# Modified Policy Iteration for Exponential Cost Risk Sensitive MDPs

**Yashaswini Murthy** YMURTHY2@ILLINOIS.EDU
*University of Illinois at Urbana-Champaign*

**Mehrdad Moharrami** MOHARAMI@ILLINOIS.EDU
*University of Illinois at Urbana-Champaign*

**R. Srikant** RSRIKANT@ILLINOIS.EDU
*University of Illinois at Urbana-Champaign*

**Editors:** N. Matni, M. Morari, G. J. Pappas

## Abstract

Modified policy iteration (MPI) also known as optimistic policy iteration is at the core of many reinforcement learning algorithms. It works by combining elements of policy iteration and value iteration. The convergence of MPI has been well studied in the case of discounted and average-cost MDPs. In this work, we consider the exponential cost risk-sensitive MDP formulation, which is known to provide some robustness to model parameters. Although policy iteration and value iteration have been well studied in the context of risk sensitive MDPs, modified policy iteration is relatively unexplored. We provide the first proof that MPI also converges for the risk-sensitive problem in the case of finite state and action spaces. Since the exponential cost formulation deals with the multiplicative Bellman equation, our main contribution is a convergence proof which is quite different than existing results for discounted and risk-neutral average-cost problems.

**Keywords:** Robust stochastic control, dynamic programming, risk-sensitive stochastic control

## 1. Introduction

We consider stochastic control problems over finite state and action spaces, also known as Markov Decision Processes (MDPs). Traditional solutions to such problems use policy iteration, value iteration or linear programming (Bertsekas (2012b), Bertsekas (2012a), Puterman (2014)). Reinforcement learning attempts to solve the control problem when the probability transition matrix is either unknown or the probability transition matrix is known but the state space is very large to obtain exact solutions (Sutton and Barto (2018)). Much of the prior work in this area focuses on discounted-cost problems or average-cost problems. In this paper, we study a robust version of the average-cost problem.

Robust control problems with linear state-space and quadratic costs have been well studied in the control theory literature (Zhou and Doyle (1998), Dullerud and Paganini (2013), Başar and Bernhard (2008)). It is also well-known that these robust control problems are closely related to the control of systems with a risk-sensitive exponential cost (Whittle (1990)). Here, we consider the finite-state, finite-action counterpart of such robust/risk-sensitive control problems Borkar (2002, 2010, 2001). Unlike, the LQG setting in Whittle (1990), the risk-sensitive MDP does not admit a closed-form solution even when the system model is known.

The reinforcement learning (RL) problem in risk-sensitive MDPs have been considered in several papers: (i) Borkar (2002) presents a Q-learning algorithm for the tabular case; (ii) Fei et al.

(2020) provide regret bounds for risk sensitive Q-learning and risk sensitive value iteration in the context of finite horizon MDPs (iii) Hai et al. (2022) address risk sensitive RL in the discounted-cost setting through the use of time dependent risk factors, (iv) Moharrami et al. (2022) provide a trajectory based policy gradient algorithm to obtain a stationary point of the risk sensitive objective function and (v) Cavazos-Cadena and Montes-de Oca (2003) consider risk sensitive value iteration for MDPs with multiple communicating classes that permit the existence of risk sensitive average cost. However, their analysis depends crucially on value function iterates being related through the optimal risk sensitive Bellman Operator, which is not the case with modified policy iteration. In general,These algorithms have one of the following limitations: they do not solve the infinite-horizon, risk-sensitive average-cost problem that we are interested in or are not computationally feasible or do not find a global optimal policy. For these reasons, we focus on problems where the model is known but obtaining the solution may be computationally infeasible. Many major successes in RL fall in this category, e.g., board game-playing AI programs such as AlphaGo, AlphaGo Zero and AlphaZero. Recently, there have several papers studying such RL problems using versions of dynamic programming techniques that are computationally more tractable compared to traditional value iteration or policy iteration (Efroni et al. (2018), Winnicki et al. (2021), Winnicki and Srikant (2022)). These algorithms use two key ideas: (i) modified policy iteration: some version of policy iteration is used, where instead of exact policy evaluation, a few iterations of fixed-point iterations are performed (Puterman (2014)), and (ii) approximate policy iteration: both the policy evaluation and the few iterations of fixed-point iterations mentioned in (i) are performed approximately (Bertsekas (2012a)). As shown in Efroni et al. (2018); Winnicki et al. (2021); Winnicki and Srikant (2022), modified and approximate policy iterations can be used to model the concepts used in practical RL algorithms such as tree search, rollout, lookahead, and function approximation, However, all the known results in this context are for the discounted-cost infinite-horizon problem.

To develop the analog of the rich theory that exists for discounted-cost problems, one has to first develop a theory for modified policy iteration and approximate policy iteration in the context of risk-sensitive exponential cost MDPs. For risk-neutral average cost problems, there exists a theory of modified policy iteration (Van der Wal (1980)) but no complete theory for approximate policy iteration exists. For risk-sensitive MDPs, we are unaware of any results for either modified policy iteration or approximate policy iteration. In this paper, as a first step towards developing a theory of RL for risk-sensitive problems with known but large probability transition matrices, we define the equivalent of modified policy iteration in the case of risk-sensitive MDPs and prove that it converges. In the case of discounted-cost problems and average-cost problems, the proof of convergence relies on the properties of the Bellman operator which is additive in those cases. Our main contribution in this paper is to show that the modified policy iteration algorithm converges in the risk-sensitive setting despite the fact that the Bellman operator has multiplicative terms instead of additive terms, which makes much of the existing theory of modified policy iteration inapplicable to our problem. We will detail the differences in the proof techniques when we present the mathematical results later in the paper.

The rest of the paper is organized as follows. In Section 2, we present a brief introduction to risk-sensitive MDPs and in Section 3, we present the modified policy iteration algorithm, including a specific normalization technique to ensure that the value function remains bounded. We note that a large class of normalizations are possible in the case of risk-neutral average-cost problems, but a specific form appears to be required in the case of the risk-sensitive cost problems. The main results

are in Section 4, and the proofs of the supporting lemmas can be found in the appendix of Murthy et al. (2023).

## 2. Preliminaries

In this section, we present our notation and briefly overview the risk-sensitive average cost formulation and the associated multiplicative Bellman Operator.

We consider a Markov decision process with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and transition kernel $\mathbb{P}$. The class of deterministic policies is denoted by $\Pi = \{f : \mathcal{S} \to \mathcal{A}\}$, where each policy assigns an action to each state. Given a policy $f \in \Pi$, the underlying Markov process is denoted by $\mathbb{P}_f : \mathcal{S} \to \mathcal{S}$, where $\mathbb{P}_f(s'|s) := \mathbb{P}(s'|s, f(s))$ is the probability of moving to state $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ upon taking action $f(s) \in \mathcal{A}$. Associated with each state-action pair $(s, f(s))$, there is a one-step cost which is denoted by $c_f(s) := c(s, f(s)) \in [\underline{c}, \bar{c}]$. We assume that the Markov process associated with each deterministic policy $f \in \Pi$ is irreducible and aperiodic. To ensure this, one can replace $\mathbb{P}$ with $\tilde{\mathbb{P}} = (1-\epsilon)\mathbb{P} + \epsilon \mathbf{1}\mathbf{1}^\top$ where $\mathbf{1}$ is the all-one column vector and $\epsilon > 0$ is a fixed constant. We summarize our assumptions below.

**Assumption 1** *We assume that the state space and the action space are finite, and the one-step cost associated with each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is deterministic and bounded. We also assume that the Markov process associated with each deterministic policy $f \in \Pi$ is irreducible and aperiodic.*

### 2.1. Risk Sensitive Average Cost Formulation

The average cost $J_f$ associated with a deterministic policy $f \in \Pi$ is given by,

$$J_f = \lim_{t\to\infty} \frac{1}{t} \mathbb{E}\left[\sum_{k=0}^{t-1} c_f(s_k)\right].$$

Here the expectation is taken with respect to the transition probability $\mathbb{P}_f$ associated with the policy $f$. Equivalently, the average cost can be written in terms of the stationary distribution $\eta_f$ associated with the policy $f$ as:

$$J_f = \mathbb{E}_{s\sim\eta_f}\left[c_f(s)\right].$$

The traditional goal of reinforcement learning with average cost criteria is to minimize $J_f$ across all policies $f \in \Pi$. An approach to robust reinforcement learning is to take into account the model uncertainties and to minimize the worst-case average cost over a *KL*-ball around the nominal model:

$$\sup_{Q:\mathbb{E}_{s\sim\eta_Q}\left(D_{KL}\left(Q(s,\cdot)\|\mathbb{P}_f(s,\cdot)\right)\right)\le\beta} \mathbb{E}_{s\sim\eta_Q}\left[c_f(s)\right],$$

where $D_{KL}$ denotes the Kullback-Leibler divergence, and $\beta > 0$ is the radius of the *KL*-ball. This is known as the robust MDP objective. The dual formulation of the robust MDP objective is:

$$\sup_{Q\ll\mathbb{P}_f} \mathbb{E}_{s\sim\eta_Q}\left[c_f(s)\right] - \frac{1}{\alpha}\mathbb{E}_{s\sim\eta_Q}\left[D_{KL}\left(Q(s,\cdot)\|\mathbb{P}_f(s,\cdot)\right)\right],$$

where the constant $\alpha = \alpha(\beta) > 0$ depends on $\beta$ and $\ll$ represents absolute continuity. Using the Donsker-Varadhan variational formula and Collatz–Wielandt formula, it can be shown that optimizing the robust MDP objective is equivalent to minimizing

$$\Lambda_f(\alpha) = \lim_{t\to\infty} \frac{1}{t} \ln \left( \mathbb{E}\left[ \exp\left( \sum_{k=0}^{t-1} \alpha c_f(s_k) \right) \middle| s_0 = i \right] \right), \tag{1}$$

where the expectation is taken with respect $\mathbb{P}_f$. The existence of the above limit is a consequence of the Perron-Frobenius theorem, whose details can be found in Moharrami et al. (2022), Basu et al. (2008). $\Lambda_f(\alpha)$ is known as the risk sensitive average cost. Similar to $J_f$, the value of $\Lambda_f(\alpha)$ does not depend on the initial state $s_0$. $\alpha$ is thus referred to as the risk factor, since larger values of $\alpha$ implies greater risk averseness. Note that in the limit as $\alpha \to 0$, the risk-sensitive average cost converges to the risk neutral average cost, i.e., $\lim_{\alpha\to 0} \Lambda_f(\alpha) = J_f$. For simplicity, from now on, we fix $\alpha > 0$ and write $\Lambda_f$ instead of $\Lambda_f(\alpha)$.
The above risk sensitive average cost can be expressed as the solution to the following multiplicative Bellman equation,

$$e^{\Lambda_f} e^{V_f(i)} = e^{\alpha c_f(i)} \sum_{j\in\mathcal{S}} \mathbb{P}_f(j|i) e^{V_f(j)}, \quad \forall\, i \in \mathcal{S}, \tag{2}$$

where the relative value function $e^{V_f}$ is the eigenvector corresponding to the Perron-Frobenius eigenvalue $\Lambda_f$ associated with the matrix $M = [M]_{i,j} = [e^{\alpha c_f(i)} \mathbb{P}(j|i, f(i))]_{i,j}$.
Consequently, the multiplicative Bellman operator corresponding to a policy $f$, is an operator $\mathsf{T}_f : \mathbb{R}_+^{|\mathcal{S}|} \to \mathbb{R}_+^{|\mathcal{S}|}$ defined as:

$$\mathsf{T}_f e^V(i) = e^{\alpha c_f(i)} \sum_{j\in\mathcal{S}} \mathbb{P}_f(j|i) e^{V(j)}.$$

The multiplicative Bellman optimality operator $\mathsf{T} : \mathbb{R}_+^{|\mathcal{S}|} \to \mathbb{R}_+^{|\mathcal{S}|}$ is defined as:

$$\mathsf{T} e^V(i) = \min_{f\in\Pi} \mathsf{T}_f e^V(i), \qquad \forall i \in \mathcal{S}.$$

The optimal risk sensitive average cost is defined as the minimum risk averse average cost across all policies, i.e.,

$$\Lambda^* = \min_{f\in\Pi} \Lambda_f = \min_{f\in\Pi} \lim_{t\to\infty} \frac{1}{t} \ln \left( \mathbb{E}\left[ \exp\left( \sum_{k=0}^{t-1} \alpha c_f(s_k) \right) \middle| s_0 = i \right] \right). \tag{3}$$

Let $f \in \Pi$ denote the deterministic policy for which $\Lambda_f = \Lambda^*$, and let $e^{V^*} = e^{V_f}$ denote its relative value function. It can be shown that the pair $(\Lambda^*, e^{V^*})$ is the unique solution (up to multiplicative constant of $e^{V^*}$) to the following equation:

$$e^{\Lambda^*} e^{V^*(i)} = \min_{f\in\Pi} e^{\alpha c_f(i)} \sum_{j\in\mathcal{S}} \mathbb{P}_f(j|i) e^{V(j)}, \quad \forall\, i \in \mathcal{S}. \tag{4}$$

## 3. Problem Formulation

The goal of robust reinforcement learning is to find a policy $f \in \Pi$ for which $\Lambda_f = \Lambda^*$. In this work, we focus on developing a modified policy iteration to find such an optimal policy. To this end, we change the dynamics of the underlying MDP by transforming its transition probability as well as the one-step cost function. It can be shown that the optimality of a policy will not be affected by this transformation. Similar ideas have been used in the case of risk neutral average cost; however, the underlying transformation is different.

More specifically, fixing a constant $\kappa \in (0,1)$, we transform the dynamics of the MDP as follows:

- The transformed cost is given by:

$$d_f(i) = \frac{1}{\alpha}\log((1-\kappa)e^{\alpha c_f(i)} + \kappa), \qquad \forall i \in \mathcal{S}.$$

- The transformed transition probabilities are given by:

$$\mathbb{Q}(j|i,a) = \frac{(1-\kappa)e^{\alpha c(i,a)}\mathbb{P}(j|i,a) + \kappa \mathbf{1}(i=j)}{(1-\kappa)e^{\alpha c(i,a)} + \kappa}, \qquad \forall (i,a) \in \mathcal{S} \times \mathcal{A},$$

  where $\mathbf{1}(i=j)$ is the indicator function. For any policy $f \in \Pi$, $\mathbb{Q}_f(j|i)$ denotes the probability of moving to state $j \in \mathcal{S}$ from state $i \in \mathcal{S}$ upon taking action $f(i)$.

Notice that for all $(i,a) \in \mathcal{S} \times \mathcal{A}$, we have $\mathbb{Q}(i|i,a) \geq \frac{\kappa}{(1-\kappa)e^{\alpha \bar{c}} + \kappa} > 0$. In particular, the probability of staying in the same state under all policies is non-zero. In literature, such a transformation is referred to as the aperiodicity transformation. Next, we state a theorem that establishes a one-to-one correspondence between the optimal risk sensitive average cost and the associated relative value function in the original MDP and the transformed MDP. Hence, finding an optimal policy for the transformed dynamics is equivalent to finding an optimal policy for the original MDP.

**Theorem 2** *Given $\kappa \in (0,1)$, we have the followings:*

1. *Given $(\Lambda^*, e^{V^*})$ satisfies (4), define*

$$\tilde{\Lambda}^* = \log((1-\kappa)e^{\Lambda^*} + \kappa)$$

   *Then $(\tilde{\Lambda}^*, e^{V^*})$ solves the following multiplicative Bellman equation:*

$$e^{\tilde{\Lambda}^*} e^{V^*(i)} = \min_{f \in \Pi} e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}_f(j|i) e^{V^*(j)}, \quad \forall\, i \in \mathcal{S}. \tag{5}$$

2. *Conversely, given $(\tilde{\Lambda}^*, e^{V^*})$ satisfies (5), then*

$$e^{\tilde{\Lambda}^*} \geq \kappa.$$

   *Define*

$$\Lambda^* = \log\left(\frac{e^{\tilde{\Lambda}^*} - \kappa}{1-\kappa}\right). \tag{6}$$

   *Then the pair $(\Lambda^*, e^{V^*})$ satisfies (4).*

**Proof** The proof of the above theorem can be found in Cavazos-Cadena and Montes-de Oca (2003). It can also be verified that both the transformed and original problems possess the same optimal policies. ■

A crucial component to the convergence of the algorithm is a source of contraction, which is obtained from any finite product of ergodic matrices. The transformation described is necessary to ensure that such a contraction exists and is a consequence of the lemma stated below.

**Lemma 3** *There exists a finite natural number $R$ such that for any sequence of policies $f_1$, $f_2$, $\cdots$, $f_R \in \Pi$,*

$$\min_{i,j \in \mathcal{S}} \mathbb{Q}_{f_1} \mathbb{Q}_{f_2} \cdots \mathbb{Q}_{f_R}(j|i) > 0. \tag{7}$$

The modified policy iteration algorithm in the context of risk sensitive exponential cost MDPs for the transformed problem is stated below.

### 3.1. Algorithm

The algorithm takes as input a sequence of natural numbers $(m_i : i \in \mathbb{N})$ such that $m_i \geq 1$ and a vector $V_0' \in \mathbb{R}^n$ such that $\sum_{i \in \mathcal{S}} e^{V_0'(i)} = 1$.

**Algorithm 1** Risk Sensitive Modified Policy Iteration

**Require:** $(m_i : i \in \mathbb{N})$, $V_0'$.

1: Set $k = 0$

2: Set $f_{k+1}(i) = \arg\min_{f \in \Pi} e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}(j \mid i, f(i)) e^{V_k'(j)} \quad \forall i \in \mathcal{S}$ ▷ Policy Improvement

Define $e^{\alpha d_{f_{k+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}(j \mid i, f_{k+1}(i)) e^{V_k'(j)} = \left(\mathsf{T}_{f_{k+1}} e^{V_k'}\right)(i)$

3: $e^{V_{k+1}(i)} \leftarrow \left(\mathsf{T}_{f_{k+1}}^{m_k} e^{V_k'}\right)(i)$ for all $i \in \mathcal{S}$. ▷ Partial Policy Evaluation

4: $e^{V_{k+1}'(i)} \leftarrow \frac{e^{V_{k+1}(i)}}{\sum_i e^{V_{k+1}(i)}}$ for all $i \in \mathcal{S}$ ▷ Normalization

Along with the partial policy evaluation and policy improvement steps, we also introduce a normalization step where the value functions are scaled in every iteration. In the case of risk-neutral average-cost modified policy iteration, the normalization step generally involves subtracting the value function at some fixed state from the rest of the states. This ensures that the value function iterates do not diverge with repeated execution of the algorithm. However, a similar normalization trick would not work for risk sensitive modified policy iteration as not only do we need to ensure that the value functions do not diverge, it is also necessary to make sure that they are uniformly bounded away from zero. The value function being bounded away from zero is crucial to the convergence of the proof as will be seen in the subsequent section.

## 4. Convergence Analysis of Algorithm

Let the risk sensitive average cost associated with policy $f_{n+1}$ for the transformed model be represented as $\tilde{\Lambda}_{f_{n+1}}$. In the context of value iteration, it is well known that the consecutive value function iterates possess a span-seminorm contraction property (Bielecki et al. (1999), Borkar and

Meyn (2002)). More precisely, let $g, h \in \mathbb{R}^n$. Then there exist constants $\tau, k, r$ such that $0 < \tau < 1$, and $\mathbb{N} \ni k, r < \infty$ such that

$$\text{sp}\left(g_k - h_k\right) \leq \tau^r \text{sp}\left(g - h\right),$$

where the span of a vector $v$ is defined as $\text{sp}(v) = \max_i v(i) - \min_i v(i)$ and

$$g_k(i) = \min_{f \in \Pi} \left\{ \alpha d_f(i) + \ln \left( \sum_{j \in \mathcal{S}} \mathbb{Q}(j|i, f(i)) e^{g_{k-1}(y)} \right) \right\}.$$

A similar contraction in the sup norm is satisfied in the discounted-cost setting, where the discount factor serves as the source of contraction. A major roadblock in the convergence analysis of modified policy iteration in the average-cost setting (both risk-neutral and risk-sensitive) is that such a property is not satisfied by consecutive value function iterates. To circumvent this issue, we exploit an alternate property associated with the ratio of iterates obtained through a single step of policy improvement. In order to explain this property, we define:

$$g_n(i) = \frac{\mathsf{T} e^{V'_n}(i)}{e^{V'_n}(i)} \tag{8}$$

and set $u_n$ and $\ell_n$ as

$$u_n = \max_{i \in \mathcal{S}} \left(g_n(i)\right) \tag{9}$$

$$\ell_n = \min_{i \in \mathcal{S}} \left(g_n(i)\right) \tag{10}$$

**Lemma 4** *Let $\tilde{\Lambda}^*$ be the optimal risk sensitive average cost associated with the MDP considered in Algorithm 1. Then $\forall n > 0$:*

$$\ell_n \leq e^{\tilde{\Lambda}^*} \leq e^{\tilde{\Lambda}_{f_{n+1}}} \leq u_n \tag{11}$$

The above lemma is crucial to the proof of convergence of modified policy iteration.

A similar relation would hold for the reward maximization problem: $\ell_n \leq e^{\tilde{\Lambda}_{f_{n+1}}} \leq e^{\tilde{\Lambda}^*} \leq u_n$. Such a relation can be obtained in the context of risk-neutral average cost (Van der Wal (1980)) as well. But since the Bellman Operator is additive in that regime, the proof is relatively straightforward. The multiplicative nature of Bellman operator combined with the exponential cost formulation, necessitates a different proof idea which hinges on the careful utilization of the Perron-Frobenius theorem.

Such an observation helps us establish a contraction necessary to prove the convergence of $u_n$ to the optimal cost. Since $u_n$ is lower bounded by $e^{\tilde{\Lambda}^*}$, it is possible to show exponential convergence of $u_n$ (and therefore consequently $e^{\tilde{\Lambda}_{f_{n+1}}}$) to $e^{\tilde{\Lambda}^*}$. This is possible since $u_n$ is monotonically decreasing and evidently lower bounded.

**Lemma 5** *The sequence $u_n$ is non-increasing, i.e. $u_n \leq u_{n-1}$ for all $n$.*

Analogously, in the case of risk sensitive reward maximization, the sequence $\ell_n$ is monotonic in nature, that is, $\ell_n \geq \ell_{n-1}$.

Value Iteration leads to monotonicity in $u_n$(non-increasing) and $\ell_n$(non-decreasing). This is a consequence of improving the policy at every iteration without any partial policy evaluation. This symmetric monotonicity leads to an overall span contraction in the value function. However, due to partial policy evaluation in modified policy iteration, such a monotonicity is observed only for the maximum of the ratio of iterates, ie., $u_n$ (or $\ell_n$ in case risk sensitive reward maximization). Consequently, there need not be a span contraction for the value functions. Hence it is necessary to rely on arguments independent of span in order to prove algorithm convergence. This approach is delineated in the theorem below.

**Theorem 6** *Let $g_n, u_n$ and $\ell_n$ be determined from Algorithm 1 as per (8), (9) and (10) respectively. Then, $u_n$ converges exponentially fast, i.e. there exist $\gamma, k$ such that $0 < \gamma < 1$ and for each $n$:*

$$\left(u_n - e^{\tilde{\Lambda}^*}\right) \leq (1-\gamma)\left(u_{n-k} - e^{\tilde{\Lambda}^*}\right).$$

*Consequently, the risk sensitive average cost iterates converge to $\tilde{\Lambda}^*$, that is,*

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} \ell_n = e^{\tilde{\Lambda}^*}. \tag{12}$$

Before proving Theorem 6, it is necessary to prove the boundedness of the value function iterates $e^{V'_n(i)}$ for all $n > 0$. The parameter $\gamma$ in Theorem 6 is obtained as a function of the product of ergodic matrices and value function vectors $e^{V'_n}$. Hence in order for $\gamma$ to be strictly positive, it is necessary that the sequence $e^{V'_n}$ is uniformly bounded away from zero. The normalization step in Algorithm 1 serves this purpose along with ensuring that the magnitude of the iterates do not diverge.

**Lemma 7** *Let $\max_k m_k < C$, where $m_k$ corresponds to the number of fixed point iterations performed during partial policy evaluation during the kth execution of the algorithm. Then, there exists $\beta$ such that $0 < \beta < 1$,*

$$e^{V'_m(i)} > \beta > 0 \; \forall \; m \geq 0. \tag{13}$$

We are now ready to present the proof of Theorem 6.

**Proof** By definition of $g_n$, we have

$$g_n(i) = \frac{\mathsf{T}e^{V'_n}(i)}{e^{V'_n(i)}} = \frac{\left(\widetilde{\mathbb{Q}}_{f_{n+1}} e^{V'_n}\right)(i)}{e^{V'_n}(i)} \overset{(a)}{\leq} \frac{\left(\widetilde{\mathbb{Q}}_{f_n} e^{V'_n}\right)(i)}{e^{V'_n}(i)} \overset{(b)}{=} \frac{\left(\widetilde{\mathbb{Q}}_{f_n} e^{V_n}\right)(i)}{e^{V_n(i)}},$$

where $\left(\widetilde{\mathbb{Q}}_{f_n} e^{V_n}\right)(i) = e^{\alpha d_{f_n}(i)} \sum_{j\in\mathcal{S}} \mathbb{Q}(j|i, f_n(i)) e^{V_n(j)}$ (a) follows from the fact that $f_{n+1}$ is the minimizing policy, and (b) is due to $e^{V'_n(i)} = \frac{e^{V_n(i)}}{\sum_{j\in\mathcal{S}} e^{V_n(j)}}$.

Using the definition of $e^{V_n}(i)$, we have

$$\begin{aligned} g_n(i) &\leq \frac{\left(\left(\widetilde{\mathbb{Q}}_{f_n}\right)\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \cdot e^{V'_{n-1}}\right)\right)(i)}{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} e^{V'_{n-1}}\right)(i)} \\ &= \frac{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \widetilde{\mathbb{Q}}_{f_n} e^{V'_{n-1}}\right)(i)}{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} e^{V'_{n-1}}\right)(i)} \\ &\leq \frac{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \cdot \widetilde{\mathbb{Q}}_{f_{n-1}} e^{V_{n-1}}\right)(i)}{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} e^{V_{n-1}}\right)(i)} \\ &= \frac{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \widetilde{\mathbb{Q}}_{f_{n-1}}^{m_{n-2}} \widetilde{\mathbb{Q}}_{f_{n-1}} e^{V'_{n-2}}\right)(i)}{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \widetilde{\mathbb{Q}}_{f_{n-1}}^{m_{n-2}} e^{V'_{n-2}}\right)(i)}. \end{aligned}$$

Continuing the above for $k$ time steps, we get

$$g_n \leq \frac{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \widetilde{\mathbb{Q}}_{f_{n-1}}^{m_{n-2}} \widetilde{\mathbb{Q}}_{f_{n-2}}^{m_{n-3}} \cdots \widetilde{\mathbb{Q}}_{f_{n-k+1}}^{m_{n-k}} \widetilde{\mathbb{Q}}_{f_{n-k+1}} e^{V'_{n-k}}\right)(i)}{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \widetilde{\mathbb{Q}}_{f_{n-1}}^{m_{n-2}} \widetilde{\mathbb{Q}}_{f_{n-2}}^{m_{n-3}} \cdots \widetilde{\mathbb{Q}}_{f_{n-k+1}}^{m_{n-k}} e^{V'_{n-k}}\right)(i)}.$$

Let $H_{n,k} := \widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} \widetilde{\mathbb{Q}}_{f_{n-1}}^{m_{n-2}} \widetilde{\mathbb{Q}}_{f_{n-2}}^{m_{n-3}} \cdots \widetilde{\mathbb{Q}}_{f_{n-k+1}}^{m_{n-k}}$. From Lemma 3, we know that $\mathbb{Q}$ induces an irreducible Markov chain for any sequence of policies, i.e.:

$$\exists\, R < \infty \text{ such that } \forall\, \pi_1, \cdots, \pi_R \in \Pi\text{: } \left(\mathbb{Q}_{\pi_1} \mathbb{Q}_{\pi_2} \cdots \mathbb{Q}_{\pi_R}\right)(j|i) > 0 \quad \forall i, j.$$

The number of time steps $k$ is determined such that $m_{n-1} + m_{n-2} + \cdots + m_{n-k} \geq R$. This implies that $H_{n,k}(j \mid i) > 0$ for all $i, j$.

Let $e^{W'_{n-k}} := \widetilde{\mathbb{Q}}_{f_{n-k+1}} e^{V'_{n-k}}$. We have

$$\begin{aligned} g_n(i) &\leq \frac{\left(H_{n,k} e^{W'_{n-k}}\right)(i)}{H_{n,k} e^{V'_{n-k}}(i)} \\ &= \frac{\sum_{j \in \mathcal{S}} H_{n,k}(j \mid i) e^{W'_{n-k}(j)}}{\sum_{\ell \in \mathcal{S}} H_{n,k}\left(\ell \mid i\right) e^{V'_{n-k}(\ell)}} \\ &= \frac{\sum_{j \in \mathcal{S}} \left(H_{n,k}(j \mid i) e^{W'_{n-k}(j)}\right)}{\sum_{\ell \in \mathcal{S}} H_{n,k}(\ell \mid i) e^{V'_{n-k}(\ell)}} \\ &= \frac{\sum_{j \in \mathcal{S}} \left(H_{n,k}(j \mid i) e^{V'_{n-k}(j)}\right) \cdot \left(\frac{e^{W'_{n-k}(j)}}{e^{V'_{n-k}(j)}}\right)}{\sum_{\ell \in \mathcal{S}} H_{n,k}\left(\ell \mid i\right) e^{V'_{n-k}(\ell)}}. \end{aligned}$$

Define a probability measure $q$ as follows:

$$q(j \mid i) := \frac{H_{n,k}(j \mid i) e^{V'_{n-k}(j)}}{\sum_{\ell \in \mathcal{S}} H_{n,k}\left(\ell \mid i\right) e^{V'_{n-k}(\ell)}}$$

Notice that $0 < q(j \mid i) < 1$ since $H_{n,k}(j \mid i) > 0$ for all $i, j$ and $0 < \beta < e^{V'_{n-k}(i)} \leq 1$ (from Lemma 7) for all $i \in \mathcal{S}$. Therefore,

$$g_n(i) \leq \sum_{j \in \mathcal{S}} q(j \mid i) \left( \frac{\left( \widetilde{\mathbb{Q}}_{f_{n-k+1}} e^{V'_{n-k}} \right)(j)}{e^{V'_{n-k}(j)}} \right) = \sum_{j \in \mathcal{S}} q(j \mid i) \left( \frac{\mathsf{T} e^{V'_{n-k}(j)}}{e^{V'_{n-k}(j)}} \right).$$

Let $\gamma := \min_{i,j} q(j \mid i) > 0$. We have

$$g_n(i) \leq \gamma \ell_{n-k} + (1-\gamma)\, u_{n-k} \quad \forall\, i.$$

$$\implies u_n \leq \gamma \ell_{n-k} + (1-\gamma)\, u_{n-k}. \tag{14}$$

Since $\ell_{n-k} \leq e^{\tilde{\Lambda}^*}$, we have

$$u_n \leq \gamma e^{\tilde{\Lambda}^*} + (1-\gamma)\, u_{n-k}.$$

Therefore,

$$\left( u_n - e^{\tilde{\Lambda}^*} \right) \leq (1-\gamma) \left( u_{n-k} - e^{\tilde{\Lambda}^*} \right) \tag{15}$$

Since $u_n \leq u_{n-1}$ from lemma 5 and $u_n \geq e^{\tilde{\Lambda}^*}$ from lemma 4, it follows from (15) that

$$u_n \longrightarrow e^{\tilde{\Lambda}^*}$$

From (14), we obtain

$$e^{\tilde{\Lambda}^*} \leq \gamma \ell_n + (1-\gamma)\, u_n.$$

Therefore,

$$e^{\tilde{\Lambda}^*} - u_n \leq \gamma \left( \ell_n - u_n \right),$$

which yields

$$0 \leq \gamma \left( u_n - \ell_n \right) \leq \left( u_n - e^{\tilde{\Lambda}^*} \right).$$

Since $u_n \to e^{\tilde{\Lambda}^*}$, we conclude that $\ell_n \to u_n \implies \ell_n \to e^{\tilde{\Lambda}^*}$ as desired. ■

From Theorem 2 we can equivalently obtain the original optimal risk sensitive $\Lambda^*$ average cost and the corresponding value function associated with it. Note that if $\frac{(\mathsf{T} e^V)(i)}{e^{V(i)}} = \delta > 0$, then the transformation in Equation (6) provides a $\Lambda$ which is in a $\delta$-scaled neighbourhood of $\Lambda^*$. More details can be found in Cavazos-Cadena and Montes-de Oca (2003).

## 5. Conclusion

We presented a modified policy iteration algorithm which can reduce the computational burden of standard policy iteration for risk-sensitive MDPs. The proof of convergence relies on techniques that are quite different from the existing literature for discounted and risk-neutral average-cost problems. As in prior work for discounted-cost problems, our results can further be used to provide performance guarantees for RL algorithms.

## Acknowledgments

The research presented here was supported by NSF Grants CCF 22-07547, CCF 19-34986, CNS 21-06801 and ONR Grant N00014-19-1-2566.

## References

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Arnab Basu, Tirthankar Bhattacharyya, and Vivek S Borkar. A learning algorithm for risk-sensitive cost. *Mathematics of operations research*, 33(4):880–898, 2008.

D. Bertsekas. *Dynamic Programming and Optimal Control: Volume II; Approximate Dynamic Programming*. Athena Scientific optimization and computation series. Athena Scientific, 2012a. ISBN 9781886529441. URL https://books.google.com/books?id=C1JEEAAAQBAJ.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012b.

Tomasz Bielecki, Daniel Hernández-Hernández, and Stanley R Pliska. Risk sensitive control of finite state markov chains in discrete time, with applications to portfolio management. *Mathematical Methods of Operations Research*, 50(2):167–188, 1999.

Vivek S Borkar. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2): 294–311, 2002.

Vivek S Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5, 2010.

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

Rolando Cavazos-Cadena and Raúl 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.

Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. Beyond the one-step greedy approach in reinforcement learning. In *International Conference on Machine Learning*, pages 1387–1396. PMLR, 2018.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Jia Lin Hai, Marek Petrik, Mohammad Ghavamzadeh, and Reazul Russel. Rasr: Risk-averse soft-robust mdps with evar and entropic risk. *arXiv preprint arXiv:2209.04067*, 2022.

Mehrdad Moharrami, Yashaswini Murthy, Arghyadip Roy, and Rayadurgam Srikant. A policy gradient algorithm for the risk-sensitive exponential cost mdp. *arXiv preprint arXiv:2202.04157*, 2022.

Yashaswini Murthy, Mehrdad Moharrami, and R Srikant. Modified policy iteration for exponential cost risk sensitive mdps. *arXiv preprint arXiv:2302.03811*, 2023.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

J Van der Wal. Successive approximations for average reward markov games. *International Journal of Game Theory*, 9(1):13–24, 1980.

Peter Whittle. *Risk-sensitive optimal control*, volume 2. Wiley, 1990.

Anna Winnicki and R Srikant. Reinforcement learning with unbiased policy evaluation and linear function approximation. *arXiv preprint arXiv:2210.07338, to appear in Proceedings of IEEE Conference on Decision and Control 2022*, 2022.

Anna Winnicki, Joseph Lubars, Michael Livesay, and R Srikant. The role of lookahead and approximate policy evaluation in policy iteration with linear value function approximation. *arXiv preprint arXiv:2109.13419*, 2021.

Kemin Zhou and John Comstock Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.