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

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. In the appendix, the proof of approximate modified policy iteration for risk sensitive MDPs is also provided.

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 or 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, and (iv) Moharrami et al. (2022) provide a trajectory based policy gradient algorithm to obtain a stationary point of the risk sensitive objective function. 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 and in the supplementary material. The key ideas presented in this paper can also be used to provide performance guarantees for approximate policy iteration but we do not include them here due to space limitations. It is worth noting that, with these results, one may not only be able to analyze RL algorithms with known models as in Efroni et al. (2018); Winnicki et al. (2021); Winnicki and Srikant (2022) but one may also be able analyze general RL problems as shown in Chen and Maguluri (2022) for discounted-cost problems.

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, with some of the proofs relegated to the supplementary material.

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 $\Pi = \{f \colon \mathcal{S} \to \mathcal{A}\}$, 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 $c_f(s) := c(s, f(s)) \in [\underline{c}, \overline{c}]$. 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 η_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:

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

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:

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

where the constant $\alpha = \alpha(\beta) > 0$ depends on β. 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. α is thus referred to as the risk factor, since larger values of α 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 Λ_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 Λ_f associated with the matrix $M = [M]_{i,j} = [e^{\alpha c_f(i)}\mathbb{P}\left(j|i, f(i)\right)]_{i,j}$.

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 (Λ^*, 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 following:*

1. *Given (Λ^*, 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 (Λ^*, 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 $f_1, f_2, \cdots, f_R \in \Pi$,*

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

Proof The proof of this lemma can be found in Appendix A.1. ■

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 $\sum_{i \in \mathcal{S}} e^{V_0'(i)} = 1$.

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

Algorithm 1 Risk Sensitive Modified Policy Iteration

```
Require: $(m_i : i \in \mathbb{N})$, $V'_0$.
Set $k = 0$
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 $e^{\alpha d_{f_{k+1}}(i)} \sum_{j\in\mathcal{S}} \mathbb{Q}(j \mid i, f_{k+1}(i)) e^{V'_k(j)} = \left(\mathsf{T}_{f_{k+1}} e^{V'_k}\right)(i)$
$e^{V_{k+1}(i)} \leftarrow \left(\mathsf{T}^{m_k}_{f_{k+1}} e^{V'_k}\right)(i)$ for all $i \in \mathcal{S}$.      ▷ Partial Policy Evaluation
$e^{V'_{k+1}(i)} \leftarrow \frac{e^{V_{k+1}(i)}}{\sum_i e^{V_{k+1}(i)}}$ for all $i \in \mathcal{S}$      ▷ Normalization
```

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 τ, 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 ℓ_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 $\forall n > 0$:*

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

Proof The proof of the above lemma can be found in Appendix A.2. ■

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.*

Proof The proof of this lemma can be found in Appendix A.3. ■

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

Value Iteration leads to monotonicity in u_n(non-increasing) and ℓ_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 ℓ_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 ℓ_n be determined from Algorithm 1 as per (8), (9) and (10) respectively. Then, u_n converges exponentially fast, i.e. there exist γ, 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 γ 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 γ 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 1. Then, there exists β such that $0 < \beta < 1$,*

$$e^{V_m'(i)} > \beta > 0 \,\forall\, m \geq 0. \tag{13}$$

Proof The proof of this lemma can be found in Appendix A.4. The proof once again relies on the ergodicity of the probability transition matrices $\mathbb{Q}$, specifically on Lemma 3. ■

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 $\left(\widetilde{\mathbb{Q}}_{f_n} e^{V_n}\right)(i) = e^{\alpha d_{f_n}(i)} \sum_{j\in\mathcal{S}} \mathbb{P}(j|i, f_n(i)) e^{V_n(j)}$ (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

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

This 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 Λ^* average cost and the corresponding value function associated with it. Note that if $\frac{\left(\mathsf{T}e^V\right)(i)}{e^{V(i)}} = \delta > 0$, then the transformation in Equation (6) provides a Λ which is in a δ-scaled neighbourhood of Λ^*. 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.

Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 11195–11214. PMLR, 2022.

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. Performance bounds for policy-based average reward reinforcement learning algorithms, 2023. URL https://arxiv.org/abs/2302.01450.

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.

Appendix A.

A.1. Proof of Lemma 3

Given finite state and action spaces, the total number of deterministic policies are given by $|\Pi|= |\mathcal{A}|^{|\mathcal{S}|}= m^n$. Since every policy induces an irreducible Markov chain, $\mathbb{P}^n_f(j|i) > 0 \, \forall \, i,j \in \mathcal{S}, \forall \, f \in \Pi$. When $R = m^n + 1$, there exists a policy that is repeated at least twice in the sequence. Hence, if $R = (n-1) \cdot m^n + 1$, there exists a policy which is repeated at least n times. Since in the transformed model, under every policy the probability of staying in the same state is non-zero, there exists a non-zero probability of traversing from any state to any other state when $R \geq (n-1) \cdot m^n + 1$ under any sequence of policies.

A.2. Proof of Lemma 4

- $u_n \geq e^{\tilde{\Lambda}_{f_{n+1}}}$. Recall by definition that

$$u_n = \max_i \left(\frac{\mathsf{T} e^{V'_n(i)}}{e^{V'_n(i)}} \right) \geq \frac{\mathsf{T} e^{V'_n(i)}}{e^{V'_n(i)}} = \frac{\mathsf{T}_{f_{n+1}} e^{V'_n}(i)}{e^{V'_n}(i)} \quad \forall i \in \mathcal{S}$$

Therefore, we have

$$u_n e^{V'_n(i)} \geq e^{d_{f_{n+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}_{f_{n+1}} \left(j \mid i \right) e^{V'_n(j)},$$

and thus

$$u_n e^{V'_n} \geq \widetilde{\mathbb{Q}}_{f_{n+1}} \left(e^{V'_n} \right).$$

where, $\widetilde{\mathbb{Q}}_{f_{n+1}}(i,j) = e^{d_{f_{n+1}}}(i) \mathbb{Q}_{f_{n+1}} \left(j \mid i \right)$

Consequently, it follows that

$$u_n \widetilde{\mathbb{Q}}^k_{f_{n+1}} \left(e^{V'_n} \right) \geq \left(\widetilde{\mathbb{Q}}_{f_{n+1}} \right)^{k+1} e^{V'_n}$$

and

$$\frac{u_n}{e^{\tilde{\Lambda}_{f_{n+1}}}} \left(\frac{\widetilde{\mathbb{Q}}_{f_{n+1}}}{e^{\tilde{\Lambda}_{f_{n+1}}}} \right)^k e^{V'_n} \geq \left(\frac{\widetilde{\mathbb{Q}}_{f_{n+1}}}{e^{\tilde{\Lambda}_{f_{n+1}}}} \right)^{k+1} e^{V'_n}$$

Consequently,

$$\lim_{k \to \infty} \frac{u_n}{e^{\tilde{\Lambda}_{f_{n+1}}}} \left(\frac{\widetilde{\mathbb{Q}}_{f_{n+1}}}{e^{\tilde{\Lambda}_{f_{n+1}}}} \right)^k e^{V'_n} \geq \lim_{k \to \infty} \left(\frac{\widetilde{\mathbb{Q}}_{f_{n+1}}}{e^{\tilde{\Lambda}_{f_{n+1}}}} \right)^{k+1} e^{V'_n}$$

Since $\mathbb{Q}$ satisfies the conditions of the Perron-Frobenius theorem, by definition, so does $\widetilde{Q}$. The vector $e^{V'_n}$ consists of all positive elements and hence, in the limit of $k \longrightarrow \infty$, due to power iteration, the following holds true:

$$\frac{u_n}{e^{\tilde{\Lambda}_{f_{n+1}}}} z \geq z,$$

As $\widetilde{\mathbb{Q}}_{f_{n+1}}$ and $e^{V_n'}$ are both non negative, the resulting z is a non-zero vector containing non negative elements. Hence we obtain,

$$u_n \geq e^{\tilde{\Lambda}_{f_{n+1}}},$$

as desired.

- $e^{\tilde{\Lambda}^*} \leq e^{\tilde{\Lambda}_{f_{n+1}}}$. Recall that

$$\begin{aligned} e^{\tilde{\Lambda}^*} e^{V^*(i)} &= \min_{f \in \Pi} e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f(i)\right) e^{V^*(j)} \\ &\leq e^{\alpha d_{f_{n+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f_{n+1}(i)\right) e^{V^*(j)}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} e^{V^*(i)} &\leq \frac{e^{\alpha d_{f_{n+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f_{n+1}(i)\right) e^{V^*(j)}}{e^{\tilde{\Lambda}^*}} \\ &= \frac{e^{\alpha d_{f_{n+1}}(i)} \mathbb{E}\left[e^{V^*(x_1)} \mid x_0 = i, f_{n+1}\right]}{e^{\tilde{\Lambda}^*}} \\ &\leq \frac{e^{\alpha d_{f_{n+1}}(i)}}{e^{\tilde{\Lambda}^*}} \cdot \mathbb{E}\left[\frac{e^{\alpha d_{f_{n+1}}(x_1)}}{e^{\tilde{\Lambda}^*}} \cdot \mathbb{E}\left[e^{V^*(x_2)} \mid x_1, f_{n+1}\right] \middle| \, x_0 = i, f_{n+1}\right] \end{aligned}$$

Iterating, we get

$$e^{V^*(x_0)} \leq \mathbb{E}\left[\frac{e^{\sum_{i=0}^{k-1} \alpha d_{f_{n+1}}(x_i)} V^*(x_k)}{\left(e^{\tilde{\Lambda}^*}\right)^k} \middle| \, x_0\right]$$

Since $V^*(x_k) \leq M < \infty$ for all $x_k \in \mathcal{S}$, we have

$$k \cdot \tilde{\Lambda}^* + V^*(x_0) \leq \ln\left(\mathbb{E}_{x_0}\left[e^{\alpha \sum_{i=0}^{k-1} d(x_i, f_{n+1}(x_i))}\right]\right) + \ln(M),$$

or equivalently

$$\tilde{\Lambda}^* + \frac{V^*(x_0)}{k} \leq \frac{1}{k} \ln\left(\mathbb{E}_{x_0}\left[e^{\alpha \sum_{i=0}^{k-1} d(x_i, f_{n+1}(x_i))}\right]\right) + \frac{M}{k}.$$

As $k \to \infty$, we obtain

$$\tilde{\Lambda}^* \leq \tilde{\Lambda}_{f_{n+1}}.$$

By monotonicity of the exponential function, we conclude as desired that

$$e^{\tilde{\Lambda}^*} \leq e^{\tilde{\Lambda}_{f_{n+1}}}.$$

- $\ell_n \leq e^{\tilde{\Lambda}^*}$.

  Recall that $e^{\tilde{\Lambda}^*}$ satisfies the following equation:

$$e^{\tilde{\Lambda}^*} e^{V^*} = \min_{f \in \Pi} e^{\alpha d_f} \mathbb{Q}_f \left(e^{V^*} \right).$$

  Let the minimising policy be f^*. We have

$$\ell_n = \min_i \frac{\mathsf{T} e^{V_n'}(i)}{e^{V_n'(i)}} \leq \frac{\mathsf{T} e^{V_n'}(i)}{e^{V_n'(i)}} = \frac{\mathsf{T}_{f_{n+1}} e^{V_n'}(i)}{e^{V_n'(i)}} \leq \frac{\mathsf{T}_{f^*} e^{V_n'(i)}}{e^{V_n'}(i)}.$$

  Therefore,

$$\ell_n e^{V_n'(i)} \leq e^{\alpha d_{f^*}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f^*(i)\right) e^{V_n'(j)} = \left(\widetilde{\mathbb{Q}}_{f^*} \left(e^{V_n'} \right) \right)(i)$$

  It follows that

$$\ell_n \widetilde{\mathbb{Q}}_{f^*}^k \left(e^{V_n'} \right) \leq \widetilde{\mathbb{Q}}_{f^*}^{k+1} e^{V_n'}.$$

  Similarly, we have

$$\frac{\ell_n}{e^{\tilde{\Lambda}^*}} \left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}} \right)^k e^{V_n'} \leq \left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}} \right)^{k+1} e^{V_n'}$$

  As a consequence of Perron-Frobenius theorem, it follows that

$$\ell_n \leq e^{\tilde{\Lambda}^*},$$

  as desired. This concludes the proof.

A.3. Proof of Lemma 5

Recall that

$$u_n = \max_i \frac{\left(\mathsf{T} e^{V_n'} \right)(i)}{e^{V_n'}(i)}$$

Let $x^* = \arg\max_i \frac{\left(\mathsf{T} e^{(V_n')}\right)(i)}{e^{V_n'}(i)}$, so that

$$\begin{aligned} u_n &= \frac{\left(\mathsf{T} e^{(V_n')} \right)(x^*)}{e^{V_n'}(x^*)} \\ &= \frac{\left(\widetilde{\mathbb{Q}}_{f_{n+1}} \left(e^{(V_n')} \right)(x^*) \right)}{e^{V_n'}(x^*)} \\ &\leq \frac{\left(\widetilde{\mathbb{Q}}_{f_n} \left(e^{(V_n')} \right)(x^*) \right)}{e^{V_n'}(x^*)}. \end{aligned}$$

Since $e^{V_n'(j)} = \frac{e^{V_n(j)}}{\sum_{j\in\mathcal{S}} e^{V_n(j)}}$, it follows that

$$
\begin{aligned}
u_n &\leq \frac{\widetilde{\mathbb{Q}}_{f_n}\left(e^{V_n}(x^*)\right)}{e^{V_n(x^*)}} \\
&= \frac{\widetilde{\mathbb{Q}}_{f_n}\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} e^{V_{n-1}'}(x^*)}{\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} e^{V_{n-1}'}(x^*)} \\
&= \frac{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}}\left(\widetilde{\mathbb{Q}}_{f_n} e^{V_{n-1}'}\right)\right)(x^*)}{\left(\widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}} e^{V_{n-1}'}\right)(x^*)}.
\end{aligned}
$$

Let $e^{W_{n-1}'} = \widetilde{\mathbb{Q}}_{f_n} e^{V_{n-1}'}$ and $H_n = \widetilde{\mathbb{Q}}_{f_n}^{m_{n-1}}$. It then follows that

$$
\begin{aligned}
u_n &\leq \frac{\left(H_n e^{W_{n-1}'}\right)(x^*)}{\left(H_n e^{V_{n-1}'}\right)(x^*)} \\
&= \frac{\sum_{j\in\mathcal{S}} H(j \mid x^*) e^{W_{n-1}'(j)}}{\sum_{j\in\mathcal{S}} H(j \mid x^*) e^{V_{n-1}'(j)}}.
\end{aligned}
$$

Let $p = \arg\max_i \frac{e^{W_{n-1}'(i)}}{e^{V_{n-1}'(i)}}$. Then, we have

$$
\frac{e^{W_{n-1}'(p)}}{e^{V_{n-1}'(p)}} \geq \frac{e^{W_{n-1}'(i)}}{e^{V_{n-1}'(i)}} \quad \forall\, i
$$

This yields

$$
u_n \leq \frac{\sum_{j\in\mathcal{S}}\left(H(j \mid x^*) e^{V_{n-1}'(j)} \cdot \frac{e^{W_{n-1}'(p)}}{e^{V_{n-1}'(p)}}\right)}{\sum_{j\in\mathcal{S}} H(j \mid x^*) e^{V_{n-1}'(j)}}.
$$

Therefore,

$$
\begin{aligned}
u_n &\leq \frac{e^{W_{n-1}'(p)}}{e^{V_{n-1}'(p)}} \\
&= \frac{\widetilde{\mathbb{Q}}_{f_n}\left(e^{V_{n-1}'}\right)(p)}{e^{V_{n-1}'(p)}} \\
&= \max_i \frac{\left(\mathsf{T} e^{V_{n-1}'}\right)(i)}{e^{V_{n-1}'(i)}} = u_{n-1},
\end{aligned}
$$

which establishes the desired monotonicity.

A.4. Proof of Lemma 7

Since $e^{V'_k(i)} = \frac{e^{V_k(i)}}{\sum_{j\in\mathcal{S}} e^{V_k(j)}}$, it follows that $\sum_{j\in\mathcal{S}} e^{V'_k(j)} = 1$. We then have

$$e^{V_k(i)} = (T^{m_{k-1}}_{f_k} e^{V'_{k-1}})(i)$$

Let $\underline{d} = \min_{f\in\Pi} d_f$ and $\overline{d} = \max_{f\in\Pi} d_f$. Then,

$$\begin{aligned} T_{f_k} e^{V'_{k-1}}(i) &= e^{\alpha d_f(i)} \sum_{j\in\mathcal{S}} \mathbb{Q}(j|i, f_k(i)) e^{V'_{k-1}(j)} \\ &\geq e^{\alpha\underline{d}} (\mathbb{Q}_{f_k} e^{V'_{k-1}})(i) \end{aligned}$$

Iterating,

$$\begin{aligned} (T^{m_{k-1}}_{f_k} e^{V'_{k-1}})(i) &\geq e^{m_{k-1}\alpha\underline{d}} (\mathbb{Q}^{m_{k-1}}_{f_k} e^{V'_{k-1}})(i) \\ &= \frac{e^{m_{k-1}\alpha\underline{d}} (\mathbb{Q}^{m_{k-1}}_{f_k} e^{V_{k-1}})(i)}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)}} \end{aligned}$$

$$e^{V_k(i)} \geq \frac{e^{m_{k-1}\alpha\underline{d}} (\mathbb{Q}^{m_{k-1}}_{f_k} e^{V_{k-1}})(i)}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)}}$$

Further iterating,

$$e^{V_k(i)} \geq \frac{e^{\sum_{l=1}^k m_{k-l}\alpha\underline{d}} (\mathbb{Q}^{m_{k-1}}_{f_k} \mathbb{Q}^{m_{k-2}}_{f_{k-1}} \dots \mathbb{Q}^{m_0}_{f_1} e^{V'_0})(i)}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)} \sum_{j\in\mathcal{S}} e^{V_{k-2}(j)} \dots \sum_{j\in\mathcal{S}} e^{V_1(j)}}$$

From Algorithm 1, for a sufficiently large k, we have $\sum_{l=0}^{k-1} m_l > R$. Defining $H_k = \mathbb{Q}^{m_{k-1}}_{f_k} \mathbb{Q}^{m_{k-2}}_{f_{k-1}} \dots \mathbb{Q}^{m_0}_{f_1}$, Lemma 3 yields $\varepsilon = \min_{i,j} H_k(i \mid j) > 0$, we continue the above sequence of inequalities:

$$\begin{aligned} e^{V_k(i)} &\geq \frac{e^{\sum_{l=1}^k m_{k-l}\alpha\underline{d}} \sum_{j\in\mathcal{S}} \varepsilon e^{V'_0(j)}}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)} \sum_{j\in\mathcal{S}} e^{V_{k-2}(j)} \dots \sum_{j\in\mathcal{S}} e^{V_1(j)}} && (16) \\ &= \frac{e^{\sum_{l=1}^k m_{k-l}\alpha\underline{d}} \varepsilon}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)} \sum_{j\in\mathcal{S}} e^{V_{k-2}(j)} \dots \sum_{j\in\mathcal{S}} e^{V_1(j)}} && (17) \end{aligned}$$

Similarly we obtain,

$$e^{V_k(i)} \leq \frac{e^{m_{k-1}\alpha\overline{d}} (\mathbb{Q}^{m_{k-1}}_{f_k} e^{V_{k-1}})(i)}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)}}$$

Further iterating,

$$e^{V_k(i)} \leq \frac{e^{\sum_{l=1}^k m_{k-l}\alpha\overline{d}} (\mathbb{Q}^{m_{k-1}}_{f_k} \mathbb{Q}^{m_{k-2}}_{f_{k-1}} \dots \mathbb{Q}^{m_0}_{f_1} e^{V'_0})(i)}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)} \sum_{j\in\mathcal{S}} e^{V_{k-2}(j)} \dots \sum_{j\in\mathcal{S}} e^{V_1(j)}}$$

This implies

$$\sum_{i\in\mathcal{S}} e^{V_k(i)} \leq \frac{ne^{\sum_{l=1}^{k} m_{k-l}\alpha\overline{d}}}{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)} \sum_{j\in\mathcal{S}} e^{V_{k-2}(j)} \dots \sum_{j\in\mathcal{S}} e^{V_1(j)}}$$

Therefore,

$$\frac{1}{\sum_{i\in\mathcal{S}} e^{V_k(i)}} \geq \frac{\sum_{j\in\mathcal{S}} e^{V_{k-1}(j)} \sum_{j\in\mathcal{S}} e^{V_{k-2}(j)} \dots \sum_{j\in\mathcal{S}} e^{V_1(j)}}{ne^{\sum_{l=1}^{k} m_{k-l}\alpha\overline{d}}} \tag{18}$$

Combining (16) and (18), since $\forall l,\ m_l \geq 1$ and $m_l < C$

$$\frac{e^{V_k(i)}}{\sum_{j\in\mathcal{S}} e^{V_k(j)}} \geq \frac{e^{\sum_{l=1}^{k} m_{k-l}\alpha\underline{d}_{\mathcal{E}}}}{ne^{\sum_{l=1}^{k} m_{k-l}\alpha\overline{d}}} > \frac{e^{k\alpha\underline{d}_{\mathcal{E}}}}{ne^{kC\alpha\overline{d}}} > 0 \tag{19}$$

We conclude that $e^{V'_m(i)} > \beta > 0$ for all m, where $\beta = \frac{e^{k\alpha\underline{d}_{\mathcal{E}}}}{ne^{kC\alpha\overline{d}}}$.

Appendix B. Approximate Risk Sensitive Modified Policy Iteration

In this appendix, we prove the convergence of risk sensitive approximate modified policy iteration for exponential cost MDPs subject to certain constraints on the approximation accuracy.
Define the Bellman Operator $\mathsf{T}_f : \mathbb{R}^+ \to \mathbb{R}^+$ as follows:

$$\left(\mathsf{T}_f e^h\right)(i) = e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j|i, f(i)\right) e^{h(j)}$$

Algorithm 2 Approximate Risk Sensitive Modified Policy Iteration

```
Require: $(m_i : i \in \mathbb{N})$, $h_0$ such that $\sum_{i \in \mathcal{S}} e^{h_0(i)} = 1$.
Set $k = 0$
Compute $f_{k+1} \in \Pi$ such that:
    $$1 \leq \left\| \frac{\mathsf{T}_{f_{k+1}} e^{h_k}}{\mathsf{T} e^{h_k}} \right\|_\infty \leq \epsilon.$$
                                                        ▷ Approximate Policy Improvement
Define $e^{h_{f_{k+1}}(i)} \leftarrow \left(\mathsf{T}^{m_k}_{f_{k+1}} e^{h_k}\right)(i)$ for all $i \in \mathcal{S}$.     ▷ Partial Policy Evaluation
$e^{h'_{f_{k+1}}(i)} \leftarrow \frac{e^{h_{f_{k+1}}(i)}}{\sum_i e^{h_{f_{k+1}}(i)}}$ for all $i \in \mathcal{S}$.                 ▷ Normalization
Compute $e^{h_{k+1}}$ such that,
    $$\delta_1 \leq \left\| \frac{e^{h'_{f_{k+1}}}}{e^{h_{k+1}}} \right\|_\infty \leq \delta_2.$$
                                                        ▷ Approximate Policy Evaluation
```

B.1. Convergence Analysis

Consider the following definitions required for the proof of convergence of the algorithm iterates:

$$g_k(i) = \frac{\mathsf{T} e^{h_k}(i)}{e^{h_k(i)}}$$
$$l_k = \min_{i \in \mathcal{S}} g_k(i)$$
$$u_k = \max_{i \in \mathcal{S}} g_k(i)$$

In order to obtain the finite time and asymptotic performance of the iterates obtained as a consequence of the above algorithm, it is necessary to prove the lemma characterising a single step of approximate policy improvement below.

Lemma 8 *Given the iterates l_k, u_k obtained from algorithm 2, $\forall k$ it is true that,*

$$l_k \leq e^{\tilde{\Lambda}^*} \leq e^{\tilde{\Lambda}_{f_{k+1}}} \leq u_k \epsilon \tag{20}$$

Proof Recall the definition of l_k,

$$\begin{aligned} l_k &= \min_i \frac{\mathsf{T}e^{h_k}(i)}{e^{h_k(i)}} \\ &\leq \frac{\mathsf{T}e^{h_k}(i)}{e^{h_k(i)}} \\ &\leq \min_{f\in\Pi} \frac{e^{\alpha d_f(i)}\sum_{j\in\mathcal{S}} \mathbb{Q}\left(j|i,f(i)\right)e^{h_k}(j)}{e^{h_k(i)}} \end{aligned}$$

Let f^* be the policy which yields the lowest risk sensitive average cost e^{Λ^*}. Recall that $e^{\tilde{\Lambda}^*}$ satisfies the following equation:

$$e^{\tilde{\Lambda}^*}e^{h^*} = \min_{f\in\Pi} e^{\alpha d_f}\mathbb{Q}_f\left(e^{h^*}\right).$$

Then,

$$l_k \leq \frac{e^{\alpha d_{f^*}(i)}\sum_{j\in\mathcal{S}} \mathbb{Q}\left(j|i,f^*(i)\right)e^{h_k}(j)}{e^{h_k(i)}}$$

Therefore,

$$\begin{aligned} l_k e^{h_k(i)} &\leq e^{\alpha d_{f^*}(i)}\sum_{j\in\mathcal{S}} \mathbb{Q}\left(j|i,f^*(i)\right)e^{h_k(j)} \\ .&= \left(\widetilde{\mathbb{Q}}_{f^*}\left(e^{h_k}\right)\right)(i) \end{aligned}$$

where $\widetilde{\mathbb{Q}}(j|i,f^*(i)) = e^{\alpha d_{f^*}(i)}\mathbb{Q}\left(j|i,f^*(i)\right)$. It follows that,

$$l_k\left(\widetilde{\mathbb{Q}}^n_{f^*}e^{h_k}\right) \leq \widetilde{\mathbb{Q}}^{n+1}_{f^*}e^{h_k}.$$

Similarly, we have

$$\begin{aligned} \frac{l_k}{e^{\tilde{\Lambda}^*}}\left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}}\right)^n e^{h_k} &\leq \left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}}\right)^{n+1} e^{h_k} \\ \lim_{n\to\infty}\frac{l_k}{e^{\tilde{\Lambda}^*}}\left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}}\right)^n e^{h_k} &\leq \lim_{n\to\infty}\left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}}\right)^{n+1} e^{h_k} \end{aligned}$$

Since $\mathbb{Q}_f$ satisfies the conditions of the Perron-Frobenius theorem, by definition, so does $\widetilde{\mathbb{Q}}_f$ for all $f \in \Pi$. The vector e^{h_k} consists of all positive elements and hence, in the limit of $n \longrightarrow \infty$, due to power iteration we obtain the following:

$$\lim_{n\to\infty}\left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^*}}\right)^n e^{h_k} = z$$

where z is a vector with all non negative entries.
Hence we obtain the following,

$$\frac{l_k}{e^{\tilde{\Lambda}^*}} z \leq z$$

which yields the following:

$$l_k \leq e^{\tilde{\Lambda}^*}$$

as desired.

Recall that

$$\begin{aligned} e^{\tilde{\Lambda}^*} e^{h^*(i)} &= \min_{f \in \Pi} e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f(i)\right) e^{h^*(j)} \\ &\leq e^{\alpha d_{f_{k+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f_{k+1}(i)\right) e^{h^*(j)}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} e^{h^*(i)} &\leq \frac{e^{\alpha d_{f_{k+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j \mid i, f_{k+1}(i)\right) e^{h^*(j)}}{e^{\tilde{\Lambda}^*}} \\ &= \frac{e^{\alpha d_{f_{k+1}}(i)} \mathbb{E}\left[e^{h^*(x_1)} \mid x_0 = i, f_{k+1}\right]}{e^{\tilde{\Lambda}^*}} \\ &\leq \frac{e^{\alpha d_{f_{k+1}}(i)}}{e^{\tilde{\Lambda}^*}} \cdot \mathbb{E}\left[\frac{e^{\alpha d_{f_{k+1}}(x_1)}}{e^{\tilde{\Lambda}^*}} \cdot \mathbb{E}\left[e^{h^*(x_2)} \mid x_1, f_{k+1}\right] \middle| x_0 = i, f_{k+1}\right] \end{aligned}$$

Iterating, we get

$$e^{h^*(x_0)} \leq \mathbb{E}\left[\frac{e^{\sum_{i=0}^{m-1} \alpha d_{f_{k+1}}(x_i)} h^*(x_m)}{\left(e^{\tilde{\Lambda}^*}\right)^m} \middle| x_0\right]$$

Since $h^*(x_m) \leq M < \infty$ for all $x_k \in \mathcal{S}$, we have

$$m \cdot \tilde{\Lambda}^* + h^*(x_0) \leq \ln\left(\mathbb{E}_{x_0}\left[e^{\alpha \sum_{i=0}^{m-1} d(x_i, f_{k+1}(x_i))}\right]\right) + \ln(M),$$

or equivalently

$$\tilde{\Lambda}^* + \frac{h^*(x_0)}{m} \leq \frac{1}{m} \ln\left(\mathbb{E}_{x_0}\left[e^{\alpha \sum_{i=0}^{m-1} d(x_i, f_{k+1}(x_i))}\right]\right) + \frac{M}{m}.$$

As $m \to \infty$, we obtain

$$\tilde{\Lambda}^* \leq \tilde{\Lambda}_{f_{k+1}}.$$

By monotonicity of the exponential function, we conclude as desired that

$$e^{\tilde{\Lambda}^*} \leq e^{\tilde{\Lambda}_{f_{n+1}}}.$$

We now present the last part of the proof.
Recall the definition of u_k,

$$\begin{aligned} u_k &= \max_i \frac{\mathsf{T}e^{h_k}(i)}{e^{h_k(i)}} \\ &\geq \frac{\mathsf{T}e^{h_k}(i)}{e^{h_k(i)}} \end{aligned}$$

From the policy improvement step of the algorithm, we obtain,

$$u_k \geq \frac{\mathsf{T}_{f_{k+1}} e^{h_k}(i)}{\epsilon e^{h_k(i)}}$$

$$\epsilon u_k e^{h_k(i)} \geq e^{\alpha d_{f_{k+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j|i, f_{k+1}(i)\right) e^{h_k(j)}$$

Since $\widetilde{\mathbb{Q}}(j|i, f_{k+1}(i)) = e^{\alpha d_{f_{k+1}}(i)} \mathbb{Q}\left(j|i, f_{k+1}(i)\right)$, it follows that,

$$u_k \epsilon \left(\widetilde{\mathbb{Q}}^n_{f_{k+1}} e^{h_k}\right) \geq \widetilde{\mathbb{Q}}^{n+1}_{f_{k+1}} e^{h_k}.$$

Similarly, we have

$$\begin{aligned} \frac{\epsilon u_k}{e^{\tilde{\Lambda}^{f_{k+1}}}} \left(\frac{\widetilde{\mathbb{Q}}_{f_{k+1}}}{e^{\tilde{\Lambda}^{f_{k+1}}}}\right)^n e^{h_k} &\geq \left(\frac{\widetilde{\mathbb{Q}}_{f_{k+1}}}{e^{\tilde{\Lambda}^{f_{k+1}}}}\right)^{n+1} e^{h_k} \\ \lim_{n \to \infty} \frac{\epsilon u_k}{e^{\tilde{\Lambda}^{f_{k+1}}}} \left(\frac{\widetilde{\mathbb{Q}}_{f_{k+1}}}{e^{\tilde{\Lambda}^{f_{k+1}}}}\right)^n e^{h_k} &\geq \lim_{n \to \infty} \left(\frac{\widetilde{\mathbb{Q}}_{f_{k+1}}}{e^{\tilde{\Lambda}^{f_{k+1}}}}\right)^{n+1} e^{h_k} \end{aligned}$$

Since $\mathbb{Q}_f$ satisfies the conditions of the Perron-Frobenius theorem, by definition, so does $\widetilde{\mathbb{Q}}_f$ for all $f \in \Pi$. The vector e^{h_k} consists of all positive elements and hence, in the limit of $n \longrightarrow \infty$, due to power iteration we obtain the following:

$$\lim_{n \to \infty} \left(\frac{\widetilde{\mathbb{Q}}_{f^*}}{e^{\tilde{\Lambda}^{f_{k+1}}}}\right)^n e^{h_k} = b$$

where b is a vector with all non negative entries.
Hence we obtain the following,

$$\frac{\epsilon u_k}{e^{\tilde{\Lambda}^{f_{k+1}}}} b \geq b$$

which yields the following:

$$u_k \epsilon \geq e^{\tilde{\Lambda}^{f_{k+1}}}$$

as desired. Hence, we obtain,

$$l_k \leq e^{\tilde{\Lambda}^*} \leq e^{\tilde{\Lambda}_{f_{k+1}}} \leq u_k \epsilon$$

■

Given the above lemma, we now present the main theorem of this section that is the performance bound of policy obtained through repeated execution of algorithm 2.

Theorem 9 *Let $\gamma \in (0,1)$. Let $n \in \mathbb{N}$ such that $m_k + m_{k-1} + \cdots + m_{k-n} \geq R$. Suppose the approximation errors δ_1, δ_2 and ϵ are such that*

$$\left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1-\gamma) < 1.$$

Then, the iterates h_k generated by algorithm 2 satisfy:

$$\left(e^{\tilde{\Lambda}_{f_{k+1}}} - e^{\tilde{\Lambda}^*}\right) \leq (\gamma')^{\frac{k}{n}} \left(u_0 \epsilon - e^{\tilde{\Lambda}^*}\right) + \frac{\sigma e^{\tilde{\Lambda}^*}}{1-\gamma'},$$

where $\sigma = \left(\left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1 + (\epsilon - 1)\gamma) - 1\right)$ and $\gamma' = \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1-\gamma)$.

Proof Recall the definition of $g_k(i)$:

$$\begin{aligned} g_k(i) &\leq \frac{\mathsf{T} e^{h_k}(i)}{e^{h_k}(i)} \\ &\leq \frac{\mathsf{T}_{f_k} e^{h_k}(i)}{e^{h_k(i)}} \end{aligned}$$

Since T is a minimizing operator. From algorithm 2, we know that

$$\delta_1 \leq \left\| \frac{e^{h'_{f_k}}}{e^{h_k}} \right\|_\infty \leq \delta_2$$

Since $e^{h_k(i)} \leq \frac{e^{h'_{f_k}(i)}}{\delta_1}$, $\forall i \in \mathcal{S}$, we obtain the following:

$$g_k(i) \leq \frac{\mathsf{T}_{f_k} e^{h'_{f_k}}(i)}{\delta_1 e^{h_k(i)}}$$

Since $\frac{1}{e^{h_k(i)}} \leq \frac{\delta_2}{e^{h'_{f_k}(i)}}$, we obtain the following:

$$g_k(i) \leq \frac{\delta_2}{\delta_1} \frac{\mathsf{T}_{f_k} e^{h'_{f_k}}(i)}{e^{h'_{f_k}}(i)}$$

An important property of $\mathsf{T}_f \forall f \in \Pi$ is its scaling with scaled inputs as shown below. Let $\beta \in \mathbb{R}$ be a constant, then for all $i \in \mathcal{S}$,

$$\begin{aligned} \left(\mathsf{T}_f \left(\beta e^h\right)\right)(i) &= e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j | i, f(i)\right) \beta e^{h(j)} \\ &= \beta e^{\alpha d_f(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}\left(j | i, f(i)\right) e^{h(j)} \\ &= \beta \left(\mathsf{T}_f \left(e^h\right)\right)(i) \end{aligned}$$

Recall that $e^{h'_{f_k}(i)} = \frac{e^{h_{f_k}(i)}}{\sum_i e^{h_{f_k}(i)}}$. Due to the multiplicative scaling property of T_f, substituting for $e^{h'_{f_k}(i)}$ yields the following,

$$g_k(i) \leq \frac{\delta_2}{\delta_1} \frac{\mathsf{T}_{f_k} e^{h_{f_k}}(i)}{e^{h_{f_k}(i)}}$$

From algorithm 2, we know that $e^{h_{f_k}(i)} = \mathsf{T}_{f_k}^{m_{k-1}} e^{h_{k-1}}(i)$,

$$\begin{aligned} g_k(i) &\leq \frac{\delta_2}{\delta_1} \frac{\mathsf{T}_{f_k}^{m_{k-1}+1} e^{h_{k-1}}(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h_{k-1}}(i)} \\ &= \frac{\delta_2}{\delta_1} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T}_{f_k} e^{h_{k-1}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h_{k-1}}(i)} \end{aligned}$$

Since $\mathsf{T}_{f_k} e^{h_{k-1}} \leq \epsilon \mathsf{T} e^{h_{k-1}}$,

$$\begin{aligned} g_k(i) &\leq \frac{\delta_2 \epsilon}{\delta_1} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T} e^{h_{k-1}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h_{k-1}}(i)} \\ &\leq \frac{\delta_2 \epsilon}{\delta_1} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T}_{f_{k-1}} e^{h_{k-1}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h_{k-1}}(i)} \\ &\leq \frac{\delta_2 \epsilon}{\delta_1^2} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T}_{f_{k-1}} e^{h_{f_{k-1}}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h_{k-1}}(i)} \\ &= \frac{\delta_2^2 \epsilon}{\delta_1^2} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T}_{f_{k-1}} e^{h'_{f_{k-1}}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h'_{f_{k-1}}}(i)} \\ &\leq \frac{\delta_2^2 \epsilon}{\delta_1^2} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T}_{f_{k-1}} e^{h_{f_{k-1}}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} e^{h_{f_{k-1}}}(i)} \\ &= \frac{\delta_2^2 \epsilon}{\delta_1^2} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \left(\mathsf{T}_{f_{k-1}}^{m_{k-2}+1} e^{h_{k-2}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} e^{h_{k-2}}(i)} \\ &= \frac{\delta_2^2 \epsilon}{\delta_1^2} \frac{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} \left(\mathsf{T}_{f_{k-1}} e^{h_{k-2}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} e^{h_{k-2}}(i)} \\ &= \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^2 \frac{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} \left(\mathsf{T} e^{h_{k-2}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} e^{h_{k-2}}(i)} \end{aligned}$$

Iterating similarly, we obtain,

$$g_k(i) \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \frac{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} \dots \mathsf{T}_{f_{k-n+1}}^{m_{k-n}} \left(\mathsf{T} e^{h_{k-n}}\right)(i)}{\mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} \dots \mathsf{T}_{f_{k-n+1}}^{m_{k-n}} e^{h_{k-n}}(i)}$$

Define the following:

$$\begin{aligned} H_{n,k} &= \mathsf{T}_{f_k}^{m_{k-1}} \mathsf{T}_{f_{k-1}}^{m_{k-2}} \dots \mathsf{T}_{f_{k-n+1}}^{m_{k-n}} \\ e^{w_{n,k}(i)} &= \mathsf{T} e^{h_{k-n}}(i) \\ e^{v_{n,k}(i)} &= e^{h_{k-n}}(i) \end{aligned}$$

$g_k(i)$ can hence be expressed as:

$$\begin{aligned} g_k(i) &\leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \frac{H_{n,k} e^{w_{n,k}}(i)}{H_{n,k} e^{v_{n,k}}(i)} \\ &= \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \frac{\sum_{j \in \mathcal{S}} H_{n,k}(j|i) e^{w_{n,k}}(j)}{\sum_{j \in \mathcal{S}} H_{n,k}(j|i) e^{v_{n,k}}(j)} \end{aligned}$$

where $H_{n,k}(j|i)$ is the element corresponding to the ith row and jth column.

$$g_k(i) \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \sum_{j \in \mathcal{S}} \left(\frac{H_{n,k}(j|i) e^{v_{n,k}(j)}}{\sum_{y \in \mathcal{S}} H_{n,k}(y|i) e^{v_{n,k}(y)}} \frac{e^{w_{n,k}(j)}}{e^{v_{n,k}(j)}} \right)$$

Define a probability measure $q_{n,k}$ as follows:

$$q_{n,k}(j|i) = \frac{H_{n,k}(j|i) e^{v_{n,k}(j)}}{\sum_{y \in \mathcal{S}} H_{n,k}(y|i) e^{v_{n,k}(y)}}$$

Let $H_{n,k} := \widetilde{\mathbb{Q}}_{f_k}^{m_{k-1}} \widetilde{\mathbb{Q}}_{f_{k-1}}^{m_{k-2}} \widetilde{\mathbb{Q}}_{f_{k-2}}^{m_{k-3}} \cdots \widetilde{\mathbb{Q}}_{f_{k-n+1}}^{m_{k-n}}$. 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 n is determined such that $m_{k-1} + m_{k-2} + \cdots + m_{k-n} \geq R$. This implies that $H_{n,k}(j \mid i) > 0$ for all i, j.
Note from Lemma 10, we obtain that $e^{v_{n,k}(i)} \geq c_1 \geq 0, \forall n, k$, where c_1 is a positive constant. Hence the minimum value of the transition measure $q_{n,k}$ is nonzero and defined as below:

$$\gamma := \min_{i,j} q(j|i)$$

Thus $g_k(i)$ can be bounded as below:

$$\begin{aligned} g_k(i) &\leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma \min_j \frac{e^{w_{n,k}(j)}}{e^{v_{n,k}(j)}} + (1-\gamma) \max_j \frac{e^{w_{n,k}(j)}}{e^{v_{n,k}(j)}} \right) \\ &= \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma \min_j \frac{\mathsf{T} e^{h_{k-n}}(j)}{e^{h_{k-n}}(j)} + (1-\gamma) \max_j \frac{\mathsf{T} e^{h_{k-n}}(j)}{e^{h_{k-n}}(j)} \right) \\ &\leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma l_{k-n} + (1-\gamma) u_{k-n} \right) \end{aligned}$$

From Lemma 8,

$$g_k(i) \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma e^{\tilde{\Lambda}^*} + (1-\gamma) u_{k-n}\right)$$

Since the above relation is true for all $i \in \mathcal{S}$,

$$u_k \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma e^{\tilde{\Lambda}^*} + (1-\gamma) u_{k-n}\right)$$

$$u_k \epsilon \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma \epsilon e^{\tilde{\Lambda}^*} + (1-\gamma) u_{k-n} \epsilon\right)$$

$$\left(u_k \epsilon - e^{\tilde{\Lambda}^*}\right) \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n \left(\gamma \epsilon e^{\tilde{\Lambda}^*} + (1-\gamma)\left(u_{k-n}\epsilon - e^{\tilde{\Lambda}^*}\right)\right) + \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1-\gamma) e^{\tilde{\Lambda}^*} - e^{\tilde{\Lambda}^*}$$

$$\left(u_k \epsilon - e^{\tilde{\Lambda}^*}\right) \leq \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1-\gamma)\left(u_{k-n}\epsilon - e^{\tilde{\Lambda}^*}\right) + e^{\tilde{\Lambda}^*}\left(\left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1 + (\epsilon - 1)\gamma) - 1\right)$$

Define the following:

$$\sigma = \left(\left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1 + (\epsilon - 1)\gamma) - 1\right)$$

$$\gamma' = \left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n (1-\gamma)$$

Important Condition: For γ' to be a valid source of contraction, γ' should be less than 1. Therefore the m_i have to be chosen such that, $\left(\frac{\delta_2 \epsilon}{\delta_1}\right)^n$ is small enough to ensure that γ' is a contraction. Note that $\gamma \leq \frac{1}{|\mathcal{S}|}$. Provided that γ' is less than 1, we have the following:

$$\begin{aligned}\left(u_k \epsilon - e^{\tilde{\Lambda}^*}\right) &\leq \gamma'\left(u_{k-n}\epsilon - e^{\tilde{\Lambda}^*}\right) + \sigma e^{\tilde{\Lambda}^*} \\ &\leq \gamma'\left(\gamma'\left(u_{k-n}\epsilon - e^{\tilde{\Lambda}^*}\right) + \sigma e^{\tilde{\Lambda}^*}\right) + \sigma e^{\tilde{\Lambda}^*}\end{aligned}$$

Note that from Lemma 8, we know that $u_k \epsilon \geq e^{\tilde{\Lambda}^*}, \forall k$. Hence it is possible to iterate the above equation to obtain,

$$\left(u_k \epsilon - e^{\tilde{\Lambda}^*}\right) \leq (\gamma')^{\frac{k}{n}}\left(u_0 \epsilon - e^{\tilde{\Lambda}^*}\right) + \frac{\sigma e^{\tilde{\Lambda}^*}}{1-\gamma'}$$

Since from Lemma 8, we know that $e^{\tilde{\Lambda}_{f_{k+1}}} \leq u_k \epsilon$, we obtain the following,

$$\left(e^{\tilde{\Lambda}_{f_{k+1}}} - e^{\tilde{\Lambda}^*}\right) \leq (\gamma')^{\frac{k}{n}}\left(u_0 \epsilon - e^{\tilde{\Lambda}^*}\right) + \frac{\sigma e^{\tilde{\Lambda}^*}}{1-\gamma'}$$

■

B.2. Some Remarks on theorem 9

- Note that as the approximation becomes perfect, that is $\delta_1 = 1, \delta_2 = 1$ and $\epsilon = 1$, the risk sensitive average cost $e^{\tilde{\Lambda}_{f_{k+1}}}$ obtained from the algorithm converges exactly to the optimal risk sensitive average cost. In other words σ will be 0.
- Note the condition on γ' for it to constitute a valid source of contraction is a consequence of the multiplicative Bellman equation associated with the risk sensitive average cost. In the case of risk neutral average cost, such a condition disappears and the approximation results hold more generally for all magnitudes of δ_1, δ_2 and ϵ. For more details refer Murthy et al. (2023).
- Since the condition on γ' requires more accurate policy evaluations and policy improvements as n increases, a possible way to circumvent this issue is by ensuring that $m_i \geq R$ for all iterations of the algorithm. This would ensure n to be equal to one. In other words, better the policy evaluation, lesser the approximation loss.

Lemma 10 *Let $\max_k m_k < C_1$, where m_k corresponds to the number of fixed point iterations performed during partial policy evaluation during the kth execution of the algorithm 2. Then, there exists τ' such that*

$$e^{h_k(i)} \geq \tau' > 0.$$

Proof Recall from the algorithm,

$$e^{h'_{f_{k+1}}(i)} = \frac{e^{h_{f_{k+1}}(i)}}{\sum_{i\in\mathcal{S}} e^{h_{f_{k+1}}(i)}}$$

$$\implies \sum_{i\in\mathcal{S}} e^{h'_{f_{k+1}}(i)} = 1$$

Then,

$$\begin{aligned} e^{h_{f_{k+1}}(i)} &= \mathsf{T}^{m_k}_{f_{k+1}} e^{h_k}(i) \\ &\geq \frac{\mathsf{T}^{m_k}_{f_{k+1}} e^{h'_{f_k}}(i)}{\delta_2} \end{aligned}$$

Let $\underline{d} = \min_{f\in\Pi} d_f$, then

$$\begin{aligned} \left(\mathsf{T}_{f_{k+1}} e^{h'_{f_k}}(i)\right) &= e^{\alpha d_{f_{k+1}}(i)} \sum_{j\in\mathcal{S}} \mathbb{Q}(j|i, f_{k+1}(i)) e^{h'_{f_k}(j)} \\ &\geq e^{\alpha\underline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h'_{f_k}}\right)(i) \end{aligned}$$

Repeating the above we obtain,

$$\left(\mathsf{T}^{m_k}_{f_{k+1}} e^{h'_{f_k}}\right)(i) \geq \frac{e^{m_k\alpha\underline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h'_{f_k}}\right)(i)}{\delta_2}$$

Substituting for h'_{f_k},

$$\left(\mathsf{T}^{m_k}_{f_{k+1}} e^{h'_{f_k}}\right)(i) \geq \frac{e^{m_k \alpha \underline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h_{f_k}}\right)(i)}{\delta_2 \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)}}$$

$$e^{h_{f_{k+1}}(i)} \geq \frac{e^{m_k \alpha \underline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h_{f_k}}\right)(i)}{\delta_2 \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)}}$$

Iterating the above equation we obtain,

$$e^{h_{f_{k+1}}(i)} \geq \frac{e^{\alpha \underline{d} \sum_{p=0}^{k} m_{k-p}} \left(\mathbb{Q}^{m_k}_{f_{k+1}} \mathbb{Q}^{m_{k-1}}_{f_k} \dots \mathbb{Q}^{m_0}_{f_1} e^{h_0}\right)(i)}{(\delta_2)^{k+1} \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)} \sum_{j \in \mathcal{S}} e^{h_{f_{k-1}}(j)} \dots \sum_{j \in \mathcal{S}} e^{h_{f_1}(j)}}$$

From Algorithm 2, for a sufficiently large k, we have $\sum_{l=0}^{k-1} m_l > R$.
Defining $H_k = \mathbb{Q}^{m_k}_{f_{k+1}} \mathbb{Q}^{m_{k-1}}_{f_k} \dots \mathbb{Q}^{m_0}_{f_1}$, Lemma 3 yields $\lambda = \min_{i,j} H_k(i \mid j) > 0$, we continue the above sequence of inequalities:

$$e^{h_{f_{k+1}}(i)} \geq \frac{e^{\alpha \underline{d} \sum_{p=0}^{k} m_{k-p}} \lambda \sum_{j \in \mathcal{S}} e^{h_0(j)}}{(\delta_2)^{k+1} \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)} \sum_{j \in \mathcal{S}} e^{h_{f_{k-1}}(j)} \dots \sum_{j \in \mathcal{S}} e^{h_{f_1}(j)}}$$

Since we know that $\sum_{j \in \mathcal{S}} e^{h_0(j)} = 1$, we obtain the following relation,

$$e^{h_{f_{k+1}}(i)} \geq \frac{e^{\alpha \underline{d} \sum_{p=0}^{k} m_{k-p}} \lambda}{(\delta_2)^{k+1} \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)} \sum_{j \in \mathcal{S}} e^{h_{f_{k-1}}(j)} \dots \sum_{j \in \mathcal{S}} e^{h_{f_1}(j)}} \tag{21}$$

Similarly,

$$e^{h_{f_{k+1}}(i)} \leq \frac{\mathsf{T}^{m_k}_{f_{k+1}} e^{h'_{f_k}}(i)}{\delta_1}$$

Let $\overline{d} = \max_{f \in \Pi} d_f$, then

$$\begin{aligned}\left(\mathsf{T}_{f_{k+1}} e^{h'_{f_k}}(i)\right) &= e^{\alpha d_{f_{k+1}}(i)} \sum_{j \in \mathcal{S}} \mathbb{Q}(j|i, f_{k+1}(i)) e^{h'_{f_k}(j)} \\ &\leq e^{\alpha \overline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h'_{f_k}}\right)(i)\end{aligned}$$

Repeating the above we obtain,

$$\left(\mathsf{T}^{m_k}_{f_{k+1}} e^{h'_{f_k}}\right)(i) \leq \frac{e^{m_k \alpha \overline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h'_{f_k}}\right)(i)}{\delta_1}$$

Substituting for h'_{f_k},

$$\left(\mathsf{T}^{m_k}_{f_{k+1}} e^{h'_{f_k}}\right)(i) \le \frac{e^{m_k \alpha \overline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h_{f_k}}\right)(i)}{\delta_1 \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)}}$$

$$e^{h_{f_{k+1}}(i)} \le \frac{e^{m_k \alpha \overline{d}} \left(\mathbb{Q}_{f_{k+1}} e^{h_{f_k}}\right)(i)}{\delta_1 \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)}}$$

Iterating the above equation we obtain,

$$e^{h_{f_{k+1}}(i)} \le \frac{e^{\alpha \overline{d} \sum_{p=0}^{k} m_{k-p}} \left(\mathbb{Q}^{m_k}_{f_{k+1}} \mathbb{Q}^{m_{k-1}}_{f_k} \dots \mathbb{Q}^{m_0}_{f_1} e^{h_0}\right)(i)}{(\delta_1)^{k+1} \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)} \sum_{j \in \mathcal{S}} e^{h_{f_{k-1}}(j)} \dots \sum_{j \in \mathcal{S}} e^{h_{f_1}(j)}}$$

$$\sum_{i \in \mathcal{S}} e^{h_{f_{k+1}}(i)} \le \frac{|\mathcal{S}| e^{\alpha \overline{d} \sum_{p=0}^{k} m_{k-p}}}{(\delta_1)^{k+1} \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)} \sum_{j \in \mathcal{S}} e^{h_{f_{k-1}}(j)} \dots \sum_{j \in \mathcal{S}} e^{h_{f_1}(j)}}$$

We thus obtain,

$$\frac{1}{\sum_{i \in \mathcal{S}} e^{h_{f_{k+1}}(i)}} \ge \frac{(\delta_1)^{k+1} \sum_{j \in \mathcal{S}} e^{h_{f_k}(j)} \sum_{j \in \mathcal{S}} e^{h_{f_{k-1}}(j)} \dots \sum_{j \in \mathcal{S}} e^{h_{f_1}(j)}}{|\mathcal{S}| e^{\alpha \overline{d} \sum_{p=0}^{k} m_{k-p}}} \tag{22}$$

Thus from appendix B.2 and appendix B.2, we obtain the following,

$$\frac{e^{h_{f_{k+1}}(i)}}{\sum_{i \in \mathcal{S}} e^{h_{f_{k+1}}(i)}} \ge \left(\frac{\delta_1}{\delta_2}\right)^{k+1} \frac{e^{\alpha\left(\underline{d} - \overline{d}\right) \sum_{p=0}^{k} m_{k-i}} \lambda}{|\mathcal{S}|}$$

Let $\tau = \frac{e^{\alpha\left(\underline{d} - \overline{d}\right) \sum_{p=0}^{k} m_{k-i}} \lambda}{|\mathcal{S}|}$, then we have,

$$e^{h'_{f_{k+1}}(i)} \ge \tau$$

From the algorithm 2, we then obtain the following,

$$e^{h_{k+1}(i)} \ge \frac{e^{h'_{f_{k+1}}}(i)}{\delta_2} \ge \frac{\tau}{\delta_2}$$

Defining $\frac{\tau}{\delta_2} = \tau'$, we obtain the statement of the lemma,

$$e^{h_{k+1}(i)} \ge \tau' > 0$$

■