\|\cdot\|_\infty$ .

Next, we define a sequence of equivalent metrics on  $\mathcal{R}$ . Fix  $\delta \in (\beta, 1)$ . For each  $t \in \{0, 1, \dots\}$ , define a metric  $\mu_t$  on  $\mathcal{R}$  by letting  $\mu_t(B_1, B_2) = \delta^t \|B_1 - B_2\|_\infty$  for  $B_1, B_2 \in \mathcal{R}$ . Note that, for each  $t > 0$ , the metric space  $(\mathcal{R}, \mu_t)$  is a compact metric space with diameter  $2\delta^t$ .

As used by [8], we characterize an MR decision rule by a row-stochastic matrix. We identify an MR decision rule  $d : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$  with the unique matrix in  $\mathcal{R}$  whose  $(i, j)$ th element is the probability of taking action  $a_j$  at state  $s_i$  under the decision rule  $d$ . It is clear that the identification just described sets up a bijection between  $\mathcal{R}$  and the set of MR decision rules. As a departure from the convention of using upper case letters for matrices, we denote the matrix corresponding to an MR decision rule  $d$  by  $d$  again, and use the suggestive notation  $d(a_j|s_i)$  denote its  $(i, j)$ th element.

Next, it is easy to see that a MR policy can be identified with a unique sequence having elements in  $\mathcal{R}$ . We therefore identify  $\Pi_{\text{MR}}$  with the set  $\mathcal{R}^\infty$  of all sequences having elements in  $\mathcal{R}$ . In the rest of the paper, we will use  $\Pi_{\text{MR}}$  interchangeably with  $\mathcal{R}^\infty$ . Using this identification, we now define a metric on  $\Pi_{\text{MR}}$  as follows. Given policies  $\pi_1 = \{d_t\}_{t=0}^\infty$  and  $\pi_2 = \{f_t\}_{t=0}^\infty$  in  $\Pi_{\text{MR}} (= \mathcal{R}^\infty)$ , define  $\mu(\pi_1, \pi_2) := \sup_{t \geq 0} \mu_t(d_t, f_t)$ . Our next result shows that  $\mu$  is a metric on  $\Pi_{\text{MR}}$ , and provides the foundation for the framework that we use to prove our main results.

**Theorem 4.1** *The map  $\mu$  is a metric on  $\Pi_{\text{MR}}$ , and  $(\Pi_{\text{MR}}, \mu)$  is a compact metric space.*

It is well-known that any Cartesian product of compact spaces is compact in the product topology. Theorem 4.1 simply follows by showing that the metric  $\mu$  constructed above metrizes the product topology on  $\mathcal{R}^\infty$ . The choice of the metric is thus crucial to Theorem 4.1. For instance, compactness fails to hold if  $\mu$  is chosen to be the more familiar  $l^\infty$  metric on  $\mathcal{R}^\infty$  obtained by setting  $\delta = 1$ .

### 4.1.3 Continuity of standard discounted cost functions

To show that the finite- and infinite-horizon standard discounted cost functions appearing in problems  $(P)$ ,  $(P_T^-)$  and  $(P_T^+)$  are continuous, we exploit well-known expressions for such a cost function which use matrix-vector notation to explicitly bring out the dependence on the decision rules that constitute the policy (see, for instance, [27, eqn. (6.1.2)]). To do so, we will need some additional notation.

First, given  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we denote by  $p(\cdot|s, a) \in \mathbb{R}^m$  the vector whose  $i$ th element is the transition probability  $p(s_i|s, a)$ . Given an immediate cost function  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ , we abuse the notation slightly and denote by  $R \in \mathbb{R}^{m \times n}$  the matrix whose  $(i, j)$ th element is given by  $R(s_i, a_j)$ . Following [27, sec. 5.6], the vector of expected costs  $R_d \in \mathbb{R}^m$  and the transition

probability matrix  $P_d \in \mathbb{R}^{m \times m}$  corresponding to a MR decision rule  $d \in \mathcal{R}$  are defined element-wise by

$$(R_d)_i := \sum_{j=1}^n R(s_i, a_j) d(a_j | s_i), \quad (P_d)_{i,j} := \sum_{k=1}^n p(s_j | s_i, a_k) d(a_k | s_i). \quad (4.2)$$

Observe that  $R_d = (R \odot d) \mathbb{1}_{n \times 1}$ .

Next, given a MR policy  $\pi = \{d_t\}_{t=0}^\infty$  define the  $t$ -step transition probability matrix  $P_t^\pi := \prod_{0 < l < t} P_{d_l}$  for each  $t = 1, 2, \dots$ , and set  $P_0^\pi$  to be the  $m \times m$  identity matrix. The vector expressions for the standard discounted costs  $v_R^\pi$  and  $v_{R,T}^\pi$  defined in (2.2) may now be written as

$$v_{R,T}^\pi = \sum_{t=0}^{T-1} \beta^t P_t^\pi R_{d_t}, \quad v_R^\pi = \sum_{t=0}^{\infty} \beta^t P_t^\pi R_{d_t}. \quad (4.3)$$

The expressions above are well known and given, for instance, in [27, Ch. 6]. Note that  $P_t^\pi R_{d_t}$ , the  $t$ th term in the summation for  $v_{R,T}^\pi$  above, is the expectation of the immediate cost incurred at time  $t$  under the policy  $\pi$ . The continuity of  $v_{R,T}^\pi$  in (4.3) thus depends on how the expected immediate cost at time  $t$  changes when the decision rules up to time  $t$  are changed. The bound in the next lemma answers this question.

**Lemma 4.2** *Let  $\pi_1 = \{d_t\}_{t=0}^\infty$  and  $\pi_2 = \{f_t\}_{t=0}^\infty$  be two policies in  $\Pi_{\text{MR}}$ , and let  $t \geq 0$ . Then,  $\|P_t^{\pi_1} R_{d_t} - P_t^{\pi_2} R_{f_t}\|_\infty \leq C \sum_{i=0}^t \|d_i - f_i\|_\infty$  holds.*

The inequality above along with the expression (4.3) leads to Lipschitz continuity of finite-horizon standard discounted cost functions stated in Theorem 4.4 below. The proof of continuity of infinite-horizon standard discounted cost functions depends on the following result which states that the finite-horizon standard discounted cost converges to the infinite-horizon standard discounted cost as the horizon  $T$  tends to infinity, uniformly in the policy  $\pi$ . While the result appears to be well known, we include a proof in A for the sake of completeness.

**Lemma 4.3 [Convergence of finite-horizon standard discounted cost:]**

*Let  $x \in \mathcal{S}$  and  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  be an immediate cost function. Then, for every  $\epsilon > 0$ , there exists  $T^* > 0$  such that  $|v_{R,T}^\pi(x) - v_R^\pi(x)| < \epsilon$  for all  $T \geq T^*$  and all  $\pi \in \Pi_{\text{MR}}$ .*

The continuity properties of finite- and infinite-horizon standard discounted costs are summarized in our next result.

**Theorem 4.4** [*Continuity of standard discounted costs*] Let  $T > 0$ . The mappings  $\pi \mapsto v_{R,T}^\pi$  and  $\pi \mapsto v_R^\pi$  from the metric space  $(\Pi_{\text{MR}}, \mu)$  to the metric space  $(\mathbb{R}^m, \|\cdot\|_\infty)$  are Lipschitz continuous with Lipschitz constants  $\frac{K(\delta^T - \beta^T)}{\delta^{T-1}(\delta - \beta)}$  and  $K\delta(\delta - \beta)^{-1}$ , respectively.

#### 4.1.4 Continuity of RS cost functions

As in the previous subsection, we begin by expressing RS cost functions in a way that makes the dependence on decision rules explicit. To this end, we introduce the risk-sensitive version of the familiar state-action value function or  $Q$ -factor. Given a finite-horizon  $T > 0$  and a time instant  $t \in \{0, \dots, T-1\}$ , the RS state-action value function  $Q_{t,T}^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  of a policy  $\pi \in \Pi_{\text{MR}}$  at the state-action pair  $(s, a) \in \mathcal{S} \times \mathcal{A}$  is the expected RS cost incurred between  $t$  to  $T$  when the action  $a$  is applied at time  $t$  in the state  $s$ , and the policy  $\pi$  is applied thereafter. More precisely, given  $(s, a) \in \mathcal{S} \times \mathcal{A}$  and  $t, T$  as above, we have

$$Q_{t,T}^\pi(s, a) := \mathbb{E}^\pi \left[ e^{\gamma \sum_{\tau=t}^{T-1} \beta^\tau R(X_\tau, A_\tau)} \middle| X_t = s, A_t = a \right]. \quad (4.4)$$

As in the case of the immediate cost function  $R$  in sub-subsection 4.1.3, it will be convenient to treat  $Q_{t,T}^\pi$  as the  $m \times n$  matrix having  $Q_{t,T}^\pi(s_i, a_j)$  as its  $(i, j)$ th entry.

The next proposition provides recursive expressions for the risk-sensitive state-action value function as well as the finite-horizon RS cost for a given policy analogous to well-known recursive expressions for the standard discounted cost as given in [27, eqn. (4.2.6)].

**Proposition 4.5** Suppose  $T \geq 1$ , and let  $\pi = \{d_t\}_{t=0}^\infty \in \Pi_{\text{MR}}$ . Then the following equations hold.

$$Q_{T-1,T}^\pi = e^{\odot \gamma \beta^{T-1} R}, \quad (4.5)$$

$$Q_{t-1,T}^\pi(s, a) = e^{\gamma \beta^{t-1} R(s,a)} p(\cdot | s, a)' (d_t \odot Q_{t,T}^\pi) \mathbb{1}_{n \times 1},$$

for  $t = 1, 2, \dots, T-1$ ,  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , (4.6)

$$J_{\gamma, R, T}^\pi = (d_0 \odot Q_{0,T}^\pi) \mathbb{1}_{n \times 1}. \quad (4.7)$$

Equations (4.5)-(4.7) provide a backward-recursion-based procedure for policy evaluation of finite-horizon RS cost similar to that for standard discounted cost (see equation (22) of [27]), and could be of independent interest.

It is clear from (4.5)-(4.7) that continuity properties of the finite-horizon RS cost is determined by those of the matrices  $Q_{0,T}^\pi, \dots, Q_{T-1,T}^\pi \in \mathbb{R}^{m \times n}$ .

Our next result is a lemma that gives explicit bounds on the elements of these matrices and their differences. Both bounds will be used for proving that the finite-horizon RS cost defined in (2.1) depends continuously on the policy. The proof is provided in A.

**Lemma 4.6** *Let  $\pi_1 = \{d_t\}_{t=0}^\infty$  and  $\pi_2 = \{f_t\}_{t=0}^\infty$  be two policies in  $\Pi_{\text{MR}}$ . Then, for every  $T \geq 1$  and every  $0 \leq t \leq T - 1$ , we have*

$$\begin{aligned} \|Q_{t,T}^{\pi_1}\|_{\max} &\leq e^{|\gamma|K(\beta^t - \beta^T)}, & (4.8) \\ \|Q_{t,T}^{\pi_1} - Q_{t,T}^{\pi_2}\|_{\max} &\leq \begin{cases} 0, & \text{for } t = T - 1, \\ e^{|\gamma|K(\beta^t - \beta^T)} \sum_{\tau=t+1}^{T-1} \|d_\tau - f_\tau\|_\infty, & \text{for } 0 \leq t \leq T - 2. \end{cases} & (4.9) \end{aligned}$$

The inequalities (4.8)-(4.9) along with the backward recursive expression given in Proposition 4.5 combine to yield continuity of the finite-horizon RS cost function. The extension to infinite-horizon RS cost comes from Theorem 1 of [23], which asserts that the finite-horizon RS cost converges exponentially to the infinite-horizon RS cost as the horizon increases. We state here a weaker, asymptotic version of the result by [23, Thm 1] for convenience.

**Lemma 4.7 [Convergence of finite-horizon RS cost:]** *Let  $x \in \mathcal{S}$ , and let  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  be an immediate cost function. Then, for every  $\epsilon > 0$  there exists  $T^* > 0$  such that  $|J_{\gamma,R,T}^\pi(x) - J_{\gamma,R}^\pi(x)| < \epsilon$  for all  $T \geq T^*$  and all  $\pi \in \Pi_{\text{MR}}$ .*

The following result summarizes the continuity properties of finite- and infinite-horizon RS costs. The proof is given in A.

**Theorem 4.8 [Continuity of RS costs:]** *Let  $T \geq 0$ . The mapping  $\pi \mapsto J_{\gamma,R,T}^\pi$  from the metric space  $(\Pi_{\text{MR}}, \mu)$  to the metric space  $(\mathbb{R}^m, \|\cdot\|_\infty)$  is Lipschitz continuous with Lipschitz constant  $\delta^{-(T-1)}(1 - \delta)^{-1} e^{|\gamma|K(1 - \beta^T)}$ . Furthermore, the mapping  $\pi \mapsto J_{\gamma,R}^\pi$  from the metric space  $(\Pi_{\text{MR}}, \mu)$  to the metric space  $(\mathbb{R}^m, \|\cdot\|_\infty)$  is uniformly continuous.*

Note that components of a vector-valued Lipschitz or uniformly continuous function are also Lipschitz or uniformly continuous, respectively. As a result, the assertions of theorems 4.4 and 4.8 also hold for the respective cost functions for a given initial condition. This is the form in which theorems 4.4 and 4.8 will be used in the sequel.

### 4.1.5 Compactness of feasible regions

Theorem 4.1 already states that  $\Pi_{\text{MR}}$  is a compact space in the topology induced by the metric  $\mu$ . Hence compactness of the feasible regions of the problems  $(P)$ ,  $(P_T^-)$  and  $(P_T^+)$  will follow if it can be shown that these sets are closed subsets of  $\Pi_{\text{MR}}$ . However, each of these sets is defined through non-strict inequalities involving finite- and infinite-horizon classical and RS cost functions. Since all of these functions have been shown to be continuous in the topology on  $\Pi_{\text{MR}}$  by theorems 4.4 and 4.8, it follows that the feasible regions are closed subsets of  $\Pi_{\text{MR}}$ , and hence compact. This conclusion is stated as the first assertion of our next lemma. The remaining two assertions of the lemma pertain to the behaviour of the feasible regions of the problems  $(P_T^-)$  and  $(P_T^+)$  for different values of  $T$ , and will be useful for the proofs of theorems 3.2 and 3.4.

**Lemma 4.9** *The following statements hold.*

- (i) *The set  $\mathcal{F}$  is a compact subset of  $\Pi_{\text{MR}}$ .*
- (ii)  *$\{\mathcal{F}_T^-\}_{T=1}^\infty$  is a non-decreasing sequence of nested compact sets contained in  $\mathcal{F}$ .*
- (iii)  *$\{\mathcal{F}_T^+\}_{T=1}^\infty$  is a non-increasing sequence of nested compact sets satisfying  $\bigcap_{T=1}^\infty \mathcal{F}_T^+ = \mathcal{F}$ .*

### 4.1.6 Proof of Theorem 3.1

The results of the preceding few sub-subsections now enable us to provide a formal proof of Theorem 3.1.

To prove the first assertion in Theorem 3.1, suppose  $\mathcal{F}$  is nonempty. By Lemma 4.9.(iii),  $\mathcal{F}_T^+$  is nonempty for all  $T$ . Theorem 4.8 implies that the objective functions of the problems  $(P)$  and  $(P_T^+)$  are continuous functions on the metric space of policies  $(\Pi_{\text{MR}}, \mu)$ . Statements (i) and (iii) of Lemma 4.9 guarantee that the respective feasible regions  $\mathcal{F}$  and  $\mathcal{F}_T^+$  are compact subsets of  $(\Pi_{\text{MR}}, \mu)$ . The first assertion now follows from the well-known fact that a continuous function achieves its infimum over a non-empty compact set.

To prove the second assertion, let  $T^* > 0$  be such that  $\mathcal{F}_{T^*}^-$  is nonempty. Statement (ii) of Lemma 4.9 implies that  $\mathcal{F}$  and  $\mathcal{F}_T^-$  are nonempty for all  $T \geq T^*$ . The rest of the assertion now follows by arguing as in the previous paragraph.  $\square$



## 4.2 Proof Outline of Theorem 3.2

The starting point for the proof of Theorem 3.2 is Lemma 4.7 which states that, for a given initial state, the finite-horizon RS and standard discounted costs converge to corresponding infinite-horizon costs uniformly in the policy as the horizon  $T \rightarrow \infty$ . In view of Lemma 4.7 and Theorem 3.1, to prove Theorem 3.2, it suffices to show that the optimal value of  $(P_T^-)$  converges to that of  $(P)$  as  $T \rightarrow \infty$ , that is,

$$\lim_{T \rightarrow \infty} \left( \inf_{\pi \in \mathcal{F}_T^-} J_{\gamma,R,T}^\pi(x) \right) = \inf_{\pi \in \mathcal{F}} J_{\gamma,R}^\pi(x). \quad (4.10)$$

Indeed, if (4.10) holds, then the proof of Theorem 3.2 may be completed as follows:

- (a). Suppose (4.10) holds. Choose  $\epsilon > 0$ . By Lemma 4.7, there exists  $T^* > 0$  such that  $|J_{\gamma,R,T}^\pi(x) - J_{\gamma,R}^\pi(x)| < \epsilon/2$  for all  $T \geq T^*$  and all  $\pi \in \Pi_{\text{MR}}$ .
- (b). By Theorem 3.1, the right hand side of (4.10) is well defined, which implies that  $\mathcal{F}_T^- \neq \emptyset$  eventually.
- (c). We may choose  $T^*$  in (a) above to be further larger, if required, so that every  $T > T^*$  also satisfies  $\mathcal{F}_T^- \neq \emptyset$  and

$$\left| \inf_{\pi \in \mathcal{F}_T^-} J_{\gamma,R,T}^\pi(x) - \inf_{\pi \in \mathcal{F}} J_{\gamma,R}^\pi(x) \right| < \epsilon/2.$$

- (d). Again by Theorem 3.1, for every  $T > T^*$ , there exists a policy  $\eta \in \Pi_{\text{MR}}$  that solves  $(P_T^-)$ , that is, a  $\eta \in \mathcal{F}_T^-$  such that  $J_{\gamma,R,T}^\eta(x) = \inf_{\pi \in \mathcal{F}_T^-} J_{\gamma,R,T}^\pi(x)$ .
- (e). Finally, the inequalities in (a) and (c) together show that every policy  $\eta \in \mathcal{F}_T^-$  that solves problem  $(P_T^-)$  for a given  $T > T^*$  also satisfies  $J_{\gamma,R}^\eta(x) - \inf_{\pi \in \mathcal{F}} J_{\gamma,R}^\pi(x) = (J_{\gamma,R}^\eta(x) - J_{\gamma,R,T}^\eta(x)) + (J_{\gamma,R,T}^\eta(x) - \inf_{\pi \in \mathcal{F}} J_{\gamma,R}^\pi(x)) < \epsilon$ , that is,  $\eta$  satisfies (3.2).

Thus, the proof of Theorem 3.2 is achieved if (4.10) is proved. Hence we next outline the steps involved in proving (4.10).

Lemma 4.7 asserts that the objective function  $J_{\gamma,R,T}^\pi(x)$  of the problem  $(P_T^-)$  appearing on the left hand side of (4.10) uniformly approximates the objective function  $J_{\gamma,R}^\pi(x)$  of the problem  $(P)$  appearing on the right hand side of (4.10) as  $T \rightarrow \infty$ . Hence one intuitively expects (4.10) to hold if the

feasible region  $\mathcal{F}_T^-$  of problem  $(P_T^-)$  converges to the feasible region  $\mathcal{F}$  of the problem  $(P)$  in some suitable sense as  $T \rightarrow \infty$ .

We already know from Lemma 4.9 that  $\{\mathcal{F}_T^-\}_{T=1}^\infty$  is a non-decreasing sequence of nested subsets of  $\mathcal{F}$ . The following result shows that the additional condition under which (4.10) holds is the equality  $\text{cl}(\cup_{T=1}^\infty \mathcal{F}_T^-) = \mathcal{F}$ , where  $\text{cl}$  denotes topological closure. The result is stated below in the more general context of topological spaces, as it may be of independent interest.

**Proposition 4.10** *Let  $\mathcal{U}$  be a topological space. Let  $\{f_T\}_{T=1}^\infty$  be a sequence of real-valued functions on  $\mathcal{U}$  converging uniformly to a continuous function  $f : \mathcal{U} \rightarrow \mathbb{R}$ . Let  $\{\mathcal{V}_T\}_{T=1}^\infty$  be a non-decreasing sequence of nested subsets of  $\mathcal{U}$ , and let  $\mathcal{V} := \text{cl}(\cup_{T=1}^\infty \mathcal{V}_T)$ . Suppose  $f$  is bounded below on the set  $\mathcal{V}$ . Then,*

$$\lim_{T \rightarrow \infty} \left( \inf_{y \in \mathcal{V}_T} f_T(y) \right) = \inf_{y \in \mathcal{V}} f(y). \quad (4.11)$$

In light of the above proposition, the limit in (4.10) holds if

$$\mathcal{F} = \text{cl}(\cup_{T=1}^\infty \mathcal{F}_T^-) \quad (4.12)$$

holds. Recall that, the sets  $\{\mathcal{F}_T^-\}_{T=1}^\infty$  are intersections of sublevel sets of finite-horizon standard discounted and RS cost functions, while the set  $\mathcal{F}$  is the intersection of sublevel sets of finite- and infinite-horizon standard discounted and RS cost functions. Moreover, by lemmas 4.3 and 4.7, the functions defining the subsets  $\{\mathcal{F}_T^-\}_{T=1}^\infty$  converge to the functions defining the set  $\mathcal{F}$ . These properties lead one to expect (4.12) to hold. However, as a counterexample below shows, (4.12) may not hold in general. The following result provides the additional condition under which the desired equality above holds. The additional condition requires the maximum constraint violation function to not have 0 as a local minimum value, and is the source of the similar-looking condition appearing in Theorem 3.2. The result is again stated in the more general context of topological spaces, as it may be of independent interest when stated as such.

**Proposition 4.11** *Let  $\mathcal{U}$  be a topological space, and let  $N > 0$ . For each  $i \in \{1, \dots, N\}$ , let  $\{g_{T,i}\}_{T=1}^\infty$  be a sequence of real-valued functions on  $\mathcal{U}$  converging pointwise to a continuous function  $g_i : \mathcal{U} \rightarrow \mathbb{R}$  as  $T \rightarrow \infty$ , and let  $\{B_{T,i}\}_{T=1}^\infty$  be a sequence of real numbers converging to  $B_i$  as  $T \rightarrow \infty$ . For each  $T$  and  $i$ , denote  $\mathcal{G}_{T,i} = \{z \in \mathcal{U} : g_{T,i}(z) \leq B_{T,i}\}$ ,  $\mathcal{H}_T = \cap_{i=1}^N \mathcal{G}_{T,i}$ ,  $\mathcal{G}_i = \{z : g_i(z) \leq B_i\}$  and  $\mathcal{H} = \cap_{i=1}^N \mathcal{G}_i$ . Assume that  $\mathcal{G}_{T,i} \subseteq \mathcal{G}_i$  for every  $i$  and  $T$ . Define the map  $h : \mathcal{U} \rightarrow \mathbb{R}$  by  $h(z) = \max_{i \in \{1, \dots, N\}} \{g_i(z) - B_i\}$ . If 0 is not a local minimum value of  $h$ , then  $\text{cl}(\cup_{T=1}^\infty \mathcal{H}_T) = \mathcal{H}$ .*

While the condition that 0 should not be a local minimum value of the function  $h$  in Proposition 4.11 appears cumbersome to check, it is easy to construct an example where the condition is not satisfied, and the conclusion of the proposition fails to hold. Indeed, let  $\mathcal{U} = \mathbb{R}$ ,  $N = 2$ ,  $B_1 = B_2 = 0$  and  $g_1(x) = -g_2(x) = x$  for all  $x \in \mathbb{R}$ . For each  $T$ , let  $g_{T,1}(x) = g_1(x)$ ,  $g_{T,2}(x) = g_2(x)$ , and  $B_{T,1} = B_{T,2} = -T^{-1}$ . The function  $h$  of Proposition 4.11 is then given by  $h(x) = |x|$ , and has 0 as a local minimum value. It is easy to see that  $\mathcal{G}_{T,1} = (-\infty, -T^{-1}]$ ,  $\mathcal{G}_1 = (-\infty, 0]$ ,  $\mathcal{G}_{T,2} = [T^{-1}, \infty)$  and  $\mathcal{G}_2 = [0, \infty)$ . As a result,  $\mathcal{H} = \{0\}$ , while  $\mathcal{H}_T = \emptyset$  for all  $T$ , showing that the assertion of Proposition 4.11 fails to hold.

The obstruction posed by the failure of the local minimum condition goes beyond the purely topological setting of Proposition 4.11. Indeed, we present a more elaborate counter-example in D in a CRSMDP setting to show that the conclusion of Theorem 3.2 can fail to hold if the local minimum condition assumed therein is violated.

#### 4.2.1 Proof of Theorem 3.2

Problem (P) involves a total of  $\widetilde{M} := M + \widehat{M} + \bar{M} + \check{M}$  constraints. For each  $T \in \{1, 2, \dots, \}$  and  $i \in \{1, \dots, \widetilde{M}\}$ , define

$$g_{T,i}(\pi) := \begin{cases} v_{C_i, T}^\pi(x), & i = 1, 2, \dots, M, \\ J_{\gamma, \widehat{C}_j, T}^\pi(x), & i = M + j, j = 1, \dots, \widehat{M}, \\ v_{\widehat{C}_k, \bar{T}_k}^\pi(x), & i = M + \widehat{M} + k, k = 1, \dots, \bar{M}, \\ J_{\gamma, \check{C}_l, \bar{T}_l}^\pi(x), & i = M + \widehat{M} + \bar{M} + l, l = 1, \dots, \check{M}, \end{cases} \quad (4.13)$$

$$B_{T,i} := \begin{cases} b_i - K\beta^T, & i = 1, 2, \dots, M, \\ \widehat{b}_j / K_T, & i = M + j, j = 1, \dots, \widehat{M}, \\ \bar{b}_k, & i = M + \widehat{M} + k, k = 1, \dots, \bar{M}, \\ \check{b}_l, & i = M + \widehat{M} + \bar{M} + l, l = 1, \dots, \check{M}. \end{cases} \quad (4.14)$$

Lemmas 4.3 and 4.7 show that, for each  $i \in \{1, \dots, \widetilde{M}\}$ , the sequence of real-valued functions  $\{g_{T,i}\}_{T=1}^\infty$  converges uniformly on  $\Pi_{\text{MR}}$  to the function  $g_i : \Pi_{\text{MR}} \rightarrow \mathbb{R}$  defined by

$$g_i(\pi) := \begin{cases} v_{C_i}^\pi(x), & i = 1, 2, \dots, M, \\ J_{\gamma, \widehat{C}_j}^\pi(x), & i = M + j, j = 1, \dots, \widehat{M}, \\ v_{\widehat{C}_k, \bar{T}_k}^\pi(x), & i = M + \widehat{M} + k, k = 1, \dots, \bar{M}, \\ J_{\gamma, \check{C}_l, \bar{T}_l}^\pi(x), & i = M + \widehat{M} + \bar{M} + l, l = 1, \dots, \check{M}. \end{cases} \quad (4.15)$$

Also, for each  $i \in \{1, \dots, \widetilde{M}\}$  the sequence  $\{B_{T,i}\}_{T=1}^\infty$  converges to  $B_i$ , where

$$B_i := \begin{cases} b_i, & i = 1, 2, \dots, M, \\ \hat{b}_j, & i = M + j, j = 1, \dots, \hat{M}, \\ \bar{b}_k, & j = M + \hat{M} + k, k = 1, \dots, \bar{M}, \\ \check{b}_l, & i = M + \hat{M} + \bar{M} + l, l = 1, \dots, \check{M}. \end{cases} \quad (4.16)$$

Next, for each  $T \in \{1, 2, \dots\}$  and  $i \in \{1, \dots, \widetilde{M}\}$ , define the sets  $\mathcal{G}_{T,i} := \{\pi \in \Pi_{\text{MR}} : g_{T,i}(\pi) \leq B_{T,i}\}$  and  $\mathcal{G}_i := \{\pi \in \Pi_{\text{MR}} : g_i(\pi) \leq B_i\}$ . Note that  $\bigcap_{i=1}^{\widetilde{M}} \mathcal{G}_{T,i} = \mathcal{F}_T^-$  for each  $T \in \{1, 2, \dots\}$ , while  $\bigcap_{i=1}^{\widetilde{M}} \mathcal{G}_i = \mathcal{F}$ . Moreover, claims 1 and 2 established in the proof of Lemma 4.9 imply that  $\mathcal{G}_{T,i} \subseteq \mathcal{G}_i$  for all  $T \in \{1, 2, \dots\}$  and  $i \in \{1, \dots, \widetilde{M}\}$ . We also observe that the function  $h : \Pi_{\text{MR}} \rightarrow \mathbb{R}$  defined in (3.1) is equivalently given by  $h(\pi) = \max_{i \in \{1, 2, \dots, \widetilde{M}\}} \{g_i(\pi) - B_i\}$ . Proposition 4.11 now applies with  $\mathcal{U} := \Pi_{\text{MR}}$ ,  $N := \widetilde{M}$ ,  $\mathcal{H} := \mathcal{F}$  and  $\mathcal{H}_T := \mathcal{F}_T^-$  for each  $T \in \{1, 2, \dots\}$ , and allows us to conclude that (4.12) holds.

Next, recall from Lemma 4.7 that the sequence  $\{J_{\gamma,R,T}^\pi(x)\}_{T=1}^\infty$  converges uniformly to  $J_{\gamma,R}^\pi(x)$  on  $\Pi_{\text{MR}}$ . Moreover, the continuous function  $\pi \mapsto J_{\gamma,R}^\pi(x)$  is bounded below on the compact metric space  $\Pi_{\text{MR}}$ . Since (4.12) holds, Proposition 4.10 applies by letting  $\mathcal{U} = \Pi_{\text{MR}}$ ,  $f_T = J_{\gamma,R,T}^\pi(x)$  and  $\mathcal{V}_T = \mathcal{F}_T^-$  for each  $T$ ,  $f = J_{\gamma,R}^\pi(x)$  and  $\mathcal{V} = \mathcal{F}$ . We conclude from Proposition 4.10 that (4.10) holds. The rest of the proof is completed by using the arguments given in the paragraph following (4.10). This completes the proof of Theorem 3.2.  $\square$

### 4.3 Proof Outline of Theorem 3.4

The proof of Theorem 3.4 follows a pattern similar to that of Theorem 3.2. The main step in the proof is to show that the value of the problem  $(P_T^+)$  approximates the value of  $(P)$  for large  $T$ , that is,

$$\lim_{T \rightarrow \infty} \left( \inf_{\pi \in \mathcal{F}_T^+} J_{\gamma,R,T}^\pi(x) \right) = \inf_{\pi \in \mathcal{F}} J_{\gamma,R}^\pi(x). \quad (4.17)$$

Theorem 3.2 then follows by arguing as in the paragraph following (4.10).

Intuitively, one expects (4.17) to hold as Lemma 4.7 guarantees that  $J_{\gamma,R,T}^\pi(x)$  converges uniformly in  $\pi$  to  $J_{\gamma,R}^\pi(x)$  with increasing  $T$ , while statement (iii) of Lemma 4.9 asserts that the sequence of sets  $\{\mathcal{F}_T^+\}_{T=1}^\infty$  decrease to  $\mathcal{F}$ . The next result formalizes this intuition. The result is stated in the more general topological setting, as it might be of independent interest.

**Proposition 4.12** *Let  $\mathcal{U}$  be a topological space, and let  $g : \mathcal{U} \rightarrow \mathbb{R}$  be a continuous function. Let  $\{\mathcal{G}_T\}_{T=1}^\infty$  be a non-increasing sequence of nested compact subsets of  $\mathcal{U}$  converging to a subset  $\mathcal{G}$  of  $\mathcal{U}$  in the sense  $\bigcap_{T=1}^\infty \mathcal{G}_T = \mathcal{G}$ . Let  $\{g_T\}_{T=1}^\infty$  be a sequence of real-valued functions on  $\mathcal{U}$  converging to  $g$  uniformly on  $\mathcal{G}_1$ . Then,*

$$\lim_{T \rightarrow \infty} \left( \inf_{y \in \mathcal{G}_T} g_T(y) \right) = \inf_{y \in \mathcal{G}} g(y). \quad (4.18)$$

### 4.3.1 Proof of Theorem 3.4

To begin, choose  $\epsilon > 0$ . It is easy to see that, for each  $i$  and  $j$ , the constants  $b_i + K\beta^T$  and  $\hat{b}_j K_T$  defining the bounds appearing in the definitions (2.9) and (2.10) of  $\mathcal{C}_T^+$  and  $\hat{\mathcal{C}}_T^+$  converge to  $b_i$  and  $\hat{b}_j$ , respectively, from above as  $T \rightarrow \infty$ . It therefore follows from lemmas 4.3 and 4.7 that there exists  $T_1$  such that, for every  $T > T_1$ , every policy in the feasible set  $\mathcal{F}_T^+$  of problem  $(P_T^+)$  is  $\epsilon$ -feasible for the problem  $(P)$ .

Next, we apply Proposition 4.12 by taking  $\mathcal{U}$ ,  $g$ ,  $\{\mathcal{G}_T\}_{T=1}^\infty$ ,  $\mathcal{G}$  and  $\{g_T\}_{T=1}^\infty$  in that proposition to be  $\Pi_{\text{MR}}$ ,  $J_{\gamma,R}^\pi(x)$ ,  $\{\mathcal{F}_T^+\}_{T=1}^\infty$ ,  $\mathcal{F}$ , and  $\{J_{\gamma,R,T}^\pi(x)\}_{T=1}^\infty$ , respectively. Lemma 4.9, Lemma 4.7 and Theorem 4.8 imply that all the hypotheses of Proposition 4.12 are satisfied, and hence (4.17) follows from (4.18).

Using arguments similar to those laid out in the paragraph following (4.10), we can conclude from (4.17) that there exists  $T^* \geq T_1$  such that, for every  $T > T^*$ , there exists a solution  $\eta \in \mathcal{F}_T^+$  of  $(P_T^+)$  such that (3.3) holds.  $\square$

## 5 Solution of Finite-Horizon CRSMDPs

Recall that, in light of Remark 3.5, theorems 3.2 and 3.4 provide a means of constructing US policies as approximate solutions to the infinite-horizon CRSMDP  $(P)$  from solutions of the finite-horizon CRSMDPs  $(P_T^-)$  or  $(P_T^+)$ , respectively. In this section, we provide a LP-based approach for computing the solution to a finite-horizon CRSMDP such as  $(P_T^-)$  or  $(P_T^+)$ . The approach involves first expressing finite-horizon RS cost functions as standard discounted cost functions of a terminal-cost MDP defined on an augmented state space.

To illustrate the approach, choose  $T > 0$ , and consider the RS objective cost function  $J_{\gamma,R,T}^\pi(x)$  defined in (2.1). For each  $t \in \{0, 1, \dots, T\}$ , define the random variable  $\Psi_t^\circ := \exp(\gamma \sum_{j=0}^{t-1} \beta^j R(X_j, A_j))$ . Note that, for each  $t$ , the random variable  $\Psi_t^\circ$  takes at most  $(mn)^t$  values. Consequently, the

augmented process  $\{Z_t\}_{t=0}^T$  defined by  $Z_t := (X_t, \Psi_t^o)$ ,  $t \in \{0, \dots, T-1\}$ , takes at most  $m(mn)^t$  values at each time  $t$ , and is the state process of an augmented time-dependent MDP. The probability that the augmented MDP transitions from state  $z = (s, \psi)$  at time  $t$  to  $z' = (s', \psi')$  at time  $t+1$  is given by  $p_{t+1}^A(z'|z, a) = 1_{\{\psi' = \psi \exp(\gamma\beta^t R(s,a))\}} p(s'|s, a)$ , where the  $1_S$  denotes the indicator function of the set  $S$ , and the transition probabilities  $p(\cdot|\cdot, \cdot)$  of the original MDP are as defined in section 2. Next, note that  $J_{\gamma, R, T}^\pi(x) = \mathbb{E}_x^\pi[\Psi_T^o]$ . The RS cost function  $J_{\gamma, R, T}^\pi(x)$  can therefore be viewed as the finite-horizon standard discounted cost of the augmented MDP with immediate and terminal cost functions given by  $c_t((s, \psi), a) := 0$  for all  $t = 0, 1, \dots, T-1$ , and  $c_T((s, \psi)) := \psi$ . The same idea may be extended to the RS cost functions appearing in the constraints in problems  $(P_T^-)$  and  $(P_T^+)$  to obtain a constrained terminal-cost MDP problem with standard discounted cost functions for the objective and constraint functions. Standard LP-based techniques for constrained standard discounted cost MDPs described, for instance in [8], may then be applied. For the sake of completeness, we explicitly provide the resulting LP formulation below.

Recall that each of the problems  $(P_T^-)$  and  $(P_T^+)$  involve  $\hat{M} + \check{M}$  finite-horizon RS constraints and  $M + \bar{M}$  finite-horizon standard discounted cost constraints. Of these,  $\check{M}$  RS constraints are finite-horizon RS constraints with horizons  $\check{T}_1, \dots, \check{T}_{\check{M}}$ ,  $\bar{M}$  constraints are finite-horizon standard discounted cost constraints with horizons  $\bar{T}_1, \dots, \bar{T}_{\bar{M}}$ , and the rest arise by truncating infinite-horizon constraints of the RS or standard discounted type. For the sake of simplicity, we consider only the problem  $(P_T^-)$ , and restrict ourselves to the case where the truncation horizon  $T$  is greater than the horizons of all the original finite-horizon constraints, that is,  $T > \max_{1 \leq l \leq \check{M}} \check{T}_l$  and  $T > \max_{1 \leq k \leq \bar{M}} \bar{T}_k$ . Define random processes  $\{\Psi_t^1\}_{t=0}^T, \dots, \{\Psi_t^{\hat{M}}\}_{t=0}^T$  corresponding to the truncated RS constraints in a manner analogous to the construction of the process  $\{\Psi_t^o\}_{t=0}^T$  described above. Additionally, define random processes  $\{\Psi_t^{\hat{M}+1}\}_{t=0}^T, \dots, \{\Psi_t^{\hat{M}+\check{M}}\}_{t=0}^T$  corresponding to the finite-horizon RS constraints by

$$\Psi_t^{\hat{M}+l} := \exp \left[ \gamma \sum_{j=0}^{(t \wedge \check{T}_l)-1} \beta^j \check{C}_l(X_j, A_j) \right], \quad l \in \{1, \dots, \check{M}\},$$

where  $a \wedge b := \min\{a, b\}$ . Finally, define the augmented state process  $\{Z_t\}_{t=0}^T$  by letting  $Z_t := (X_t, \Psi_t^o, \Psi_t^1, \dots, \Psi_t^{\hat{M}+\check{M}})$  for each  $t \in \{0, \dots, T\}$ . For each  $t$ , denote by  $\mathcal{S}_t^A$  the set of all possible values that the random tuple  $Z_t$  can take, and note that  $\mathcal{S}_t^A$  is finite. Also, given  $t \in \{0, \dots, T\}$ ,  $i \in \{1, \dots, \hat{M} + \check{M}\}$  and  $z \in \mathcal{S}_t^A$ , we abuse notation slightly to denote the second and  $(i+2)$ th

components of the tuple  $z$  by  $\Psi_t^o(z)$  and  $\Psi_t^i(z)$ , respectively.

Next, in order to state the LP formulation, we introduce a variable  $y(t, z, a)$  for each  $t \in \{0, \dots, T-1\}$ ,  $z \in \mathcal{S}_t^A$  and  $a \in \mathcal{A}$ , and denote the tuple of all such variables by  $\mathbf{y} = \{y(t, z_t, a) : t \leq T-1, z_t \in \mathcal{S}_t^A, a \in \mathcal{A}\}$ . It can be observed that the variable  $y(t, z, a)$  is the occupation measure which is interpreted as the probability of being in state  $z$  and taking action  $a$  at time  $t$  ([8, 1]). The required LP formulation is then given by

$$\begin{aligned} \min_{\mathbf{y}} \quad & \sum_{\substack{a \in \mathcal{A}, z \in \mathcal{S}_{T-1}^A \\ z' \in \mathcal{S}_T^A}} \Psi_T^o(z') y(T-1, z, a) p_T^A(z'|z, a) \text{ subject to} \\ & \sum_{a \in \mathcal{A}} y(0, z, a) = 1_{\{z=(x,1,\dots,1) \in \mathcal{S}_0^A\}}, \\ \sum_{a \in \mathcal{A}} y(t, z', a) = \sum_{a \in \mathcal{A}} \sum_{z \in \mathcal{S}_{t-1}^A} & p_t^A(z'|z, a) y(t-1, z, a) \text{ for all } 1 \leq t \leq T-1, z' \in \mathcal{S}_t^A, \\ \sum_{a \in \mathcal{A}, z \in \mathcal{S}_{T-1}^A, z' \in \mathcal{S}_T^A} & \Psi_T^i(z') y(T-1, z, a) p_T^A(z'|z, a) \leq b_i^R \text{ for all } i \in \{1, \dots, \check{M} + \hat{M}\}, \\ \sum_{t \leq T-1, z \in \mathcal{S}_t^A, a \in \mathcal{A}} & \beta^t c^i(t, z, a) y(t, z, a) \leq b_i^L \text{ for all } i \in \{1, \dots, M + \bar{M}\}, \end{aligned}$$

where  $x$  is the initial condition of the finite-horizon CRSMDP,

$$b_i^L = \begin{cases} b_i - K\beta^T & \text{for } i = 1, \dots, M, \\ \bar{b}_k & \text{for } i = M + k \text{ where } k = 1, \dots, \bar{M}, \end{cases} \quad (5.1)$$

$$b_i^R = \begin{cases} \frac{\hat{b}_i}{K_T} & \text{for } i = 1, \dots, \hat{M}, \\ \check{b}_k & \text{for } i = \hat{M} + k \text{ where } k = 1, \dots, \check{M}, \text{ and} \end{cases} \quad (5.2)$$

$$c^i(t, z, a) = \begin{cases} C_i(z[1], a) & \text{for } 1 \leq i \leq M \text{ and for all } 0 \leq t \leq T-1, \\ \bar{C}_k(z[1], a) & \text{for } i = M + k \text{ where } 1 \leq k \leq \bar{M}, \text{ and } t \leq T_k - 1, \\ 0 & \text{for } i = M + k \text{ where } 1 \leq k \leq \bar{M}, \text{ and } T_k \leq t \leq T-1, \end{cases} \quad (5.3)$$

with  $z[1] \in \mathcal{S}$  denoting the first component of vector  $z$ .

The optimal policy  $\pi^*$  for the CRSMDP may be constructed from the optimal solution  $\mathbf{y}^*$  of the LP problem described above. For this purpose, let  $\mathcal{S}_t^A(x)$  denote the set of tuples in  $\mathcal{S}_t^A$  having  $x$  as their first component, for each  $x \in \mathcal{X}$  and  $t \in \{0, \dots, T-1\}$ . Then the probability of taking action

$a \in \mathcal{A}$  in state  $s \in \mathcal{S}$  at time  $t$  under the time- $t$  decision rule  $d_t^*$  of the optimal policy  $\pi^*$  is given by [22]

$$d_t^*(a|s) = \frac{y^*(t, s, a)}{\sum_{a' \in \mathcal{A}} y^*(t, s, a')}, \quad \text{where } y^*(t, s, a) = \sum_{z \in \mathcal{S}_t^\Delta(s)} y^*(t, z, a). \quad (5.4)$$

It is possible to obtain a LP formulation for a finite-horizon CRSMDP without first converting it to the standard discounted setting [22]. This alternative LP formulation from [22] may also be applied to the finite-horizon CRSMDPs ( $P_T^-$ ) or ( $P_T^+$ ).

## 6 Conclusions

We have shown that a CRSMDP involving a fairly general set of constraints in the form of upper bounds on finite- and infinite-horizon RS and standard discounted costs possesses a solution as long as it is feasible. Moreover, near-optimal and near-feasible US policies for the CRSMDP may be found by solving two approximating finite-horizon CRSMDPs obtained by time truncating the original objective cost and constraint functions and suitably perturbing the bounding constants in the constraints. The approximating finite-horizon CRSMDPs may be solved by using a LP-based approach.

## A Proofs for Results from Subsection 4.1

*Proof of Theorem 4.1.* For each  $t$ , let  $\mathcal{Y}_t$  denote the set  $\mathcal{R}$  equipped with the topology induced by the metric  $\mu_t$  defined as in sub-section 4.1.2. The topology on  $\mathcal{Y}_t$  agrees with the subspace topology on  $\mathcal{R} \subset \mathbb{R}^{m \times n}$ . Since  $\mathcal{R}$  is a compact subset of  $\mathbb{R}^{m \times n}$ , it follows that  $\mathcal{Y}_t$  is a compact topological space for each  $t$ . Tychonoff's theorem (refer [28, p. 245], [9, p. 224] implies that the Cartesian product  $\prod_{t=0}^{\infty} \mathcal{Y}_t$  is compact under the product topology. A direct application of Theorem 7.2 from [9, p. 190] now shows that  $\mu$  is a metric which metrizes the product topology on  $\prod_{t=0}^{\infty} \mathcal{Y}_t$ . The last two assertions show that  $(\prod_{t=0}^{\infty} \mathcal{Y}_t, \mu)$  is a compact metric space. The result now follows by noting that, as a set, the Cartesian product above equals  $\mathcal{R}^\infty$ , which is identifiable with  $\Pi_{\text{MR}}$ .  $\square$

The proofs of lemmas 4.2 and 4.6, and Theorem 4.8 below make use of the following inequalities for  $A, B \in \mathbb{R}^{m \times n}$  and  $y, z \in \mathbb{R}^n$ .

$$\|A \odot B\|_\infty \leq \|A\|_{\max} \|B\|_\infty \leq \|A\|_\infty \|B\|_\infty, \quad (\text{A.1})$$

$$\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty, \quad \|Ay\|_\infty \leq \|A\|_\infty \|y\|_\infty, \quad (\text{A.2})$$

$$|z'y| \leq \|z\|_1 \|y\|_\infty. \quad (\text{A.3})$$



We also note that, if  $A, B \in \mathbb{R}^{m \times m}$  are nonnegative and row-stochastic, then  $\|A\|_{\max} \leq 1 = \|A\|_{\infty}$ , and  $AB$  is nonnegative and row-stochastic.

*Proof of Lemma 4.2.* First, we claim that the inequalities

$$\|R_d - R_f\|_{\infty} \leq C\|d - f\|_{\infty} \quad \text{and} \quad \|P_d - P_f\|_{\infty} \leq \|d - f\|_{\infty} \quad (\text{A.4})$$

hold for every  $d, f \in \mathcal{R}$ . Using  $R_d = (R \odot d)\mathbb{1}_{n \times 1}$ , (A.2), (A.1) and  $\|\mathbb{1}_{n \times 1}\|_{\infty} = 1$  in that order gives  $\|R_d - R_f\|_{\infty} = \|(R \odot d - R \odot f)\mathbb{1}_{n \times 1}\|_{\infty} \leq \|R \odot (d - f)\|_{\infty} \|\mathbb{1}_{n \times 1}\|_{\infty} \leq \|R\|_{\max} \|d - f\|_{\infty}$ . Next, using  $\|R\|_{\max} \leq C$  yields the first equation of (A.4).

The sum of the absolute values of the entries of the  $i$ th row of the matrix  $P_d - P_f$  satisfies

$$\begin{aligned} \sum_{j=1}^m |(P_d)_{i,j} - (P_f)_{i,j}| &= \sum_{j=1}^m \left| \sum_{k=1}^n p(s_j | s_i, a_k) \{d(a_k | s_i) - f(a_k | s_i)\} \right| \\ &\leq \sum_{j=1}^m \sum_{k=1}^n |p(s_j | s_i, a_k)| \cdot |d(a_k | s_i) - f(a_k | s_i)| \\ &= \sum_{k=1}^n \left( \sum_{j=1}^m p(s_j | s_i, a_k) \right) |d(a_k | s_i) - f(a_k | s_i)| \\ &= \sum_{k=1}^n |d(a_k | s_i) - f(a_k | s_i)| \leq \|d - f\|_{\infty}. \end{aligned} \quad (\text{A.5})$$

Thus every absolute row sum of the  $m \times m$  matrix  $P_d - P_f$  is bounded above by  $\|d - f\|_{\infty}$ . It now follows that the maximum absolute row sum norm  $\|P_d - P_f\|_{\infty}$  satisfies the second inequality in (A.4).

Let  $\pi_1, \pi_2 \in \Pi_{\text{MR}}$  be as in the statement of the lemma. We claim that

$$\|P_t^{\pi_1} - P_t^{\pi_2}\|_{\infty} \leq \sum_{i=0}^{t-1} \|d_i - f_i\|_{\infty}, \quad (\text{A.6})$$

for every  $t > 0$ . To prove our claim, we first note that, for every  $t > 0$ , the time- $t$  transition probability matrix  $P_{d_t}$  and the  $t$ -step transition probability matrix  $P_{t-1}^{\pi_1}$  are both nonnegative and row-stochastic. Next, for  $t > 0$ , we have

$$\begin{aligned} \|P_t^{\pi_1} - P_t^{\pi_2}\|_{\infty} &= \|P_{t-1}^{\pi_1} P_{d_{t-1}} - P_{t-1}^{\pi_2} P_{f_{t-1}}\|_{\infty} \\ &\leq \|P_{t-1}^{\pi_1} (P_{d_{t-1}} - P_{f_{t-1}})\|_{\infty} + \|(P_{t-1}^{\pi_1} - P_{t-1}^{\pi_2}) P_{f_{t-1}}\|_{\infty} \\ &\leq \|P_{t-1}^{\pi_1}\|_{\infty} \|P_{d_{t-1}} - P_{f_{t-1}}\|_{\infty} + \|P_{t-1}^{\pi_1} - P_{t-1}^{\pi_2}\|_{\infty} \|P_{f_{t-1}}\|_{\infty} \\ &\leq \|d_{t-1} - f_{t-1}\|_{\infty} + \|P_{t-1}^{\pi_1} - P_{t-1}^{\pi_2}\|_{\infty}, \end{aligned} \quad (\text{A.7})$$

where we have used the first inequality in (A.2), the second inequality in (A.4), and the fact that  $\|P_{t-1}^{\pi_1}\|_\infty = \|P_{f_{t-1}}\|_\infty = 1$ . Solving the recursion (A.7) and noting that  $\|P_0^{\pi_1} - P_0^{\pi_2}\|_\infty = 0$  shows that (A.6) holds.

Next, for  $t \geq 0$ , we have

$$\begin{aligned} \|P_t^{\pi_1} R_{d_t} - P_t^{\pi_2} R_{f_t}\|_\infty &\leq \|P_t^{\pi_1}(R_{d_t} - R_{f_t})\|_\infty + \|(P_t^{\pi_1} - P_t^{\pi_2})R_{f_t}\|_\infty \\ &\leq \|P_t^{\pi_1}\|_\infty \|R_{d_t} - R_{f_t}\|_\infty + \|P_t^{\pi_1} - P_t^{\pi_2}\|_\infty \|R_{f_t}\|_\infty \\ &\leq C \sum_{i=0}^t \|d_i - f_i\|_\infty, \end{aligned} \quad (\text{A.8})$$

where we have used the second inequality in (A.2) along with the first inequality in (A.4),  $\|P_t^{\pi_1}\|_\infty = 1$  and  $\|R_{f_t}\|_\infty \leq C$ . This proves the lemma.  $\square$

*Proof of Lemma 4.3.* Let  $\pi \in \Pi_{\text{MR}}$ ,  $T > 0$  and  $x \in \mathcal{S}$  be arbitrary. We have  $|v_R^\pi(x) - v_{R,T}^\pi(x)| = |\mathbb{E}_x^\pi [\sum_{t=T}^\infty \beta^t R(X_t, A_t)]| \leq \mathbb{E}_x^\pi [\sum_{t=T}^\infty \beta^t |R(X_t, A_t)|] \leq \beta^T K$ . Since the bound in the previous inequality is independent of the policy  $\pi$ , the result follows by letting  $T \rightarrow \infty$ .  $\square$

*Proof of Theorem 4.4.* Let  $\pi_1 = \{d_t\}_{t=0}^\infty$  and  $\pi_2 = \{f_t\}_{t=0}^\infty$  be two policies in  $\Pi_{\text{MR}}$ . Using the vector expression in (4.3) for the finite-horizon standard discounted cost and applying Lemma 4.2 gives

$$\begin{aligned} \|v_{R,T}^{\pi_1} - v_{R,T}^{\pi_2}\|_\infty &= \left\| \sum_{t=0}^{T-1} \beta^t P_t^{\pi_1} R_{d_t} - \sum_{t=0}^{T-1} \beta^t P_t^{\pi_2} R_{f_t} \right\|_\infty \\ &= \left\| \sum_{t=0}^{T-1} \beta^t (P_t^{\pi_1} R_{d_t} - P_t^{\pi_2} R_{f_t}) \right\|_\infty \\ &\leq \sum_{t=0}^{T-1} \beta^t \|P_t^{\pi_1} R_{d_t} - P_t^{\pi_2} R_{f_t}\|_\infty \leq \sum_{t=0}^{T-1} \beta^t C \sum_{i=0}^t \|d_i - f_i\|_\infty \\ &= C \sum_{i=0}^{T-1} \sum_{t=i}^{T-1} \beta^t \|d_i - f_i\|_\infty = C \sum_{i=0}^{T-1} \frac{\beta^i (1 - \beta^{T-i})}{1 - \beta} \|d_i - f_i\|_\infty \\ &\leq K \sum_{i=0}^{T-1} \beta^i \|d_i - f_i\|_\infty = K \sum_{i=0}^{T-1} \left(\frac{\beta}{\delta}\right)^i \delta^i \|d_i - f_i\|_\infty \\ &\leq K \sum_{i=0}^{T-1} \left(\frac{\beta}{\delta}\right)^i \mu(\pi_1, \pi_2) = \frac{K(\delta^T - \beta^T)}{\delta^{T-1}(\delta - \beta)} \mu(\pi_1, \pi_2). \end{aligned} \quad (\text{A.9})$$

Thus, the vector valued map  $\pi \rightarrow v_{R,T}^\pi$  on  $\Pi_{\text{MR}}$  is Lipschitz continuous with Lipschitz constant  $\frac{K(\delta^T - \beta^T)}{\delta^{T-1}(\delta - \beta)}$ .

Next, choose  $\epsilon > 0$ . We know from Lemma 4.3 that there exists  $T^* > 0$  such that  $|v_{R,T}^{\pi_1}(x) - v_{R,T}^{\pi_2}(x)| < \epsilon/2$  for all  $T \geq T^*$ ,  $\pi \in \Pi_{MR}$  and  $x \in \mathcal{S}$ . Choose  $T \geq T^*$ . We then have

$$\begin{aligned} \|v_R^{\pi_1} - v_R^{\pi_2}\|_\infty &\leq \|v_R^{\pi_1} - v_{R,T}^{\pi_1}\|_\infty + \|v_{R,T}^{\pi_1} - v_{R,T}^{\pi_2}\|_\infty + \|v_{R,T}^{\pi_2} - v_R^{\pi_2}\|_\infty \\ &\leq \frac{\epsilon}{2} + \frac{K(\delta^T - \beta^T)}{\delta^{T-1}(\delta - \beta)}\mu(\pi_1, \pi_2) + \frac{\epsilon}{2} \leq \epsilon + \frac{K\delta}{(\delta - \beta)}\mu(\pi_1, \pi_2). \end{aligned}$$

Letting  $\epsilon \rightarrow 0$  now shows that the map  $\pi \rightarrow v_R^\pi$  is Lipschitz continuous with Lipschitz constant  $\frac{K\delta}{(\delta - \beta)}$ .  $\square$

*Proof of Proposition 4.5:* Note that (4.5) directly follows by letting  $t = T - 1$  in (4.4). Next, choose  $t \in \{1, \dots, T - 1\}$  and  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . Applying (4.4), we have

$$\begin{aligned} Q_{t-1,T}^\pi(s, a) &= \mathbb{E}^\pi \left[ e^{\gamma \sum_{\tau=t-1}^{T-1} \beta^\tau R(X_\tau, A_\tau)} \middle| X_{t-1} = s, A_{t-1} = a \right] \\ &= e^{\gamma \beta^{t-1} R(s, a)} \mathbb{E}^\pi \left[ \mathbb{E}^\pi \left[ e^{\gamma \sum_{\tau=t}^{T-1} \beta^\tau R(X_\tau, A_\tau)} \middle| X_t, A_t \right] \middle| X_{t-1} = s, A_{t-1} = a \right] \\ &= e^{\gamma \beta^{t-1} R(s, a)} \mathbb{E}^\pi [Q_{t,T}^\pi(X_t, A_t) | X_{t-1} = s, A_{t-1} = a] \\ &= e^{\gamma \beta^{t-1} R(s, a)} \sum_{i=1}^m \sum_{j=1}^n p(s_i | s, a) d_t(a_j | s_i) Q_{t,T}^\pi(s_i, a_j). \end{aligned} \tag{A.10}$$

The double summation above equals  $p(\cdot | s, a)'(d_t \odot Q_{t,T}^\pi) \mathbb{1}_{n \times 1}$ , and (4.6) follows.

To prove (4.7), choose  $i \in \{1, \dots, m\}$ , and note that

$$\begin{aligned} J_{\gamma, R, T}^\pi(s_i) &= \mathbb{E}^\pi \left[ \mathbb{E}^\pi \left[ e^{\gamma \sum_{\tau=0}^{T-1} \beta^\tau R(X_\tau, A_\tau)} \middle| X_0, A_0 \right] \middle| X_0 = s_i \right] \\ &= \mathbb{E}^\pi \left[ Q_{0,T}^\pi(X_0, A_0) \middle| X_0 = s_i \right] = \sum_{j=1}^n d_0(a_j | s_i) Q_{0,T}^\pi(s_i, a_j). \end{aligned}$$

The summation above is exactly the  $i$ -th element of the vector  $(d_0 \odot Q_{0,T}^\pi) \mathbb{1}_{n \times 1}$ . This completes the proof.  $\square$

*Proof of Lemma 4.6.* To begin, note that (4.8) follows directly from (4.5) in the case  $t = T - 1$ . Next, choose  $(s, a) \in \mathcal{S} \times \mathcal{A}$  and  $1 \leq t \leq T - 1$ . From

(4.6), we have

$$\begin{aligned} |Q_{t-1,T}^{\pi_1}(s, a)| &= e^{\gamma\beta^{t-1}R(s,a)} |p(\cdot|s, a)'(d_t \odot Q_{t,T}^{\pi_1}) \mathbb{1}_{n \times 1}| \\ &\leq e^{|\gamma|\beta^{t-1}C} \|p(\cdot|s, a)\|_1 \cdot \|(d_t \odot Q_{t-1,T}^{\pi_1}) \mathbb{1}_{n \times 1}\|_\infty \end{aligned} \quad (\text{A.11})$$

$$\leq e^{|\gamma|\beta^{t-1}C} \|d_t \odot Q_{t,T}^{\pi_1}\|_\infty \cdot \|\mathbb{1}_{n \times 1}\|_\infty \quad (\text{A.12})$$

$$\leq e^{|\gamma|\beta^{t-1}C} \|d_t\|_\infty \cdot \|Q_{t,T}^{\pi_1}\|_{\max} \leq e^{|\gamma|\beta^{t-1}C} \|Q_{t,T}^{\pi_1}\|_{\max} \quad (\text{A.13})$$

Inequalities (A.11), (A.12) and (A.13) follow by using (A.3), (A.2) and (A.1), respectively, along with  $|R(s, a)| \leq C$  and  $\|p(\cdot|s, a)\|_1 = \|\mathbb{1}_{n \times 1}\|_\infty = \|d_t\|_\infty = 1$ . Since the bound in (A.13) is independent of  $s$  and  $a$ , it follows that  $\|Q_{t-1,T}^{\pi_1}\|_{\max} \leq e^{|\gamma|\beta^{t-1}C} \|Q_{t,T}^{\pi_1}\|_{\max}$  for all  $1 \leq t \leq T-1$ . Solving this recursion and noting from (4.5) that  $\|Q_{T-1,T}^{\pi_1}\|_{\max} \leq e^{|\gamma|\beta^{T-1}C}$  gives  $\|Q_{t,T}^{\pi_1}\|_{\max} \leq \exp\left(|\gamma|C\beta^t \sum_{i=0}^{T-t-1} \beta^i\right) = \exp[|\gamma|K(\beta^t - \beta^T)]$  for every  $t$  satisfying  $0 \leq t < T-1$ . This proves (4.8).

To prove (4.9), we first derive an upper bound on the term  $\|d_t \odot Q_{t,T}^{\pi_1} - f_t \odot Q_{t,T}^{\pi_2}\|_\infty$  for a given  $t$  satisfying  $0 \leq t \leq T-1$ . We have

$$\begin{aligned} \|d_t \odot Q_{t,T}^{\pi_1} - f_t \odot Q_{t,T}^{\pi_2}\|_\infty &\leq \|d_t \odot (Q_{t,T}^{\pi_1} - Q_{t,T}^{\pi_2})\|_\infty + \|Q_{t,T}^{\pi_2} \odot (d_t - f_t)\|_\infty \\ &\leq \|d_t\|_\infty \cdot \|Q_{t,T}^{\pi_1} - Q_{t,T}^{\pi_2}\|_{\max} + \|Q_{t,T}^{\pi_2}\|_{\max} \cdot \|d_t - f_t\|_\infty \end{aligned} \quad (\text{A.14})$$

$$\leq \|Q_{t,T}^{\pi_1} - Q_{t,T}^{\pi_2}\|_{\max} + \|Q_{t,T}^{\pi_2}\|_{\max} \cdot \|d_t - f_t\|_\infty, \quad (\text{A.15})$$

where the inequality (A.14) follows from (A.1).

Next, let  $(s, a) \in \mathcal{S} \times \mathcal{A}$  and  $1 \leq t \leq T-1$ . From (4.6), we have

$$\begin{aligned} &|Q_{t-1,T}^{\pi_1}(s, a) - Q_{t-1,T}^{\pi_2}(s, a)| \\ &= |e^{\gamma\beta^{t-1}R(s,a)} p(\cdot|s, a)'(d_t \odot Q_{t,T}^{\pi_1} - f_t \odot Q_{t,T}^{\pi_2}) \mathbb{1}_{n \times 1}| \\ &\leq e^{|\gamma|\beta^{t-1}C} \|p(\cdot|s, a)\|_1 \cdot \|(d_t \odot Q_{t,T}^{\pi_1} - f_t \odot Q_{t,T}^{\pi_2}) \mathbb{1}_{n \times 1}\|_\infty \end{aligned} \quad (\text{A.16})$$

$$\leq e^{|\gamma|\beta^{t-1}C} \|d_t \odot Q_{t,T}^{\pi_1} - f_t \odot Q_{t,T}^{\pi_2}\|_\infty \cdot \|\mathbb{1}_{n \times 1}\|_\infty \quad (\text{A.17})$$

$$= e^{|\gamma|\beta^{t-1}C} \|d_t \odot Q_{t,T}^{\pi_1} - f_t \odot Q_{t,T}^{\pi_2}\|_\infty \quad (\text{A.18})$$

$$\leq e^{|\gamma|\beta^{t-1}C} (\|Q_{t,T}^{\pi_1} - Q_{t,T}^{\pi_2}\|_{\max} + \|Q_{t,T}^{\pi_2}\|_{\max} \cdot \|d_t - f_t\|_\infty). \quad (\text{A.19})$$

The inequalities (A.16), (A.17) and (A.18) follow by applying (A.3), (A.2) and (A.1), respectively, along with  $|R(s, a)| \leq C$  and  $\|p(\cdot|s, a)\|_1 = \|\mathbb{1}_{n \times 1}\|_\infty = 1$ . The last inequality (A.19) follows from (A.15).

Note that the bound in (A.19) is independent of  $s$  and  $a$ . Also, we had chosen  $1 \leq t \leq T-1$  arbitrarily. Therefore, it follows that

$$\|Q_{t-1,T}^{\pi_1} - Q_{t-1,T}^{\pi_2}\|_{\max} \leq e^{|\gamma|\beta^{t-1}C} (\|Q_{t,T}^{\pi_1} - Q_{t,T}^{\pi_2}\|_{\max} + \|Q_{t,T}^{\pi_2}\|_{\max} \cdot \|d_t - f_t\|_\infty), \quad (\text{A.20})$$

for all  $1 \leq t \leq T - 1$ .

On recalling from (4.5) that  $Q_{T-1,T}^{\pi_1} = Q_{T-1,T}^{\pi_2}$ , we see that (4.9) holds for  $t = T - 1$ . To complete the proof by induction, suppose (4.9) holds for  $t = k$ , where  $1 \leq k \leq T - 1$ . Letting  $t = k$  in (A.20) and using the induction hypothesis along with (4.8) gives

$$\begin{aligned}
\|Q_{k-1,T}^{\pi_1} - Q_{k-1,T}^{\pi_2}\|_{\max} &\leq e^{|\gamma|\beta^{k-1}C} (\|Q_{k,T}^{\pi_1} - Q_{k,T}^{\pi_2}\|_{\max} + \|Q_{k,T}^{\pi_2}\|_{\max} \cdot \|d_k - f_k\|_{\infty}) \\
&\leq e^{|\gamma|\beta^{k-1}C} e^{|\gamma|K(\beta^k - \beta^T)} \left( \sum_{i=k+1}^{T-1} \|d_i - f_i\|_{\infty} + \|d_k - f_k\|_{\infty} \right) \\
&= e^{|\gamma|K(\beta^{k-1} - \beta^T)} \sum_{i=k}^{T-1} \|d_i - f_i\|_{\infty}. \tag{A.21}
\end{aligned}$$

In other words, (4.9) holds for  $t = k - 1$ . It now follows by induction that (4.9) holds for all  $t$  satisfying  $0 \leq t \leq T - 1$ .  $\square$

*Proof of Theorem 4.8.* Let  $\pi_1 = \{d_t\}_{t=0}^{\infty}$  and  $\pi_2 = \{f_t\}_{t=0}^{\infty}$  be policies in  $\Pi_{\text{MR}}$ . Recall from the proof of Lemma 4.6 that (A.15) holds for every  $t$  satisfying  $0 \leq t \leq T - 1$ . Letting  $t = 0$  in (A.15) and using (4.8) and (4.9) for  $t = 0$  gives

$$\begin{aligned}
\|d_0 \odot Q_{0,T}^{\pi_1} - f_0 \odot Q_{0,T}^{\pi_2}\|_{\infty} &\leq \|Q_{0,T}^{\pi_1} - Q_{0,T}^{\pi_2}\|_{\max} + \|Q_{0,T}^{\pi_2}\|_{\max} \cdot \|d_0 - f_0\|_{\infty} \\
&\leq e^{|\gamma|K(1 - \beta^T)} \sum_{t=0}^{T-1} \|d_t - f_t\|_{\infty}. \tag{A.22}
\end{aligned}$$

Next, starting from (4.7) and using (A.2), (A.1) and (A.22) along with  $\|\mathbb{1}_{n \times 1}\|_{\infty} = 1$  and  $\|R\|_{\max} \leq C$ , we have

$$\begin{aligned}
\|J_{\gamma,R,T}^{\pi_1} - J_{\gamma,R,T}^{\pi_2}\|_{\infty} &= \|(d_0 \odot Q_{0,T}^{\pi_1})\mathbb{1}_{n \times 1} - (f_0 \odot Q_{0,T}^{\pi_2})\mathbb{1}_{n \times 1}\|_{\infty} \\
&\leq \|d_0 \odot Q_{0,T}^{\pi_1} - f_0 \odot Q_{0,T}^{\pi_2}\|_{\infty} \|\mathbb{1}_{n \times 1}\|_{\infty} \\
&\leq e^{|\gamma|K(1 - \beta^T)} \sum_{t=0}^{T-1} \|d_t - f_t\|_{\infty} = e^{|\gamma|K(1 - \beta^T)} \sum_{t=0}^{T-1} \frac{\delta^t \|d_t - f_t\|_{\infty}}{\delta^t} \\
&\leq e^{|\gamma|K(1 - \beta^T)} \mu(\pi_1, \pi_2) \sum_{t=0}^{T-1} \delta^{-t} \\
&\leq e^{|\gamma|K(1 - \beta^T)} (1 - \delta)^{-1} \cdot \delta^{-(T-1)} \mu(\pi_1, \pi_2). \tag{A.23}
\end{aligned}$$

This proves the first assertion of the theorem.

To prove the second assertion, recall from Lemma 4.7 that  $J_{\gamma,R,T}^{\pi}$  converges to  $J_{\gamma,R}^{\pi}$  uniformly in  $\pi$  as  $T \rightarrow \infty$ . As shown above, for each  $T$ , the

function  $\pi \mapsto J_{\gamma, R, T}^\pi$  is Lipschitz continuous, and hence uniformly continuous, on the metric space  $(\Pi_{\text{MR}}, \mu)$ . The second assertion now follows from the fact that the uniform limit of a sequence of uniformly continuous functions on a compact metric space is uniformly continuous. Refer the following literature by [9, Thm. 4.6, p. 234], [25, Thm. 21.6], [28, Prop. 23, p. 202] for more details.  $\square$

*Proof of Lemma 4.9.* We begin by showing that the sets  $\mathcal{F}$ ,  $\mathcal{F}_T^-$  and  $\mathcal{F}_T^+$  are compact, which also serves to prove (i). Fix  $T \in \{1, 2, \dots\}$ . By theorems 4.4 and 4.8, each of the sets defined in (2.3)-(2.10) is a finite intersection of inverse images of closed intervals under continuous functions, and are hence closed subsets of the compact space  $\Pi_{\text{MR}}$ . This proves statement (i) as well as compactness in statements (ii) and (iii).

(ii) For each  $i \in \{1, \dots, M\}$  and  $T \in \{1, 2, \dots\}$ , define  $\mathcal{C}_{i,T}^- := \{\pi \in \Pi_{\text{MR}} : v_{\mathcal{C}_{i,T}^-}^\pi(x) \leq b_i - K\beta^T\}$  and  $\mathcal{C}_i := \{\pi \in \Pi_{\text{MR}} : v_{\mathcal{C}_i}^\pi(x) \leq b_i\}$ . Similarly, define  $\hat{\mathcal{C}}_{j,T}^- := \{\pi \in \Pi_{\text{MR}} : J_{\gamma, \hat{\mathcal{C}}_{j,T}^-}^\pi(x) \leq \frac{\hat{b}_j}{K_T}\}$  and  $\hat{\mathcal{C}}_j := \{\pi \in \Pi_{\text{MR}} : J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x) \leq \hat{b}_j\}$  for each  $j \in \{1, \dots, \hat{M}\}$  and  $T \in \{1, 2, \dots\}$ . We can observe from (2.3), (2.4), (2.7) and (2.8) that  $\mathcal{C} = \bigcap_{i=1}^M \mathcal{C}_i$ ,  $\hat{\mathcal{C}} = \bigcap_{j=1}^{\hat{M}} \hat{\mathcal{C}}_j$  and, for each  $T = 1, 2, \dots$ ,  $\mathcal{C}_T^- = \bigcap_{i=1}^M \mathcal{C}_{i,T}^-$  and  $\hat{\mathcal{C}}_T^- = \bigcap_{j=1}^{\hat{M}} \hat{\mathcal{C}}_{j,T}^-$ . It is now easy to see from the definitions of  $\mathcal{F}_T^-$  and  $\mathcal{F}$  that statement (ii) follows if the following two claims are established.

**Claim 1:** For each  $i \in \{1, \dots, M\}$ ,  $\{\mathcal{C}_{i,T}^-\}_{T=1}^\infty$  is a non-decreasing sequence of nested sets contained in  $\mathcal{C}_i$ .

**Claim 2:** For each  $j \in \{1, \dots, \hat{M}\}$ ,  $\{\hat{\mathcal{C}}_{j,T}^-\}_{T=1}^\infty$  is a non-decreasing sequence of nested sets contained in  $\hat{\mathcal{C}}_j$ .

To prove Claim 1, note that

$$|v_{\mathcal{C}_{i,T+1}^-}^\pi(x) - v_{\mathcal{C}_{i,T}^-}^\pi(x)| = |\mathbb{E}_x^\pi[\beta^T R(X_T, A_T)]| \leq \beta^T C, \quad (\text{A.24})$$

holds for every  $i \in \{1, \dots, M\}$ ,  $T \in \{1, 2, \dots\}$  and  $\pi \in \Pi_{\text{MR}}$ . Hence, if  $v_{\mathcal{C}_{i,T}^-}^\pi(x) \leq b_i - \beta^T K$  holds, then  $v_{\mathcal{C}_{i,T+1}^-}^\pi(x) \leq v_{\mathcal{C}_{i,T}^-}^\pi(x) + \beta^T C \leq b_i - \beta^T K + \beta^T C = b_i - \beta^{T+1} K$  also holds (recall that  $C = K(1 - \beta)$ ). Thus,  $\mathcal{C}_{i,T}^- \subseteq \mathcal{C}_{i,T+1}^-$  for each  $T$  and  $i$ .

Furthermore, applying (A.24) repeatedly yields  $|v_{\mathcal{C}_{i,T+k}^-}^\pi(x) - v_{\mathcal{C}_{i,T}^-}^\pi(x)| \leq \beta^T (1 + \beta + \dots + \beta^{k-1})C$ , for each  $k \geq 1$ .

Lemma 4.3 implies that  $\lim_{k \rightarrow \infty} v_{\mathcal{C}_{i,T+k}^-}^\pi(x) = v_{\mathcal{C}_i}^\pi(x)$ . Letting  $k \rightarrow \infty$  in the last inequality thus leads to

$$|v_{\mathcal{C}_i}^\pi(x) - v_{\mathcal{C}_{i,T}^-}^\pi(x)| \leq \beta^T K. \quad (\text{A.25})$$

Hence, if  $v_{\mathcal{C}_{i,T}}^\pi(x) \leq b_i - \beta^T K$  holds, then  $v_{\mathcal{C}_i}^\pi(x) \leq v_{\mathcal{C}_{i,T}}^\pi(x) + \beta^T K \leq b_i$  holds as well. It immediately follows that  $\mathcal{C}_{i,T}^- \subseteq \mathcal{C}_i$ . This proves Claim 1.

To prove Claim 2, fix  $j \in \{1, \dots, \hat{M}\}$  and  $\pi \in \Pi_{\text{MR}}$ . Note that  $e^{-|\gamma|\beta^T C} \leq e^{\gamma\beta^T R(s,a)} \leq e^{|\gamma|\beta^T C}$  holds for every  $T \in \{1, 2, \dots\}$  and every  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . The last inequality yields

$$e^{-|\gamma|\beta^T C} J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq J_{\gamma, \hat{\mathcal{C}}_{j,T+1}}^\pi(x) \leq e^{|\gamma|\beta^T C} J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x), \quad (\text{A.26})$$

for every  $T = 1, 2, \dots$ . It is easy to see from (A.26) that if  $J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq \frac{\hat{b}_j}{K_T}$  holds for some  $T$ , then  $J_{\gamma, \hat{\mathcal{C}}_{j,T+1}}^\pi(x) \leq e^{|\gamma|\beta^T C} J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq e^{|\gamma|\beta^T C} \frac{\hat{b}_j}{K_T} = \frac{\hat{b}_j}{K_{T+1}}$  also holds. We immediately conclude that  $\hat{\mathcal{C}}_{j,T}^- \subseteq \hat{\mathcal{C}}_{j,T+1}^-$  for each  $T$  and  $j$ .

To complete the proof of Claim 2, fix  $T \in \{1, 2, \dots\}$ . Applying (A.26) repeatedly yields  $e^{-|\gamma|\beta^T K(1-\beta^k)} J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq J_{\gamma, \hat{\mathcal{C}}_{j,T+k}}^\pi(x) \leq e^{|\gamma|\beta^T K(1-\beta^k)} J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x)$  for each  $k \geq 1$ . Lemma 4.7 implies that  $\lim_{k \rightarrow \infty} J_{\gamma, \hat{\mathcal{C}}_{j,T+k}}^\pi(x) = J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x)$ . Letting  $k \rightarrow \infty$  in the last inequality thus leads to

$$\frac{1}{K_T} = e^{-|\gamma|\beta^T K} \leq \frac{J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x)}{J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x)} \leq e^{|\gamma|\beta^T K} = K_T. \quad (\text{A.27})$$

Hence, if  $J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq \frac{\hat{b}_j}{K_T}$  then  $J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x) \leq K_T J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq \hat{b}_j$ . It immediately follows that  $\hat{\mathcal{C}}_{j,T}^- \subseteq \hat{\mathcal{C}}_j$  for all  $T$  and  $j$ . Claim 2 now follows. This proves (ii) of the lemma.

(iii) For each  $i \in \{1, 2, \dots, M\}$  and  $T \in \{1, 2, \dots\}$ , define,  $\mathcal{C}_{i,T}^+ = \{\pi \in \Pi_{\text{MR}} : v_{\mathcal{C}_{i,T}}^\pi(x) \leq b_i + K\beta^T\}$ . Similarly for each  $j \in \{1, 2, \dots, \hat{M}\}$  and  $T \in \{1, 2, \dots\}$ , define,  $\hat{\mathcal{C}}_{j,T}^+ = \{\pi \in \Pi_{\text{MR}} : J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq \hat{b}_j K_T\}$ . We can observe from (2.9) and (2.10) that for each  $T = 1, 2, \dots$ ,  $\mathcal{C}_T^+ = \bigcap_{i=1}^M \mathcal{C}_{i,T}^+$  and  $\hat{\mathcal{C}}_T^+ = \bigcap_{j=1}^{\hat{M}} \hat{\mathcal{C}}_{j,T}^+$ . It is now easy to see from the definitions of  $\mathcal{F}_T^+$  and  $\mathcal{F}$  that statement (iii) follows if the following two claims are established.

**Claim 3:** For each  $i \in \{1, \dots, M\}$ ,  $\{\mathcal{C}_{i,T}^+\}_{T=1}^\infty$  is a non-increasing sequence of nested sets converging to  $\mathcal{C}_i$ .

**Claim 4:** For each  $j \in \{1, \dots, \hat{M}\}$ ,  $\{\hat{\mathcal{C}}_{j,T}^+\}_{T=1}^\infty$  is a non-increasing sequence of nested sets converging to  $\hat{\mathcal{C}}_j$ .

To prove Claim 3, fix  $i \in \{1, \dots, M\}$  and  $\pi \in \Pi_{\text{MR}}$ . It can be seen from (A.24) that, if  $v_{\mathcal{C}_{i,T+1}}^\pi(x) \leq b_i + K\beta^{T+1}$  holds for some  $T$ , then  $v_{\mathcal{C}_{i,T}}^\pi(x) \leq v_{\mathcal{C}_{i,T+1}}^\pi(x) + C\beta^T \leq b_i + K\beta^{T+1} + C\beta^T = b_i + K\beta^T$  also holds. Thus,  $\mathcal{C}_{i,T+1}^+ \subseteq \mathcal{C}_{i,T}^+$  for each  $T$  and  $i$ .

To complete the proof of Claim 3, fix  $T \in \{1, 2, \dots\}$ . It is easy to see from (A.25) that if  $v_{\hat{\mathcal{C}}_i}^\pi(x) \leq b_i$  holds, then  $v_{\hat{\mathcal{C}}_{i,T}}^\pi(x) \leq v_{\hat{\mathcal{C}}_i}^\pi(x) + K\beta^T \leq b_i + K\beta^T$  also holds. It immediately follows that  $\mathcal{C}_i \subseteq \hat{\mathcal{C}}_{i,T}^+$  for any fixed  $T$ , implying that  $\mathcal{C}_i \subseteq \bigcap_T \hat{\mathcal{C}}_{i,T}^+$ . On the other hand, if  $v_{\hat{\mathcal{C}}_{i,T}}^\pi(x) \leq b_i + K\beta^T$  holds for all  $T = 1, 2, \dots$ , then letting  $T \rightarrow \infty$  and applying Lemma 4.3 shows that  $v_{\hat{\mathcal{C}}_i}^\pi(x) \leq b_i$  holds. We can immediately conclude that  $\bigcap_T \hat{\mathcal{C}}_{i,T}^+ \subseteq \mathcal{C}_i$ . Thus Claim 3 is established.

To prove Claim 4, fix  $j \in \{1, \dots, \hat{M}\}$  and  $\pi \in \Pi_{\text{MR}}$ . It can be seen from (A.26) that if  $J_{\gamma, \hat{\mathcal{C}}_{j,T+1}}^\pi(x) \leq \hat{b}_j K_{T+1}$  for some  $T$ , then  $J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq J_{\gamma, \hat{\mathcal{C}}_{j,T+1}}^\pi(x) e^{|\gamma|C\beta^T} \leq \hat{b}_j K_{T+1} e^{|\gamma|C\beta^T} = \hat{b}_j K_T$  also holds. We therefore have  $\hat{\mathcal{C}}_{j,T+1}^+ \subseteq \hat{\mathcal{C}}_{j,T}^+$  for each  $T \in \{1, 2, \dots\}$ .

To complete the proof of Claim 4, fix  $T \in \{1, 2, \dots\}$ . It is easy to see from (A.27) that if  $J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x) \leq \hat{b}_j$  holds, then  $J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x) K_T \leq \hat{b}_j K_T$  also holds. Hence,  $\hat{\mathcal{C}}_j \subseteq \hat{\mathcal{C}}_{j,T}^+$  for every  $T \in \{1, 2, \dots\}$ . As a consequence, it follows that  $\hat{\mathcal{C}}_j \subseteq \bigcap_T \hat{\mathcal{C}}_{j,T}^+$ . On the other hand, if  $J_{\gamma, \hat{\mathcal{C}}_{j,T}}^\pi(x) \leq \hat{b}_j K_T$  holds for all  $T$ , then letting  $T \rightarrow \infty$  and applying Lemma 4.7 shows that  $J_{\gamma, \hat{\mathcal{C}}_j}^\pi(x) \leq \hat{b}_j$  also holds. Hence  $\bigcap_T \hat{\mathcal{C}}_{j,T}^+ \subseteq \hat{\mathcal{C}}_j$ . Thus Claim 4 is established and (iii) is proved.  $\square$

## B Proofs for Results from Subsection 4.2

*Proof of Proposition 4.10.* Choose  $\epsilon > 0$ . By uniform convergence, there exists  $L_1 \in \mathbb{Z}^+$  such that  $|f(z) - f_T(z)| < \frac{\epsilon}{3}$  for all  $z \in \mathcal{V}$  and  $T > L_1$ . For every  $z \in \mathcal{V}$  and  $T > L_1$ , we have  $\inf_{y \in \mathcal{V}} f(y) \leq f(z) \leq f_T(z) + \epsilon/3$ . Thus, for every  $T > L_1$ ,  $f_T$  is bounded below on  $\mathcal{V}$ , and hence on  $\mathcal{V}_T \subseteq \mathcal{V}$  as well. The last inequality now yields

$$\inf_{z \in \mathcal{V}} f(z) \leq \inf_{z \in \mathcal{V}} f_T(z) + \frac{\epsilon}{3} < \inf_{z \in \mathcal{V}_T} f_T(z) + \epsilon, \quad (\text{B.1})$$

for every  $T > L_1$ .

Next, let  $z_1 \in \mathcal{V}$  be such that  $f(z_1) < \inf_{z \in \mathcal{V}} f(z) + \frac{\epsilon}{3}$ . Define the set  $\mathcal{O} := \{z \in \mathcal{U} : |f(z) - f(z_1)| < \frac{\epsilon}{3}\}$ , and note that  $z_1 \in \mathcal{O}$ . By the continuity of  $f$ ,  $\mathcal{O}$  is an open set in  $\mathcal{U}$ . Since  $\mathcal{V}$  is the closure of the union  $\bigcup_{T=1}^\infty \mathcal{V}_T$  and  $z_1 \in \mathcal{V}$ , it follows from the definition of closure that every open neighborhood of  $z_1$  has a nonempty intersection with the union  $\bigcup_{T=1}^\infty \mathcal{V}_T$ . In particular, we may conclude that there exists  $L_2 \in \mathbb{Z}^+$  and  $z_2 \in \mathcal{U}$  such that  $z_2 \in \mathcal{O} \cap \mathcal{V}_{L_2}$ . Since the sequence of sets  $\{\mathcal{V}_T\}_{T=1}^\infty$  is non-decreasing, it follows that  $z_2 \in \mathcal{O} \cap \mathcal{V}_T$  for all  $T > L_2$ . For every  $T > \max\{L_1, L_2\}$ , we now have

$$\inf_{z \in \mathcal{V}_T} f_T(z) \leq f_T(z_2) < f(z_2) + \frac{\epsilon}{3} < f(z_1) + \frac{2\epsilon}{3} < \inf_{z \in \mathcal{V}} f(z) + \epsilon. \quad (\text{B.2})$$



Note that the first, second, third and last inequalities in (B.2) follow from  $z_2 \in \mathcal{V}_T$ , our choice of  $L_1$ ,  $z_2 \in \mathcal{O}$ , and our choice of  $z_1$ , respectively.

The inequalities (B.1) and (B.2) together imply that

$$|\inf_{z \in \mathcal{V}} f(z) - \inf_{z \in \mathcal{V}_T} f_T(x)| < \epsilon, \text{ for all } T > \max\{L_1, L_2\}.$$

Since we chose  $\epsilon > 0$  arbitrarily, (4.11) follows.  $\square$

*Proof of Proposition 4.11.* For each  $z \in \mathcal{H}$ , define  $I(z) := \{i : 1 \leq i \leq N, g_i(z) < B_i\}$  and  $J(z) := \{i : 1 \leq i \leq N, g_i(z) = B_i\}$ , and note that  $I(z) \cup J(z) = \{1, \dots, N\}$ . For every  $i \in I(z)$ , our convergence assumptions on the sequences  $\{g_{T,i}\}_{T=1}^\infty$  and  $\{B_{T,i}\}_{T=1}^\infty$  imply that the sequence  $\{g_{T,i}(z) - B_{T,i}\}_{T=1}^\infty$  converges to  $g_i(z) - B_i < 0$ . Consequently, for every  $i \in I(z)$ , there exists  $\tau_i$  such that  $g_{T,i}(z) - B_{T,i} < 0$  for all  $T > \tau_i$ . On letting  $\tau(z) = \max_{i \in I(z)} \tau_i$ , it follows that  $z \in \cap_{i \in I(z)} \mathcal{G}_{T,i}$  for all  $T > \tau(z)$ .

To prove that  $\mathcal{H} \subseteq \text{cl}(\cup_{T=1}^\infty \mathcal{H}_T)$ , choose  $y \in \mathcal{H}$ . Consider the case where  $J(y) = \emptyset$ . In this case,  $I(y) = \{1, \dots, N\}$ , and we conclude from the previous paragraph that  $y \in \cap_{i=1}^N \mathcal{G}_{T,i} = \mathcal{H}_T$  for every  $T > \tau(y)$ . In particular,  $y \in \cup_{T=1}^\infty \mathcal{H}_T \subseteq \text{cl}(\cup_{T=1}^\infty \mathcal{H}_T)$ .

Next, suppose 0 is not a local minimum value of  $h$ . Consider the case where  $J(y) \neq \emptyset$ , and let  $\mathcal{V} \subseteq \mathcal{U}$  be an open set containing  $y$ . Recall that, for every  $j \in J(y)$ ,  $g_j(y) = B_j$ . Hence it follows from  $J(y) \neq \emptyset$  that  $h(y) = 0$ . Since 0 is not a local minimum value of  $h$ , there exists  $z \in \mathcal{V}$  such that  $h(z) < 0$ . The condition  $h(z) < 0$  implies that  $z \in \mathcal{H}$  and  $J(z) = \emptyset$ . Applying the arguments of the previous paragraph to the point  $z$  shows that  $z \in \cup_{T=1}^\infty \mathcal{H}_T$ . We have thus shown that the arbitrarily chosen open neighborhood  $\mathcal{V}$  of  $y \in \mathcal{H}$  has a nonempty intersection with  $\cup_{T=1}^\infty \mathcal{H}_T$ . It follows that  $y \in \text{cl}(\cup_{T=1}^\infty \mathcal{H}_T)$ .

Since  $y \in \mathcal{H}$  above was chosen arbitrarily, it follows from the conclusions of the previous two paragraphs that  $\mathcal{H} \subseteq \text{cl}(\cup_{T=1}^\infty \mathcal{H}_T)$ . To show the reverse inclusion, we first deduce from elementary set-theoretic arguments that  $\cup_{T=1}^\infty \mathcal{H}_T = \cup_{T=1}^\infty \cap_{i=1}^N \mathcal{G}_{T,i} \subseteq \cap_{i=1}^N \cup_{T=1}^\infty \mathcal{G}_{T,i}$ . Next, we have

$$\begin{aligned} \text{cl}(\cup_{T=1}^\infty \mathcal{H}_T) &\subseteq \text{cl}(\cap_{i=1}^N \cup_{T=1}^\infty \mathcal{G}_{T,i}) \\ &\subseteq \cap_{i=1}^N \text{cl}(\cup_{T=1}^\infty \mathcal{G}_{T,i}) \subseteq \cap_{i=1}^N \text{cl}(\mathcal{G}_i) = \cap_{i=1}^N \mathcal{G}_i = \mathcal{H}. \end{aligned} \quad (\text{B.3})$$

The first inclusion in (B.3) follows from the fact that the closure operation preserves inclusion, the second from the fact that the closure of an intersection is contained in the intersection of the closures, and the third from our assumption that  $\mathcal{G}_{T,i} \subseteq \mathcal{G}_i$  for each  $T$  and  $i$ . The first equality in (B.3) follows from the fact that, for each  $i$ ,  $\mathcal{G}_i$  is the inverse image of the closed interval  $(-\infty, B_i]$  under the continuous function  $g_i$ , and hence closed. This completes the proof.  $\square$

## C Proofs for Results from Subsection 4.3

*Proof of Proposition 4.12.* Choose  $\epsilon > 0$ . By uniform convergence, there exists  $N > 0$  such that  $|g_T(y) - g(y)| \leq \epsilon$  for all  $y \in \mathcal{G}_1$  and all  $T \geq N$ . Since a continuous function attains its infimum on a compact set, for every  $T \geq 1$ , there exists  $y_T \in \mathcal{G}_T$  such that  $g_T(y_T) = \inf_{y \in \mathcal{G}_T} g_T(y)$ .

For every  $T \geq N$  and every  $z \in \mathcal{G} \subseteq \mathcal{G}_T \subseteq \mathcal{G}_1$ , we have  $g_T(y_T) \leq g_T(z) \leq g(z) + \epsilon$ . It follows that

$$\limsup_{T \rightarrow \infty} g_T(y_T) \leq \inf_{y \in \mathcal{G}} g(y) + \epsilon. \quad (\text{C.1})$$

There exists a strictly increasing sequence  $\{T_l\}_{l=1}^{\infty}$  of integers such that the sequence  $\{g_{T_l}(y_{T_l})\}_{l=1}^{\infty}$  converges to  $\liminf_{T \rightarrow \infty} g_T(y_T)$ . We may choose  $L$  such that  $T_L > N$  and  $g_{T_l}(y_{T_l}) \leq \liminf_{T \rightarrow \infty} g_T(y_T) + \frac{\epsilon}{2}$  for all  $l \geq L$ .

Next, for each  $T \geq 1$ , define the set  $\mathcal{D}_T := \{y_t : t \geq T\} \subseteq \mathcal{G}_T$ . Note that, for every  $T \geq 1$ ,  $\text{cl}(\mathcal{D}_T)$  is a closed subset of the compact set  $\mathcal{G}_T$ , and hence compact. The sequence of sets  $\{\text{cl}(\mathcal{D}_T)\}_{T=1}^{\infty}$  is thus a non-increasing sequence of non-empty compact sets. By the Frechet Intersection, Theorem given by [28, p. 236], [9, p. 253],  $\cap_{T \geq 1} \text{cl}(\mathcal{D}_T)$  is non-empty. Also, note that  $\cap_{T \geq 1} \text{cl}(\mathcal{D}_T) \subseteq \cap_{T \geq 1} \mathcal{G}_T = \mathcal{G}$ .

Next, choose,  $z \in \cap_{T \geq 1} \text{cl}(\mathcal{D}_T) \subseteq \mathcal{G}$ . By continuity, there exists an open neighbourhood  $\mathcal{V} \subseteq \mathcal{U}$  of  $z$  such that  $g(y) > g(z) - \frac{\epsilon}{2}$  for all  $y \in \mathcal{V}$ . Since  $z \in \cap_{T \geq 1} \text{cl}(\mathcal{D}_T) \subseteq \text{cl}(\mathcal{D}_{T_L})$ , the set  $\mathcal{V} \cap \mathcal{D}_{T_L}$  is nonempty. In other words, there exists  $l \geq L$  such that  $y_{T_l} \in \mathcal{V}$ . Note that  $T_l \geq T_L > N$ . We now have,

$$\inf_{y \in \mathcal{G}} g(y) - 2\epsilon \leq g(z) - 2\epsilon \leq g(y_{T_l}) - \frac{3\epsilon}{2} \leq g_{T_l}(y_{T_l}) - \frac{\epsilon}{2} \leq \liminf_{T \rightarrow \infty} g_T(y_T). \quad (\text{C.2})$$

Comparing (C.1) and (C.2) shows that  $\inf_{y \in \mathcal{G}} g(y) - 2\epsilon \leq \liminf_{T \rightarrow \infty} g_T(y_T) \leq \limsup_{T \rightarrow \infty} g_T(y_T) \leq \inf_{y \in \mathcal{G}} g(y) + \epsilon$ . Since  $\epsilon > 0$  was chosen arbitrarily, we conclude that  $\inf_{y \in \mathcal{G}} g(y) = \liminf_{T \rightarrow \infty} g_T(y_T) = \limsup_{T \rightarrow \infty} g_T(y_T)$ . In other words, (4.18) holds. This completes the proof.  $\square$

## D Counter Example

**Example D.1** Consider an MDP with  $\mathcal{S} = \{s\}$  and  $\mathcal{A} = \{a_1, a_2\}$ . Define two immediate cost functions  $C_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ ,  $i \in \{1, 2\}$  by  $C_1(s, a_1) = 1$ ,  $C_1(s, a_2) = 0$ , and  $C_2 := 1 - C_1$ . Let  $\beta = 2^{-1}$ . Note that  $C = 1$ , and thus  $K = C/(1 - \beta) = 2$ .

Next, for each  $i \in \{1, 2\}$ , recall the definitions (2.2) of the infinite-horizon standard discounted costs  $\pi \mapsto v_{C_i}^{\pi}$  and the finite-horizon standard discounted

costs  $\pi \mapsto v_{C_i, T}^\pi$  for  $T \in \{1, 2, \dots\}$ . It is easy to see that for all  $\pi \in \Pi_{\text{MR}}$  and  $T \in \{1, 2, \dots\}$ ,

$$v_{C_1, T}^\pi(s) + v_{C_2, T}^\pi(s) = 2 - \frac{1}{2^{T-1}}, \quad \text{and} \quad v_{C_1}^\pi(s) + v_{C_2}^\pi(s) = 2. \quad (\text{D.1})$$

Next, define  $b_1 = b_2 = 1$ , and fix a finite horizon  $T > 0$ . Recall the set  $\mathcal{C}_T^-$  defined by (2.7) with  $x = s$ . If  $\pi \in \mathcal{C}_T^-$ , then  $\pi$  must satisfy  $v_{C_1, T}^\pi(s) + v_{C_2, T}^\pi(s) \leq b_1 + b_2 - 2K\beta^T = 2 - (\frac{1}{2})^{T-2}$ , which contradicts (D.1). The contradiction shows that the feasible region  $\mathcal{F}_T^- = \mathcal{C}_T^-$  of the problem  $(P_T^-)$  is empty for each  $T \in \{1, 2, \dots\}$ .

We claim that the feasible region  $\mathcal{F}$  of the problem  $(P)$  is nonempty. Indeed, let  $\phi \in \Pi_{\text{MR}}$  denote the stationary policy that selects from the two actions uniformly at random at each decision epoch. It is easy to check that  $v_{C_1}^\phi(s) = v_{C_2}^\phi(s) = 1$ . It follows that  $\phi \in \mathcal{F} = \mathcal{C}$ , where  $\mathcal{C}$  is defined by (2.3) with  $x = s$ .

We conclude from the previous paragraph that the assertion of Theorem 3.2 does not hold for the MDP considered in this example. Next we claim that the local minimum condition assumed in the theorem fails to hold. Indeed, let  $h : \Pi_{\text{MR}} \rightarrow \mathbb{R}$  denote the maximum constraint violation function defined in the theorem, that is,  $h(\pi) = \max\{v_{C_1}^\pi(s) - 1, v_{C_2}^\pi(s) - 1\}$  for all  $\pi \in \Pi_{\text{MR}}$ . In light of the second equality in (D.1), we have  $h(\pi) = \max\{v_{C_1}^\pi(s) - 1, 1 - v_{C_1}^\pi(s)\} \geq 0$  for all  $\pi \in \Pi_{\text{MR}}$ . Also, from the previous paragraph, we have  $h(\phi) = 0$ . Thus 0 is a global (and hence a local) minimum value for  $h$ . This example shows that the assertion of Theorem 3.2 may not hold if the local minimum condition assumed in the theorem is violated.

## References

- [1] Altman, Eitan. Constrained Markov decision processes: stochastic modeling. Routledge, 1999.
- [2] Borkar, Vivek, and Rahul Jain. "Risk-constrained Markov decision processes." IEEE Transactions on Automatic Control 59, no. 9 (2014): 2574-2579.
- [3] Bäuerle, Nicole, and Ulrich Rieder. "More risk-sensitive Markov decision processes." Mathematics of Operations Research 39, no. 1 (2014): 105-120.
- [4] Coraluppi, Stefano Paolo. Optimal control of Markov decision processes for performance and robustness. University of Maryland, College Park, 1997.

- [5] Coraluppi, Stefano P., and Steven I. Marcus. "Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes." *Automatica* 35, no. 2 (1999): 301-309.
- [6] Moreira, Daniel Augusto De Melo, Karina Valdivia Delgado, and Leliane Nunes de Barros. "Risk-sensitive Markov decision process with limited budget." In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 109-114. IEEE, 2017.
- [7] Denardo, Eric V., Haechurl Park, and Uriel G. Rothblum. "Risk-sensitive and risk-neutral multiarmed bandits." *Mathematics of Operations Research* 32, no. 2 (2007): 374-394.
- [8] Derman, Cyrus, and Morton Klein. "Some remarks on finite horizon Markovian decision models." *Operations research* 13, no. 2 (1965): 272-278.
- [9] Dugundji, J. 1993. *Topology Topology* (Third Indian Reprint ed.). Wm. C. Brown Publishers
- [10] Feinberg, Eugene A., and Adam Schwartz. "Constrained discounted dynamic programming." *Mathematics of Operations Research* 21, no. 4 (1996): 922-945.
- [11] Geibel, Peter, and Fritz Wyszotzki. "Risk-sensitive reinforcement learning applied to control under constraints." *Journal of Artificial Intelligence Research* 24 (2005): 81-108.
- [12] Golabi, Kamal, Ram B. Kulkarni, and George B. Way. "A statewide pavement management system." *Interfaces* 12, no. 6 (1982): 5-21.
- [13] Gosavi, Abhijit. "Variance-penalized Markov decision processes: Dynamic programming and reinforcement learning techniques." *International Journal of General Systems* 43, no. 6 (2014): 649-669.
- [14] Haskell, William B., and Rahul Jain. "A convex analytic approach to risk-aware Markov decision processes." *SIAM Journal on Control and Optimization* 53, no. 3 (2015): 1569-1598.
- [15] Horn, Roger A., and Charles R. Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [16] Howard, Ronald A., and James E. Matheson. "Risk-sensitive Markov decision processes." *Management science* 18, no. 7 (1972): 356-369.

- [17] Iyer, Krishnamurthy, and Nandyala Hemachandra. "Sensitivity analysis and optimal ultimately stationary deterministic policies in some constrained discounted cost models." *Mathematical Methods of Operations Research* 71, no. 3 (2010): 401-425.
- [18] Jaquette, Stratton C. "Markov decision processes with a new optimality criterion: Discrete time." *The Annals of Statistics* 1, no. 3 (1973): 496-505.
- [19] Jaquette, Stratton C. "A utility criterion for Markov decision processes." *Management Science* 23, no. 1 (1976): 43-49.
- [20] Filar, Jerzy A., Lodewijk CM Kallenberg, and Huey-Miin Lee. "Variance-penalized Markov decision processes." *Mathematics of Operations Research* 14, no. 1 (1989): 147-161.
- [21] Kallenberg, Lodewijk CM. "Linear programming and finite Markovian control problems." *MC Tracts* (1983).
- [22] Kumar, Atul, Veeraruna Kavitha, and Nandyala Hemachandra. "Finite horizon risk sensitive MDP and linear programming." In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 7826-7831. IEEE, 2015.
- [23] Kumar M, Uday, Sanjay P. Bhat, Veeraruna Kavitha, and Nandyala Hemachandra. "Ultimately Stationary Policies to Approximate Risk-Sensitive Discounted MDPs." In *Proceedings of the 12th EAI International Conference on Performance Evaluation Methodologies and Tools*, pp. 63-70. 2019.
- [24] Mannor, Shie, and John Tsitsiklis. "Mean-variance optimization in Markov decision processes." *arXiv preprint [arXiv:1104.5601](https://arxiv.org/abs/1104.5601)* (2011).
- [25] Munkres, J. R. 1999. *Topology Topology (Second Edition ed.)*. Prentice Hall
- [26] Piunovskiy, Alexey B. "Dynamic programming in constrained Markov decision processes." *Control and Cybernetics* 35, no. 3 (2006): 645.
- [27] Puterman, Martin L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [28] Royden, H., & Fitzpatrick, P. 2010. *Real Analysis Real analysis (Vol. Fourth edition)*. Prentice Hall.

- [29] Xia, Li. "Risk-Sensitive Markov Decision Processes with Combined Metrics of Mean and Variance." *Production and Operations Management* 29, no. 12 (2020): 2808-2827.
- [30] Yang, Buheerdun. "Conditional value-at-risk minimization in finite state markov decision processes: Continuity and compactness." *Journal of Uncertain Systems* 7, no. 1 (2013): 50-57.