Approximate Solution of Infinite-horizon Risk-sensitive Markov Decision Processes

Jack Zhang
Operations Research Center, Massachusetts Institute of Technology

Saumya Sinha
Department of Industrial and Systems Engineering, University of Minnesota-Twin Cities

Mehdi Hemmati
School of Industrial and Systems Engineering, University of Oklahoma

Andrew J. Schaefer
Department of Computational Applied Mathematics & Operations Research, Rice University

Abstract

Infinite-horizon risk-sensitive Markov decision processes (MDPs) under the discounted cost criterion are challenging to solve because the optimal policy may be non-stationary. Existing solution methods reformulate the problem as a continuous-state (risk-neutral) MDP and solve it using state-discretization or value function approximation. Such approaches typically lack explicit stopping conditions or error bounds. In this paper, we present approximate variants of value iteration, linear programming (LP), and policy iteration approaches that can directly solve risk-sensitive MDPs without the need for reformulation. We show that the approximate value functions obtained by value iteration and LP converge to the optimal value function, and a near-optimal policy can be obtained. The policies generated by policy iteration also converge in value to the optimal policy. Importantly, in addition to proving asymptotic convergence, we provide explicit stopping conditions to achieve arbitrarily close approximations to the optimal value function.

1 Introduction

Markov decision processes (MDPs) are widely used to model sequential decision-making problems. In a classical MDP, an agent seeks to find optimal actions for every decision epoch so as to minimize the expected total cost (or, equivalently, maximize the expected total reward) accrued over the planning horizon [19]. Such an agent may be viewed as "risk-neutral," in that the perceived value of a policy is linear with respect to the incurred cost. In many applications, however, the decision maker may exhibit risk-averse or risk-seeking preferences [3, 17]. This is modeled via a nonlinear utility function, which leads to RMDP, where the agent seeks to minimize a (nonlinear) utility function of the total cost.

Among such utility functions, those with *hyperbolic absolute risk aversion* (HARA) have received special attention because of their flexibility and tractability [12]. They were popularized by their use for risk-sensitive control in the financial decision-making literature, such as utility-based portfolio control [14], hedging in incomplete markets [9], and consumption-investment with stochastic volatility [10]. In this paper, we focus on the exponential utility function, a special case of the HARA class and the only risk-sensitive utility function with constant local risk aversion and a wealth-invariant risk premium [6]. Exponential utility is a popular choice in the RMDP literature because it allows for a tractable recursive structure of the optimality equations, while still meaningfully capturing risk sensitivity. An exponential utility function is parameterized by a non-zero risk parameter γ; positive values of γ correspond to risk-seeking behavior, whereas negative values imply risk aversion. We emphasize that $\gamma = 0$ should not be interpreted as the risk-neutral case, as we explain in Section 2.

Howard and Matheson [11] were the first to study infinite-horizon RMDPs with exponential utility functions under a total cost criterion. They used the concept of certainty equivalents to evaluate the performance of RMDPs and proposed policy iteration for this class of problems. Chung and Sobel [6] formalized the theory of and derived optimality equations for RMDPs with discounted costs and general utility functions. More recently, Di Masi and Stettner [8] derived Bellman equations for discounted-cost RMDPs with continuous state spaces. Bäuerle and Rieder [2] reformulated RMDPs as risk-neutral MDPs with an extended continuous state-space, which can then be solved via value and policy iteration. Wu and Xu [23] studied RMDPs (with general utility functions) through the lens of reinforcement learning, where dynamic programming (DP) and the resulting solution approaches are recovered by augmenting the state with the cumulative reward. Kumar et al. [15] introduced an LP approach for solving finite-horizon stationary RMDPs. See Bäuerle and Jaśkiewicz [1] for a recent survey on RMDPs.

Many papers have focused on the solution of RMDPs in the average-cost setting, including variations of value and policy iterations with provable convergence results. Bielecki et al. [3] studied finite-state, discrete-time risk-sensitive control with exponential cost and established DP results under weaker irreducibility-type assumptions, with an application to portfolio management. Cavazos-Cadena and Montes-De-Oca [5] further analyzed the value-iteration algorithm for finite-state risk-sensitive average-cost Markov decision chains, showing that it can be used to approximate the optimal average cost and obtain near-optimal stationary policies. Borkar and Meyn [4] extended this to countable-state models with unbounded costs, showing existence of optimal stationary feedback policies under (near-)monotonicity conditions and providing corresponding value and policy-iteration algorithms. Murthy et al. [18] presented provably-convergent modified policy iteration for finite-state, average-cost exponential-cost MDPs. Unlike these works, our paper focuses on the discounted-cost setting, where the objective, Bellman recursion, and technical issues differ substantially from the average-cost framework.

A fundamental challenge in solving infinite-horizon RMDPs with discounted costs is their inherent non-stationarity—the optimal value of a state for am RMDP parameterized by the risk parameter γ depends on the optimal value function of the same problem parameterized by $\gamma\beta$, where $\beta \in (0, 1)$ is the discount factor. This was first noted by Jaquette [13], who considered the expected exponential utility of the total discounted return, thereby incorporating constant risk sensitivity directly into the control criterion. A central insight of the paper was that unlike the risk-neutral case, discounted-cost RMDPs with exponential utilities are not guaranteed to have stationary optimal policies, even when the underlying model itself is stationary (see Chung and Sobel [6] for a more detailed discussion). As such, risk-sensitive control is not merely a minor variation of the risk-neutral theory, but a setting with genuinely different structural and analytical features. To our knowledge, the only existing methods of solving infinite-horizon discounted-cost RMDPs are the value and policy iteration algorithms proposed by Bäuerle and Rieder [2], which employ continuous-state (risk-neutral) MDPs.

In this paper, we extend the frameworks of Bäuerle and Rieder [2] and Di Masi and Stettner [8] to propose approximate solution methods for infinite-horizon discounted-cost RMDPs with exponential utility functions and finite state and action spaces. We first adapt the Bellman equations from Di Masi and Stettner [8] to discrete state-spaces and extend them to risk-averse RMDPs. In Di Masi and Stettner [8], only risk-seeking RMDPs were discussed, and the optimal value function is expressed as a two-variable function that depends on the risk parameter in addition to the state of the MDP. Subsequently, we propose three methods to approximately solve the RMDP to arbitrary accuracy, namely, value iteration, LP, and policy iteration. In each case, we prove convergence and derive an explicit stopping condition. Here we highlight the fact that we also include the risk parameter, γ, in the value functions as a parameter taking values in a countably infinite set. This is different from the approach of Bäuerle and Rieder [2], where γ was treated as an additional continuous state variable. While Di Masi and Stettner [8] also used γ as a discrete parameter in the value function, they only proved the existence of a unique optimal value function under this scheme, but did not provide a solution algorithm. On the other hand, the papers that provide solution techniques for computing optimal value function or optimal policies include the risk parameter γ within the augmented state space, which transforms the problem into a standard risk-neutral continuous-state MDP. These approaches require state-space discretization, and importantly, do not provide a stropping condition.

Contributions: In this paper, we study infinite-horizon RMDPs with finite state and action spaces under the discounted cost criterion. Our main contributions are as follows:

- We develop approximate solution methods for infinite-horizon discounted RMDPs that work directly with the original risk-sensitive problem. This is in contrast to existing works, which typically employ reformulation or state-augmentation. We present approximate variants of three classical DP approaches: value iteration, LP, and policy iteration, adapted to the risk-sensitive setting.

- We prove that the approximate value functions produced by the proposed value iteration and LP methods converge to the optimal value function. Further, we show that the policy iteration procedure yields policies whose values converge to the optimal value, and in particular allows one to obtain near-optimal policies.

- Beyond asymptotic convergence, we provide explicit stopping conditions for the proposed algorithms. This allows arbitrarily accurate approximations to the optimal value function. To the best of our knowledge, this is the first work that computes stopping criteria for the approximate solution techniques of discounted-cost infinite-horizon RMDPs with exponential utility functions.

In summary, our results provide a direct algorithmic framework for solving discounted RMDPs, together with convergence guarantees and actionable stopping criteria that support both theoretical analysis and practical computation.

2 Infinite-horizon RMDPs

In this section, we introduce infinite-horizon RMDPs and present the corresponding Bellman equations that characterize the optimal value function.

Table 1: Table of Notation

Symbol	**Description**
γ	Risk-parameter, non-zero
$d_{\max}$	Maximum value of $\lvert\gamma\rvert$
Γ^+	$(0, d_{\max}]$; $\gamma \in \Gamma^+$ corresponds to risk-seeking decision-maker
Γ^-	$[-d_{\max}, 0)$; $\gamma \in \Gamma^-$ corresponds to risk-averse decision-maker
Γ	Set of possible risk-parameters, $\Gamma^+ \cup \Gamma^- = [-d_{\max}, d_{\max}] \setminus \{0\}$
β	Discount factor, a fixed parameter in $[0, 1)$
$\mathcal{S}$	State space (discrete, finite)
x, y	States in $\mathcal{S}$
x_i	State of the MDP in period i
$\mathcal{A}$	Action space (discrete, finite)
Π	Set of single-period decision rules, i.e., $\Pi \coloneqq \{\pi : \mathcal{S} \to \mathcal{A}\}$
$\boldsymbol{\Pi}^n$	Set of all possible (non-stationary) policies in the **first** n periods
$\boldsymbol{\Pi}^\infty$	Set of all possible (non-stationary) infinite-horizon policies
$\boldsymbol{\pi}^n$	A feasible element (a sequence of policies) in $\boldsymbol{\Pi}^n$
$\boldsymbol{\pi}^\infty$	A feasible element (an infinite sequence of policies) in $\boldsymbol{\Pi}^\infty$
π_i^n (π_i^∞)	Policy for the ith period in $\boldsymbol{\pi}^n$ ($\boldsymbol{\pi}^\infty$)
$c(x, a)$	The incurred (nonnegative) cost for taking action $a \in \mathcal{A}$ when in state $x \in \mathcal{S}$
$P^a(x, y)$	Transition probability from state $x \in \mathcal{S}$ to $y \in \mathcal{S}$ under action $a \in \mathcal{A}$
$v^{\boldsymbol{\pi}^\infty}(x; \gamma)$	Value function of an infinite-horizon non-stationary policy $\boldsymbol{\pi}^\infty$ for a given γ when starting in state x
$v^*(x; \gamma)$	Optimal value function for state $x \in \mathcal{S}$ for a given γ

Note: Bold symbols (e.g., $\boldsymbol{\pi}$) refer to sequences of elements (possibly vectors), where elements in the sequence (e.g., π_i) are indexed by time $i = 1, \ldots, n$ with $n \in \mathbb{N} \cup \{\infty\}$.

2.1 Problem Definition

Consider an infinite-horizon MDP with a finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. In any epoch, given a state $x \in \mathcal{S}$, let $P^a(x, y)$ denote the probability of transitioning to state $y \in \mathcal{S}$ and $c(x,a) \geq 0$ be the immediate cost incurred when action a is chosen. In the discounted-cost setting, the cost in period n is multiplied by β^n, where $\beta \in (0,1)$ is a (fixed and known) discount factor. Let $\boldsymbol{\pi}^\infty = (\pi_1^\infty, \pi_2^\infty, \ldots)$ denote an infinite-horizon (non-stationary) deterministic policy, where $\pi_i^\infty : \mathcal{S} \to \mathcal{A}$ is the decision rule in period $i \in \mathbb{N}$. Then, the expected total discounted cost is given by

$$\sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)),$$

where x_i is the state in period i and $\pi_i^\infty(x_i)$ is the action chosen in state x_i in period $i \in \mathbb{N}$. Table 1 summarizes the notation used throughout the paper.

For RMDPs, the total cost is defined via a utility function. A popular choice for a risk-sensitive utility function is the exponential utility $u(z) = \mathrm{sgn}(\gamma)\exp(\gamma z)$, where γ is a fixed risk parameter and sgn(.) is the sign function, so that $\mathrm{sgn}(z) = 1$ if $z \geq 0$ and -1, otherwise. A positive (negative) risk parameter denoted by γ models the risk-seeking (risk-averse) behavior of the decision maker. We assume that $|\gamma| \leq d_{\max}$ (for some $d_{\max} \in \mathbb{R}^+$) and define $\Gamma^+ = (0, d_{\max}]$ and $\Gamma^- = [-d_{\max}, 0)$. In this paper, we study both risk-seeking and risk-averse cost criteria for $\gamma \in \Gamma = \Gamma^+ \cup \Gamma^-$.

The expected total discounted (risk-sensitive) cost is defined as

$$v^{\boldsymbol{\pi}^\infty}(x;\gamma) = \mathbb{E}\left[\mathrm{sgn}(\gamma)\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i))\right)\right], \quad \text{for all } x \in \mathcal{S}, \tag{1}$$

where $x_0 = x$ is the initial state. The expectation in (1) is taken with respect to future states $\{x_i\}_{i=1}^\infty$ and policy $\boldsymbol{\pi}^\infty$. Because this is assumed throughout the paper, we omit subscripts for all expectations unless the expectation is taken with respect to a different set of parameters, in which case we will be explicit with the subscript. We also note that in this representation, γ is a fixed parameter and not part of the state; we separate the state x and the parameter γ with a semicolon to emphasize the distinction.

Under the cost criteria defined in (1), the optimal value function of the RMDP is defined as

$$v^*(x;\gamma) = \inf_{\boldsymbol{\pi}^\infty \in \boldsymbol{\Pi}^\infty} \left\{\mathbb{E}\left[\mathrm{sgn}(\gamma)\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i))\right)\right]\right\}, \quad \text{for all } x \in \mathcal{S}, \tag{2}$$

where $x_0 = x$ is the initial state. Define $\|c\| = \max_{x,a}\{c(x,a)\}$ as the maximum cost for any state-action pair. Note that v^* is well defined because the operand in inf(.) is non-empty and bounded. Specifically, since $\beta \in [0,1)$, we have

$$\mathbb{E}\left[\mathrm{sgn}(\gamma)\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i))\right)\right] \in \left[\mathrm{sgn}(\gamma), \mathrm{sgn}(\gamma)\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\right].$$

However, it is not immediate whether the infimum in (2) is attained. In the remainder of this section, we derive Bellman equations for this problem setting and establish the existence of an optimal policy. We remark that our results are partially adapted from Di Masi and Stettner [8].

Henceforth, we omit the class of policies that we are optimizing over in min and max expressions as it is clear in the context. For example, in (2), we write $\min_{\boldsymbol{\pi}^\infty}$ instead of $\min_{\boldsymbol{\pi}^\infty \in \boldsymbol{\Pi}^\infty}$. We will be more specific in instances with ambiguity.

Remark 1. Note that we exclude $\gamma = 0$ from Γ. This is because when we let $\gamma \uparrow 0$ or $\gamma \downarrow 0$, the utility function respectively becomes a constant 1 or -1. Therefore, contrary to what one might expect, setting γ to 0 does not directly give us a risk-neutral MDP in this case, even though $\gamma = 0$ separates the risk-averse and risk-seeking regions.

2.2 Bellman Equations

Let $\mathcal{C}(\mathcal{S} \times \Gamma)$ be the space of all real-valued continuous bounded functions with the domain $\mathcal{S} \times \Gamma$. Then, $\mathcal{C}(\mathcal{S} \times \Gamma)$ is a Banach space under the uniform norm $\|\cdot\|_\infty$ [20]. Define the operator T on $\mathcal{C}(\mathcal{S} \times \Gamma)$ as

$$\begin{aligned} Tf(x,\gamma) &= \min_\pi \left\{ \exp\left(\gamma c(x,\pi(x))\right) \sum_{y \in \mathcal{S}} f(y, \gamma\beta) P^{\pi(x)}(x,y) \right\} \\ &= \min_\pi \left\{ \mathbb{E}_y \left[\exp\left(\gamma c(x,\pi(x))\right) f(y,\gamma\beta) \right] \right\}, \quad \forall\, f \in \mathcal{C}(\mathcal{S} \times \Gamma). \end{aligned} \tag{3}$$

It is easy to see that T is well defined. We can see that T is a continuous operator since the minimum of finitely many (since there are only finitely many deterministic policies π) continuous functions is still continuous.

It is easy to see that the operator T is well defined. Further, since Π is finite, then the operator T, the minimum of finitely many deterministic decision rules, remains continuous. Lemmas 1–4 establish some properties of T. Proofs not included in the text are provided in the appendix.

Note that we use a comma instead of a semicolon as the delimiter here because f is a function in $\mathcal{C}(\mathcal{S} \times \Gamma)$, for which both x and γ are treated as function arguments.

Lemma 1. *Suppose $f(x,\gamma) \in \mathcal{C}(\mathcal{S} \times \Gamma)$ satisfies the condition*

$$\lim_{\gamma \to 0} \max_{x \in \mathcal{S}} |f(x,\gamma) - \operatorname{sgn}(\gamma)| = 0. \tag{4}$$

Then, (4) *is also satisfied by $Tf(x,\gamma)$.*

Lemma 2. *The operator T is an order-preserving mapping, i.e., given functions $f_1, f_2 \in \mathcal{C}(\mathcal{S} \times \Gamma)$ such that $f_1(x,\gamma) \geq f_2(x,\gamma)$ $\forall\, x \in \mathcal{S}$ and $\gamma \in \Gamma$, it follows that*

$$Tf_1(x,\gamma) \geq Tf_2(x,\gamma), \quad \forall\, x \in \mathcal{S},\ \gamma \in \Gamma.$$

Because T maps $\mathcal{C}(\mathcal{S} \times \Gamma)$ to itself, it can be applied successively to any function f in its domain. Then, for any integer n and for $f \in \mathcal{C}(\mathcal{S} \times \Gamma)$, $T^n f$ may be expressed as

$$T^n f(x,\gamma) = \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x)) \right) f(x_n, \gamma\beta^n) \right] \right\},\ \forall\, x \in \mathcal{S},$$

where $\boldsymbol{\pi}^n$ is an n-period policy and $\boldsymbol{\Pi}^n$ is the space of all such policies. Note that $\boldsymbol{\pi}^n$ may be viewed as a truncation of an infinite-horizon policy $\boldsymbol{\pi}^\infty \in \boldsymbol{\Pi}^\infty$.

Lemma 3. *For every $x \in \mathcal{S}, \gamma \in \Gamma$, define $\underline{f}(x,\gamma) = \operatorname{sgn}(\gamma)$ and*

$$\overline{f}(x,\gamma) = \operatorname{sgn}(\gamma) \exp\left(\frac{\gamma \|c\|}{1-\beta}\right).$$

Then, $\underline{f}$, $\overline{f} \in \mathcal{C}(\mathcal{S} \times \Gamma)$ and

$$\underline{f}(x,\gamma) \le T^n \underline{f}(x,\gamma) \le T^n \overline{f}(x,\gamma) \le \overline{f}(x,\gamma), \quad \forall\ n \ge 1.$$

Moreover, the sequences $T^n \underline{f}(x,\gamma)$ and $T^n \overline{f}(x,\gamma)$ are increasing and decreasing, respectively.

Lemma 4. *Let $\underline{f}(x,\gamma)$, $\overline{f}(x,\gamma)$ be as defined in Lemma 3. Then, $\left|T^n \overline{f}(x,\gamma) - T^n \underline{f}(x,\gamma)\right| \to 0$ as $n \to \infty$. Moreover, the convergence is uniform in both $x \in \mathcal{S}$ and $\gamma \in \Gamma^+ \cup \Gamma^-$.*

Next, we define the Bellman equations for this problem and prove their optimality. Theorem 1 and its proof for $\gamma > 0$ are adapted from Di Masi and Stettner [8, Proposition 3.1] to the special case of finite, discrete state and action spaces, and then extended to the risk-seeking setting ($\gamma < 0$).

Theorem 1. *There exists a unique function $w \in \mathcal{C}(\mathcal{S} \times \Gamma)$ that satisfies the equations*

$$w(x,\gamma) = \min_{\pi} \left\{ \exp\left(\gamma c(x,\pi(x))\right) \sum_{y \in \mathcal{S}} w(y;\gamma\beta) P^{\pi(x)}(x,y) \right\}, \quad \forall\ x \in \mathcal{S}, \gamma \in \Gamma, \tag{5}$$

with boundary condition

$$\lim_{\gamma \to 0} \left(\max_{x \in \mathcal{S}} |w(x,\gamma) - \operatorname{sgn}(\gamma)| \right) = 0. \tag{6}$$

Moreover, w equals the optimal value function v^* defined in (2)*.*

We outline the main idea of the proof here; the complete proof is provided in Appendix A.5.

By Lemma 3, $T^n \underline{f}(x,\gamma)$ and $T^n \overline{f}(x,\gamma)$ are bounded monotonic sequences and therefore convergent for every $x \in \mathcal{S}$, $\gamma \in \Gamma$. Lemma 4 establishes that $\left|T^n \overline{f}(x,\gamma) - T^n \underline{f}(x,\gamma)\right|$ converges uniformly to zero as $n \to \infty$, which implies both $T^n \underline{f}(x,\gamma)$ and $T^n \overline{f}(x,\gamma)$ must converge uniformly to a common limit, say, $w(x;\gamma)$. It follows that $w \in C(\mathcal{S} \times \Gamma)$ is the unique solution to (5) and satisfies the boundary condition (6).

Remark 2. We can use (5) to construct a deterministic policy whose value equals the optimal value function v^*, by always assigning a minimizing action of (5) to the corresponding $x \in \mathcal{S}, \gamma \in \Gamma$. This demonstrates the existence of a deterministic optimal policy. However, unlike the risk-neutral case, the optimal policy may no longer be stationary.

In the rest of the paper, we present algorithms for finding an approximately optimal nonstationary deterministic policy. All of our methods directly incorporate the risk parameter within the iteration scheme, by iterating over pairs $(x, \gamma\beta^i)$, for $x \in \mathcal{S}$ and $i \ge 0$. While it may be possible to view this setting as a countable-state MDP by augmenting the state space to include $\gamma\beta^i$, that reformulation would not directly lead to applicable solution methods. This is because solution methods for countable-state MDPs typically rely on the Bellman operator being a contraction, an assumption that fails in our setting.

3 Value Iteration for RMDPs

In this section, we introduce an approximate value iteration algorithm for RMDPs and establish its convergence. The standard value iteration algorithm solves risk-neutral MDPs by repeatedly applying the Bellman operator to an initial guess $V_0(\cdot)$ through the update rule

$$V_{k+1}(x) = \min_{a\in\mathcal{A}} \sum_{y\in\mathcal{S}} P^a(x,y)\big[c(x,a) + \gamma V_k(y)\big], \quad k = 0, 1, 2, \ldots.$$

Each iteration k provides a new estimate of state values $V_k(\cdot)$ by looking one step ahead and choosing the best action, asymptotically converging to the optimal value function V^*. Given V^*, a stationary optimal policy π^* is obtained from the Bellman equations.

To adapt value iteration to the risk-sensitive setting, we incorporate the risk-parameter γ into the iteration scheme. Subsequently, we develop a value iteration algorithm presented in Algorithm 1 that produces an approximation to the optimal value function of the RMDP. We initialize the algorithm with v^0 as a starting approximation of $v^*(x;\gamma)$ (Line 2 of Algorithm 1) defined as

$$v^0(x;\gamma) = \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\|c\|}{1-\beta}\right). \tag{7}$$

The algorithm then generates a sequence of approximations $v^n, n = 1, \ldots, N$, by successively applying the operator T (defined in (3)) to the previous iterate v^{n-1} (Line 6 of Algorithm 1). The algorithm concludes by selecting $d^N(x)$, the minimizing actions with respect to v^N (Line 11 of Algorithm 1). Algorithm 1 contains two types of superscripts: the first is the exponent i in β^i, whereas the second is the iteration counter n in v^n and N in d^N. We emphasize this distinction, and note that the same types of superscripts are also used in subsequent sections.

3.1 Stopping condition.

Given a tolerance $\epsilon > 0$, we define the number of iterations N_ϵ that guarantees that the final iterate of the algorithm yields an ϵ-approximation to the optimal value function. The parameter N_ϵ depends on the sign of γ, so that $N_\epsilon = n^+(\epsilon)$ if $\gamma > 0$ and $N_\epsilon = n^-(\epsilon)$ if $\gamma < 0$, where:

$$n^+(\epsilon) = \begin{cases} \left\lceil \log_\beta\left(\frac{1-\beta}{\gamma\|c\|}\ln\left(1+\epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right)\right)\right)\right\rceil, & \text{if } \epsilon < \left[\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)-1\right]\exp\left(\frac{\gamma\|c\|}{1-\beta}\right), \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

$$n^-(\epsilon) = \begin{cases} \left\lceil \log_\beta\left(\frac{1-\beta}{\gamma\|c\|}\ln(1-\epsilon)\right)\right\rceil, & \text{if } \epsilon < 1-\exp\left(\frac{\gamma\|c\|}{1-\beta}\right), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

We examine (8) and (9) to make several key observations. For a fixed tolerance ϵ, as $\gamma \downarrow 0$, the threshold specified in (8),

$$\left[\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)-1\right]\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)$$

becomes smaller than ϵ. Similarly, as $\gamma \uparrow 0$, the threshold $1 - \exp\left(\frac{\gamma\|c\|\beta}{1-\beta}\right)$ in (9) becomes smaller than ϵ as well. Thus, for a fixed $\epsilon > 0$, when $|\gamma|$ is sufficiently small, $N_\epsilon = 0$. In other words, when the magnitude of the risk parameter is small, the initial iterate v^0 is already an ϵ-approximation to the optimal value function without requiring any additional iterations of the algorithm. A parallel argument applies when γ is fixed and ϵ increases: for a fixed $\gamma \in \Gamma$, if ϵ is sufficiently large, then $N_\epsilon = 0$ because the initial value function v^0 is a close-enough approximation to v^*.

On the other hand, standard arguments show that $n^+(\epsilon)$ and $n^-(\epsilon)$ increase monotonically to ∞ as ϵ approaches 0. That is, more iterations of the algorithm are needed to get more accurate approximations of the optimal value function. The derivations are straightforward and therefore omitted.

Finally, a simple asymptotic expansion of (8) and (9) shows that the number of iterations needed is logarithmic in the inverse tolerance. Indeed, for fixed nonzero γ, the arguments inside the natural logarithms in both $n^+(\epsilon)$ and $n^-(\epsilon)$ are proportional to ϵ as $\epsilon \downarrow 0$. Since $\beta \in (0, 1)$, taking $\log_\beta$ of such a term yields a quantity of order $\ln(1/\epsilon)$. Therefore,

$$n^+(\epsilon) = \Theta\left(\ln \frac{1}{\epsilon}\right) \quad \text{for } \gamma > 0, \qquad n^-(\epsilon) = \Theta\left(\ln \frac{1}{\epsilon}\right) \quad \text{for } \gamma < 0.$$

Thus, the required number of iterations grows logarithmically in $1/\epsilon$ as $\epsilon \downarrow 0$.

Collectively, these observations clarify the relationship between γ, ϵ, and the number of iterations N_ϵ required to achieve a specified approximation accuracy in the value function, thereby illustrating how the algorithm's convergence behavior depends jointly on the risk parameter γ and the desired precision level ϵ.

3.2 Convergence Analysis

To establish convergence of Algorithm 1, we define the approximation error $\mathcal{E}_n(x)$, $x \in \mathcal{S}$, as the absolute gap between the optimal value function and the approximation v^n obtained by the n^{th} iteration of the algorithm, i.e.,

$$\mathcal{E}_n(x) = |v^n(x;\gamma) - v^*(x;\gamma)| = \left|T^n v^0(x;\gamma) - v^*(x;\gamma)\right|, \ \forall\, x \in \mathcal{S}. \tag{10}$$

Lemma 5 provides upper bounds on the approximation error $\mathcal{E}_n(x)$.

Lemma 5. *Let $\mathcal{E}_n(x)$ be the approximation error defined in* (10)*. Then,*

(a) If $\gamma > 0$, $\mathcal{E}_n(x) < \exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\left[\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right) - 1\right]$ for all $x \in \mathcal{S}$.

(b) If $\gamma < 0$, $\mathcal{E}_n(x) < \left[1 - \exp\left(\frac{\gamma\|c\|\beta^n}{1-\beta}\right)\right]$ for all $x \in \mathcal{S}$.

We use Lemma 5 to derive the stopping conditions in Algorithm 1 as summarized in Theorem 2.

Theorem 2. *Given $\epsilon > 0$, the approximation error after $N = N_\epsilon$ iterations satisfies $\mathcal{E}_N(x) \leq \epsilon$ for all $x \in \mathcal{S}$.*

Algorithm 1 Approximate Value Iteration for RMDPs

Input: $\epsilon > 0$

Set $N = N_\epsilon$.

Define

$$v^0(x;\gamma\beta^i) = \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\beta^i\|c\|}{1-\beta}\right), \qquad \text{for all } x \in \mathcal{S},\ 0 \le i \le N.$$

for $n = 1, 2, \ldots, N$ **do**

  for $i = 0, 1, \ldots, N-n$ **do**

    for all $x \in \mathcal{S}$ **do**

      Compute

$$v^n(x;\gamma\beta^i) = \min_{a\in\mathcal{A}}\left\{\exp\left(\gamma\beta^i c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)v^{n-1}(y;\gamma\beta^{i+1})\right\}.$$

    end for

  end for

end for

for all $x \in \mathcal{S}$ and $n = 1, \ldots, N$ **do**

  Choose

$$d^N(x;\gamma\beta^n) \in \arg\min_{a\in\mathcal{A}}\left\{\exp\left(\gamma\beta^n c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)v^{n-1}(y;\gamma\beta^{n+1})\right\}.$$

end for

Output: d^N

Proof. Proof. To compute $n^+(\epsilon)$ and $n^-(\epsilon)$, we find the smallest number of iterations n needed to ensure that the approximation error bounds obtained in Lemma 5 are smaller than ϵ. In particular, when $\gamma > 0$, we choose n so that

$$\mathcal{E}_n(x) < \exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\left[\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right) - 1\right] < \epsilon.$$

That is,

$$\begin{aligned}
\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right) < 1 + \epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right) \iff\ & \frac{\gamma\beta^n\|c\|}{1-\beta} < \ln\left(1 + \epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right)\right) \\
\iff\ & \beta^n < \frac{1-\beta}{\gamma\|c\|}\cdot\ln\left(1 + \epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right)\right) \qquad (11) \\
\iff\ & n > \log_\beta\left(\frac{1-\beta}{\gamma\|c\|}\cdot\ln\left(1 + \epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right)\right)\right). \qquad (12)
\end{aligned}$$

Such an n always exists because the right-hand side (RHS) of (11) is always positive and $\beta \in (0,1)$. We choose $n^+(\epsilon)$ to be the smallest n that satisfies (12). In particular, $n^+(\epsilon) = 0$

if the RHS of (12) is negative, which is equivalent to

$$\frac{1-\beta}{\gamma\|c\|}\cdot\ln\left(1+\epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right)\right)>1 \iff \epsilon>\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\left[\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)-1\right].$$

That is, no value iteration is needed and the initial guess $v^0(x;\gamma)$ is an ϵ-approximation to the optimal value function when ϵ is large enough. Otherwise, we define

$$n^+(\epsilon)=\left\lceil \log_\beta\left(\frac{1-\beta}{\gamma\|c\|}\cdot\ln\left(1+\epsilon\exp\left(-\frac{\gamma\|c\|}{1-\beta}\right)\right)\right)\right\rceil.$$

Next, for $\gamma<0$, we again invoke Lemma 5 and choose $n^-(\epsilon)$ as the smallest number of iterations needed to achieve $\mathcal{E}_n(x)<\epsilon$ for all $x\in\mathcal{S}$. In particular, we let

$$\mathcal{E}_n\le 1-\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)<\epsilon.$$

Without loss of generality, let $\epsilon<1$. Then,

$$\begin{aligned}1-\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)<\epsilon \iff 1-\epsilon<\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right) &\iff \ln(1-\epsilon)<\frac{\gamma\beta^n\|c\|}{1-\beta}\\ &\iff \beta^n<\frac{1-\beta}{\gamma\|c\|}\ln(1-\epsilon) \qquad (13)\\ &\iff n>\log_\beta\left(\frac{1-\beta}{\gamma\|c\|}\ln(1-\epsilon)\right). \qquad (14)\end{aligned}$$

As in the previous case, we choose $n^-(\epsilon)$ to be the smallest n that satisfies (14). Such an n exists because the RHS of (13) is always positive (as both γ and $\ln(1-\epsilon)<0$). Moreover, we can choose $n^-(\epsilon)=0$ when the RHS of (14) is negative, which is equivalent to

$$\frac{1-\beta}{\gamma\|c\|}\ln(1-\epsilon)>1 \iff \ln(1-\epsilon)<\frac{\gamma\|c\|}{1-\beta} \iff 1-\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)<\epsilon.$$

Therefore, when $\epsilon\le\epsilon^-=1-\exp\left(1-\frac{\gamma\|c\|}{1-\beta}\right)$, we define

$$n^-(\epsilon)=\left\lceil \log_\beta\left(\frac{1-\beta}{\gamma\|c\|}\ln(1-\epsilon)\right)\right\rceil.$$

Finally, we define $N_\epsilon=n^+(\epsilon)$ when $\gamma>0$ and $N_\epsilon=n^-(\epsilon)$ when $\gamma<0$. □

Theorem 2 establishes that the approximate value function, v^N, obtained via Algorithm 1 is an ϵ-approximation to the optimal value function. However, the (approximately) optimal decision-rules computed in Line 11 of Algorithm 1 only prescribe actions for the first N periods. Theorem 3 shows that any infinite-horizon policy constructed by appending arbitrary actions for periods $n>N$ to the policy computed in Algorithm 1, is approximately optimal with the same error tolerance ϵ.

Formally, let $\boldsymbol{d}^N \in \boldsymbol{\Pi}^N$ be the truncated policy from Line 11 of Algorithm 1. Construct an infinite-horizon policy $\bar{\boldsymbol{d}}^N \in \boldsymbol{\Pi}^\infty$ as follows:

$$\bar{d}^N(x;\gamma\beta^n) = \begin{cases} d^N(x;\gamma\beta^n) & n \le N \\ a & n > N, \end{cases} \tag{15}$$

where a is an arbitrary action in $\mathcal{A}$.

Theorem 3. *The policy $\bar{\boldsymbol{d}}^N \in \boldsymbol{\Pi}^\infty$ computed via* (15) *is an ϵ-optimal policy, i.e.,*

$$\left| v^{\bar{d}^N}(x;\gamma) - v^*(x;\gamma) \right| \le \epsilon, \quad \forall\ x \in \mathcal{S},$$

where $v^{\bar{d}^N}(x;\gamma)$ is the value function of the infinite-horizon policy $\bar{\boldsymbol{d}}^N \in \boldsymbol{\Pi}^\infty$ as defined in (1)*.*

Proof. Proof. First, note that $\bar{d}^N$ is a feasible policy. Therefore,

$$v^{\bar{d}^N}(x;\gamma) \ge v^*(x;\gamma). \tag{16}$$

Next, we use induction to show that $v^{\bar{d}^N}(x;\gamma) \le v^N(x;\gamma)$, $\forall\ x \in \mathcal{S}$. By construction, we have

$$v^{\bar{d}^N}(x;\gamma\beta^N) \le \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\beta^N \|c\|}{1-\beta}\right) = v^0(x;\gamma\beta^N).$$

Then,

$$\begin{aligned} v^{\bar{d}^N}(x;\gamma\beta^{N-1}) &= \exp\left(\gamma\beta^N c\left(x, \bar{d}^N\left(x,\gamma\beta^N\right)\right)\right) \sum_{y\in\mathcal{S}} v^{\bar{d}^N}(y;\gamma\beta^N) \\ &= \exp\left(\gamma\beta^N c\left(x, d^N\left(x,\gamma\beta^N\right)\right)\right) \sum_{y\in\mathcal{S}} v^{\bar{d}^N}(y;\gamma\beta^N) \\ &\le \exp\left(\gamma\beta^N c\left(x, d^N\left(x,\gamma\beta^N\right)\right)\right) \sum_{y\in\mathcal{S}} v^0(y;\gamma\beta^N) \\ &= v^1(x;\gamma\beta^{N-1}), \end{aligned}$$

where the first equality holds because $d^N(x;\gamma\beta^i) = \overline{d}^N(x;\gamma\beta^i)$ for $i \le N$ by the definition of $\overline{d}^N$. Repeating this step gives us

$$v^{\bar{d}^N}(x;\gamma) \le v^N(x;\gamma), \quad \forall\ x \in \mathcal{S}. \tag{17}$$

Combining inequalities (16) and (17), we have

$$\begin{aligned} \left| v^{\bar{d}^N}(x;\gamma) - v^*(x;\gamma) \right| &= v^{\bar{d}^N}(x;\gamma) - v^*(x;\gamma) \\ &\le v^N(x;\gamma) - v^*(x;\gamma) = \left| v^N(x;\gamma) - v^*(x;\gamma) \right| \\ &\le \epsilon. \end{aligned}$$

□

Remark 3. An important consequence of Theorem 3 is that the approximate policy generated by Algorithm 1 can be used (as in (15)) to construct a near-optimal policy that is ultimately stationary. That is, although the policy may be nonstationary over a finite initial segment of the horizon, it eventually settles on a stationary decision rule and remains stationary thereafter. This mirrors the classical result in Jaquette [13], which shows that, in discounted exponential-utility MDPs, an infinite-horizon optimal policy need not be stationary from the outset. Furthermore, recall that this initial nonstationarity is quantitatively limited: the length of the nonstationary prefix grows only logarithmically in the inverse tolerance, that is, as $O(\ln(1/\epsilon))$. Thus, after a finite and slowly growing (as $\epsilon \downarrow 0$) number of stages, the approximate policy becomes stationary.

4 LP for RMDPs

In this section, we introduce an LP for RMDPs. It is well known that a risk-neutral MDP can be modeled as an LP with state values as decision variables and the Bellman optimality condition enforced as constraints [7]. The formulation maximizes the sum of (weighted) state values subject to constraints ensuring that each state's value is no more than the immediate cost plus the discounted future value for every possible action. Furthermore, while this primal LP has a natural value-function interpretation, the dual LP admits an elegant interpretation in terms of occupation measures for state-action pairs.

The LP formulation of RMDPs is less transparent than its risk-neutral counterpart. Under exponential utility, the Bellman equations become multiplicative, so the familiar value-function/occupation-measure duality from risk-neutral MDPs no longer applies directly [15]. Nevertheless, Kumar et al. [15] show that, for finite-horizon RMDPs, the LP variable can still be interpreted as the value function.

In the discounted infinite-horizon setting, an analogous interpretation holds but only for the finite-dimensional approximation. Because the horizon is infinite, the exact problem cannot be represented by the same finite LP used in the finite-horizon setting; rather, one truncates the problem at an approximation level N, leading to an LP whose size grows with N. An optimal dual solution can therefore be interpreted as identifying actions that are greedy with respect to the approximate value functions computed by the primal LP. However, two additional steps are required to turn this observation into a guarantee for the original infinite-horizon problem. First, one must upper-bound the approximation error between the value functions obtained from the primal LP and the true optimal value function. Second, one must show that the policy obtained by following the greedy actions from the LP, together with an appended tail policy for the remaining infinite horizon, inherits the same approximation guarantee. This is the key distinction between the finite-horizon and discounted infinite-horizon settings.

In this section, we present our LP formulation for discounted infinite-horizon RMDPs. We first show that the LP variables indexed by states coincide with the approximate value functions generated by Algorithm 1. Consequently, an optimal solution of the LP yields an ϵ-approximation of the optimal value function v^*. Therefore, by Theorem 3, the greedy policy induced by this optimal LP solution is itself ϵ-optimal.

Given $\epsilon > 0$, let N_ϵ be the number of iterations in Algorithm 1. Fix $N \leq N_\epsilon$. For every

$x \in \mathcal{S}$, define decision variables $h^n(x;\gamma\beta^i)$ for $0 \le i \le N-n$, $1 \le n \le N$. Next, consider the following LP.

$$\bar{Z}_N = \max \sum_{x\in\mathcal{S}} h^N(x;\gamma) \tag{18a}$$

$$\text{s.t. } h^0(x;\gamma\beta^i) = \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\beta^i\|c\|}{1-\beta}\right), \quad \forall x \in \mathcal{S},\ 0 \le i \le N, \tag{18b}$$

$$h^n(x;\gamma\beta^i) \le \exp\big(\gamma\beta^i c(x,a)\big)\sum_{y\in S} P^a(x,y)h^{n-1}(y;\gamma\beta^{i+1}),$$

$$\forall x \in \mathcal{S}, a \in \mathcal{A},\ 1 \le n \le N,\ 0 \le i \le N-n. \tag{18c}$$

Theorem 4. *Given $\epsilon > 0$ and $N \le N_\epsilon$, let v^n, $0 \le n \le N$ be the sequence of approximate value functions computed from Algorithm 1 in N iterations. Define*

$$\tilde{h}^n(x;\gamma\beta^i) = v^n(x;\gamma\beta^i),\ \forall x \in \mathcal{S},\ 0 \le n \le N,\ 0 \le i \le N-n. \tag{19}$$

Then, $\tilde{h}$ is an optimal solution of (18) *and*

$$\bar{Z}_N = \sum_{x\in\mathcal{S}} v^N(x;\gamma).$$

Proof. Proof. Given $\epsilon > 0$, fix $N \le N_\epsilon$. Suppose $\bar{h}^n$, $0 \le n \le N$ is an optimal solution of the LP. We first use induction on n to prove that $\bar{h}^n(x;\gamma\beta^i) \le v^n(x;\gamma\beta^i)$ for every $x \in \mathcal{S}$, $1 \le n \le N$ and $0 \le i \le N-n$.

Base case: We will show that

$$\bar{h}^1(x;\gamma\beta^i) \le v^1(x;\gamma\beta^i), \quad \forall\ x \in \mathcal{S}, 0 \le i \le N-1.$$

By definition, $\bar{h}^0(x;\gamma\beta^i) = v^0(x;\gamma\beta^i)$ for all $x \in \mathcal{S}$ and $0 \le i \le N$. For every $x \in \mathcal{S}$ and $0 \le i \le N-1$, we have

$$\bar{h}^1(x;\gamma\beta^i) \le \exp\left(\gamma c(x,a)\right)\sum_{y\in S} P^a(x,y)\bar{h}^0(y;\gamma\beta^{i+1}), \quad \forall\ a \in \mathcal{A},$$

which implies that

$$\begin{aligned}\bar{h}^1(x;\gamma\beta^i) &\le \min_{a\in\mathcal{A}}\left\{\exp\big(\gamma\beta^i c(x,a)\big)\sum_{y\in\mathcal{S}} P^a(x,y)\bar{h}^0(y;\gamma\beta^{i+1})\right\}\\ &= \min_{a\in\mathcal{A}}\left\{\exp\big(\gamma\beta^i c(x,a)\big)\sum_{y\in\mathcal{S}} P^a(x,y)v^0(y;\gamma\beta^{i+1})\right\} = v^1(x;\gamma\beta^i).\end{aligned}$$

Inductive step: Suppose that the inequality holds for some n such that $1 < n < N$, i.e.,

$$\bar{h}^n(x;\gamma\beta^i) \le v^n(x;\gamma\beta^i)\ \forall x \in \mathcal{S},\ 0 \le i \le N-1.$$

Then, for every $x \in \mathcal{S}$ and $0 \le i \le N-1$,

$$\bar{h}^{n+1}(x;\gamma\beta^i) \le \exp\left(\gamma c(x,a)\right)\sum_{y\in S} P^a(x,y)\bar{h}^n(y;\gamma\beta^{i+1}), \quad \forall\ a \in \mathcal{A},$$

which is equivalent to

$$\begin{aligned}\bar{h}^{n+1}(x;\gamma\beta^i) &\le \min_{a\in\mathcal{A}}\left\{\exp\left(\gamma\beta^i c(x,a)\right)\sum_{y\in\mathcal{S}}P^a(x,y)\bar{h}^n(y;\gamma\beta^{i+1})\right\}\\ &\le \min_{a\in\mathcal{A}}\left\{\exp\left(\gamma\beta^i c(x,a)\right)\sum_{y\in\mathcal{S}}P^a(x,y)v^n(y;\gamma\beta^{i+1})\right\} = v^{n+1}(x;\gamma\beta^i).\end{aligned}$$

Thus, $\bar{h}^n(x;\gamma\beta^i) \le v^n(x;\gamma\beta^i)$ for every $x \in \mathcal{S}$, $0 \le i \le N-n$, $1 \le n \le N$. This further implies that $\bar{Z}_N = \sum_{x\in\mathcal{S}}\bar{h}^N(x;\gamma) \le \sum_{x\in\mathcal{S}} v^N(x;\gamma)$.

Now, we prove that $\tilde{h}$ defined in (19) is a feasible solution of the LP.

First, by definition (Line 2 of Algorithm 1), we know that $v^0(x;\gamma\beta^i) = \mathrm{sgn}(\gamma)\exp\left(\frac{\gamma\beta^i\|c\|}{1-\beta}\right)$ $\forall\ x \in \mathcal{S}$ and $0 \le i \le N$. Therefore, constraint (18b) is satisfied. Then, from Line 6 of Algorithm 1, we also have for every $x \in \mathcal{S}$ that

$$v^n(x;\gamma\beta^i) = \min_{a\in\mathcal{A}}\left\{\exp\left(\gamma\beta^i c(x,a)\right)\sum_{y\in\mathcal{S}}P^a(x,y)v^{n-1}(y;\gamma\beta^{i+1})\right\},\ \forall 1 \le n \le N, 0 \le i \le N-n.$$

This implies that $\tilde{h}^n$ satisfies constraint (18c). Thus, $\tilde{h}$ is a feasible solution of the LP. It follows that $\sum_{x\in\mathcal{S}} v^N(x;\gamma) = \sum_{x\in\mathcal{S}}\tilde{h}^N(x;\gamma) \le \bar{Z}_N$.

Therefore,

$$\bar{Z}_N = \sum_{x\in\mathcal{S}} v^N(x;\gamma) = \sum_{x\in\mathcal{S}}\tilde{h}^N(x;\gamma),$$

and $\tilde{h}$ defined in (19) is an optimal solution of (18). □

Corollary 1. *Given $\epsilon > 0$ and $N = N_\epsilon$, the optimal solution $\bar{h}$ of* (18) *is an ϵ-approximation of the optimal value function. That is,*

$$|\bar{h}^N(x;\gamma) - v^*(x;\gamma)| \le \epsilon\ \forall x \in \mathcal{S}.$$

Proof. Proof. The result follows from Theorems 2 and 4. □

5 Policy Iteration for RMDPs

In this section, we introduce an approximate policy iteration algorithm for RMDPs and establish its convergence. The standard policy iteration algorithm for risk-neutral MDPs alternates between two main steps: policy evaluation and policy improvement. During policy evaluation, the algorithm calculates the value function associated with the current policy by solving a system of linear equations derived from the Bellman equation. The policy improvement step then constructs a new policy by selecting the action that maximizes the expected immediate reward plus the discounted future value at each state. In every iteration, the policy improvement step either finds a strictly better policy (improving the value function) or confirms that the current policy is already optimal. Because there are only finitely many deterministic policies, the algorithm terminates with an optimal policy. Finite convergence is guaranteed, typically taking many fewer iterations than the total number of possible policies, making policy iteration both theoretically sound and practically efficient [19].

As in value iteration, we adapt policy iteration to our RMDP framework by incorporating the risk parameter into the iteration scheme. Algorithm 2 presents our proposed policy iteration method. The algorithm initializes a random policy d_0 and sets the initial choice for value functions u^0 in Line 2. Each iteration comprises approximate variants of of policy evaluation (Line 5) and policy improvement (Line 8). The algorithm terminates with a near-optimal policy that satisfies a prescribed error tolerance.

Algorithm 2 may be interpreted as an instance of generalized policy iteration (Sutton and Barto [21, Chapter 4.6]). A generalized policy iteration algorithm maintains both a policy and a value function, updating the policy to be greedy with respect to the current value estimate while driving the value estimate toward the value function associated with the current policy. Specifically, in each iteration of Algorithm 2, the current policy is approximately evaluated, and the next policy is then chosen to be greedy with respect to the resulting approximate value function.

Algorithm 2 Approximate Policy Iteration for RMDPs

Input: $\epsilon > 0$

1: Set $N = N_\epsilon^{\text{pol}}$.

2: Initialize a random policy d_0 and a value function

$$u^0(x;\gamma\beta^i) = \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\beta^i\|c\|}{1-\beta}\right), \quad \text{for all } x \in \mathcal{S},\ 0 \le i \le N.$$

3: **for** $k = 1, 2, \dots, N$ **do**

4:     **for all** $x \in \mathcal{S}$ **do**

5:         Compute

$$u^k(x;\gamma\beta^i) = \begin{cases} \exp\big(\gamma\beta^i c\big(x, d_{k-1}(x,\gamma\beta^i)\big)\big)\sum_{y\in\mathcal{S}} P^{d_{k-1}(x,\gamma\beta^i)}(x,y)u^k(y;\gamma\beta^{i+1}), & i \le k, \\ u^0(x;\gamma\beta^i), & i > k. \end{cases}$$

6:     **end for**

7:     **for all** $x \in \mathcal{S}$ **do**

8:         Choose d_k as follows.

$$\begin{aligned} i > k: &\quad d_k(x;\gamma\beta^i) = d_0(x;\gamma\beta^i), \\ i \le k: &\quad d_k(x;\gamma\beta^i) \in \arg\min\nolimits_{a\in\mathcal{A}}\left\{\exp\left(\gamma\beta^i c(x,a)\right)\sum\nolimits_{y\in\mathcal{S}} P^a(x,y)u^k(y;\gamma\beta^{i+1}),\right\}, \\ &\qquad \text{with } d_k(x;\gamma\beta^i) = d_{k-1}(x;\gamma\beta^i) \text{ if the latter belongs to the argmin set.} \end{aligned}$$

9:     **end for**

10: **end for**

Output: d_N

Note that the RMDP can be reformulated as a countably infinite risk-neutral MDP with state space $(x, \gamma\beta^i) : x \in \mathcal{S}, i \in \mathbb{N} \cup \{\infty\}$. However, related algorithms for countably infinite risk-neutral MDPs require either a contraction property of the Bellman operator or the primal-dual correspondence available in risk-neutral MDPs [16, 19, 22]. Neither property holds for general RMDPs. As such, these methods are not directly applicable in our setting.

5.1 Stopping Condition.

Similar to Algorithm 1, we define the number of iterations N_ϵ^{pol} that will be performed given a tolerance $\epsilon > 0$. N_ϵ^{pol} depends on the sign of γ, so that $N_\epsilon^{\text{pol}} = N^+(\epsilon)$ if $\gamma > 0$ and $N_\epsilon^{\text{pol}} = N^-(\epsilon)$ if $\gamma < 0$, where

$$N^+(\epsilon) = \begin{cases} 2\left\lceil \log_\beta \left(\ln \left(\epsilon \exp\left(\frac{-\gamma\|c\|}{1-\beta} \right) + 1 \right) \frac{1-\beta}{\gamma\|c\|} \right) - 1 \right\rceil, & \text{if } \epsilon < \left[\exp\left(\frac{\gamma\|c\|\beta}{1-\beta} \right) - 1 \right] \exp\left(\frac{\gamma\|c\|}{1-\beta} \right), \\ 0, & \text{otherwise,} \end{cases} \tag{20}$$

$$N^-(\epsilon) = \begin{cases} 2\left\lceil \log_\beta \left(\ln \left(1 - \epsilon \exp\left(\frac{-\gamma\|c\|}{1-\beta} \right) \right) \frac{1-\beta}{\gamma\|c\|} \right) - 1 \right\rceil, & \text{if } \epsilon < \left[1 - \exp\left(\frac{\gamma\|c\|}{1-\beta} \right) \right] \exp\left(\frac{\gamma\|c\|}{1-\beta} \right), \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Several observations arise upon inspecting (20) and (21). For a fixed ϵ, as $\gamma \downarrow 0$, the threshold

$$\left[\exp\left(\frac{\gamma\|c\|\beta}{1-\beta} \right) - 1 \right] \exp\left(\frac{\gamma\|c\|}{1-\beta} \right)$$

in (20) becomes smaller than ϵ. Similarly, as $\gamma \uparrow 0$, the threshold

$$\left[1 - \exp\left(\frac{\gamma\|c\|}{1-\beta} \right) \right] \exp\left(\frac{\gamma\|c\|}{1-\beta} \right)$$

in (21) falls below ϵ. Therefore, for $|\gamma|$ small enough, $N_\epsilon^{\text{pol}} = 0$. In other words, the value function corresponding to the random policy d_0 initialized in Line 2 is sufficient to satisfy the prescribed error tolerance ϵ without the need for further improvement. A parallel conclusion holds when γ is fixed and ϵ is increased.

As in Section 3, a simple asymptotic expansion of (20) and (21) shows that the number of iterations needed is only logarithmic in the inverse tolerance. Indeed, for fixed nonzero γ, the arguments inside the natural logarithms in both $N^+(\epsilon)$ and $N^-(\epsilon)$ are proportional to ϵ as $\epsilon \downarrow 0$. Since $\beta \in (0, 1)$, taking $\log_\beta$ of such a term yields a quantity of order $\ln(1/\epsilon)$. Therefore,

$$N^+(\epsilon) = \Theta\left(\ln \frac{1}{\epsilon} \right) \quad \text{for } \gamma > 0, \qquad N^-(\epsilon) = \Theta\left(\ln \frac{1}{\epsilon} \right) \quad \text{for } \gamma < 0.$$

Thus, the required number of iterations grows only logarithmically in $1/\epsilon$ as $\epsilon \downarrow 0$.

5.2 Convergence Analysis

In our policy iteration algorithm, we implicitly transform the MDP into one with a countably infinite state space: $\mathcal{S}' = \{\{(x, \gamma\beta^i)\}_{x\in\mathcal{S}}\}_{i\geq 0}$. For convenience, we denote $\mathcal{S}'_j = \{\{(x, \gamma\beta^i)\}_{x\in\mathcal{S}}\}_{i\leq j}$ the finite, truncated state space. We start with an initial guess policy d_0 defined over $\mathcal{S}'$. In every iteration k, we first compute a function u^k defined over $\mathcal{S}'$, such that u^k approximates the value of policy d_k. For $s \in \mathcal{S}'_k$, we compute $u^k(s)$ according to Line 5 of Algorithm 2. Then, for $s \in \mathcal{S}' \setminus \mathcal{S}'_k$, we set $u^k(s) = u^0(s)$.

To establish convergence of the algorithm, we first show that the sequence of approximate value functions u^k is non-increasing and bounded below by the optimal value function v^*. We then find an explicit upper bound on the error $|u^k(s) - v^*(s)|$ that vanishes aymptotically as $k \to \infty$ and is used to derive the stopping condition from Section 5.1. As such, Algorithm 2 produces a sequence of policies whose values converge to the optimal value function. We note, however, that the policies are not guaranteed to be monotonically improving, as the monotonicity result only holds for the approximate value functions u^k.

Theorem 5. *Let u^k be the approximate value function computed in iteration k of Algorithm 2. The approximation error is bounded above as follows.*

$$|u^k(x;\gamma) - v^*(x;\gamma)| \leq \exp\left(\gamma\|c\|\frac{1-\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right)\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right) - 1\right|, \quad \forall x \in \mathcal{S}.$$

Proof. Proof. See Appendix C.1. □

As a consequence of the error bound derived in Theorem 5, Corollary 2 asserts that the approximate value functions calculated in Algorithm 2 converge to the optimal value function v^*.

Corollary 2. *Let u^k be the approximate value function computed in iteration k of Algorithm 2. Then, for every $x \in \mathcal{S}$, $u^k(x;\gamma) \to v^*(x;\gamma)$ as $k \to \infty$.*

Proof. Proof. In the upper bound on the approximation error derived in Theorem 5, the first term $\exp\left(\gamma\|c\|\frac{1-\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right)$ is bounded above by $\exp\left(\gamma\|c\|\frac{1}{1-\beta}\right)$ when $\gamma > 0$, and by 1 when $\gamma < 0$ because $\beta \in [0, 1)$. The second term $\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right) - 1\right|$ converges to 0 as $k \to \infty$ for all $\gamma \in \Gamma$. Therefore, the error bound approaches 0 and u^k converges to v^* as $k \to \infty$. □

Theorem 5 provides an upper bound on the error as a function of the number of iterations. By inverting this bound, we can determine a sufficient number of iterations needed to ensure that the error is below a prescribed tolerance.

Theorem 6. *Given $\epsilon > 0$ and for $N = N_\epsilon^{\mathrm{pol}}$, the approximate value function u^N from Algorithm 2 is an ϵ-approximation of the optimal value function. That is,*

$$|u^N(x;\gamma) - v^*(x;\gamma)| \leq \epsilon, \quad \forall x \in \mathcal{S}.$$

Proof. Proof. From Theorem 5, we have that $\left|u^N(x;\gamma) - v^*(x;\gamma)\right| \leq \epsilon$ for all $x \in \mathcal{S}$ if

$$\exp\left(\gamma\|c\|\frac{1-\beta^{\lfloor \frac{N}{2}\rfloor+1}}{1-\beta}\right)\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor \frac{N}{2}\rfloor+1}}{1-\beta}\right)-1\right| \leq \epsilon. \tag{22}$$

Note that Equation (22) holds if N satisfies

$$\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor \frac{N}{2}\rfloor+1}}{1-\beta}\right)-1\right| \leq \epsilon,$$

because $\beta \in [0,1)$. Hence, it will be sufficient to show that N satisfies

$$\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor \frac{N}{2}\rfloor+1}}{1-\beta}\right)-1\right| \leq \epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right). \tag{23}$$

We proceed by discussing the two cases of positive and negative γ separately.

Part I: $\gamma > 0$. We need

$$\exp\left(\gamma\|c\|\frac{\beta^{\lfloor \frac{N}{2}\rfloor+1}}{1-\beta}\right)-1 \leq \epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right).$$

Rearranging and taking the natural log on both sides, this is equivalent to

$$\gamma\|c\|\frac{\beta^{\lfloor \frac{N}{2}\rfloor+1}}{1-\beta} \leq \ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right).$$

Again, rearranging gives

$$\beta^{\lfloor \frac{N}{2}\rfloor+1} \leq \ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}.$$

Finally, taking the log with base β on both sides, we have

$$\left\lfloor\frac{N}{2}\right\rfloor + 1 \geq \log_\beta\left(\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\right).$$

Simplifying gives us a lower-bound on N as

$$N \geq 2\left\lceil\log_\beta\left(\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\right)-1\right\rceil.$$

Therefore, we set N at the lower-bound of $2\lceil\log_\beta\left(\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\right)-1\rceil$ to satisfy (23)

Moreover, we recognize that the RHS of (23) becomes nonpositive when ϵ exceeds a certain threshold, in which case we set N to 0. Next, we identify this threshold.

Observe that if

$$2\left\lceil\log_\beta\left(\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\right)-1\right\rceil \leq 0,$$

we have

$$\log_\beta\left(\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\right)\le 1,$$

or equivalently,

$$\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\le \beta.$$

Multiplying both sides by $\frac{\gamma\|c\|}{1-\beta}$ and applying the exponential function, we obtain

$$\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\le\exp\left(\frac{\gamma\|c\|\beta}{1-\beta}\right).$$

Therefore,

$$\epsilon\le\left[\exp\left(\frac{\gamma\|c\|\beta}{1-\beta}\right)-1\right]\exp\left(\frac{\gamma\|c\|}{1-\beta}\right).$$

To summarize, we let

$$N^+(\epsilon)=\begin{cases}2\left\lceil\log_\beta\left(\ln\left(\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)+1\right)\frac{1-\beta}{\gamma\|c\|}\right)-1\right\rceil, & \epsilon<\left[\exp\left(\frac{\gamma\|c\|\beta}{1-\beta}\right)-1\right]\exp\left(\frac{\gamma\|c\|}{1-\beta}\right),\\ 0, & \text{otherwise.}\end{cases}$$

This concludes the proof for $\gamma>0$.

Part II: $\gamma<0$. We need

$$1-\exp\left(\gamma\|c\|\frac{\beta^{\lfloor\frac{N}{2}\rfloor+1}}{1-\beta}\right)\le\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right).$$

Rearranging gives

$$\exp\left(\gamma\|c\|\frac{\beta^{\lfloor\frac{N}{2}\rfloor+1}}{1-\beta}\right)\ge 1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right). \tag{24}$$

Notice that the inequality holds trivially if the RHS of (24) is nonpositive, i.e., $1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\le 0$. Simplifying the inequality, we get $\epsilon\ge\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)$. In this case, it suffices to set $N=0$. Therefore, we assume $\epsilon<\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)$. Taking natural log in Inequality (24), we get that (24) is equivalent to

$$\gamma\|c\|\frac{\beta^{\lfloor\frac{N}{2}\rfloor+1}}{1-\beta}\ge\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right).$$

Rearranging the terms, we obtain

$$\beta^{\lfloor\frac{N}{2}\rfloor+1}\le\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right)\frac{1-\beta}{\gamma\|c\|}. \tag{25}$$

Again, (25) holds trivially if the RHS is at least 1, i,e., $\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right)\frac{1-\beta}{\gamma\|c\|}\ge 1$. Simplifying the expression, we obtain $\epsilon\ge\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)-\exp\left(\frac{2\gamma\|c\|}{1-\beta}\right)$. In this case, it suffices to set $N=0$. Therefore, we assume $\epsilon < \exp\left(\frac{\gamma\|c\|}{1-\beta}\right) - \exp\left(\frac{2\gamma\|c\|}{1-\beta}\right)$. Then, taking log with respect to β in (25), the inequality holds if

$$\left\lfloor \frac{N}{2} \right\rfloor + 1 \geq \log_\beta\left(\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right)\frac{1-\beta}{\gamma\|c\|}\right).$$

After simplification, we obtain

$$N \geq 2\left\lceil \log_\beta\left(\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right)\frac{1-\beta}{\gamma\|c\|}\right) - 1\right\rceil.$$

Therefore, we can set N to the smallest value, which is $2\left\lceil \log_\beta\left(\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right)\frac{1-\beta}{\gamma\|c\|}\right) - 1\right\rceil$. Notice that

$$\begin{aligned}&\min\left(\exp\left(\frac{\gamma\|c\|}{1-\beta}\right) - \exp\left(\frac{2\gamma\|c\|}{1-\beta}\right), \exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\right)\\ =&\exp\left(\frac{\gamma\|c\|}{1-\beta}\right) - \exp\left(\frac{2\gamma\|c\|}{1-\beta}\right) = \left[1-\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\right]\exp\left(\frac{\gamma\|c\|}{1-\beta}\right).\end{aligned}$$

This gives us

$$N^-(\epsilon) = \begin{cases} 2\left\lceil \log_\beta\left(\ln\left(1-\epsilon\exp\left(\frac{-\gamma\|c\|}{1-\beta}\right)\right)\frac{1-\beta}{\gamma\|c\|}\right) - 1\right\rceil, \epsilon < \left[1-\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\right]\exp\left(\frac{\gamma\|c\|}{1-\beta}\right), \\ 0, \text{otherwise.}\end{cases}$$

This concludes the proof for negative γ. □

Recall that u^N is only an approximation of the value of policy d_N. Theorem 7 shows that the true value function of d_N also satisfies the same error bound from Theorem 6.

Theorem 7. *Given $\epsilon > 0$ and $N = N_\epsilon^{\text{pol}}$, let d_N be the policy obtained from Algorithm 2 in iteration N. Let $\tilde{u}^N$ be the value function of policy d_N. Then, $\tilde{u}^N$ is an ϵ-approximation to the optimal value function. That is,*

$$\left|\tilde{u}^N(x;\gamma) - v^*(x;\gamma)\right| \leq \epsilon \ \forall x \in \mathcal{S}.$$

Proof. Proof. Recall from Theorem 6 that

$$\left|u^N(x;\gamma) - v^*(x;\gamma)\right| \leq \epsilon, \qquad \forall x \in \mathcal{S}.$$

For any $x \in \mathcal{S}$, $j \geq 0$,

$$\begin{aligned}\tilde{u}^N(x;\gamma\beta^j) &= \mathbb{E}\left[\exp\left(\gamma\beta^j\sum_{i=0}^{\infty}\beta^i c(x_i, d(x_i;\gamma\beta^{j+i}))\right)\right]\\ &\geq \min_{(\boldsymbol{\pi}^\infty)}\left\{\mathbb{E}\left[\exp\left(\gamma\beta^j\sum_{i=0}^{\infty}\beta^i c(x_i, \pi_i^\infty(x_i))\right)\right]\right\} = v^*(x;\gamma\beta^j).\end{aligned}$$

Next, we show that $\tilde{u}^N(x;\gamma\beta^j) \leq u^N(x;\gamma\beta^j)$, $\forall\ x \in \mathcal{S}$, $j \geq 0$.

When $j > N$ and $\gamma > 0$,

$$\begin{aligned}
\tilde{u}^N(x;\gamma\beta^j) &= \mathbb{E}\left[\exp\left(\gamma\beta^j\sum_{i=0}^{\infty}\beta^i c(x_i, d(x_i;\gamma\beta^{j+i}))\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\gamma\beta^j\sum_{i=0}^{\infty}\beta^i\|c\|\right)\right] \\
&= \exp\left(\gamma\|c\|\frac{\beta^j}{1-\beta}\right) = u^N(x;\gamma\beta^j).
\end{aligned}$$

Similarly, when $\gamma < 0$,

$$\begin{aligned}
\tilde{u}^N(x;\gamma\beta^j) &= \mathbb{E}\left[-\exp\left(\gamma\beta^j\sum_{i=0}^{\infty}\beta^i c(x_i, d(x_i,\gamma\beta^{j+i}))\right)\right] \\
&\leq -\mathbb{E}\left[\exp\left(\gamma\beta^j\sum_{i=0}^{\infty}\beta^i\|c\|\right)\right] \\
&= -\exp\left(\gamma\|c\|\frac{\beta^j}{1-\beta}\right) = u^N(x;\gamma\beta^j).
\end{aligned}$$

When $j \leq N$, we prove the inequality using a recursive argument.

First, consider $j = N$. For any $x \in \mathcal{S}$, we have

$$\begin{aligned}
\tilde{u}^N(x;\gamma\beta^N) &= \exp\left(\gamma\beta^N c\left(x, d_N\left(x,\gamma\beta^N\right)\right)\right)\sum_{y\in\mathcal{S}}\tilde{u}^N(y;\gamma\beta^{N+1}) \\
&\leq \exp\left(\gamma\beta^N c\left(x, d_N\left(x,\gamma\beta^N\right)\right)\right)\sum_{y\in\mathcal{S}}u^N(y;\gamma\beta^{N+1}) \\
&= u^N(x;\gamma\beta^N).
\end{aligned}$$

We then prove the inequality for $j-1$.

$$\begin{aligned}
\tilde{u}^N(x;\gamma\beta^{j-1}) &= \exp\left(\gamma\beta^{j-1} c\left(x, d_N\left(x,\gamma\beta^{j-1}\right)\right)\right)\sum_{y\in\mathcal{S}}\tilde{u}^N(y;\gamma\beta^j) \\
&\leq \exp\left(\gamma\beta^{j-1} c\left(x, d_N\left(x,\gamma\beta^{j-1}\right)\right)\right)\sum_{y\in\mathcal{S}}u^N(y;\gamma\beta^j) \\
&= u^N(x;\gamma\beta^j).
\end{aligned}$$

Repeatedly applying this argument proves that $\tilde{u}^N(x;\gamma\beta^j) \leq u^N(x;\gamma\beta^j)$ for $j \leq N$.

It follows that

$$\begin{aligned}
\left|\tilde{u}^N(x;\gamma) - v^*(x;\gamma)\right| &= \tilde{u}^N(x;\gamma) - v^*(x;\gamma) \\
&\leq u^N(x;\gamma) - v^*(x;\gamma) \\
&= \left|u^N(x;\gamma) - v^*(x;\gamma)\right| \leq \epsilon.
\end{aligned}$$

□

Similar to the value iteration case, Theorem 7 shows that the policy produced by Algorithm 2 is eventually stationary. Specifically, the number of iterations required before the policy stabilizes is bounded by a term of order $O(\ln(1/\epsilon))$, after which Algorithm 2 repeatedly selects the same stationary decision rule.

6 Conclusion

In this paper, we developed approximate solution methods for RMDPs. We presented a risk-sensitive variant of value iteration to solve RMDPs and provided convergence results along with performance guarantees. We further developed an LP formulation of RMDPs and proved its equivalence to optimal value functions obtained from the proposed value iteration method. Finally, we developed a policy iteration algorithm to solve RMDPs and provided performance guarantees for the resulting value functions. Future research directions include empirical evaluation of the proposed algorithms for applications of RMDPs and extending the results to more general settings, such as risk-sensitive partially observable MDPs or decentralized MDPs with agent-specific risk parameters. Another promising research direction is to investigate similar theoretical results under alternative risk measures, such as conditional value-at-risk, among others.

References

[1] N. Bäuerle and A. Jaśkiewicz. Markov decision processes with risk-sensitive criteria: An overview. *Mathematical Methods of Operations Research*, 99(1):141–178, 2024.

[2] N. Bäuerle and U. Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

[3] T. Bielecki, D. Hernández-Hernández, and S. R. Pliska. Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. *Mathematical Methods of Operations Research*, 50:167–188, 1999.

[4] V. S. Borkar and S. P. Meyn. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

[5] R. Cavazos-Cadena and R. Montes-De-Oca. The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space. *Mathematics of Operations Research*, 28(4):752–776, 2003.

[6] K.-J. Chung and M. J. Sobel. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

[7] F. d'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10 (1):98–108, 1963.

[8] G. B. Di Masi and L. Stettner. Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, 38(1):61–78, 1999.

[9] D. Duffie, W. Fleming, H. M. Soner, and T. Zariphopoulou. Hedging in incomplete markets with HARA utility. *Journal of Economic Dynamics and Control*, 21(4-5):753–782, 1997.

[10] W. H. Fleming and D. Hernández-Hernández. An optimal consumption model with stochastic volatility. *Finance and Stochastics*, 7(2):245–262, 2003.

[11] R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

[12] J. E. Ingersoll. *Theory of Financial Decision Making*. Bloomsbury Publishing PLC, 1987.

[13] S. C. Jaquette. A utility criterion for Markov decision processes. *Management Science*, 23(1): 43–49, 1976.

[14] I. Karatzas, J. P. Lehoczky, S. P. Sethi, and S. E. Shreve. Explicit solution of a general consumption/investment problem. *Mathematics of Operations Research*, 11(2):261–294, 1986.

[15] A. Kumar, V. Kavitha, and N. Hemachandra. Finite horizon risk sensitive MDP and linear programming. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 7826–7831. IEEE, 2015.

[16] I. Lee, M. A. Epelman, H. E. Romeijn, and R. L. Smith. Simplex algorithm for countable-state discounted Markov decision processes. *Operations Research*, 65(4):1029–1042, 2017.

[17] S. I. Marcus, E. Fernández-Gaucherand, D. Hernández-Hernández, S. Coraluppi, and P. Fard. Risk sensitive Markov decision processes. In *Systems and Control in the Twenty-first Century*, pages 263–279. Springer, 1997.

[18] Y. Murthy, M. Moharrami, and R. Srikant. On the convergence of modified policy iteration in risk-sensitive exponential cost Markov decision processes. *Operations Research*, 74(3):1425–1436, 2025.

[19] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

[20] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.

[21] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press Cambridge, 1998.

[22] D. J. White. Finite state approximations for denumerable-state infinite horizon contracted Markov decision processes: The policy space method. *Journal of Mathematical Analysis and Applications*, 72(2):512–523, 1979.

[23] Z. Wu and R. Xu. Risk-sensitive Markov decision process and learning under general utility functions. *arXiv preprint arXiv:2311.13589*, 2023.

A Proofs for Section 2

A.1 Proof of Lemma 1

Lemma 1. *Suppose $f(x,\gamma) \in \mathcal{C}(\mathcal{S}\times\Gamma)$ satisfies the condition*

$$\lim_{\gamma\to 0}\max_{x\in\mathcal{S}} |f(x,\gamma) - \mathrm{sgn}(\gamma)| = 0. \tag{4}$$

Then, (4) is also satisfied by $Tf(x,\gamma)$.

Proof. Proof. We prove the lemma for $\gamma \geq 0$. The proof is similar when $\gamma < 0$ and is omitted for brevity. Suppose $f(x,\gamma) \in \mathcal{C}(\mathcal{S}\times\Gamma)$ satisfies (4). For a fixed $x \in \mathcal{S}$ and $\gamma \in \Gamma^+$, we obtain

$$\begin{aligned}
&|Tf(x,\gamma) - \mathrm{sgn}(\gamma)| = \ |Tf(x,\gamma) - 1| \\
&= \Big| \min_{\pi\in\Pi} \Big\{ \Big[\exp\left(\gamma c(x,\pi(x))\right) \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi(x)}(x,y) \Big] - 1 \Big\} \Big| \\
&= \Big| \min_{\pi\in\Pi} \Big\{ \big(\exp\left(\gamma c(x,\pi(x))\right) - 1 \big) \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi(x)}(x,y) + \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi(x)}(x,y) - 1 \Big\} \Big| \\
&= \Big| \min_{\pi\in\Pi} \Big\{ \big(\exp\left(\gamma c(x,\pi(x))\right) - 1 \big) \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi(x)}(x,y) + \sum_{y\in\mathcal{S}} \big(f(y,\gamma\beta) - 1 \big) P^{\pi(x)}(x,y) \Big\} \Big|,
\end{aligned}$$

where the last equality holds because $\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y) = 1$. Let π^* be a minimizing policy in the last equation. It follows that

$$\begin{aligned}
&|Tf(x,\gamma) - \mathrm{sgn}(\gamma)| \\
&= \Big| \big(\exp\left(\gamma c(x,\pi^*(x))\right) - 1 \big) \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi^*(x)}(x,y) + \sum_{y\in\mathcal{S}} \big(f(y,\gamma\beta) - 1 \big) P^{\pi^*(x)}(x,y) \Big| \\
&\leq \Big| \big(\exp\left(\gamma c(x,\pi^*(x))\right) - 1 \big) \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi^*(x)}(x,y) \Big| + \Big| \sum_{y\in\mathcal{S}} \big(f(y,\gamma\beta) - 1 \big) P^{\pi^*(x)}(x,y) \Big| \\
&\leq \max_{\pi\in\Pi} \Big\{ \Big| \big(\exp\left(\gamma c(x,\pi(x))\right) - 1 \big) \sum_{y\in\mathcal{S}} f(y,\gamma\beta) P^{\pi(x)}(x,y) \Big| \Big\} \\
&\quad + \max_{\pi\in\Pi} \Big\{ \Big| \sum_{y\in\mathcal{S}} \big(f(y,\gamma\beta) - 1 \big) P^{\pi(x)}(x,y) \Big| \Big\} \\
&\leq \max_{\pi\in\Pi} \Big\{ \big| \exp\left(\gamma c(x,\pi(x))\right) - 1 \big| \sum_{y\in\mathcal{S}} \big| f(y,\gamma\beta) \big| P^{\pi(x)}(x,y) \Big\} \\
&\quad + \max_{\pi\in\Pi} \Big\{ \sum_{y\in\mathcal{S}} \big| f(y,\gamma\beta) - 1 \big| P^{\pi(x)}(x,y) \Big\}.
\end{aligned} \tag{26}$$

Next we show that the final expression in (26) vanishes as $\gamma \downarrow 0$. First, consider the first term. Fix an arbitrary $\pi \in \Pi$. As $\gamma \to 0^+$, we have $\big| \exp\left(\gamma c(x,\pi(x))\right) - 1 \big| \to 0$, for all $x \in \mathcal{S}$ because $c(x,\pi(x))$ is bounded. Moreover, as $\gamma \to 0^+$, $f(y,\gamma\beta) \to 1$ for all $y \in \mathcal{S}$ by the assumption that f satisfies (4). Then,

$$\sum_{y\in\mathcal{S}} \big| f(y,\gamma\beta) \big| P^{\pi(x)}(x,y) \to \sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y) = 1 \quad \text{as } \gamma \to 0^+.$$

Therefore,

$$\lim_{\gamma\to 0^+} \Big| \exp\left(\gamma c(x,\pi(x))\right) - 1 \Big| \cdot \sum_{y\in\mathcal{S}} \big|f(y,\gamma\beta)\big| P^{\pi(x)}(x,y) = 0\cdot 1 = 0.$$

Because this holds for every $\pi \in \Pi$, we have

$$\lim_{\gamma\to 0^+} \max_{\pi\in\Pi} \left\{ \Big| \exp\left(\gamma c(x,\pi(x))\right) - 1 \Big| \cdot \sum_{y\in\mathcal{S}} \big|f(y,\gamma\beta)\big| P^{\pi(x)}(x,y) \right\} = 0.$$

Next, we use a similar argument to show that the second term in (26) also converges to 0 when $\gamma \to 0^+$. Fix an arbitrary policy $\pi \in \Pi$. Again, because f satisfies (4), we must have $\big|f(y,\gamma\beta) - 1\big| \to 0$ for all $y \in \mathcal{S}$. Consequently,

$$\sum_{y\in\mathcal{S}} \Big| f(y,\gamma\beta) - 1 \Big| \cdot P^{\pi(x)}(x,y) \to \sum_{y\in\mathcal{S}} 0 \cdot P^{\pi(x)}(x,y) = 0 \quad \text{as } \gamma \to 0^+.$$

Therefore,

$$\lim_{\gamma\to 0^+} \max_{\pi\in\Pi} \left\{ \sum_{y\in\mathcal{S}} \Big| f(y,\gamma\beta) - 1 \Big| \cdot P^{\pi(x)}(x,y) \right\} = 0.$$

We thus conclude that the expression in (26) converges to 0 when $\gamma \to 0^+$, which proves the lemma for $\gamma > 0$. $\square$

A.2 Proof of Lemma 2

Lemma 2. *The operator T is an order-preserving mapping, i.e., given functions $f_1, f_2 \in \mathcal{C}(\mathcal{S}\times\Gamma)$ such that $f_1(x,\gamma) \geq f_2(x,\gamma)$ $\forall\, x \in \mathcal{S}$ and $\gamma \in \Gamma$, it follows that*

$$Tf_1(x,\gamma) \geq Tf_2(x,\gamma), \quad \forall\, x \in \mathcal{S},\ \gamma \in \Gamma.$$

Proof. Proof. Let $f_1, f_2 \in \mathcal{C}(\mathcal{S}\times\Gamma)$ such that $f_1(x,\gamma) \geq f_2(x,\gamma)$ for all $x \in \mathcal{S}$ and $\gamma \in \Gamma$. We need to show that $Tf_1 \geq Tf_2$.

Fix $x \in \mathcal{S}$ and $\gamma \in \Gamma$. By the definition of T,

$$Tf_1(x,\gamma) = \min_{\pi\in\Pi} \left\{ \exp\left(\gamma c(x,\pi(x))\right) \sum_{\mathcal{S}} f_1(y,\gamma\beta) P^{\pi(x)}(x,y) \right\}. \tag{27}$$

Let $\pi_1 \in \Pi$ be a minimizer in (27). Similarly, let $\pi_2 \in \Pi$ be a minimizing decision-rule in the definition of $Tf_2(x;\gamma)$. Then,

$$\begin{aligned} Tf_1(x,\gamma) &= \exp\left(\gamma c(x,\pi_1(x))\right) \sum_{\mathcal{S}} f_1(y,\gamma\beta) P^{\pi_1(x)}(x,y) \\ &\geq \exp\left(\gamma c(x,\pi_1(x))\right) \sum_{\mathcal{S}} f_2(y,\gamma\beta) P^{\pi_1(x)}(x,y) \\ &\geq \exp\left(\gamma c(x,\pi_2(x))\right) \sum_{\mathcal{S}} f_2(y,\gamma\beta) P^{\pi_2(x)}(x,y) \ = Tf_2(x,\gamma). \end{aligned}$$

$\square$

A.3 Proof of Lemma 3

Lemma 3. *For every $x \in \mathcal{S}, \gamma \in \Gamma$, define $\underline{f}(x,\gamma) = \operatorname{sgn}(\gamma)$ and*

$$\overline{f}(x,\gamma) = \operatorname{sgn}(\gamma) \exp\left(\frac{\gamma\|c\|}{1-\beta}\right).$$

Then, $\underline{f}$, $\overline{f} \in \mathcal{C}(\mathcal{S} \times \Gamma)$ and

$$\underline{f}(x,\gamma) \leq T^n \underline{f}(x,\gamma) \leq T^n \overline{f}(x,\gamma) \leq \overline{f}(x,\gamma), \quad \forall\, n \geq 1.$$

Moreover, the sequences $T^n \underline{f}(x,\gamma)$ and $T^n \overline{f}(x,\gamma)$ are increasing and decreasing, respectively.

Proof. Proof. First, note that the sign function, and therefore $\underline{f}$ and $\overline{f}$, are continuous on $\mathcal{S} \times \Gamma$ (because $0 \notin \Gamma$).

We prove the inequalities by induction. Note that when f is a constant function of its first argument, so that $f(x,\gamma) = h(\gamma)$ for some function h, then

$$Tf(x,\gamma) = \min_{\pi \in \Pi} \Big\{ h(\gamma\beta) \cdot \exp\left(\gamma c(x,\pi(x))\right) \Big\} \;\forall\, x \in \mathcal{S}, \gamma \in \Gamma.$$

Therefore, for $n = 1$, using the fact that $\operatorname{sgn}(\gamma\beta) = \operatorname{sgn}(\gamma)$, we have

$$T\underline{f}(x,\gamma) = \min_{\pi \in \Pi} \Big\{ \operatorname{sgn}(\gamma) \exp\left(\gamma c(x,\pi(x))\right) \Big\} \geq \operatorname{sgn}(\gamma).$$

For $\gamma \in \Gamma^+$, the inequality holds because $\operatorname{sgn}(\gamma) \exp\left(\gamma c(x,\pi(x))\right) = \exp\left(\gamma c(x,\pi(x))\right) \geq 1 = \operatorname{sgn}(\gamma)$. For $\gamma \in \Gamma^-$, we have $\operatorname{sgn}(\gamma) \exp\left(\gamma c(x,\pi(x))\right) = -\exp\left(\gamma c(x,\pi(x))\right) \geq -1 = \operatorname{sgn}(\gamma)$. Note that we assume that the cost function c is nonnegative.

Therefore, $\underline{f} \leq T\underline{f}$. Then, by Lemma 2, $T\underline{f} \leq T^2\underline{f}$. Repeating this argument gives us that $T^n\underline{f}$ is an increasing sequence and $\underline{f} \leq T^n\underline{f}$ for all $n \geq 1$. Similarly, for $\overline{f}$,

$$\begin{aligned} T\overline{f}(x,\gamma) =& \min_{\pi \in \Pi} \Big\{ \operatorname{sgn}(\gamma) \exp\left(\gamma\beta \frac{||c||}{1-\beta}\right) \cdot \exp\left(\gamma c(x,\pi(x))\right) \Big\} \\ \leq& \operatorname{sgn}(\gamma) \exp\left(\gamma \frac{||c||}{1-\beta}\right) = \overline{f}(x,\gamma). \end{aligned}$$

Again, for $\gamma \in \Gamma^+$, the inequality holds because

$$\exp\left(\gamma c(x,\pi(x))\right) \exp\left(\gamma\beta \frac{||c||}{1-\beta}\right) = \exp\left(\gamma c(x,\pi(x)) + \gamma\beta \frac{||c||}{1-\beta}\right)$$

and $\gamma c(x,\pi(x)) + \gamma\beta \frac{||c||}{1-\beta} \leq \gamma||c|| + \gamma\beta\frac{||c||}{1-\beta} = \gamma\frac{||c||}{1-\beta}$ when $\gamma > 0$. Similar reasoning with the inequalities reversed can be used when $\gamma \in \Gamma^-$.

As before, using Lemma 2, $T\overline{f} \leq \overline{f}$ implies that $T^2\overline{f} \leq T\overline{f}$, and so on. Therefore, $T^n\overline{f}$ is a decreasing sequence, and $T^n\overline{f} \leq \overline{f}$ for all $n \geq 1$.

Finally, $\underline{f} = \operatorname{sgn}(\gamma) \leq \operatorname{sgn}(\gamma)\exp\left(\gamma\frac{||c||}{1-\beta}\right) = \overline{f}$ for all $\gamma \in \Gamma$. Lemma 2 implies that $T\underline{f} \leq T\overline{f}$. Repeatedly applying the operator T and using Lemma 2 give us that $T^n\underline{f} \leq T^n\overline{f}$ for all $n \geq 1$. □

A.4 Proof of Lemma 4

Lemma 4. *Let* $\underline{f}(x,\gamma)$, $\overline{f}(x,\gamma)$ *be as defined in Lemma 3. Then,* $\left|T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma)\right| \to 0$ *as* $n \to \infty$*. Moreover, the convergence is uniform in both* $x \in \mathcal{S}$ *and* $\gamma \in \Gamma^+ \cup \Gamma^-$*.*

Proof. Proof. Fix $n \geq 1$, $x \in \mathcal{S}$ and $\gamma \in \Gamma^+$. We know from Lemma 3 that $T^n\underline{f}(x,\gamma) \leq T^n\overline{f}(x,\gamma)$, so we may remove the absolute value. Then,

$$\begin{aligned}
T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma) &= \min_{\boldsymbol{\pi}^n}\Big\{\mathbb{E}\Big[\exp\Big(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\Big)\exp\Big(\frac{\gamma\beta^n\|c\|}{1-\beta}\Big)\Big]\Big\} \\
&\quad - \min_{\boldsymbol{\pi}^n}\Big\{\mathbb{E}\Big[\exp\Big(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\Big)\Big]\Big\} \\
&= \Big[\exp\Big(\frac{\gamma\beta^n\|c\|}{1-\beta}\Big) - 1\Big]\min_{\boldsymbol{\pi}^n}\Big\{\mathbb{E}\Big[\exp\Big(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\Big)\Big]\Big\}. \quad (28)
\end{aligned}$$

The term inside the minimization in (28) is uniformly bounded above for every n, because

$$\exp\Big(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\Big) \leq \exp\Big(\gamma\sum_{i=0}^{n-1}\beta^i\|c\|\Big) \leq \exp\Big(\gamma\frac{\beta\|c\|}{1-\beta}\Big).$$

In fact,

$$\begin{aligned}
0 \leq T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma) &\leq \left[\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right) - 1\right]\cdot\exp\left(\gamma\frac{\beta\|c\|}{1-\beta}\right) \\
&\leq \left[\exp\left(\frac{d\beta^n\|c\|}{1-\beta}\right) - 1\right]\cdot\exp\left(d\frac{\beta\|c\|}{1-\beta}\right).
\end{aligned}$$

As $n \to \infty$, the RHS of the last inequality vanishes, and so does $|T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma)|$. Moreover, because the upper bound is independent of x and γ, the convergence is uniform.

Next, consider $\gamma \in \Gamma^-$, and fix $n \geq 1$, $x \in \mathcal{S}$.

$$\begin{aligned}
T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma) &= \min_{\boldsymbol{\pi}^n}\left\{-\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\right)\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)\right]\right\} \\
&\quad - \min_{\boldsymbol{\pi}^n}\left\{-\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\right)\right]\right\} \\
&= -\max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\right)\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)\right]\right\} \\
&\quad + \max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\right)\right]\right\} \\
&= \left[1 - \exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)\right]\max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x))\right)\right]\right\}.
\end{aligned}$$

Here, the term inside the maximization is uniformly bounded above by 1. Therefore, we get

$$0 \le T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma) \le \left[1 - \exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)\right] \cdot 1 \le 1 - \exp\left(\frac{-d\beta^n\|c\|}{1-\beta}\right).$$

As $n \to \infty$, the rightmost term vanishes, Then, by a similar reasoning as before, $|T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma)|$ converges uniformly to zero as $n \to \infty$. □

A.5 Proof of Theorem 1

A.5.1 Uniform Convergence of $T^n(\overline{f})$ and $T^n(\underline{f})$

Theorem 1. *There exists a unique function $w \in \mathcal{C}(\mathcal{S} \times \Gamma)$ that satisfies the equations*

$$w(x,\gamma) = \min_{\pi}\left\{\exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} w(y;\gamma\beta)P^{\pi(x)}(x,y)\right\}, \quad \forall\, x \in \mathcal{S}, \gamma \in \Gamma, \tag{5}$$

with boundary condition

$$\lim_{\gamma\to 0}\left(\max_{x\in\mathcal{S}} |w(x,\gamma) - \operatorname{sgn}(\gamma)|\right) = 0. \tag{6}$$

Moreover, w equals the optimal value function v^* defined in (2)*.*

Proof. Proof. For any $x \in \mathcal{S}$ and $\gamma \in \Gamma$, it follows from Lemma 3 that $T^n\underline{f}(x,\gamma)$ and $T^n\overline{f}(x,\gamma)$ are bounded monotonic sequences and therefore convergent. Then, Lemma 4 implies that they converge to a common limit, say, $w(x;\gamma)$. The monotonicity of the sequences further implies that $T^n\underline{f}(x,\gamma) \le w(x;\gamma) \le T^n\overline{f}(x,\gamma)$. Therefore,

$$\begin{aligned} \left|w(x;\gamma) - T^n\underline{f}(x,\gamma)\right| &\le \left|T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma)\right|, \\ \left|w(x;\gamma) - T^n\overline{f}(x,\gamma)\right| &\le \left|T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma)\right|. \end{aligned} \tag{29}$$

Lemma 4 established that $\left|T^n\overline{f}(x,\gamma) - T^n\underline{f}(x,\gamma)\right|$ converges uniformly to 0. Using this in (29) gives us that both $T^n\underline{f}$ and $T^n\overline{f}$ converge uniformly to w in both x and γ. It follows that $w \in \mathcal{C}(\mathcal{S} \times \Gamma)$.

A.5.2 Satisfiability of the Bellman Equations (5)

We first show that $w(x;\gamma) \in C(\mathcal{S} \times \Gamma)$ and is a solution to (5). Fix $x \in \mathcal{S}$ and $\gamma \in \Gamma$. Because $T^n\overline{f}(x,\gamma)$ converges to $w(x;\gamma)$, we have

$$\begin{aligned} w(x;\gamma) &= \lim_{n\to\infty} T^n\overline{f}(x,\gamma) = \lim_{n\to\infty} T^{n+1}\overline{f}(x,\gamma) = \lim_{n\to\infty} T\left(T^n\overline{f}(x,\gamma)\right) \\ &= T\left(\lim_{n\to\infty} T^n\overline{f}(x,\gamma)\right) = T(w(x;\gamma)), \end{aligned}$$

where the penultimate equality holds because T is continuous. Therefore, the function w satisfies the fixed-point equation $T(w) = w$. That is,

$$w(x;\gamma) = \min_{\pi\in\Pi}\left\{\exp\left(\gamma c(x,\pi(x))\right)\sum_{\mathcal{S}} w(y;\gamma\beta)P^{\pi(x)}(x,y)\right\}, \quad \forall x \in \mathcal{S}, \gamma \in \Gamma,$$

which is exactly (5).

A.5.3 Satisfiability of Boundary Condition (6)

Recall that $\underline{f}(x,\gamma) = \operatorname{sgn}(\gamma)$ and $\overline{f}(x,\gamma) = \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\|c\|}{1-\beta}\right)$. Therefore,

$$\lim_{\gamma\to 0}\left(\max_{x\in\mathcal{S}}|\overline{f}(x,\gamma)-\operatorname{sgn}(\gamma)|\right)=0 \quad \text{and} \quad \lim_{\gamma\to 0}\left(\max_{x\in\mathcal{S}}|\underline{f}(x,\gamma)-\operatorname{sgn}(\gamma)|\right)=0.$$

Moreover, because $\underline{f}(x,\gamma)\le w(x;\gamma)\le \overline{f}(x,\gamma)$ for all $x\in\mathcal{S}$, we have that

$$\begin{aligned}
& 0 = \underline{f}(x,\gamma)-\operatorname{sgn}(\gamma) \le w(x;\gamma)-\operatorname{sgn}(\gamma) \le \overline{f}(x,\gamma)-\operatorname{sgn}(\gamma)\ \forall\ x\in\mathcal{S},\\
\implies\ & 0 = \lim_{\gamma\to 0}\max_{x\in\mathcal{S}}\ |\underline{f}(x,\gamma)-\operatorname{sgn}(\gamma)| \le \lim_{\gamma\to 0}\max_{x\in\mathcal{S}}\ |w(x;\gamma)-\operatorname{sgn}(\gamma)|\\
& \quad \le \lim_{\gamma\to 0}\max_{x\in\mathcal{S}}\ |\overline{f}(x,\gamma)-\operatorname{sgn}(\gamma)| = 0.
\end{aligned}$$

Thus, w satisfies the boundary condition (6).

A.5.4 v^* Satisfies the Bellman equation

We show that the optimal value function v^* defined in Equation (2) satisfies the Bellman equation:

$$v^*(x;\gamma)=\min_{\pi}\left\{\exp\Big(\gamma c(x,\pi(x))\Big)\sum_{y\in\mathcal{S}}P^{\pi(x)}(x,y)\,v^*(y;\gamma\beta)\right\}.$$

We prove the claim by establishing two inequalities:

$$v^*(x;\gamma)\ge\min_{\pi}\left\{\exp\Big(\gamma c(x,\pi(x))\Big)\sum_{y\in\mathcal{S}}P^{\pi(x)}(x,y)\,v^*(y;\gamma\beta)\right\}, \tag{30}$$

$$v^*(x;\gamma)\le\min_{\pi}\left\{\exp\Big(\gamma c(x,\pi(x))\Big)\sum_{y\in\mathcal{S}}P^{\pi(x)}(x,y)\,v^*(y;\gamma\beta)\right\}. \tag{31}$$

Step 1: "$\ge$" inequality. We first show that (30) holds. Fix any policy $\pi^\infty=(\pi_0,\pi_1,\dots)$. Split the exponential in (1) as

$$\gamma\sum_{i=0}^{\infty}\beta^i c_i=\gamma c_0+\gamma\beta\sum_{j=0}^{\infty}\beta^j c_{j+1},$$

where c_i is shorthand for $c(x_i,\pi_i^\infty(x_i))$. Hence

$$\operatorname{sgn}(\gamma)\exp\!\left(\gamma\sum_{i=0}^{\infty}\beta^i c_i\right)=\exp\left(\gamma c_0\right)\,\operatorname{sgn}(\gamma\beta)\exp\!\left(\gamma\beta\sum_{j=0}^{\infty}\beta^j c_{j+1}\right).$$

Taking expectation in (1) and conditioning on $x_0=x$ gives

$$v^{\boldsymbol{\pi}^\infty}(x;\gamma)=\exp\left(\gamma c(x,\pi_0(x))\right)\sum_{y\in\mathcal{S}}P^{\pi_0(x)}(x,y)\,\mathbb{E}_y^{\pi^\infty,+}\!\left[\operatorname{sgn}(\gamma\beta)\exp\!\left(\gamma\beta\sum_{j=0}^{\infty}\beta^j c_j\right)\right],$$

where $\pi^{\infty,+} = (\pi_1, \pi_2, \dots)$ is the continuation policy.

By definition of v^*,

$$\mathbb{E}_y^{\pi^{\infty,+}}\left[\operatorname{sgn}(\gamma\beta)\exp\left(\gamma\beta\sum_{j=0}^{\infty}\beta^j c_j\right)\right] \geq v^*(y;\gamma\beta).$$

Therefore

$$v^{\boldsymbol{\pi}^\infty}(x;\gamma) \geq \exp\left(\gamma c(x,\pi_0(x))\right)\sum_{y\in\mathcal{S}} P^{\pi_0(x)}(x,y)\, v^*(y;\gamma\beta).$$

Since π_0 was arbitrary,

$$v^{\boldsymbol{\pi}^\infty}(x;\gamma) \geq \min_{\pi}\left\{\exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y)\, v^*(y;\gamma\beta)\right\}.$$

Taking $\inf_{\pi^\infty}$ on the left-hand side yields

$$v^*(x;\gamma) \geq \min_{\pi}\left\{\exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y)\, v^*(y;\gamma\beta)\right\}.$$

Step 2: "$\leq$" inequality. We now show that (31) holds. Let π^* attain the minimum on the RHS. Choose a continuation policy $\pi^{\infty,+}$ satisfying

$$v^{\boldsymbol{\pi}^{\infty,+}}(y;\gamma\beta) \leq v^*(y;\gamma\beta) + \varepsilon,\ \ \forall y \in \mathcal{S},\ \text{ for some } \varepsilon > 0.$$

Note that such a continuation policy exists because the finite-horizon optimal values converge uniformly to $w(\cdot,\gamma\beta)$, hence for large enough horizon N, an N-horizon optimal policy is ε-optimal for all initial states.

Construct a concatenated policy π^∞, such that it equals an arbitrary policy π at time 0, and equals $\pi^{\infty,+}$ in subsequent periods.

Then,

$$\begin{aligned} v^{\boldsymbol{\pi}^\infty}(x;\gamma) &= \exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y) v^{\boldsymbol{\pi}^{\infty,+}}(y;\gamma\beta) \\ &\leq \exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y)\big(v^*(y;\gamma\beta)+\varepsilon\big). \end{aligned}$$

Thus

$$v^*(x;\gamma) \leq \exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y) v^*(y;\gamma\beta) + \varepsilon\exp\left(\gamma c(x,\pi(x))\right).$$

Because this inequality holds for any $\epsilon > 0$, we have

$$v^*(x;\gamma) \leq \min_{\pi}\left\{\exp\left(\gamma c(x,\pi(x))\right)\sum_{y\in\mathcal{S}} P^{\pi(x)}(x,y)\, v^*(y;\gamma\beta)\right\}.$$

Combining the two inequalities proves the claim.

A.5.5 Uniqueness

To prove uniqueness, we need to show that any function $f \in \mathcal{C}(\mathcal{S} \times \Gamma)$ that satisfies (5) and (6) must equal v^*. To do so, we first define

$$\begin{aligned}\mathcal{E}_n(x) = \Bigg| &\underbrace{\min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\}}_{A} \\ &- \underbrace{\min_{\boldsymbol{\pi}^\infty} \mathbb{E}\left[\operatorname{sgn}(\gamma) \exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right]}_{B} \Bigg|,\end{aligned} \tag{32}$$

where f is an arbitrary function in $\mathcal{C}(\mathcal{S} \times \Gamma)$ that satisfies (6) and $\underline{f}(x, \gamma) \le f(x, \gamma) \le \overline{f}(x, \gamma)$ for all $x \in \mathcal{S}$, $\gamma \in \Gamma$. Let A and B be the first and second terms inside the absolute value in (32) respectively. Note in B that we replaced $\inf_{\boldsymbol{\pi}^\infty}$ with $\min_{\boldsymbol{\pi}^\infty}$. This is because we showed in A.5.4 that v^* satisfies the Bellman equation (5). Therefore, we can use (5) to construct a deterministic policy whose value equals the optimal value function v^*, by always assigning a minimizing action of (5) to the corresponding $x \in \mathcal{S}, \gamma \in \Gamma$. This demonstrates the existence of a deterministic optimal policy.

We show that $\mathcal{E}_n(x) \to 0$ as $n \to \infty$. We divide the proof into four cases based on the sign of γ and the relation between A and B.

Case a.I: $\gamma > 0$ and $A \ge B$. Let $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}^\infty$ be a minimizing policy for B in (32). Let $\{\pi_i^*\}_{i=0}^{n-1}$ be the truncation of $\boldsymbol{\pi}^*$ for the first n periods, which is a feasible policy for the minimization term in the first inequality below. Therefore,

$$\begin{aligned}\min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} &\le \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^*(x_i)) \right) \right] \\ &\le \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^*(x_i)) \right) \right] \\ &= \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\}.\end{aligned}$$

For any state $x \in \mathcal{S}$,

$$
\begin{aligned}
\mathcal{E}_n(x) \ &= \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n; \gamma\beta^n) \right] \right\} \\
&\quad - \min_{\boldsymbol{\pi}^\infty} \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \\
&\leq \left[\max_{x_n \in \mathcal{S}} f(x_n, \gamma\beta^n) \right] \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \\
&\quad - \min_{\boldsymbol{\pi}^\infty} \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \\
&\leq \left[\max_{x_n \in \mathcal{S}} f(x_n, \gamma\beta^n) \right] \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} \\
&\quad - \min_{\boldsymbol{\pi}^\infty} \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \left[\max_{x_n \in \mathcal{S}} f(x_n, \gamma\beta^n) - 1 \right] \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} \\
&\leq \left[\max_{x_n \in \mathcal{S}} f(x_n, \gamma\beta^n) - 1 \right] \exp\left(\frac{\gamma \|c\|}{1-\beta} \right).
\end{aligned}
$$

Because f satisfies the boundary condition (6), it follows that $\mathcal{E}_n(x) \to 0$ as $n \to \infty$.

Case a.II: $\gamma > 0$ **and** $A < B$**.**

$$
\begin{aligned}
&\mathcal{E}_n(x) \\
&= \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty) \right) \right] \right\} - \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\} \\
&\leq \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} - \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\},
\end{aligned} \tag{33}
$$

where the inequality holds because $\gamma > 0$ and $f(x_n, \gamma\beta^n) \geq \underline{f}(x, \gamma\beta^n) = \operatorname{sgn}(\gamma) = 1$. Next, let $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}^n$ be an optimal policy for the following minimization problem

$$
\min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) + \gamma \sum_{i=n}^{\infty} \beta^i \|c\| \right) \right] \right\}.
$$

Define $\bar{\pi}^*$ as an infinite-horizon policy, where $\bar{\pi}_i^* = \pi_i^*$ for $i = 0, \ldots, n-1$ and $\{\pi_i^*\}_{i=n}^\infty$ is set arbitrarily. It follows that

$$
\begin{aligned}
&\min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) + \gamma \sum_{i=n}^{\infty} \beta^i \|c\| \right) \right] \right\} \\
&= \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \cdot \exp\left(\gamma \sum_{i=n}^{\infty} \beta^i \|c\| \right) \\
&= \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^*(x_i)) \right) \right] \exp\left(\gamma \sum_{i=n}^{\infty} \beta^i \|c\| \right) \\
&\geq \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \bar{\pi}_i^*(x_i)) + \gamma \sum_{i=n}^{\infty} \beta^i c(x_i, \bar{\pi}_i^*(x_i)) \right) \right] \\
&= \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \bar{\pi}_i^*(x_i)) \right) \right] \geq \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\},
\end{aligned}
$$

where the first inequality holds because $\|c\|$ is an upper bound on the costs, and the second inequality holds because $\bar{\pi}^* \in \boldsymbol{\Pi}^\infty$ is a feasible policy for the last minimization problem. In summary, we showed that

$$
\begin{aligned}
&\min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} \\
&\leq \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \exp\left(\gamma \sum_{i=n}^{\infty} \beta^i \|c\| \right).
\end{aligned} \tag{34}
$$

Using (34), we can continue (33) as follows.

$$
\begin{aligned}
\mathcal{E}_n(x) &\leq \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} \\
&\quad - \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \\
&\leq \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \exp\left(\gamma \sum_{i=n}^{\infty} \beta^i \|c\| \right) \\
&\quad - \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \\
&\leq \exp\left(\frac{\gamma(1-\beta^n)\|c\|}{1-\beta} \right) \left(\exp\left(\frac{\gamma \beta^n \|c\|}{1-\beta} \right) - 1 \right) \\
&< \exp\left(\frac{\gamma \|c\|}{1-\beta} \right) \left(\exp\left(\frac{\gamma \beta^n \|c\|}{1-\beta} \right) - 1 \right).
\end{aligned}
$$

Taking limits as $n \to \infty$ on both sides, we obtain $\mathcal{E}_n(x) \to 0$.

Case b.I: $\gamma < 0$ **and** $A > B$. In this case, we have

$$\begin{aligned}
\mathcal{E}_n(x) &= \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\} \\
&\quad - \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[-\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} \\
&= \max_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right\} \\
&\quad - \max_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[-\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\} \\
&\leq \max_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \\
&\quad - \max_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[-\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\},
\end{aligned}$$

where the last inequality holds because $\gamma < 0$. Combining terms, we have

$$\begin{aligned}
\mathcal{E}_n(x) &\leq \max_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \max_{x_n \in \mathcal{S}} \Big[(1 - f(x_n, \gamma\beta^n) \Big] \\
&\leq \max_{x_n \in \mathcal{S}} \Big[(1 - f(x_n, \gamma\beta^n) \Big],
\end{aligned}$$

where the inequality holds because $\gamma < 0$, which gives

$$\max_{\boldsymbol{\pi}^n} E_x \left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] < 1.$$

Again, because f satisfies the boundary condition (6), we get that $\lim_{n\to\infty} \mathcal{E}_n(x) = 0$.

Case b.II: $\gamma < 0$ **and** $A < B$**.** In this case, (32) reduces to

$$\begin{aligned}
\mathcal{E}_n(x) &= \min_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\left[-\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^\infty(x_i)\right)\right]\right\} \\
&- \min_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x_i))\right)f(x_n,\gamma\beta^n)\right]\right\} \\
&= \max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[-\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x_i))\right)f(x_n,\gamma\beta^n)\right]\right\} \\
&- \max_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^\infty(x_i))\right)\right]\right\} \\
&\le \max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x_i))\right)\right]\right\} \\
&- \max_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^\infty(x_i))\right)\right]\right\},
\end{aligned}$$

where the inequality holds because $-f(x_n,\gamma\beta^n) \le \underline{f}(x_n,\gamma\beta^n) = -\operatorname{sgn}(\gamma) = 1$ by assumption. Therefore,

$$\begin{aligned}
\mathcal{E}_n(x) \le \Bigg(&\max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x_i))\right)\right]\right\} \\
&- \max_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^\infty(x_i))\right)\right]\right\}\Bigg).
\end{aligned} \tag{35}$$

Next, let $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}^n$ be a maximizing policy for the following problem

$$\max_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x_i))\right)\right]\right\},$$

and define $\bar{\pi}^* \in \boldsymbol{\Pi}^\infty$ as an infinite-horizon policy, where $\bar{\pi}_i^* = \pi_i^*$ for $i = 0,\ldots,n-1$ and $\{\pi_i^*\}_{i=n}^\infty$ is set arbitrarily. Then, $\bar{\pi}^*$ is a feasible policy for

$$\max_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\left[\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^\infty(x_i))\right)\right]\right\},$$

which implies

$$\begin{aligned}
&\max_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\bar{\pi}_i^\infty(x_i))\right)\bigg]\right\}\\
&\geq \mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\bar{\pi}_i^*(x_i))\right)\bigg]\\
&= \mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\bar{\pi}_i^*(x_i))+\gamma\sum_{i=n}^{\infty}\beta^i c(x_i,\bar{\pi}_i^*(x_i))\right)\bigg]\\
&\geq \mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^*(x_i))\right)\bigg]\exp\left(\gamma\sum_{i=n}^{\infty}\beta^i\|c\|\right),
\end{aligned}\tag{36}$$

where the last inequality holds because γ is negative and $c(x,a)\leq\|c\|$ for all $x\in\mathcal{S}$, $a\in\mathcal{A}$. Using (36) in (35), we obtain

$$\mathcal{E}_n(x)\leq\mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^*(x_i))\right)\bigg]\bigg[1-\exp\left(\frac{\gamma\|c\|\beta^n}{1-\beta}\right)\bigg]<\bigg[1-\exp\left(\frac{\gamma\|c\|\beta^n}{1-\beta}\right)\bigg],$$

where the last two inequalities hold because $\gamma<0$. It follows that $\mathcal{E}_n(x)$ vanishes as $n\to\infty$.

We can hereby conclude that $\lim_{n\to\infty}\mathcal{E}_n=0$, where $\mathcal{E}_n$ is defined in (32). Equivalently,

$$\begin{aligned}
&\lim_{n\to\infty}\min_{\boldsymbol{\pi}^n}\left\{\mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,\pi_i^n(x_i))\right)f(x_n,\gamma\beta^n)\bigg]\right\}\\
&=\min_{\boldsymbol{\pi}^\infty}\mathbb{E}\bigg[\operatorname{sgn}(\gamma)\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^\infty(x_i))\right)\bigg]
\end{aligned}\tag{37}$$

Now, we show that any function f that satisfies (5) and (6) also satisfies

$$f(x,\gamma)=\min_{\boldsymbol{\pi}^\infty}\left\{\mathbb{E}\bigg[\operatorname{sgn}(\gamma)\exp\left(\gamma\sum_{i=0}^{\infty}\beta^i c(x_i,\pi_i^n(x_i))\right)\bigg]\right\}.$$

Because f satisfies (5), we have that

$$f(x,\gamma)=\min_{\pi}\left\{\exp\Big(\gamma c(x,\pi(x))\Big)\sum_{y\in\mathcal{S}}f(y,\gamma\beta)P^{\pi(x)}(x,y)\right\}\ \forall\ x\in\mathcal{S},\gamma\in\Gamma.\tag{38}$$

Because $f(y,\gamma\beta)$ on the RHS also satisfies (5) by assumption, we can express (38) as

$$\begin{aligned}
f(x,\gamma)&=\min_{\pi_1}\bigg\{\exp\Big(\gamma c(x,\pi_1(x))\Big)\\
&\sum_{y\in\mathcal{S}}\min_{\pi_2}\left\{\exp\left(\gamma\beta c(y,\pi_2(y))\right)\sum_{z\in\mathcal{S}}f(z;\gamma\beta^2)P^{\pi_2(y)}(y,z)\right\}P^{\pi_1(x)}(x,y)\bigg\}\\
&=\min_{\boldsymbol{\pi}^2}\left\{\mathbb{E}\bigg[\exp\left(\gamma\sum_{i=0}^{1}\beta^i c(x_i,\pi_i^2(x_i))\right)f(x_2,\gamma\beta^2)\bigg]\right\}.
\end{aligned}$$

This relation holds for all $x \in \mathcal{S}$ and $\gamma \in \Gamma$. Recursively applying this procedure on the RHS, we obtain that for any n:

$$f(x,\gamma) = \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\}.$$

Note that f satisfies the boundary condition (6) and that $\underline{f}(x,\gamma) \leq f(x,\gamma) \leq \overline{f}(x,\gamma)$, $\forall x \in \mathcal{S}$ and $\gamma \in \Gamma$. Letting $n \to \infty$, and using (37), we obtain that

$$\begin{aligned} f(x,\gamma) &= \lim_{n\to\infty} \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) f(x_n, \gamma\beta^n) \right] \right\} \\ &= \min_{\boldsymbol{\pi}^\infty} \left\{ \mathbb{E}\left[\operatorname{sgn}(\gamma) \exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\}. \end{aligned}$$

This proves uniqueness. □

B Proofs for Section 3

B.1 Proof of Lemma 5

Lemma 5. *Let $\mathcal{E}_n(x)$ be the approximation error defined in* (10)*. Then,*

(a) If $\gamma > 0$, $\mathcal{E}_n(x) < \exp\left(\frac{\gamma\|c\|}{1-\beta}\right)\left[\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right) - 1\right]$ for all $x \in \mathcal{S}$.

(b) If $\gamma < 0$, $\mathcal{E}_n(x) < \left[1 - \exp\left(\frac{\gamma\|c\|\beta^n}{1-\beta}\right)\right]$ for all $x \in \mathcal{S}$.

Proof. Proof. Fix $x \in \mathcal{S}$. Plugging in $v^0(x;\gamma) = \operatorname{sgn}(\gamma)\exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta}\right)$ from (7) into the definition of $\mathcal{E}_n(x)$ in (10), we obtain

$$\begin{aligned} \mathcal{E}_n(x) &= \left| T^n v^0(x;\gamma) - v^*(x;\gamma) \right| \\ &= \left| \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\operatorname{sgn}(\gamma) \exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta} \right) \right] \right\} \right. \\ &\left. - \min_{\boldsymbol{\pi}^\infty} \mathbb{E}\left[\operatorname{sgn}(\gamma) \exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right| \\ &= \left| \min_{\boldsymbol{\pi}^n} \left\{ \mathbb{E}\left[\operatorname{sgn}(\gamma) \exp\left(\gamma \sum_{i=0}^{n-1} \beta^i c(x_i, \pi_i^n(x_i)) \right) \right] \right\} \cdot \exp\left(\frac{\gamma\beta^n\|c\|}{1-\beta} \right) \right. \\ &\left. - \min_{\boldsymbol{\pi}^\infty} \mathbb{E}\left[\operatorname{sgn}(\gamma) \exp\left(\gamma \sum_{i=0}^{\infty} \beta^i c(x_i, \pi_i^\infty(x_i)) \right) \right] \right|. \end{aligned}$$

Part (a) of the lemma is proven with **Case a.I** and **Case a.II**, while part (b) of the lemma is proven with **Case b.I** and **Case b.II** in the proof for (32) in A.5.5. □

C Proofs for Section 5

C.1 Proof of Lemma 5

Theorem 5. *Let u^k be the approximate value function computed in iteration k of Algorithm 2. The approximation error is bounded above as follows.*

$$|u^k(x;\gamma) - v^*(x;\gamma)| \le \exp\left(\gamma\|c\|\frac{1-\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right)\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right)-1\right|, \quad \forall x \in \mathcal{S}.$$

Proof. Proof. Fix k. We first use an inductive argument to prove that for every $x \in \mathcal{S}$,

$$u^{k-1}(x;\gamma\beta^i) \ge u^k(x;\gamma\beta^i), \quad \forall 0 \le i \le k.$$

When $i = k$ and $\gamma > 0$, as $u^{k-1}(x;\gamma\beta^k) = u^0(x;\gamma\beta^k) = \exp\left(\gamma\beta^k\frac{\|c\|}{1-\beta}\right)$, $\forall\ x$, and

$$\begin{aligned}
u^k(x;\gamma\beta^k) &= \exp\left(\gamma\beta^k c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)u^k(y;\gamma\beta^{k+1})\\
&= \exp\left(\gamma\beta^k c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)\exp\left(\gamma\beta^{k+1}\frac{\|c\|}{1-\beta}\right)\\
&\le \exp\left(\gamma\beta^k\|c\|\right)\exp\left(\gamma\beta^{k+1}\frac{\|c\|}{1-\beta}\right) = \exp\left(\gamma\beta^k\frac{\|c\|}{1-\beta}\right),
\end{aligned}$$

where a is the action prescribed by the policy at the kth iteration.

When $i = k$ and $\gamma < 0$, $u^{k-1}(x;\gamma\beta^k) = u^0(x;\gamma\beta^k) = -\exp\left(\gamma\beta^k\frac{\|c\|}{1-\beta}\right)$, $\forall\ x$, and

$$\begin{aligned}
u^k(x;\gamma\beta^k) &= \exp\left(\gamma\beta^k c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)u^k(y;\gamma\beta^{k+1})\\
&= -\exp\left(\gamma\beta^k c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)\exp\left(\gamma\beta^{k+1}\frac{\|c\|}{1-\beta}\right)\\
&\le -\exp\left(\gamma\beta^k\|c\|\right)\exp\left(\gamma\beta^{k+1}\frac{\|c\|}{1-\beta}\right) = -\exp\left(\gamma\beta^k\frac{\|c\|}{1-\beta}\right),
\end{aligned}$$

where a is the action prescribed by the policy at the kth iteration.

When $i \le k-1$, assume $u^{k-1}(x;\gamma\beta^{i+1}) \ge u^k(x;\gamma\beta^{i+1})$, $\forall\ x$. Then,

$$\begin{aligned}
u^{k-1}(x;\gamma\beta^i) &= \exp\left(\gamma\beta^i c(x,a_1)\right)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^{k-1}(y;\gamma\beta^{i+1})\\
u^k(x;\gamma\beta^i) &= \exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)u^k(y;\gamma\beta^{i+1}),
\end{aligned}$$

where

$$a_2 \in \arg\min\{\exp\left(\gamma\beta^i c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)u^{k-1}(y;\gamma\beta^{i+1})\}.$$

Therefore,

$$\begin{aligned}
u^k(x;\gamma\beta^i) &= \exp\big(\gamma\beta^i c(x,a_2)\big)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)u^k(y;\gamma\beta^{i+1}) \\
&\le \exp\big(\gamma\beta^i c(x,a_2)\big)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)u^{k-1}(y;\gamma\beta^{i+1}) \\
&\le \exp\big(\gamma\beta^i c(x,a_1)\big)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^{k-1}(y;\gamma\beta^{i+1}) = u^{k-1}(x;\gamma\beta^i), \quad \forall\, x\in\mathcal{S}.
\end{aligned}$$

Thus, the inequality holds for all $0 \le i \le k$.

A similar inductive argument also shows that $u^k(x;\gamma\beta^i) \ge v^*(x;\gamma\beta^i)$, $\forall\, x \in \mathcal{S}$, $0 \le i \le k$. We know that $u^k(x;\gamma\beta^k) \ge v^*(x;\gamma\beta^k)$. Then, suppose $u^k(x;\gamma\beta^{i+1}) \ge v^*(x;\gamma\beta^{i+1})$ for some $i \le k-1$, then

$$\begin{aligned}
v^*(x;\gamma\beta^i) &= \min\{\exp\big(\gamma\beta^i c(x,a)\big)\sum_{y\in\mathcal{S}} P^{a}(x,y)v^*(y;\gamma\beta^{i+1})\} \\
&\le \exp\big(\gamma\beta^i c(x,a_1)\big)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)v^*(y;\gamma\beta^{i+1}) \\
&\le \exp\big(\gamma\beta^i c(x,a_1)\big)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^k(y;\gamma\beta^{i+1}) = u^k(x;\gamma\beta^i), \quad \forall\, x\in\mathcal{S},
\end{aligned}$$

where a_1 is the action prescribed by the kth policy on $u^k(x;\gamma\beta^i)$. Again, this implies the inequality holds for all $0 \le i \le k$.

Then, for any $x \in \mathcal{S}$, $k > 0$ and $0 \le i \le k$, we have

$$\begin{aligned}
|u^k(x;\gamma\beta^i) - v^*(x;\gamma\beta^i)| &= u^k(x;\gamma\beta^i) - v^*(x;\gamma\beta^i) \\
&= \exp\big(\gamma\beta^i c(x,a_1)\big)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^k(y;\gamma\beta^{i+1}) \\
&\quad - \min\{\exp\big(\gamma\beta^i c(x,a)\big)\sum_{y\in\mathcal{S}} P^{a}(x,y)v^*(y;\gamma\beta^{i+1})\} \\
&= \exp\big(\gamma\beta^i c(x,a_1)\big)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^k(y;\gamma\beta^{i+1}) \\
&\quad - \exp\big(\gamma\beta^i c(x,a_2)\big)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)v^*(y;\gamma\beta^{i+1}),
\end{aligned}$$

where a_1 is the action prescribed by the kth policy, and

$$a_2 \in \arg\min\{\exp\big(\gamma\beta^i c(x,a)\big)\sum_{y\in\mathcal{S}} P^{a}(x,y)v^*(y;\gamma\beta^{i+1})\}.$$

Then,

$$\begin{aligned}
&\exp\left(\gamma\beta^i c(x,a_1)\right)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^k(y;\gamma\beta^{i+1})\\
&\quad-\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)v^*(y;\gamma\beta^{i+1})\\
&\leq\exp\left(\gamma\beta^i c(x,a_1)\right)\sum_{y\in\mathcal{S}} P^{a_1}(x,y)u^{k-1}(y;\gamma\beta^{i+1})\\
&\quad-\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)v^*(y;\gamma\beta^{i+1})\\
&\leq\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)u^{k-1}(y;\gamma\beta^{i+1})\\
&\quad-\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)v^*(y;\gamma\beta^{i+1}),
\end{aligned}$$

the last inequality holds because $a_1\in\arg\min\{\exp\left(\gamma\beta^i c(x,a)\right)\sum_{y\in\mathcal{S}} P^a(x,y)u^{k-1}(y;\gamma\beta^{i+1})\}$.

Then,

$$\begin{aligned}
&\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)u^{k-1}(y;\gamma\beta^{i+1})\\
&-\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)v^*(y;\gamma\beta^{i+1})
\end{aligned}$$

$$\begin{aligned}
&=\exp\left(\gamma\beta^i c(x,a_2)\right)\sum_{y\in\mathcal{S}} P^{a_2}(x,y)[u^{k-1}(y;\gamma\beta^{i+1})-v^*(y;\gamma\beta^{i+1})]\\
&\leq\exp\left(\gamma\beta^i\|c\|\right)\max_{y\in\mathcal{S}}\left|u^{k-1}(y;\gamma\beta^{i+1})-v^*(y;\gamma\beta^{i+1})\right|.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&|u^k(x;\gamma)-v^*(x;\gamma)|\\
\leq&\exp\left(\gamma\|c\|\right)\max_{y\in\mathcal{S}}\left|u^{k-1}(y;\gamma\beta)-v^*(y;\gamma\beta)\right|\\
\leq&\exp\left(\gamma\|c\|\right)\exp\left(\gamma\beta\|c\|\right)\max_{y\in\mathcal{S}}\left|u^{k-2}(y;\gamma\beta^2)-v^*(y;\gamma\beta^2)\right|\\
&...\\
\leq&\exp\left(\gamma\|c\|\right)\exp\left(\gamma\beta\|c\|\right)...\exp\left(\gamma\beta^{\lfloor\frac{k}{2}\rfloor+1}\|c\|\right)\max_{y\in\mathcal{S}}\left|u^{k-\lfloor\frac{k}{2}\rfloor-1}(y;\gamma\beta^{\lfloor\frac{k}{2}\rfloor+1})-v^*(y;\gamma\beta^{\lfloor\frac{k}{2}\rfloor+1})\right|\\
\leq&\exp\left(\gamma\|c\|\frac{1-\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right)\left|\exp\left(\gamma\|c\|\frac{\beta^{\lfloor\frac{k}{2}\rfloor+1}}{1-\beta}\right)-1\right|.
\end{aligned}$$

□