Planning and Learning in Average Risk-aware MDPs

Weikai Wang[1,2] **Erick Delage**[1,2]
[1] GERAD & HEC Montréal [2]Mila - Québec AI Institute
{weikai.wang, erick.delage}@hec.ca

Abstract

For continuing tasks, average cost Markov decision processes have well-documented value and can be solved using efficient algorithms. However, it explicitly assumes that the agent is risk-neutral. In this work, we extend risk-neutral algorithms to accommodate the more general class of dynamic risk measures. Specifically, we propose a relative value iteration (RVI) algorithm for planning and design two model-free Q-learning algorithms, namely a generic algorithm based on the multi-level Monte Carlo (MLMC) method, and an off-policy algorithm dedicated to utility-based shortfall risk measures. Both the RVI and MLMC-based Q-learning algorithms are proven to converge to optimality. Numerical experiments validate our analysis, confirm empirically the convergence of the off-policy algorithm, and demonstrate that our approach enables the identification of policies that are finely tuned to the intricate risk-awareness of the agent that they serve.

1 Introduction

For continuing tasks where there is a need to optimize a long term periodic payoff, such as network control, supply chain designs, or maintenance problems (Puterman, 1994), average cost (or reward) Markov decision processes (MDPs) serve as a crucial model in reinforcement learning (Sutton and Barto, 2018; Naik et al., 2019) and can be solved using efficient algorithms. In the risk-neutral setting, different forms of value iteration algorithms co-exist (Puterman, 1994; Bertsekas, 2007) and some have been extended to a model-free setting using Q-learning style algorithms (Abounadi et al., 2001; Wan et al., 2021). The question of how to formulate and solve average-cost MDPs however becomes challenging when the agent is considered risk sensitive. It originates from the pioneering work of Howard and Matheson (1972) and is covered in recent surveys such as Biswas and Borkar (2023) and Bäuerle and Jaskiewicz (2024).

This work focuses on average risk-aware MDPs, a general formulation introduced in Shen et al. (2013) that attempts to find a policy π that minimizes the long-term average of the risk of the cost process generated by π. While the theoretical foundations of this framework are well studied, finding efficient solution techniques to these problems remains a challenging task. To solve the average risk-aware MDP problem, one can apply the classic value iteration algorithm (Cavazos-Cadena and Montes-de Oca, 2003; Ruszczyński, 2010; Shen et al., 2015). This approach relies on the iteration of the risk-aware Bellman operator and computes the average to obtain the optimal average risk. However, it is known to suffer from overflow issues when the number of iterations is large. In the risk-neutral setting, the relative value iteration (RVI) algorithm is widely used (Bertsekas, 2007; Gupta et al., 2015), as it mitigates overflow issues during long iterations by subtracting a reference value for each state at every step. However, while some studies have explored RVI algorithms based on the entropic risk measure (Borkar, 2010; Arapostathis and Borkar, 2019; Hmedi et al., 2023), a general formulation of the RVI algorithm for risk-aware MDPs remains missing in the literature.

Meanwhile, in practical applications, the environment is rarely known in full, highlighting the importance of developing model-free learning algorithms. To the best of our knowledge, extensions of the risk-neutral Q-learning algorithms to the average risk-aware setting appear to only exist for

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

the case of entropic risk measure (see Borkar (2002), Borkar (2010), Moharrami et al. (2024) and the reference therein). This is in sharp contrast to the extensive literature on algorithms for discounted or finite-horizon risk-aware MDPs, where many studies exist: for instance, see Chow and Ghavamzadeh (2014), Tamar et al. (2015), or Chow et al. (2018) for conditional value-at-risk, see Huang and Haskell (2017), Köse and Ruszczyński (2021), or Lam et al. (2023) for general coherent risk measures, Shen et al. (2014) or Marzban et al. (2023) for utility-based shortfall risk (UBSR), and see Hau et al. (2025) for quantiles. To conclude, the design of a model-free learning algorithm for average risk-aware MDPs with a general risk measure remains an open research field.

The literature most closely related to addressing this gap has focused on planning and learning algorithms for distributionally robust MDPs, leveraging the fact that coherent risk measures admit a worst-case expected value representation. Studies have explored the discounted case (Liu et al., 2022; Wang et al., 2023a, 2024) as well as the average case (Wang et al., 2023b,c), where an ambiguity set is constructed around the transition kernel to safeguard against potential distributional shifts. These results, however, do not apply to general classes of possibly non-coherent dynamic risk measures.

This paper presents planning and learning algorithms for average risk-aware MDPs with a general dynamic risk measure. We describe our contributions as follows:

1. **Planning:** We propose a model-based RVI algorithm for average risk-aware MDPs, which produces a policy that provably converges to the optimal policy for a general class of dynamic risk measures. While existing studies on model-based algorithms for this problem focus either on the risk-neutral setting or the case of entropic risk, our work appears to be the first to consider such a general class of dynamic risk measures.
2. **Learning:** We introduce two novel model-free Q-learning algorithms for average risk-aware MDPs. The first one generalizes the multi-level Monte Carlo (MLMC) based Q-learning algorithm introduced in Wang et al. (2023c) for robust average MDPs to a broader class of dynamic risk measures, which may not necessarily be coherent, while ensuring provable convergence to optimality. The conditions that we impose for convergence are weak and are satisfied by many popular risk measures such as UBSR, optimized certainty equivalent, and spectral risk measures. Additionally, we propose an asynchronous algorithm that is specialized for UBSR and amenable to off-policy learning by waiving the need for a resampling procedure. While the theoretical convergence remains open, we validate it empirically under different loss functions.
3. **Empirics:** We confirm empirically the convergence of all algorithms under different choice of risk measures and practically relevant sampling rates for MLMC Q-learning, and compare the sample efficiency. We also showcase how average risk-aware MDPs identify policies that are tuned to the agents' preferences in popular environments from the literature.

The structure of the paper is as follows. Section 2 introduces average risk-aware MDPs. Section 3 introduces the model-based algorithms, including the risk-aware RVI and its generalization to Q-factors. Section 4 describes the model-free algorithms, featuring a general Q-learning algorithm with MLMC and an asynchronous Q-learning algorithm specifically designed for UBSR. Section 5 presents the numerical experiments. Section 6 concludes the paper and proposes further research. Pseudo-codes, proofs, and additional experiment details and results are provided in the appendix.

2 Preliminaries

Notations: Given any finite probability space $(\Omega, \sigma(\Omega), P(\cdot))$, abbreviated as $(\Omega, P(\cdot))$, with Ω a finite set of outcomes, $\sigma(\Omega)$ the power set of Ω (sigma-algebra) and $P(\cdot)$ a probability mass function in the probability simplex $\mathscr{P}(\Omega)$, we denote by $\mathcal{L}(\Omega)$ the set of finite real-valued functions (a.k.a. random variables) on Ω and $|\Omega|$ the cardinality of Ω. For $v, w \in \mathcal{L}(\Omega)$, the notation $v \geq w$ refers to $v(\omega) \geq w(\omega)$ for all $\omega \in \Omega$, and $v \geq w$ almost surely (a.s.) refers to $v(\omega) \geq w(\omega)$ for all $\omega \in \Omega$ such that $P(\omega) > 0$. The infinity norm of $v \in \mathcal{L}(\Omega)$ is $\|v\|_\infty := \sup_{\omega\in\Omega} |v(\omega)|$, while its span-seminorm is: $\|v\|_{sp} := \max_{\omega\in\Omega} v(\omega) - \min_{\omega\in\Omega} v(\omega)$. For $A \subseteq \Omega$, the indicator function $\mathbf{1}\{\omega \in A\}$ equals 1 if $\omega \in A$ and 0 otherwise. Finally, e and $\mathbf{0}$ represent the constant functions of one and zero respectively, while e denotes the base of the natural logarithm.

2.1 Risk Maps

We begin by defining the notion of a risk measure following Shapiro et al. (2021).

Definition 2.1. *Given a finite probability space $(\Omega, P(\cdot))$, a risk measure $\rho : \mathcal{L}(\Omega) \to \mathbb{R}$ that maps a random cost to a real value capturing its risk is said to be monetary if it satisfies the following properties:*

(1) (Monotonicity) $\rho(v) \leq \rho(w)$ for all $v, w \in \mathcal{L}(\Omega)$ such that $v \leq w$ a.s.;
(2) (Translation invariance) $\rho(v + \lambda) = \rho(v) + \lambda$ for any $\lambda \in \mathbb{R}$, $v \in \mathcal{L}(\Omega)$;
(3) (Normalization) $\rho(0) = 0$;

further called convex if:

(4) (Convexity) For all $\alpha \in [0, 1]$, $v, w \in \mathcal{L}(\Omega)$, $\rho(\alpha v + (1 - \alpha)w) \leq \alpha\rho(v) + (1 - \alpha)\rho(w)$;

and coherent if

(5) (Positive homogeneity) For all $\lambda \geq 0$, $v \in \mathcal{L}(\Omega)$, $\rho(\lambda v) = \lambda\rho(v)$.

In the following, we introduce some popular kinds of risk measures that will be of interest.

Definition 2.2 (Definition 4.112, Föllmer and Schied (2016))**.** *A risk measure on $(\Omega, P(\cdot))$ is called a utility-based shortfall risk (UBSR) measure if it can be represented as:*

$$\mathrm{SR}(v) := \inf \{m \in \mathbb{R} : \mathbb{E}[\ell(v - m)] \leq 0\}, \quad \forall v \in \mathcal{L}(\Omega),$$

for some continuous non-decreasing convex loss function $\ell : \mathbb{R} \to \mathbb{R}$ such that $\ell(0) = 0$.[1]

Example 2.3 (Expected value)**.** When $\ell(x) = x$, the UBSR measure reduces to the expected value, which we refer as the risk-neutral measure.

Example 2.4 (Entropic risk measure)**.** When $\ell(x) = \mathrm{e}^{\beta x} - 1$, with $\beta > 0$ representing risk sensitivity, the resulting UBSR measure is the entropic risk measure $\mathrm{SR}(v) = \frac{1}{\beta}\log(\mathbb{E}[\mathrm{e}^{\beta v}])$.

Example 2.5 (Expectile)**.** Following Bellini and Bignozzi (2015), the expectile is the only coherent UBSR, defined using the loss function $\ell(x) = \tau x^+ - (1 - \tau)x^-$, where $\tau \in [0, 1]$ represents the degree of risk aversion. This measure spans from the essential infimum of the random cost at $\tau = 0$ to its essential supremum at $\tau = 1$, passing through the expected value at $\tau = 0.5$.

Definition 2.6 (Definition 2.1, Ben-Tal and Teboulle (2008))**.** *A risk measure on $(\Omega, P(\cdot))$ is called an optimized certainty equivalent (OCE) risk measure if it can be represented as*

$$\mathrm{OCE}(v) := \inf_{\xi \in \mathbb{R}} \{\xi + \mathbb{E}[\ell(v - \xi)]\}, \quad \forall v \in \mathcal{L}(\Omega),$$

for some nondecreasing convex loss function $\ell : \mathbb{R} \to \mathbb{R}$ such that $\ell(0) = 0$ and $1 \in \partial\ell(0)$, where $\partial\ell(0)$ is the subgradient of ℓ at 0.

Definition 2.7 (Definition 3.1, Acerbi (2002))**.** *A risk measure on $(\Omega, P(\cdot))$ is called a spectral risk measure associated to a risk spectrum function $\phi : [0, 1] \to [0, \infty)$ such that $\int_0^1 \phi(\beta)d\beta = 1$, if it can be represented as*

$$M^\phi(v) := \int_0^1 \phi(\beta)F_v^{-1}(\beta)d\beta, \quad \forall v \in \mathcal{L}(\Omega),$$

where F_v is the cumulative distribution function of v and $F_v^{-1}(\beta) := \inf\{m \in \mathbb{R} : F_v(m) \geq \beta\}$.

Example 2.8 (Conditional Value-at-Risk)**.** When $\ell(x) = (1-\alpha)^{-1}x^+$ for $\alpha \in (0, 1)$, the OCE risk is the conditional Value-at-Risk (CVaR) at level α. CVaR is coherent and it is also a spectral risk measure with risk spectrum $\phi(\beta) = (1 - \alpha)^{-1}\mathbf{1}\{\beta \geq \alpha\}$.

Example 2.9 (Mean-CVaR)**.** When $\phi(\beta) = \eta + (1 - \eta)(1 - \alpha)^{-1}\mathbf{1}\{\beta \geq \alpha\}$ for some $\eta \in (0, 1)$, the spectral risk measure defined as $M^\phi(v) = \eta\mathbb{E}[v] + (1 - \eta)\mathrm{CVaR}_\alpha(v)$ is referred to as the mean-CVaR risk measure.

[1] Shen et al. (2014) employs equivalently $\mathbb{E}[\bar{\ell}(v - m)] \leq m_0$ using the replacement $\ell(z) := \bar{\ell}(z) - m_0$. We also focus on normalized UBSR.

2.2 Average Risk-aware MDPs

We consider a finite MDP defined through the tuple $(\mathcal{X}, \mathcal{A}, P, c, x_0)$, where $\mathcal{X}$ and $\mathcal{A}$ are finite state and action spaces, denoting $\mathcal{K} := \mathcal{X} \times \mathcal{A}$ for short. The transition kernel $P : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$ specifies the probability $P(y|x,a)$ of transitioning from state x to state y given action a. The bounded cost function is defined as $c : \mathcal{X} \times \mathcal{A} \to [-\bar{C}, \bar{C}]$. For time $t = 0, 1, \cdots$, the state and action are x_t and a_t, governed by a Markov policy $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$, with each $\pi_t \in \Pi := \{\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})\}$, where $\pi_t(\cdot|x_t)$ denotes the probability of choosing a_t given x_t. A policy is called deterministic if it assigns a probability of one to a specific action for each state, and is called stationary if $\pi_t \equiv \pi$ for all t for some $\pi \in \Pi$.

In a risk-neutral setting, the infinite horizon average cost MDP problem takes the form:

$$\text{(ACMDP)} \quad \bar{J}^* := \inf_{\boldsymbol{\pi} \in \Pi^\infty} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T} c^\pi(X_t)\right],$$

where X_t is the state at stage t and $c^\pi(x) := \sum_{a \in \mathcal{A}} \pi(a|x) c(x,a)$. One seeks to identify a stationary policy that minimizes the long term average expected total cost generated by the MDP, when starting from some initial state X_0 and following policy $\boldsymbol{\pi}$.

Following Shen et al. (2013), we consider the risk-aware version of the average cost MDP by replacing $\mathbb{E}[\cdot]$ with a class of dynamic risk measures that is specially designed for MDPs.

Definition 2.10. *A risk map $\mathcal{R}$ is a function that maps each state $(x,a) \in \mathcal{K}$ to a monetary risk measure on the space $(\mathcal{X}, P(\cdot|x,a))$. Furthermore, for any $\pi \in \Pi$ we define $\mathcal{R}^\pi(v|x) := \sum_{a \in \mathcal{A}} \pi(a|x) \mathcal{R}(v|x,a)$. To simplify notation, we sometimes write $\mathcal{R}_{x,a}(v) := \mathcal{R}(v|x,a)$ and $\mathcal{R}^\pi_x(v) := \mathcal{R}^\pi(v|x)$.*

We first consider a risk-aware T-stage total cost problem and define our risk-aware objective as follows: $J_T(\boldsymbol{\pi}) := c^{\pi_0}(X_0) + \mathcal{R}^{\pi_0}_{X_0}(c^{\pi_1}(X_1) + \cdots + \mathcal{R}^{\pi_{T-1}}_{X_{T-1}}(c^{\pi_T}(X_T)) \cdots)$. The infinite horizon average risk-aware MDP problem therefore seeks to find a policy $\boldsymbol{\pi}$ that minimizes :

$$\text{(ARMDP)} \quad J^* := \inf_{\boldsymbol{\pi} \in \Pi^\infty} J_\infty(\boldsymbol{\pi}),$$

where $J_\infty(\boldsymbol{\pi}) := \limsup_{T \to \infty} \frac{1}{T} J_T(\boldsymbol{\pi})$. It is easy to see that ARMDP reduces to ACMDP when $\mathcal{R}_{x,a}(v) = \mathbb{E}_{x,a}[v] := \mathbb{E}[v(y)]$ with $y \sim P(\cdot|x,a)$.

Remark 2.11. As argued in Shen et al. (2013), preserving the Markov property is essential to guarantee stationary optimal policies for infinite horizon objectives (Ruszczyński and Shapiro, 2006; Shen et al., 2013). Therefore, we restrict our attention to Markovian risk measures that depend only on the current state. Readers can refer to Ruszczyński (2010) for a broader framework.

2.3 Average Risk Optimality Equation

Shen et al. (2013) establishes several assumptions on the risk maps of an MDP to guarantee the existence and uniqueness of the optimal average risk for ARMDP. Here, we modify and adapt these assumptions to suit our setting of a finite MDP.

Assumption 2.12 (Doeblin type condition, Assumption 5.4, Shen et al. (2013))**.** *There exists a coherent risk measure $\nu : \mathcal{L}(\mathcal{X}) \to \mathbb{R}$, and some constant $\bar{\alpha} \in (0,1)$ such that for all $v \geq v' \in \mathcal{L}(\mathcal{X})$, we have $\min_{(x,a) \in \mathcal{K}} \{\mathcal{R}(v|x,a) - \bar{\alpha}\nu(v) - \mathcal{R}(v'|x,a) + \bar{\alpha}\nu(v')\} \geq 0$.*

Assumption 2.12 defines a form of ergodicity property of each state under the risk map. In Shen et al. (2013), ν is not necessarily required to be coherent, whereas we impose this condition here for simplicity in the subsequent derivations.

From Shen et al. (2013), if the risk maps satisfy Assumption 2.12, then there exists an optimal stationary deterministic Markov policy $\boldsymbol{\pi}^*$ such that $J^* = J_\infty(\boldsymbol{\pi}^*)$. We restate the result as follows.

Theorem 2.13 (Theorem 5.9, 5.10, Shen et al. (2013))**.** *Under Assumption 2.12, there exists a unique $g^* \in \mathbb{R}$ and an $h^* \in \mathcal{L}(\mathcal{X})$ satisfying the average risk optimality equation (AROE):*

$$g + h(x) = \min_{a \in \mathcal{A}} \{c(x,a) + \mathcal{R}(h|x,a)\}. \tag{2.1}$$

Moreover, $g^* = J^* = J_\infty(\boldsymbol{\pi}^*)$, for the stationary deterministic policy $\pi^*_t(a|x) = \mathbf{1}\{a = a^*(x)\}$, where $a^*(x)$ minimizes $c(x,a) + \mathcal{R}(h^*|x,a)$, and g^* is independent of x_0.

Remark 2.14. In Assumption 5.4 of Shen et al. (2013), an additional Lyapunov-type condition is introduced, which imposes a growth constraint using a nonnegative weight function W. This condition can be dropped in a finite MDP, see Appendix B.1.

Remark 2.15. Assumption 2.12 is a sufficient condition for the existence of an optimal average risk independent of the initial state and is stronger than the unichain assumption commonly used in risk-neutral average MDPs. It is well-known that for risk-aware MDPs, the unichain assumption alone does not guarantee this independence. For specific risk measures, such as the entropic risk measure, this condition can be relaxed to require only that the Markov chain is irreducible and aperiodic under all stationary policies (see Cavazos-Cadena and Fernández-Gaucherand (1999)). For the three types of risk measures considered in this work, Assumption 2.12 is restricted to an ergodicity condition on the underlying Markov chain, together with a condition on the loss function for UBSR and OCE, and on the risk spectrum for the spectral risk measure (see Theorem 4.10).

3 Model-based Algorithms

In this section, we propose a risk-aware version of the RVI algorithm to solve the ARMDP problem. Additionally, to lay the foundation for the Q-learning algorithm in the next section, we introduce a risk-aware relative Q-factor iteration algorithm as a generalization of the risk-aware RVI algorithm.

3.1 Risk-aware Relative Value Iteration

Following Abounadi et al. (2001), the risk-neutral RVI algorithm is defined as

$$V_{n+1}(x) := \min_{a\in\mathcal{A}} \mathbb{E}\left[c(x,a) + V_n\right] - f(V_n), \quad \forall x \in \mathcal{X}, \tag{3.1}$$

where $V_n \in \mathcal{L}(\mathcal{X})$ and V_0 is arbitrarily initialized, and $f : \mathcal{L}(\mathcal{X}) \to \mathbb{R}$ is a function that satisfies conditions discussed below, e.g. $f(v) := v(x_0)$.[2] Using a general function $f(V_n)$ instead of $V(x_0)$ allows the RVI algorithm to eliminate the need for a reference state, making it more flexible and efficient for computation (also see the discussion in Wan et al. (2021) for the risk-neutral case). It is known that under the unichain assumption, the risk-neutral RVI algorithm converges to a unique V^*, which solves the risk-neutral version of the AROE using $h := V^*$ and $g^* = f(V^*)$.

We propose extending the RVI algorithm to the risk-aware setting by replacing the expected value operator with an appropriate risk map. This gives rise to the following risk-aware RVI algorithm:

$$V_{n+1}(x) = \mathcal{G}(V_n)(x) - f(V_n), \quad \forall x \in \mathcal{X}, \tag{3.2}$$

where $\mathcal{G} : \mathcal{L}(\mathcal{X}) \to \mathcal{L}(\mathcal{X})$ is the risk-aware Bellman optimality operator, defined as $\mathcal{G}(v)(x) := \min_{a\in\mathcal{A}} \mathcal{R}_{x,a}(c(x,a) + v)$ for all $x \in \mathcal{X}$ and $v \in \mathcal{L}(\mathcal{X})$.

To guarantee convergence of algorithm (3.2), we impose the following conditions on f.

Assumption 3.1. *The function $f : \mathcal{L}(\mathcal{X}) \to \mathbb{R}$ satisfies:*

(1) For any $\lambda \in \mathbb{R}$ and $v \in \mathcal{L}(\mathcal{X})$, $f(\mathbf{0}) = 0$, $f(v + \lambda) = f(v) + \lambda$.
(2) f is Lipschitz, i.e., $\exists \tilde{L} \geq 0$, such that $\|f(v) - f(w)\|_\infty \leq \tilde{L}\|v - w\|_\infty$, $\forall v, w \in \mathcal{L}(\mathcal{X})$.

Assumption 3.1 is equivalent to imposing that f is translation invariant and is naturally satisfied by $f(v) := v(x_0)$. Such extension for the RVI seems to be first proposed in Abounadi et al. (2001) for the risk-neutral case, although the authors did not include proofs of their validity.

The following theorem confirms that convergence of RVI remains valid in the risk-aware setting.

Theorem 3.2. *Under assumptions 2.12 and 3.1, the risk-aware RVI algorithm (3.2) converges to a unique fixed point V^*, which identifies, using $h^*(x) := V^*(x)$ and $g^* := f(V^*)$, a solution to the AROE (2.1).*

Clearly, the risk-aware RVI algorithm (3.2) reduces to RVI algorithm (3.1) in the risk-neutral setting.

[2]The RVI algorithm in Bertsekas (2007), for example, replaces $f(V_n)$ with $f(V_{n+1}) := V_{n+1}(x_0)$. We adopt the formulation from Abounadi et al. (2001) as it is better suited for designing a Q-learning algorithm.

3.2 Risk-aware Relative Q-factor Iteration

The risk-aware RVI algorithm (3.2) suggests that, when letting $Q_{n+1}(x,a) := \mathcal{R}_{x,a}(c(x,a)+V_n) - f(V_n)$, (3.2) can be reformulated as the following risk-aware relative Q-factor iteration (RQI):

$$Q_{n+1}(x,a) := \mathcal{R}_{x,a}\left(c(x,a) + \min_{a'\in\mathcal{A}} Q_n(\cdot,a')\right) - f\left(\min_{a'\in\mathcal{A}} Q_n(\cdot,a')\right), \quad \forall (x,a)\in\mathcal{K},$$

where $Q_n \in \mathcal{L}(\mathcal{K})$ and Q_0 is arbitrarily initialized. As suggested in Abounadi et al. (2001) for the risk-neutral case, this can be more generally defined as:

$$Q_{n+1}(x,a) = \mathcal{H}(Q_n)(x,a) - f(Q_n), \quad \forall (x,a)\in\mathcal{K}, \tag{3.3}$$

where $\mathcal{H} : \mathcal{L}(\mathcal{K}) \to \mathcal{L}(\mathcal{K})$ is the risk-aware Bellman optimality operator for Q-factors, defined as $\mathcal{H}(q)(x,a) := \mathcal{R}_{x,a}(c(x,a) + \min_{a'\in\mathcal{A}} q(\cdot,a'))$, for all $(x,a)\in\mathcal{K}$ and $q\in\mathcal{L}(\mathcal{K})$. With a slight abuse of notation, here we define $f : \mathcal{L}(\mathcal{K}) \to \mathbb{R}$ and impose the following assumptions.

Assumption 3.3. *The function $f : \mathcal{L}(\mathcal{K}) \to \mathbb{R}$ satisfies:*

(1) For any $\lambda\in\mathbb{R}$ and $q\in\mathcal{L}(\mathcal{K})$, $f(\mathbf{0}) = 0$, $f(q+\lambda) = f(q) + \lambda$.
(2) f is Lipschitz, i.e., $\exists \tilde{L} \geq 0$ such that $\|f(p) - f(q)\|_\infty \leq \tilde{L}\|p-q\|_\infty$, $\forall p,q\in\mathcal{L}(\mathcal{K})$.

Common choices for f can be $f(q) = q(x_0,a_0)$, $f(q) = \min_a q(x_0,a)$, $f(q) = \frac{1}{|\mathcal{X}||\mathcal{A}|}\sum_{x,a} q(x,a)$. Similar to Theorem 3.2, we have the following convergence and optimality result for the risk-aware RQI algorithm.

Theorem 3.4. *Under assumptions 2.12 and 3.3, the risk-aware RQI algorithm (3.3) converges to a unique fixed point Q^*, which identifies, using $h^*(x) := \min_{a\in\mathcal{A}} Q^*(x,a)$ and $g^* := f(Q^*)$, a solution to the AROE (2.1).*

Theorem 3.4 also suggests that a solution to the AROE (2.1) can be identified by solving the following average risk optimality equation based on the Q-factor:

$$q(x,a) = \mathcal{R}_{x,a}\left(c(x,a) + \min_{a'\in\mathcal{A}} q(\cdot,a')\right) - f(q), \quad \forall (x,a)\in\mathcal{K}, q\in\mathcal{L}(\mathcal{K}). \tag{3.4}$$

4 Model-free Algorithms

In this section, we propose model-free Q-learning algorithms for solving the ARMDP. We begin by showing that almost sure convergence can be achieved when an unbiased estimator of the risk-aware Bellman optimality operator is available. In particular, we describe how such an estimator can be constructed using the multi-level Monte Carlo method. In addition, we introduce a Q-learning algorithm tailored to UBSR that avoids the need for an estimator.

4.1 Risk-aware RVI Q-learning

Motivated by the risk-aware RQI algorithm (3.3), we can propose the following model-free risk-aware RVI Q-learning algorithm:

$$Q_{n+1}(x,a) = Q_n(x,a) + \gamma(n)\left(\hat{\mathcal{H}}(Q_n)(x,a) - f(Q_n) - Q_n(x,a)\right), \quad (x,a)\in\mathcal{K}, \tag{4.1}$$

where $\hat{\mathcal{H}}$ is an estimator for the risk-aware Bellman optimality operator $\mathcal{H}$ and $\gamma(n)$ is some step size. We construct $\hat{\mathcal{H}}(q)$ as an estimator of $\mathcal{H}(q)$ satisfying the following assumption.

Assumption 4.1. *The estimator $\hat{\mathcal{H}}$ is unbiased and has controllable variance: $\mathbb{E}[\hat{\mathcal{H}}(q)] = \mathcal{H}(q)$ and there exists a $C > 0$ such that $\mathrm{Var}[\hat{\mathcal{H}}(q)(x,a)] \leq C(1 + \|q\|_\infty^2)$, $\forall (x,a)\in\mathcal{K}, q\in\mathcal{L}(\mathcal{K})$.*

To guarantee the convergence for the risk-aware RVI Q-learning algorithm (4.1), we require that the risk map satisfies an assumption called "asymptotic coherence" and the function f is homogeneous. We also impose the Robbins-Monro condition on the step size $\gamma(n)$.

Assumption 4.2 (Asymptotic coherence)**.** *The risk map $\mathcal{R}$ is asymptotically coherent, i.e., there exists a risk map $\mathcal{R}^\infty$ such that for all $(x,a)\in\mathcal{K}$, we have that $\lim_{s\to\infty}\frac{1}{s}\mathcal{R}_{x,a}(sv) = \mathcal{R}^\infty_{x,a}(v)$ for all $v\in\mathcal{L}(\mathcal{X})$ and uniformly on all compact subsets of $\mathcal{L}(\mathcal{X})$.*

Assumption 4.3. *The function f is homogeneous, i.e., $f(\lambda v) = \lambda f(v)$, $\forall \lambda \in \mathbb{R}$, $v \in \mathcal{L}(\mathcal{K})$.*

Assumption 4.4. *The step size $\{\gamma(n)\}_{n=0}^{\infty}$ satisfies $\sum_{n=0}^{\infty} \gamma(n) = \infty$ and $\sum_{n=0}^{\infty} \gamma(n)^2 < \infty$.*

We then have the following convergence result.

Theorem 4.5. *Under assumptions 2.12, 3.3, 4.1, 4.2, 4.3, 4.4, then almost surely, Q_n converges to some Q^*, and $h^*(x) := \min_{a \in \mathcal{A}} Q^*(x, a)$, $g^* := f(Q^*)$ identify a solution to the AROE (2.1). The greedy policy, $\pi_n(a|x) := \mathbf{1}\{a = a_n^*(x)\}$ with $a_n^* \in \arg\min_{a \in \mathcal{A}} Q_n(x, a)$, also converges almost surely to an optimal stationary deterministic policy of ARMDP.*

If the risk map $\mathcal{R}$ is coherent, Assumption 4.2 is automatically satisfied. This assumption is made for technical reasons, as the convergence proof of the Q-learning algorithm for the average MDP relies on ODE-based stochastic approximation (see Abounadi et al. (2001) and Wan et al. (2024) for a recent review). This approach requires the limit in Assumption 4.2 to exist for analyzing the almost sure boundedness of the iteration sequence. In the next subsection, we demonstrate that this assumption can be achieved for many risk maps that are not necessarily coherent.

4.2 Construction of an Unbiased Estimator $\hat{\mathcal{H}}$

In this subsection, we present an estimator $\hat{\mathcal{H}}$ that satisfies Assumption 4.1 for specific risk maps using the multi-level Monte Carlo (MLMC) method, an approach for unbiased statistical estimation using stochastic simulation (Blanchet and Glynn, 2015; Blanchet et al., 2019; Liu et al., 2022; Wang et al., 2023a,c). We first impose the following assumption on $\mathcal{R}$, which enables the possibility of estimating a random variable using its empirical distribution.

Assumption 4.6 (Hölder continuity)**.** *There exists an $\mathfrak{L} > 0$ such that for all $v, w \in \mathcal{L}(\mathcal{X})$, we have $|\mathcal{R}_{x,a}(v) - \mathcal{R}_{x,a}(w)| \leq \mathfrak{L} d_W(\mu_v, \mu_w)$, $\forall (x, a) \in \mathcal{K}$, where μ_v, μ_w are the probability distributions of v and w on $(\mathcal{X}, P(\cdot|x, a))$ and $d_W(\cdot, \cdot)$ is the 1-Wasserstein distance between two distributions.*

We first generate N according to a geometric distribution with parameter $r \in (0, 1)$. Then, for each $(x, a) \in \mathcal{K}$, we take action a at state x for 2^{N+1} times and observe the i.i.d. transitions $\{x_i'\}_{i=1}^{2^{N+1}}$. These 2^{N+1} samples are then divided into two groups: samples with odd indices and samples with even indices. We calculate the empirical distribution of x' using the even-index samples, odd-index samples, all the samples, and the first sample: $\hat{P}_{N+1}^E(y|x, a) := \frac{1}{2^N} \sum_{i=1}^{2^N} \mathbf{1}\{x_{2i}' = y\}$, $\hat{P}_{N+1}^O(y|x, a) := \frac{1}{2^N} \sum_{i=1}^{2^N} \mathbf{1}\{x_{2i-1}' = y\}$, $\hat{P}_{N+1}(y|x, a) := \frac{1}{2^{N+1}} \sum_{i=1}^{2^{N+1}} \mathbf{1}\{x_i' = y\}$, $\hat{P}_{N+1}^1(y|x, a) := \mathbf{1}\{x_1' = y\}$. For notation simplicity, we denote the resulting empirical transition kernels as $\hat{P}_{N+1}^E$, $\hat{P}_{N+1}^O$, $\hat{P}_{N+1}$ and $\hat{P}_{N+1}^1$, respectively. Then, we use these estimated transition kernels as nominal kernels to calculate $\mathcal{H}$. Namely, $\mathcal{H}_{\hat{P}_N}$ is the Bellman optimality operator under the empirical transition kernel $\hat{P}_N$. The multi-level estimator of $\mathcal{H}$ is then defined as

$$\hat{\mathcal{H}}(q) := \mathcal{H}_{\hat{P}_{N+1}^1}(q) + \frac{1}{p_N} \left\{ \mathcal{H}_{\hat{P}_{N+1}}(q) - \frac{1}{2} \left(\mathcal{H}_{\hat{P}_{N+1}^E}(q) + \mathcal{H}_{\hat{P}_{N+1}^O}(q) \right) \right\}, \quad \forall q \in \mathcal{L}(\mathcal{K}), \tag{4.2}$$

where $p_N := r(1 - r)^N$.

We present the following result on the unbiasedness and controllable variance for risk maps satisfying Assumption 4.6.

Theorem 4.7. *Assumption 4.1 holds if the risk map satisfies Assumption 4.6 and $r \in (0, 1/2)$.*

As shown in Section 3 of Prashanth and Bhat (2022), several popular risk measures, including UBSR, OCE and spectral risk measures, satisfy Assumption 4.6 with proper parametrization. Below, we demonstrate that under suitable conditions, the average risk-aware MDP that incorporates UBSR, OCE, or spectral risk measure satisfies all the assumptions required for Theorem 4.5 to apply.

Assumption 4.8 (Strong ergodicity)**.** *Under any stationary policy, the resulting Markov chain is irreducible and there exists a state $\bar{x} \in \mathcal{X}$ such that $P(\bar{x}|x, a) > 0$, $\forall (x, a) \in \mathcal{K}$.*

Assumption 4.9 (Bounded slope)**.** *The loss function $\ell(x)$ is strictly increasing on $\mathbb{R}$ and there exist $L_1, \epsilon_1 > 0$ such that $0 < \epsilon_1 \leq \frac{\ell(x) - \ell(y)}{x - y} \leq L_1$, $\forall x \neq y \in \mathbb{R}$.*

Theorem 4.10. *Under Assumption 4.8, if the risk map employs a UBSR or OCE satisfying Assumption 4.9, or a spectral risk measure with $0 < \epsilon_2 \leq \phi(\cdot) \leq L_2 < \infty$, then assumptions 2.12, 4.2, and 4.6 holds. Consequently, Theorem 4.5 applies.*

We note that CVaR does not satisfy the condition in Theorem 4.10. However, the mean-CVaR risk measure, which mixes expectation and CVaR, does and hence Theorem 4.5 applies for mean-CVaR. Also, although our definitions of UBSR and OCE assume a convex loss function, Theorem 4.10 holds more generally for loss functions that are convex (concave) for $x > 0$ and concave (convex) for $x < 0$, reflecting different risk attitudes toward gains and losses (see Appendix B.4). Finally, the entropic risk measure does not satisfy Assumption 4.9, but it still meets Assumption 2.12 (Proposition 5.7, Shen et al. (2013)). Borkar (2002) proposed a Q-learning algorithm for average risk-aware MDPs with an entropic risk measure, which is derived from the multiplicative Poisson equation and does not rely on MLMC. For further details, readers may refer to Borkar (2002).

4.3 An Off-policy Q-learning Algorithm for UBSR Measures

An important practical concern of our MLMC Q-learning algorithm is the necessity for a resampling procedure for each (x, a) pair, which prevents the algorithm from being adapted for off-policy learning. This can be addressed when the risk map employs a UBSR measure using an approach proposed in Shen et al. (2014) for risk-aware discounted MDPs. Namely, Proposition 4.113 in Föllmer and Schied (2016) establishes that for any $v \in \mathcal{L}(\mathcal{X})$, the risk map $\mathrm{SR}_{x,a}(v)$ is the unique solution of $\mathbb{E}_{x,a}[\ell(v - \mathrm{SR}_{x,a}(v))] = 0$. This implies that the AROE (3.4) can be equivalently rewritten as:

$$\mathbb{E}\left[\ell\left(c(x,a) + \min_{a' \in \mathcal{A}} q(\cdot, a') - f(q) - q(x,a)\right)\right] = 0, \quad \forall (x,a) \in \mathcal{K}.$$

This motivates the following off-policy algorithm that seeks to identify the root of this AROE using stochastic approximation (see Borkar (2008)). Specifically, given any sequence $\{(x_n, a_n, x'_n)\}$, with $x'_n \sim P(\cdot|x, a)$, the asynchronous UBSR-based Q-learning consists in applying the updates:

$$Q_{n+1}(x_n, a_n) = Q_n(x_n, a_n) + \gamma(n)\ell\Big(c(x_n, a_n) + \min_{a' \in \mathcal{A}} Q_n(x'_n, a') - f(Q_n) - Q_n(x_n, a_n)\Big). \quad (4.3)$$

where each subsequence $\{\gamma(n)\}_{n:(x_n,a_n)=(x,a)}$, indexed by $(x, a) \in \mathcal{K}$, satisfies Assumption 4.4. The synchronous algorithm can also be derived if all (x, a) pairs are updated within one iteration.

It shall be noticed that the theoretical convergence of this algorithm remains an open question. Drawing from the ODE analysis used in the risk-neutral case, the corresponding ODE for (4.3) involves a high-dimensional, nonlinear system that is difficult to analyze for stability. Additional discussion on this problem can be found in Appendix C.2.

5 Experiments

In this section, we provide numerical experiments confirming the convergence of our MLMC-based Q-learning algorithm (MLMC Q-learning), comparisons to the off-policy Q-learning algorithm for UBSR (UBSR Q-learning), and apply our algorithm (see pseudo-codes in Appendix A) to real-life problems to showcase its potential. Further details and experimental investigations are also presented in Appendix C, namely regarding the sensitivity of MLMC Q-learning to r, the convergence of UBSR Q-learning, and the effect of risk-awareness in long term performance of policies.

5.1 Convergence of MLMC Q-learning

We begin by validating the convergence of the risk-aware RVI Q-learning algorithm (4.1) using a randomly generated MDP with 10 states and 5 actions per state. The nominal transition kernel P is generated from a uniform distribution over $[0, 1]$ and subsequently normalized. The cost function is sampled from a normal distribution $\mathcal{N}(1, 1)$. We choose $\gamma(n) := (1/(n + 1))^{2/3}$ and $f(q) := \frac{1}{|\mathcal{X}||\mathcal{A}|}\sum_{x,a} q(x, a)$. Due to space limit, we only show the convergence results for two special cases of UBSR and OCE: the expectile with $\tau = 0.75$ and OCE with loss function $\ell(x) := \gamma_1 x^+ - \gamma_2 x^-$ where $\gamma_1 = 2$ and $\gamma_2 = 0.5$.

We run the MLMC Q-learning algorithm 100 times independently with $r = 0.49$ and plot the mean value of $f(Q_n)$ in Figure 5.1, with the 95th and 5th percentiles as the confidence interval (CI). For comparison, trajectories from value iteration and the risk-aware RVI algorithm (3.2) are shown together with the true optimal risk (via value iteration). It is evident that the MLMC Q-learning

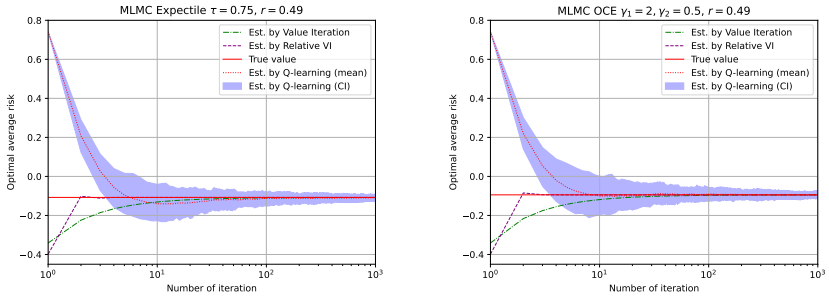

Figure 5.1: Convergence experiments for risk-aware RVI (3.2) and MLMC Q-learning(4.1).

algorithm converges to the true optimal average risk almost surely. As a model-based approach, the risk-aware RVI achieves convergence to the optimum at a significantly faster rate.

It is worth noting that selecting $r \in (0, 1/2)$ does not ensure finite sample guarantees based on Theorem 4.7, as each iteration requires an average of infinitely many samples when $r \leq 1/2$. However, for some $r \in (1/2, 3/4)$, both asymptotic and finite sample guarantees may still be achieved as observed empirically in additional experiments presented in Appendix C.1. These empirical findings are coherent with the guarantees identified in Wang et al. (2023a) for a special class of distributionally robust discounted MDPs.

5.2 Comparisons of MLMC and UBSR Q-learning Algorithms

Figure 5.2 presents the convergence of the synchronous and asynchronous UBSR Q-learning algorithms for the expectile in the same setting outlined in Section 5.1. For comparison, we also include the results of the MLMC Q-learning algorithm (4.1) with $r = 0.6$, which uses an expected 6,000 samples per state-action pair over 1,000 iterations, corresponding to 300,000 iterations of the asynchronous UBSR Q-learning algorithm (4.3) and 6,000 iterations for the synchronous version.

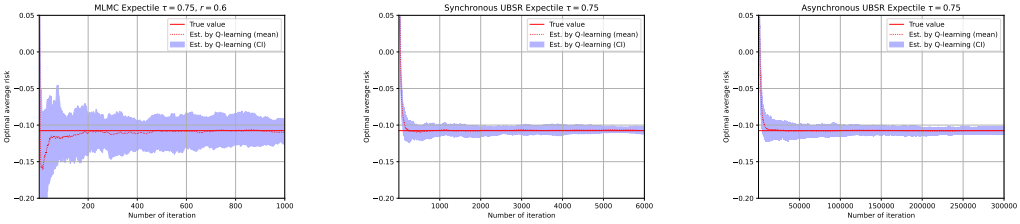

Figure 5.2: Comparison of MLMC and UBSR Q-learning with equivalent number of samples.

Further experiments with different loss functions along with a detailed discussion and comparison to the MLMC Q-learning algorithm (4.1) are provided in Appendix C.2. While the MLMC Q-learning algorithm offers provable convergence to optimality, the two UBSR Q-learning algorithms (4.3) demonstrate both faster convergence and lower variance compared to the MLMC Q-learning algorithm (4.1). Investigating the almost sure convergence and optimality of this algorithm remains an interesting direction for future research.

5.3 Applications

To illustrate the practicality of our risk-aware algorithms, we tested them on three popular average-cost MDP problems: machine replacement (MR), water reservoir management (WR), and inventory management (IM) (e.g. Puterman (1994), Hernández-Lerma (1989)). Each problem is evaluated under four risk measures: expectile (EX, $\tau = 0.9$), OCE ($\gamma_1 = 2$, $\gamma_2 = 0.5$), mean-CVaR ($\eta = 0.1, \alpha = 0.2$) and risk-neutral (RN). The optimal policies for each risk measure are obtained using the risk-aware RVI algorithm. Experimental details are provided in Appendix C.3.

Table 5.1 reports the optimal average risk for the three applications across the four risk measures, as well as the average risk of the four risk-aware policies when evaluated under the expectile risk measure with $\tau = 0.9$, i.e., the average risk computed under the expectile risk measure using the optimal policies derived for the other risk measures. It can be observed that the policies produced by our algorithms successfully attain their respective optimal average risks under difference application settings. Appendix C.4 explores risk differences across τ values, showcasing UBSR's flexibility in risk preference design. These results confirm our theory, proving the effectiveness of our algorithms in computing optimal risk-aware policies tailored to an agent's risk preferences.

Table 5.1: Average risk under different risk measures for three experimental setups: machine replacement (MR), water reservoir management (WR), and inventory management (IM)

Risk Measures	**Optimal Risk**			**Expectile Risk**		
	MR	**WR**	**IM**	**MR**	**WR**	**IM**
EX	68.7499	20.2541	24.8694	**68.7499**	**20.2541**	**24.8694**
OCE	63.9291	14.1389	23.7330	68.9323	20.3239	25.2908
Mean-CVaR	59.5244	9.9319	22.8955	69.3343	20.6413	26.5721
RN	54.4233	7.6174	20.1345	69.9359	20.6413	28.1901

6 Conclusion and Future Research

In this paper, we introduced the first risk-aware RVI algorithm and two novel model-free risk-aware RVI Q-learning algorithms for average-cost MDPs. MLMC Q-learning can be applied with a more general class of risk measures, while requiring access to repeated samples of transitions from a given state-action pair. UBSR Q-learning more closely aligns with the traditional setting of Q-learning yet is dedicated to the class of UBSR measures.

Several research directions are worth exploring. First, we conjecture that the strong ergodicity Assumption 4.8 could be weakened. Second, establishing finite-sample guarantees for MLMC Q-learning remains an open problem. It would be interesting to examine whether the conditions in Wang et al. (2023a) for a specific class of distributionally robust discounted MDPs and the variance reduction techniques in Wang et al. (2024) can be extended to our framework. Third, the almost sure convergence of UBSR Q-learning should be addressed. Finally, exploring the applicability to large-scale problems and designing other types of reinforcement learning algorithms for average risk-aware reinforcement learning constitute important directions for future research.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable feedback and insightful comments. Erick Delage was partially supported by the Canadian Natural Sciences and Engineering Research Council [Grant RGPIN-2022-05261] and by the Canada Research Chair program [950-230057]. We are also thankful to Esther Derman, Marek Petrik and Xian Chen for valuable discussions on related topics.

References

J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

A. Arapostathis and V. S. Borkar. On the relative value iteration with a risk-sensitive criterion. *arXiv preprint arXiv:1912.08758*, 2019.

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

F. Bellini and V. Bignozzi. On elicitable risk measures. *Quantitative Finance*, 15(5):725–733, 2015.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(2):449–476, 2008.

D. P. Bertsekas. *Dynamic Programming and Optimal Control, 3rd Edition, Volume II*. Athena Scientific, Belmont, MA, 2007.

A. Biswas and V. S. Borkar. Ergodic risk-sensitive control—a survey. *Annual Reviews in Control*, 55:118–141, 2023.

J. H. Blanchet and P. W. Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

J. H. Blanchet, P. W. Glynn, and Y. Pei. Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*, 2019.

V. S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2): 294–311, 2002.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

V. S. Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, pages 1327–1332, 2010.

V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

V. S. Borkar and K. Soumyanatha. An analog scheme for fixed point computation. I. Theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(4):351–355, 1997.

R. Cavazos-Cadena and E. Fernández-Gaucherand. Controlled Markov chains with risk-sensitive criteria: Average cost, optimality equations, and optimal solutions. *Mathematical Methods of Operations Research*, 49:299–324, 1999.

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

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time, 4th Edition*. Walter de Gruyter & Co., Berlin, 2016.

A. Gupta, R. Jain, and P. W. Glynn. An empirical algorithm for relative value iteration for average-cost MDPs. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5079–5084, 2015.

J. L. Hau, E. Delage, E. Derman, M. Ghavamzadeh, and M. Petrik. Q-learning for quantile MDPs: A decomposition, performance, and convergence analysis. In *International Conference on Artificial Intelligence and Statistics*, 2025.

O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer Science & Business Media, 1989.

H. Hmedi, A. Arapostathis, and G. Pang. On the global convergence of relative value iteration for infinite-horizon risk-sensitive control of diffusions. *Systems & Control Letters*, 171:105413, 2023.

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

W. Huang and W. B. Haskell. Risk-aware Q-learning for Markov decision processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4928–4933. IEEE, 2017.

U. Köse and A. Ruszczyński. Risk-averse learning by temporal difference methods with Markov risk measures. *Journal of Machine Learning Research*, 22(38):1–34, 2021.

T. Lam, A. Verma, B. K. H. Low, and P. Jaillet. Risk-aware reinforcement learning with coherent risk measures and non-linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023.

Z. Liu, Q. Bai, J. Blanchet, P. Dong, W. Xu, Z. Zhou, and Z. Zhou. Distributionally robust Q-learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.

S. Marzban, E. Delage, and J. Y. Li. Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. *Quantitative Finance*, 23(10):1411–1430, 2023.

M. Moharrami, Y. Murthy, A. Roy, and R. Srikant. A policy gradient algorithm for the risk-sensitive exponential cost MDP. *Mathematics of Operations Research*, 0(0), 2024.

A. Naik, R. Shariff, N. Yasui, H. Yao, and R. S. Sutton. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*, 2019.

L. A. Prashanth and S. P. Bhat. A Wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23(238):1–61, 2022.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994.

A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming, Series B*, 125:235–261, 2010.

A. Ruszczyński and A. Shapiro. Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561, 2006.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, 3rd Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.

Y. Shen, W. Stannat, and K. Obermayer. Risk-sensitive Markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 2014.

Y. Shen, K. Obermayer, and W. Stannat. On average risk-sensitive Markov control processes. *arXiv preprint arXiv:1403.3321*, 2015.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2018.

A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Y. Wan, A. Naik, and R. S. Sutton. Learning and planning in average-reward Markov decision processes. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10653–10662. PMLR, 2021.

Y. Wan, H. Yu, and R. S. Sutton. On convergence of average-reward Q-learning in weakly communicating Markov decision processes. *arXiv preprint arXiv:2408.16262*, 2024.

S. Wang, N. Si, J. Blanchet, and Z. Zhou. A finite sample complexity bound for distributionally robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR, 2023a.

S. Wang, N. Si, J. Blanchet, and Z. Zhou. Sample complexity of variance-reduced distributionally robust Q-learning. *Journal of Machine Learning Research*, 25(341):1–77, 2024.

Y. Wang, A. Velasquez, G. K. Atia, A. Prater-Bennette, and S. Zou. Robust average-reward Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15215–15223, 2023b.

Y. Wang, A. Velasquez, G. K. Atia, A. Prater-Bennette, and S. Zou. Model-free robust average-reward reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 36431–36469. PMLR, 2023c.

Appendices

A Algorithm Pseudo-codes

Algorithm 1 Risk-aware Relative Value Iteration

Input: V_0, f, T, $n \leftarrow 0$;
while $n < T$ **do**
  for all $x \in \mathcal{X}$ **do**
    $V_{n+1}(x) \leftarrow \mathcal{G}(V_n)(x) - f(V_n)$, where $\mathcal{G}$ defined in (3.2);
  end for
  $n \leftarrow n + 1$;
end while

Algorithm 2 Risk-aware Relative Q-factor Iteration

Input: Q_0, f, T, $n \leftarrow 0$;
while $n < T$ **do**
  for all $x \in \mathcal{X}$, $a \in \mathcal{A}$ **do**
    $Q_{n+1}(x,a) \leftarrow \mathcal{H}(Q_n)(x,a) - f(Q_n)$, where $\mathcal{H}$ defined in (3.3);
  end for
  $n \leftarrow n + 1$;
end while

Algorithm 3 Risk-aware RVI Q-learning with MLMC (MLMC Q-learning)

Input: Q_0, f, $\gamma(n)$, $r \in (0,1)$, T, $n \leftarrow 0$;
while $n < T$ **do**
  for all $x \in \mathcal{X}$, $a \in \mathcal{A}$ **do**
    Sample $N \sim \text{Geo}(r)$;
    Independently draw 2^{N+1} samples $x_i' \sim P(\cdot|x,a)$;
    $Q_{n+1}(x,a) \leftarrow Q_n(x,a) + \gamma(n) \left(\hat{\mathcal{H}}(Q_n)(x,a) - f(Q_n) - Q_n(x,a) \right)$, where $\hat{\mathcal{H}}$ is defined in (4.2);
  end for
  $n \leftarrow n + 1$;
end while

Algorithm 4 Off-policy (asynchronous) RVI Q-learning for UBSR (A-UBSR Q-learning)

Input: Q_0, ℓ, f, $\gamma(n)$, T, $n \leftarrow 0$;
while $n < T$ **do**
  Observe one transition (x, a, x');
  $Q_{n+1}(x,a) \leftarrow Q_n(x,a) + \gamma(n)\ell\left(c(x,a) + \min_{a' \in \mathcal{A}} Q_n(x',a') - f(Q_n) - Q_n(x,a)\right)$;
  $n \leftarrow n + 1$;
end while

Algorithm 5 Synchronous RVI Q-learning for UBSR (S-UBSR Q-learning)

Input: Q_0, ℓ, f, $\gamma(n)$, T, $n \leftarrow 0$;
while $n < T$ **do**
  for all $x \in \mathcal{X}$, $a \in \mathcal{A}$ **do**
    Observe one sample x';
    $Q_{n+1}(x,a) \leftarrow Q_n(x,a) + \gamma(n)\ell\left(c(x,a) + \min_{a' \in \mathcal{A}} Q_n(x',a') - f(Q_n) - Q_n(x,a)\right)$;
  end for
  $n \leftarrow n + 1$;
end while

B Proofs

B.1 Proof of Theorem 3.2 and 3.4

From the definition of the risk-aware RQI algorithm (3.3), it is evident that the risk-aware RVI algorithm (3.2) can be considered a special case of RQI by defining $V(x) := \min_{a\in\mathcal{A}} Q(x,a)$. Consequently, if the risk-aware RQI algorithm converges, then it follows that the risk-aware RVI algorithm also converges.

In order to study the Q-factor iteration, we make use of an augmented risk map $\widetilde{\mathcal{R}}$, on the $\mathcal{K}$ outcome space, using

$$\widetilde{\mathcal{R}}_{x,a}(q) := \mathcal{R}\left(\min_{a'\in\mathcal{A}} q(\cdot,a')\Big|x,a\right), \quad \forall q\in\mathcal{L}(\mathcal{K}). \tag{B.1}$$

The risk map $\widetilde{\mathcal{R}}$ implicitly reduces the average risk-aware control problem to an average risk evaluation on a cost generating Markov chain. We thus invoke a general Doeblin type condition for average risk evaluation on a Markov chain as follows, where we see $\mathcal{K}$ as the set of states of the Markov chain.

Assumption B.1 (Assumption 3.1, Shen et al. (2013))**.** *There exists a function $\tilde{w}:\mathcal{K}\to[0,+\infty)$, a monetary risk measure $\tilde{\nu}:\mathcal{L}(\mathcal{K})\to\mathbb{R}$, and some constants $\tilde{K}>0$, $\tilde{\gamma}\in(0,1)$, and $\tilde{\alpha}\in(0,1)$ such that:*

(1) Let $\widetilde{\mathcal{R}}^{\#}_{x,a}(q) := \sup_{p\in\mathcal{L}(\mathcal{X})}\{\widetilde{\mathcal{R}}_{x,a}(q+p)-\widetilde{\mathcal{R}}_{x,a}(p)\}$ and $\overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}(q) := \sup_{\lambda\neq 0}\frac{\widetilde{\mathcal{R}}^{\#}_{x,a}(\lambda q)}{\lambda}$. We have that

$$\overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}(\tilde{w}) \le \tilde{\gamma}\tilde{w}(x,a)+\tilde{K}, \quad \forall (x,a)\in\mathcal{K}.$$

(2) For all $q\ge p\in\mathcal{L}(\mathcal{K})$, we have that:

$$\inf_{(x,a)\in\mathcal{K}:\tilde{w}(x,a)\le\tilde{R}}\{\widetilde{\mathcal{R}}(q|x,a)-\tilde{\alpha}\tilde{\nu}(q)-\widetilde{\mathcal{R}}(p|x,a)+\tilde{\alpha}\tilde{\nu}(p)\}\ge 0,$$

for some $\tilde{R}>2\tilde{K}/(1-\tilde{\gamma})$.

Lemma B.2. *If $\mathcal{R}$ satisfies Assumption 2.12, then $\widetilde{\mathcal{R}}$ satisfies Assumption B.1.*

Proof. For the $\bar{\alpha}$ and ν satisfying Assumption 2.12, define $\tilde{w}:=\mathbf{0}$, $\tilde{K}:=1$, $\tilde{\gamma}:=0.5$, $\tilde{\alpha}:=0.5$, $\tilde{R}:=5>2\tilde{K}/(1-\tilde{\gamma})$, and monetary risk measure $\widetilde{\nu}(q):=\nu(\min_{a\in\mathcal{A}} q(\cdot,a))$. We have

$$\overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}(\mathbf{0}) = \sup_{\lambda\neq 0}\frac{\widetilde{\mathcal{R}}^{\#}_{x,a}(\lambda\mathbf{0})}{\lambda} = \sup_{\lambda\neq 0}\frac{\widetilde{\mathcal{R}}^{\#}_{x,a}(\mathbf{0})}{\lambda} = \sup_{\lambda\neq 0}\frac{0}{\lambda} = 0 \le 0.5\cdot 0+1 = \tilde{\gamma}\tilde{w}(x,a)+\tilde{K},$$

where we exploited the fact that:

$$\widetilde{\mathcal{R}}^{\#}_{x,a}(\mathbf{0}) = \sup_{q\in\mathcal{L}(\mathcal{K})}\{\widetilde{\mathcal{R}}_{x,a}(q)-\widetilde{\mathcal{R}}_{x,a}(q)\}=0, \quad \forall (x,a)\in\mathcal{K}.$$

Moreover, for all $q\ge p\in\mathcal{L}(\mathcal{K})$, we have

$$\begin{aligned}
&\min_{(x,a)\in\mathcal{K},\tilde{w}(x,a)\le\tilde{R}}\{\widetilde{\mathcal{R}}(q|x,a)-\bar{\alpha}\tilde{\nu}(q)-\widetilde{\mathcal{R}}(p|x,a)+\bar{\alpha}\tilde{\nu}(p)\}\\
&=\min_{(x,a)\in\mathcal{K}}\{\widetilde{\mathcal{R}}(q|x,a)-\bar{\alpha}\tilde{\nu}(q)-\widetilde{\mathcal{R}}(p|x,a)+\bar{\alpha}\tilde{\nu}(p)\}\\
&=\min_{(x,a)\in\mathcal{K}}\{\mathcal{R}_{x,a}(\min_{a'} q(\cdot,a'))-\bar{\alpha}\nu(\min_{a'} q(\cdot,a'))-\mathcal{R}_{x,a}(\min_{a'} p(\cdot,a'))+\bar{\alpha}\nu(\min_{a'} p(\cdot,a'))\}\\
&\ge 0,
\end{aligned}$$

where the last inequality follows from Assumption 2.12 using $v(\cdot):=\min_{a'\in\mathcal{A}} q(\cdot,a')$ and $v'(\cdot):=\min_{a'\in\mathcal{A}} p(\cdot,a')$. The result follows. □

We now list the properties of span-seminorm contractive operators that we will use later on.

Lemma B.3 (Theorem 6.6.2, Puterman (1994)). *Let $\mathcal{T} : \mathcal{L}(\Omega) \to \mathcal{L}(\Omega)$, for some finite space Ω, be an operator that is span-seminorm contractive, i.e., there exists an $\bar{\alpha} \in [0, 1)$ such that $\|\mathcal{T}(v) - \mathcal{T}(w)\|_{sp} \leq \bar{\alpha}\|v - w\|_{sp}$, for all $v, w \in \mathcal{L}(\Omega)$. Then the followings are true:*

(1) There exists a $v^* \in \mathcal{L}(\Omega)$ such that $\|\mathcal{T}(v^*) - v^*\|_{sp} = 0$. Such v^* is called the span-seminorm fixed point of the operator $\mathcal{T}$.
(2) For all $n \geq 0$, $\|\mathcal{T}^n(v) - v^*\|_{sp} \leq \bar{\alpha}^n\|v - v^*\|_{sp}$.
(3) For any $v \in \mathcal{L}(\Omega)$, we have $\lim_{n\to\infty} \|\mathcal{T}^n(v) - v^*\|_{sp} = 0$.
(4) Any two span-seminorm fixed points of $\mathcal{T}$ must differ by a constant.

In our proofs we will exploit the fact that $\widetilde{\mathcal{R}}$ is non-expansive and span-seminorm contractive.

Lemma B.4. *If the risk map $\mathcal{R}$ satisfies Assumption 2.12, then both $\widetilde{\mathcal{R}}$ and $\mathcal{H}$ (from equation (3.3)) are non-expansive under the infinity norm and span-seminorm contractive.*

Proof. From Proposition 3.6 in Shen et al. (2013), for any risk map $\widetilde{\mathcal{R}}$, which satisfies Assumption B.1 based on Lemma B.2, we have $|\widetilde{\mathcal{R}}_{x,a}(q + p) - \widetilde{\mathcal{R}}_{x,a}(p)| \leq \overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}(|q|)$, where $|q|(x) := |q(x)|$. Since $\overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}$ is a coherent risk measure (see Proposition 3.5 in Shen et al. (2013)), we have for all $q, p \in \mathcal{L}(\mathcal{K})$:

$$|\widetilde{\mathcal{R}}_{x,a}(q) - \widetilde{\mathcal{R}}_{x,a}(p)| \leq \overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}(|q - p|) \leq \overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}(\|q - p\|_\infty e) = \|q - p\|_\infty,$$

where the second inequality follows from monotonicity of $\overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}$, while the last equality comes from translation invariance and normalization of $\overline{\widetilde{\mathcal{R}}^{\#}}_{x,a}$.

The span-seminorm contraction straightforwardly follows from Theorem 3.11 in Shen et al. (2013) given the fact that $\widetilde{\mathcal{R}}$ satisfies Assumption B.1 as established in Lemma B.2.

These properties carry directly to $\mathcal{H}$ since

$$\|\mathcal{H}(q) - \mathcal{H}(p)\|_\infty = \|\widetilde{\mathcal{R}}(q|\cdot) - \widetilde{\mathcal{R}}(p|\cdot)\|_\infty \leq \|q - p\|_\infty,$$

and

$$\|\mathcal{H}(q) - \mathcal{H}(p)\|_{sp} = \|\widetilde{\mathcal{R}}(q|\cdot) - \widetilde{\mathcal{R}}(p|\cdot)\|_{sp} \leq \bar{\alpha}\|q - p\|_{sp}$$

for some $\bar{\alpha} \in [0, 1)$. This completes the proof. □

We are now ready prove the convergence of the risk-aware RQI algorithm (3.3).

Proof of Theorem 3.4. Define $\bar{V}_n(x) := \min_{a\in\mathcal{A}} Q_n(x, a)$, $\forall x \in \mathcal{X}$. Taking minimum over a on both sides of (3.3), we obtain

$$\bar{V}_{n+1}(x) = \min_{a\in\mathcal{A}}\{c(x, a) + \mathcal{R}_{x,a}(\bar{V}_n)\} - f(Q_n) = \mathcal{G}(\bar{V}_n)(x) - f(Q_n),\ \forall x \in \mathcal{X},$$

where $\mathcal{G}$ is the risk-aware Bellman optimality operator. If Q_n converges to some fixed point Q_∞ of (3.3) under the infinity norm, we have

$$\min_{a\in\mathcal{A}} Q_\infty(x, a) = \min_{a\in\mathcal{A}} \left\{c(x, a) + \mathcal{R}_{x,a}\left(\min_{a'\in\mathcal{A}} Q_\infty(\cdot, a')\right)\right\} - f(Q_\infty), \quad \forall x \in \mathcal{X}. \tag{B.2}$$

Notice that $\min_{a\in\mathcal{A}} Q_\infty(\cdot, a) \in \mathcal{L}(\mathcal{X})$, $f(Q_\infty) \in \mathbb{R}$ and Q_∞ satisfies B.2, we conclude that $(\min_{a\in\mathcal{A}} Q_\infty(\cdot, a), f(Q_\infty))$ is a pair of solution to the AROE (2.1). By Theorem 2.13, $f(Q_\infty) = g^*$. Therefore, We are left with the task to show that Q_n converges to some unique fixed point Q_∞ of (3.3).

To analyze the convergence, consider the augmented risk map defined in (B.1). Then algorithm (3.3) can be equivalently written as:

$$Q_{n+1}(x, a) = c(x, a) + \widetilde{\mathcal{R}}_{x,a}(Q_n) - f(Q_n), \quad \forall (x, a) \in \mathcal{K}.$$

given the translation invariance of $\mathcal{R}_{x,a}$. Its convergence can be associated to the convergence of an average risk estimator on a Markov chain, with $\mathcal{K}$ as the state space, under the risk map $\widetilde{\mathcal{R}}$ on

the $\mathcal{K}$ outcome space, which is studied in Section 3 of Shen et al. (2013). Indeed, given that $\widetilde{\mathcal{R}}$ satisfies Assumption B.1, Theorem 3.14 (i) in Shen et al. (2013) already establishes that the Poisson equation,

$$c(x,a) + \widetilde{\mathcal{R}}_{x,a}(q) = g + q(x,a), \quad \forall (x,a) \in \mathcal{K}, \tag{B.3}$$

has a solution $(q^*, \tilde{g}^*)$, where $\tilde{g}^*$ is unique.

By Lemma B.4, the risk-aware Bellman optimality operator $\mathcal{H}$ (see (3.3)) is span-seminorm contractive. This implies, based on Lemma B.3, that $\mathcal{H}$ has a span-seminorm fixed point, i.e, there exists $q^* \in \mathcal{L}(\mathcal{K})$ and $\tilde{g}^* \in \mathbb{R}$ such that $\|\mathcal{H}(q^*) - q^*\|_{sp} = 0$ and $q^* + \tilde{g}^* = \mathcal{H}(q^*)$. The latter implies that $(q^*, \tilde{g}^*)$ satisfies the Poisson equation (B.3) and that $\lim_{n\to\infty} \|\mathcal{H}^n(q) - q^*\|_{sp} = 0$, for any $q \in \mathcal{L}(\mathcal{K})$, due to the span-seminorm contraction property of $\mathcal{H}$ (Lemma B.3).

One can further show that $\mathcal{H}^{n+1}(q) - \mathcal{H}^n(q) \to \tilde{g}^*$, for any $q \in \mathcal{L}(\mathcal{K})$, using

$$\begin{aligned}
&\|\mathcal{H}^{n+1}(q) - \mathcal{H}^n(q) - \tilde{g}^*\|_\infty \\
&= \inf_g \|\mathcal{H}(q^* + g + \mathcal{H}^n(q) - g - q^*) - (q^* + g + \mathcal{H}^n(q) - g - q^*) - \tilde{g}^*\|_\infty \\
&= \inf_g \|\mathcal{H}(q^* + \mathcal{H}^n(q) - g - q^*) + g - q^* - g - \mathcal{H}^n(q) + g + q^* - \tilde{g}^*\|_\infty \\
&\le \inf_g \{\|\mathcal{H}(q^* + \mathcal{H}^n(q) - g - q^*) - \mathcal{H}(q^*)\|_\infty + \|\mathcal{H}(q^*) - q^* - \tilde{g}^*\|_\infty + \|\mathcal{H}^n(q) - g - q^*\|_\infty\} \\
&\le \inf_g \{\|\mathcal{H}^n(q) - g - q^*\|_\infty + \|q^* + \tilde{g}^* - q^* - \tilde{g}^*\|_\infty + \|\mathcal{H}^n(q) - g - q^*\|_\infty\} \\
&= 2 \inf_g \|\mathcal{H}^n(q) - g - q^*\|_\infty = \|\mathcal{H}^n(q) - q^*\|_{sp}.
\end{aligned}$$

where the second equality comes from translation invariance, the first inequality comes from the triangular inequality, the second inequality follows from $\mathcal{H}$ being non-expansive (see Lemma B.4), and finally the last equality is proved as Lemma 3.9 in Shen et al. (2013). Hence, we must have that $\lim_{n\to 0} \|\mathcal{H}^{n+1}(q) - \mathcal{H}^n(q) - \tilde{g}^*\|_\infty \le \lim_{n\to\infty} \|\mathcal{H}^n(q) - q^*\|_{sp} = 0$.

We now wish to analyze the convergence of the process $\{Q_n\}_{n=0}^\infty$ produced by our algorithm. To do so, consider the process $U_{n+1} := \mathcal{H}(U_n)$ with $U_0 := Q_0$, for which we know that $\|\tilde{g}_n - \tilde{g}^*\|_\infty \to 0$ as $n \to \infty$, where $\tilde{g}_n := U_{n+1} - U_n$, for $n \ge 0$. One can actually establish by induction that $Q_n = U_n - f(U_{n-1})$ for all $n \ge 1$. Namely, start at $n = 1$ where

$$Q_1 = \mathcal{H}(Q_0) - f(Q_0) = \mathcal{H}(U_0) - f(U_0) = U_1 - f(U_0).$$

Then iteratively assuming that $Q_n = U_n - f(U_{n-1})$, one can confirm that:

$$\begin{aligned}
Q_{n+1} &= \mathcal{H}(Q_n) - f(Q_n) = \mathcal{H}(U_n - f(U_{n-1})) - f(U_n - f(U_{n-1})) \\
&= \mathcal{H}(U_n) - f(U_{n-1}) - f(U_n) + f(U_{n-1}) = \mathcal{H}(U_n) - f(U_n) = U_{n+1} - f(U_n).
\end{aligned}$$

This relation can be used to establish that

$$\begin{aligned}
\|Q_{n+1} - Q_n\|_\infty &= \|U_{n+1} - f(U_n) - U_n + f(U_{n-1})\|_\infty \\
&= \|\tilde{g}_n - f(U_n) + f(U_{n-1} + \tilde{g}^*) - f(U_{n-1} + \tilde{g}^*) + f(U_{n-1})\|_\infty \\
&= \|\tilde{g}_n - \tilde{g}^* - f(U_n) + f(U_{n-1} + \tilde{g}^*)\|_\infty \\
&\le \|\tilde{g}_n - \tilde{g}^*\|_\infty + \|f(U_n) - f(U_{n-1} + \tilde{g}^*)\|_\infty \\
&\le \|\tilde{g}_n - \tilde{g}^*\|_\infty + \tilde{L}\|U_n - U_{n-1} - \tilde{g}^*\|_\infty \\
&= \|\tilde{g}_n - \tilde{g}^*\|_\infty + \tilde{L}\|\tilde{g}_{n-1} - \tilde{g}^*\|_\infty.
\end{aligned}$$

where the third equality follows from the translation invariance property of f imposed in Assumption 3.3 (i), while the final inequality arises from the Lipschitz property of f in in Assumption 3.3 (ii), with $\tilde{L} \ge 0$ as the Lipschitz constant. When $n \to \infty$, we have shown that $\|\tilde{g}_n - \tilde{g}^*\|_\infty = \|\mathcal{H}^{n+1}(Q_0) - \mathcal{H}^n(Q_0) - \tilde{g}^*\|_\infty \to 0$. Therefore we can conclude that Q_n converges to some Q_∞ and $f(Q_n)$ converges to $\tilde{g}^*$ as $n \to \infty$, i.e., $(Q_\infty, f(Q_\infty))$ satisfies the Poisson equation (B.3).

Finally, we show that such Q_∞ is independent of Q_0 for a fixed f. Since $Q_n \to Q_\infty$ as $n \to \infty$, from (3.3), we obtain that

$$Q_\infty(x,a) = \mathcal{H}(Q_\infty)(x,a) - f(Q_\infty), \quad \forall (x,a) \in \mathcal{K}. \tag{B.4}$$

Notice that $f(Q_\infty) = \tilde{g}^*$ is a constant, this implies that Q_∞ is a span-seminorm fixed point of $\mathcal{H}$. Suppose $\widetilde{Q}_\infty$ is another solution to (B.4). Then by Lemma B.3(iv), Q_∞ and $\widetilde{Q}_\infty$ only differs by a constant. Yet, we know that $\tilde{g}^* = f(\widetilde{Q}_\infty) = f(Q_\infty + r) = f(Q_\infty) + r = \tilde{g}^* + r$, which implies that $r = 0$ and that $Q_\infty = \widetilde{Q}_\infty$. We therefore conclude that Q_∞ is unique.

From the analysis in the first part, we conclude that $\tilde{h}^* := \min_{a\in\mathcal{A}} Q_\infty(\cdot, a)$ and $\tilde{g}^* := f(Q_\infty)$ identify a solution pair to the AROE (2.1) and thus $f(Q_\infty) = g^*$. □

We now turn to establishing Theorem 3.2.

Proof of Theorem 3.2. Let $\tilde{f} : \mathcal{L}(\mathcal{X}) \to \mathbb{R}$ be a function satisfying Assumption 3.1. Define a function $\hat{f} : \mathcal{L}(\mathcal{K}) \to \mathbb{R}$ as $\hat{f}(Q_n) := \tilde{f}(\min_{a\in\mathcal{A}} Q_n(\cdot, a))$. It is easy to verify that $\hat{f}$ satisfies Assumption 3.3.

Consider the risk-aware RQI (3.3) with function $\hat{f}$ starting with Q_0. Define $\bar{V}_n(x) := \min_{a\in\mathcal{A}} Q_n(x, a)$, for all $x \in \mathcal{X}$. Taking minimum over a on both sides of (3.3), we obtain

$$\begin{aligned}\bar{V}_{n+1}(x) &= \min_{a\in\mathcal{A}}\{c(x, a) + \mathcal{R}_{x,a}(\bar{V}_n)\} - \hat{f}(Q_n) = \mathcal{G}(\bar{V}_n)(x) - \hat{f}(Q_n)\\ &= \mathcal{G}(\bar{V}_n)(x) - \tilde{f}(\bar{V}_n), \quad \forall x \in \mathcal{X},\end{aligned}$$

where $\mathcal{G}$ is the risk-aware Bellman optimality operator defined in (3.2). This is exactly the risk-aware RVI algorithm with initial value function $V_0(x) = \min_{a\in\mathcal{A}} Q_0(x, a)$, for all $x \in \mathcal{X}$.

By Theorem 3.4, the RQI algorithm converges, i.e., Q_n converges to some unique fixed point of (3.3) called Q^*. Hence, $\bar{V}_n$ converges to some $\bar{V}^* := \min_{a\in\mathcal{A}} Q^*(\cdot, a)$. Since we can always design a Q_0 such that $\min_{a\in\mathcal{A}} Q_0(x, a) = \bar{V}_0(x)$ for any $\bar{V}_0 \in \mathcal{L}(\mathcal{X})$. We conclude that for any initial value V_0, the risk-aware RVI algorithm (3.2) converges to some V^*.

Using the same reasoning as in Lemma B.4, we can conclude that the risk-aware Bellman optimality operator $\mathcal{G}$ is non-expansive under the infinity norm and contractive with respect to the span-seminorm. Based on Assumption 3.1 and the preceding derivation, taking the limit on both sides of (3.2) yields the equation $V^*(x) = \mathcal{G}(V^*)(x) - \tilde{f}(V^*)$ for all $x \in \mathcal{X}$. This implies that $(V^*, \tilde{f}(V^*))$ identifies a pair of solution to the AROE (2.1), leading to $\tilde{f}(V^*) = g^*$ and V^* serves as a fixed point of (3.2). The uniqueness of V^* follows from the same argument used to establish the uniqueness of Q^* in the proof of Theorem 3.4. This completes the proof. □

B.2 Proof of Theorem 4.5

In this section, we use the ODE analysis of stochastic approximation to prove the convergence of the risk-aware RVI Q-learning algorithm.

Define an operator $H : \mathcal{L}(\mathcal{K}) \to \mathcal{L}(\mathcal{K})$ as

$$H(q)(x, a) := \mathcal{R}_{x,a}\left(c(x, a) + \min_{a'\in\mathcal{A}} q(\cdot, a')\right) - f(q) - q(x, a), \quad \forall q \in \mathcal{L}(\mathcal{K}).$$

Then the update of the risk-aware RVI Q-learning can be written as

$$\begin{aligned}Q_{n+1} &= Q_n + \gamma(n)(\hat{\mathcal{H}}(Q_n) - f(Q_n) - Q_n)\\ &= Q_n + \gamma(n)(H(Q_n) + \hat{\mathcal{H}}(Q_n) - H(Q_n) - f(Q_n) - Q_n).\end{aligned}$$

Hence we have the stochastic approximation iteration:

$$Q_{n+1} = Q_n + \gamma(n)(H(Q_n) + M_{n+1}), \tag{B.5}$$

where $M_{n+1} := \hat{\mathcal{H}}(Q_n) - \mathcal{H}(Q_n)$ is the noise term.

The classical approach to analyzing stochastic approximation using ODEs involves examining the stability of the equilibrium of a corresponding ODE related to (B.5):

$$\dot{p}_t = H(p_t). \tag{B.6}$$

If the ODE (B.6) has a unique globally asymptotically stable equilibrium point p^*, then under certain conditions, the stochastic approximation (B.5) converges, with $Q_n \to Q^* = p^*$ almost surely (see Theorem 2.2 of Borkar and Meyn (2000)). Notice that if such p^* exists, then Q^* is a solution to the AROE (3.4), which implies that $(\min_{a' \in \mathcal{A}} p^*(\cdot, a'), f(p^*))$ is a pair of solution to the AROE (2.1). Then the result of Theorem 4.5 follows easily.

Analyzing the stability of the equilibrium point of ODE (B.6) can sometimes be challenging. A common approach is to employ a time-averaging technique to smooth out perturbations and examine the stability of the origin in the limiting ODE. Namely, we define an operator $H_s : \mathcal{L}(\mathcal{K}) \to \mathcal{L}(\mathcal{K})$ as $H_s(Q) := \frac{1}{s} H(sQ)$, with $s \geq 1$ and consider the ODE:

$$\dot{\phi}_t = H_s(\phi_t). \tag{B.7}$$

Following Borkar and Meyn (2000), to establish the convergence of the stochastic approximation (B.5), we outline the sufficient conditions that are needed to be verified:

(i) The function H is Lipschitz.

(ii) The sequence $\{M_n, \mathcal{F}_n : n \geq 1\}$ with $\mathcal{F}_n := \sigma(Q_i, M_i, i \leq n)$ is a martingale difference sequence. Moreover, there exists some $C_0 < \infty$ and for any initial condition $Q_0 \in \mathcal{L}(\mathcal{K})$ we have almost surely,

$$\mathbb{E}[\|M_{n+1}\|_\infty^2 | \mathcal{F}_n] \leq C_0(1 + \|Q_n\|_\infty^2), \quad n \geq 0.$$

(iii) The step size satisfies the Robbins-Monro condition (see Assumption 4.4).

(iv) For any initial condition $Q_0 \in \mathcal{L}(\mathcal{K})$, the iteration is bounded almost surely, i.e., $\sup_n \|Q_n\|_\infty < \infty$, almost surely.

(v) The ODE (B.6) has a unique globally asymptotically stable equilibrium point.

(vi) The limit $H_\infty(Q) := \lim_{s \to \infty} H_s(Q)$ exists and the convergence is uniform on compact sets, and the ODE

$$\dot{\phi}_t = H_\infty(\phi_t), \tag{B.8}$$

has the origin as an asymptotically stable equilibrium.

Following Theorem 2.2 of Borkar and Meyn (2000), if conditions (i), (ii), (iii), (iv) and (v) hold, then the stochastic approximation (B.5) converges almost surely to the unique globally stable equilibrium point of the ODE (B.6), which is a solution to the AROE (3.4), thus confirming our theorem.

The Lipschitz property (i) is straightforward to verify, as the risk measure $\mathcal{R}_{x,a}$ and the function f are both Lipschitz (see Lemma B.4 and Assumption 3.3). As stated in Theorem 2.1 of Borkar and Meyn (2000), the almost sure boundedness condition (iv) follows from conditions (i), (vi), (ii) and (iii), where condition (iii) is automatically satisfied by Assumption 4.4. Thus, the remainder of this section focuses on verifying conditions (i), (ii), (v) and (vi).

B.2.1 Condition (i)

Lemma B.5. *H, H_s and H_∞, if it exists, are Lipschitz and have the same Lipschitz constant.*

Proof. Following Lemma B.4, $\widetilde{\mathcal{R}}$ is non-expansive. Hence, for any $Q_1, Q_2 \in \mathcal{L}(\mathcal{K})$, with Assumption 3.3, we have

$$\begin{aligned} H(Q_1)(x,a) - H(Q_2)(x,a) &= \widetilde{\mathcal{R}}_{x,a}(Q_1) - \widetilde{\mathcal{R}}_{x,a}(Q_2) - f(Q_1) + f(Q_2) - Q_1(x,a) - Q_2(x,a) \\ &\leq \|Q_1 - Q_2\|_\infty + \tilde{L}\|Q_1 - Q_2\|_\infty + \|Q_1 - Q_2\|_\infty \\ &= (2 + \tilde{L})\|Q_1 - Q_2\|_\infty, \end{aligned}$$

where $\tilde{L}$ is the Lipschitz constant for f. Similarly, we obtain $H(Q_2)(x,a) - H(Q_1)(x,a) \leq (2 + \tilde{L})\|Q_2 - Q_1\|_\infty$. Hence H is Lipschitz with Lipschitz constant $2 + \tilde{L}$.

Meanwhile,

$$\begin{aligned}
&H_s(Q_1)(x,a) - H_s(Q_2)(x,a) \\
&= \frac{1}{s}\Big(\mathcal{R}_{x,a}\big(c(x,a) + \min_{a'\in\mathcal{A}} sQ_1(x,a)\big) - \mathcal{R}_{x,a}\big(c(x,a) + \min_{a'\in\mathcal{A}} sQ_2(x,a)\big) \\
&\qquad - f(sQ_1) + f(sQ_2) - sQ_1(x,a) + sQ_2(x,a)\Big) \\
&\le \frac{1}{s}(s\|Q_1 - Q_2\|_\infty + s\|Q_1 - Q_2\|_\infty + s\|Q_1 - Q_2\|_\infty) \\
&= (2 + \tilde{L})\|Q_1 - Q_2\|_\infty.
\end{aligned}$$

Similarly, we obtain $H_s(Q_2)(x,a) - H_s(Q_1)(x,a) \le (2+\tilde{L})\|Q_2 - Q_1\|_\infty$. Hence H_s is Lipschitz with Lipschitz constant $2+\tilde{L}$. Similarly, if $H_\infty(Q) := \lim_{s\to\infty} H_s(Q)$ exists, H_∞ is also Lipschitz with Lipschitz constant $2 + \tilde{L}$. □

B.2.2 Condition (ii)

We check that M_n is a martingale difference sequence that satisfies (ii).

Lemma B.6. *Under Assumption 4.1, for all $n = 0, 1, ...$, we have $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$ almost surely and there exists some $C_0 < \infty$ such that for any initial condition $Q_0 \in \mathcal{L}(\mathcal{K})$ we have*

$$\mathbb{E}[\|M_{n+1}\|_\infty^2|\mathcal{F}_n] \le C_0(1 + \|Q_n\|_\infty^2), \quad n \ge 0, \quad \textit{a.s.}$$

Proof. By Assumption 4.1, it is easy to see that

$$\mathbb{E}[\hat{\mathcal{H}}(Q_n)|\mathcal{F}_n] = \mathcal{H}(Q_n), \text{ a.s}, \quad \mathrm{Var}[\hat{\mathcal{H}}(Q_n)(x,a)|\mathcal{F}_n] \le C(1 + \|Q_n\|_\infty^2), \ \forall (x,a) \in \mathcal{K}, \text{a.s.},$$

for some constant $C > 0$. Then by definition,

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = \mathbb{E}[\hat{\mathcal{H}}(Q_n) - \mathcal{H}(Q_n)|\mathcal{F}_n] = \mathbb{E}[\hat{\mathcal{H}}(Q_n)|\mathcal{F}_n] - \mathcal{H}(Q_n) = 0, \quad \text{a.s.}$$

Meanwhile, for any $(x,a) \in \mathcal{K}$, from the definition of variance, we have for all $(x,a) \in \mathcal{K}$ that almost surely

$$\begin{aligned}
\mathbb{E}[(M_{n+1}(x,a))^2|\mathcal{F}_n] &= \mathbb{E}[(\hat{\mathcal{H}}(Q_n)(x,a) - \mathcal{H}(Q_n)(x,a))^2|\mathcal{F}_n] \\
&= \mathbb{E}[(\hat{\mathcal{H}}(Q_n)(x,a) - \mathbb{E}[\hat{\mathcal{H}}(Q_n)(x,a)|\mathcal{F}_n])^2|\mathcal{F}_n] \\
&= \mathrm{Var}[\hat{\mathcal{H}}(Q_n)(x,a)|\mathcal{F}_n] \\
&\le C(1 + \|Q_n\|_\infty^2).
\end{aligned}$$

This implies that

$$\mathbb{E}[\|M_{n+1}\|_2^2|\mathcal{F}_n] \le |\mathcal{K}|C(1 + \|Q_n\|_\infty^2), \quad \text{a.s.}$$

From the L_p-norm inequality, we have

$$\mathbb{E}[\|M_{n+1}\|_\infty^2|\mathcal{F}_n] \le \mathbb{E}[\|M_{n+1}\|_2^2|\mathcal{F}_n] \le |\mathcal{K}|C(1 + \|Q_n\|_\infty^2) =: C_0(1 + \|Q_n\|_\infty^2), \quad \text{a.s.},$$

for some constant $C_0 < \infty$. This completes the proof. □

B.2.3 Condition (v)

In this subsection, we always assume Assumption 3.3 holds for all the lemmas. To prove (v), we need to analyze the stability of the equilibrium of ODE (B.6), which is quite difficult as there exists nonlinear terms $\mathcal{R}_{x,a}$ and f. Following Abounadi et al. (2001), we also analyze the behavior of an ODE where we replace $f(Q)$ with a constant g^*:

$$\dot{q}_t = \tilde{H}(q_t), \tag{B.9}$$

where $\tilde{H}(q) := \mathcal{H}(q) - g^* - q, \forall q \in \mathcal{L}(\mathcal{K})$. Clearly, the fixed point q^* of ODE (B.9), together with g^*, is a solution to the AROE (3.4). Hence, under Assumption 2.12, the set of fixed point of (B.9) is not empty and by lemmas B.3 and B.4, the fixed points differ by a constant. We conclude on the characteristic of the set of the equilibrium points of ODE (B.9) using the following lemma.

Lemma B.7. *The set G of equilibrium of ODE (B.9) satisfies $G = \{q : q = \bar{q}^* + r, r \in \mathbb{R}\}$, where $\bar{q}^*$ is the only solution to the AROE (3.4) that satisfies $f(\bar{q}^*) = g^*$.*

Proof. It is evident that any solution q^* to the AROE (3.4) satisfies $0 = \tilde{H}(q^*)$, indicating that q^* is an equilibrium point for the ODE (B.9). According to Theorem 2.13, under Assumption 2.12, the set of equilibrium points is non-empty. Similarly, for any equilibrium point $\tilde{q} \in G$, we have $0 = \tilde{H}(\tilde{q})$, which satisfies the AROE (3.4), implying that $\tilde{q}$ is a solution to the AROE (3.4), i.e., a span-seminorm fixed point of $\mathcal{H}$. By Lemma B.3, each fixed point differs only by a constant. Therefore, we conclude that $G = \{q : q = \tilde{q} + r, r \in \mathbb{R}\}$ for some equilibrium point $\tilde{q}$.

Now suppose $f(\tilde{q}) = m$ for some constant m. Then by Assumption 3.3, $f(\tilde{q} + g^* - m) = f(\tilde{q}) + g^* - m = g^*$. Hence, we can always find a $\bar{q}^* := \tilde{q} + g^* - m$ satisfying $f(\bar{q}^*) = g^*$. By definition, $\bar{q}^* \in G$, therefore is a solution to the AROE (3.4). □

For notation simplicity, define

$$\bar{\mathcal{H}}(q) := \mathcal{H}(q) - f(q), \quad \tilde{\mathcal{H}}(q) := \mathcal{H}(q) - g^*, \quad q \in \mathcal{L}(\mathcal{K}).$$

Then for the two ODEs (B.6) and (B.9), we have

$$\dot{p}_t = H(p_t) = \bar{\mathcal{H}}(p_t) - p_t, \quad \dot{q}_t = \tilde{H}(q_t) = \tilde{\mathcal{H}}(q_t) - q_t.$$

Since $\mathcal{H}$ is non-expansive (see Lemma B.4), $\tilde{\mathcal{H}}$ is also non-expansive. From Theorem 3.1 of Borkar and Soumyanatha (1997) (also see Lemma 3.1 of Abounadi et al. (2001)), the ODE (B.9) has a unique trajectory that may depend on the initial point q_0 and converges to some equilibrium point q^*. We conclude as the following lemma.

Lemma B.8. *Let q_t be a solution of ODE (B.9). Then $q_t \to q^*$ as $t \to \infty$ for some equilibrium point q^* of (B.9) that may depend on Q_0. Moreover, $q^* = \bar{q}^* + \bar{r}$ for some $\bar{r} \in \mathbb{R}$, where $\bar{q}^*$ is defined in Lemma B.7.*

Proof. The convergence result follows from Abounadi et al. (2001) Lemma 3.1. Then the result follows by applying Lemma B.7. □

Following the property of f, we can show that the equilibrium point of ODE (B.6) is unique and is also included in the set of equilibrium points of ODE (B.9).

Lemma B.9. *The point $\bar{q}^*$ is the unique equilibrium point of ODE (B.6).*

Proof. Based on Lemma B.7, since $f(\bar{q}^*) = g^*$, we have $\bar{\mathcal{H}}(\bar{q}^*) = \tilde{\mathcal{H}}(\bar{q}^*) = \bar{q}^*$, which means that $\bar{q}^*$ is an equilibrium point for (B.6). Conversely, if there exists some $\tilde{p}$ such that $\bar{\mathcal{H}}\tilde{p} = \tilde{p}$, by definition, the solution of the above equation satisfies the AROE (3.4). By Theorem 3.4, $f(\tilde{p}) = g^*$. Therefore, we have $\tilde{p} = \bar{\mathcal{H}}(\tilde{p}) = \tilde{\mathcal{H}}(\tilde{p})$, which means $\tilde{p}$ is also an equilibrium for (B.9). By Lemma B.7, $\tilde{p} = \bar{q}^* + \bar{r}$ for some $\bar{r} \in \mathbb{R}$. Then we have $g^* = f(\tilde{p}) = f(\bar{q}^* + r) = g^* + r$. This implies that $r = 0$. Therefore, $\bar{q}^*$ is a unique equilibrium point for ODE (B.6). □

The next result shows that the trajectory of ODE (B.6) and ODE (B.9) differs only by a constant function.

Lemma B.10. *Let p_t and q_t be the solutions to the ODEs (B.6) and (B.9), with the same initial value $p_0(x, a) = q_0(x, a) = Q_0(x, a)$. Then we have*

$$p_t(x, a) = q_t(x, a) + r_t, \quad \forall (x, a) \in \mathcal{K},$$

where r_t is a scalar function satisfying

$$\dot{r}_t = -r_t + g^* - f(q_t).$$

Proof. Notice that $\bar{\mathcal{H}}(Q) = \tilde{\mathcal{H}}(Q) + (g^* - f(Q))$. Then from the variation of constants formula, we have that

$$p_t(x,a) = q_0(x,a)\mathrm{e}^{-t} + \int_0^t \mathrm{e}^{-(t-s)}\tilde{\mathcal{H}}(p_s(x,a))ds + \int_0^t \mathrm{e}^{-(t-s)}(g^* - f(p_s)ds,$$
$$q_t(x,a) = q_0(x,a)\mathrm{e}^{-t} + \int_0^t \mathrm{e}^{-(t-s)}\tilde{\mathcal{H}}(q_s(x,a))ds.$$

The maximal and minimal components of $p_t - q_t$ can be bounded by

$$\begin{aligned}\max_{(x,a)\in\mathcal{K}}\{p_t(x,a) - q_t(x,a)\} &\le \int_0^t \mathrm{e}^{-(t-s)} \max_{(x,a)\in\mathcal{K}}\{\tilde{\mathcal{H}}(p_s)(x,a) - \tilde{\mathcal{H}}(q_s)(x,a)\}ds \\ &\quad + \int_0^t \mathrm{e}^{-(t-s)}(g^* - f(p_s))ds, \\ \min_{(x,a)\in\mathcal{K}}\{p_t(x,a) - q_t(x,a)\} &\ge \int_0^t \mathrm{e}^{-(t-s)} \min_{(x,a)\in\mathcal{K}}\{\tilde{\mathcal{H}}(p_s)(x,a) - \tilde{\mathcal{H}}(q_s)(x,a)\}ds \\ &\quad + \int_0^t \mathrm{e}^{-(t-s)}(g^* - f(p_s))ds.\end{aligned}$$

Hence, we have

$$\begin{aligned}\|p_t - q_t\|_{sp} &\le \int_0^t \mathrm{e}^{-(t-s)}\|\tilde{\mathcal{H}}(p_s) - \tilde{\mathcal{H}}(q_s)\|_{sp}ds \\ &\le \int_0^t \mathrm{e}^{-(t-s)}\|p_s - q_s\|_{sp}ds.\end{aligned}$$

The inequality is from the fact that $\mathcal{H}$ is span-seminorm contractive (see Lemma B.4). By the Gronwall inequality, we have $\|p_t - q_t\|_{sp} = 0$. This implies that there exists some scalar function r_t such that $p_t(x,a) = q_t(x,a) + r(t)$ for all $(x,a) \in \mathcal{K}$, with $r(0) = 0$.

Since $\tilde{\mathcal{H}}(p_t) = \tilde{\mathcal{H}}(q_t + r_t) = \tilde{\mathcal{H}}(q_t) + r_t$ and $f(p_t) = f(q_t + r_t) = f(q_t) + r_t$. Then the differential of r_t is

$$\dot{r}_t e = \dot{p}_t - \dot{q}_t = \tilde{\mathcal{H}}(p_t) + g^* - f(p_t) - p_t - \tilde{\mathcal{H}}(q_t) + q_t = (-r_t + g^* - f(q_t))e.$$

This completes the proof. □

The following lemma shows that $\bar{q}^*$ is the unique globally asymptotically stable equilibrium point of ODE (B.6).

Lemma B.11. *$\bar{q}^*$ is the unique globally asymptotically stable equilibrium point of ODE (B.6).*

Proof. From Lemma B.10, by the variation of constant formula, we have $r_t = \int_0^t \mathrm{e}^{-(t-s)}(g^* - f(q_t))ds$. By Lemma B.8, we have $q_t \to q^* \in G$. Then we have $r_t \to g^* - f(q^*)$ so that $p_t \to q^* + (g^* - f(q^*))$, which must coincide with $\bar{q}^*$ since by Lemma B.9, it is the only equilibrium point of ODE (B.6). Next we show the Lyapunov stability of $\bar{q}^*$. Notice that

$$\begin{aligned}\|p_t - \bar{q}^*\|_\infty &\le \|q_t - \bar{q}^*\|_\infty + |r_t| \\ &\le \|q_0 - \bar{q}^*\|_\infty + \int_0^t \mathrm{e}^{-(t-s)}|g^* - f(q_s)|ds \\ &\le \|p_0 - \bar{q}^*\|_\infty + \int_0^t \mathrm{e}^{-(t-s)}|f(\bar{q}^*) - f(q_s)|ds \\ &\le (1 + \tilde{L}(1 - \mathrm{e}^{-t}))\|p_0 - \bar{q}^*\|_\infty.\end{aligned}$$

Hence for any fixed $t > 0$ and any $\epsilon > 0$, we can always make $\|p_0 - \bar{q}^*\|_\infty < \delta$ where $\delta < \frac{\epsilon}{1+\tilde{L}(1-\mathrm{e}^{-t})}$ to guarantee that $\|p_t - \bar{q}^*\|_\infty < \epsilon$. The Lyapunov stability holds, completing the proof. □

B.2.4 Condition (vi)

We now look at condition (vi).

Lemma B.12. *Under assumptions 2.12 and 4.2 on $\mathcal{R}$, the risk map $\mathcal{R}^\infty$ also satisfies Assumption 2.12.*

Proof. Following Assumption 2.12, there exists a coherent risk measure ν and $\bar{\alpha} \in (0,1)$ such that for any $v \geq v' \in \mathcal{L}(\mathcal{X})$, we have

$$\min_{(x,a)\in\mathcal{K}} \{\mathcal{R}(v|x,a) - \bar{\alpha}\nu(v) - \mathcal{R}(v'|x,a) + \bar{\alpha}\nu(v')\} \geq 0.$$

Substituting v and v' with sv and sv' respectively, where $s > 0$, and then dividing both sides by s, we obtain

$$\frac{1}{s} \min_{(x,a)\in\mathcal{K}} \{\mathcal{R}(sv|x,a) - \bar{\alpha}\nu(sv) - \mathcal{R}(sv'|x,a) + \bar{\alpha}\nu(sv')\} \geq 0.$$

Since ν is coherent and by Assumption 4.2, $\lim_{s\to\infty} \frac{1}{s}\mathcal{R}_{x,a}(sv) = \mathcal{R}^\infty_{s,a}(v)$, taking the limit, we obtain

$$\begin{aligned} 0 &\leq \lim_{s\to\infty} \min_{(x,a)\in\mathcal{K}} \left\{ \frac{1}{s}\mathcal{R}(sv|x,a) - \bar{\alpha}\nu(v) - \frac{1}{s}\mathcal{R}(sv'|x,a) + \bar{\alpha}\nu(v') \right\} \\ &= \min_{(x,a)\in\mathcal{K}} \{\mathcal{R}^\infty_{x,a}(v) - \bar{\alpha}\nu(v) - \mathcal{R}^\infty_{x,a}(v') + \bar{\alpha}\nu(v')\}. \end{aligned}$$

This implies that $\mathcal{R}^\infty$ satisfies Assumption 2.12 with coherent risk measure ν and $\bar{\alpha} \in (0,1)$. □

Lemma B.13. *Under assumptions 2.12, 3.3, 4.2 and 4.3, the limit $H_\infty(q) := \lim_{s\to\infty} H_s(q)$ exists for all $Q \in \mathcal{L}(\mathcal{K})$, and convergence is uniform on any compact sets. Furthermore, the ODE (B.8) has the origin as a unique globally asymptotically stable equilibrium.*

Proof. Under Assumption 4.3, we have

$$\begin{aligned} H_s(q)(x,a) &= \frac{1}{s}\left\{ \mathcal{R}_{x,a}\left(c(x,a) + \min_{a'\in\mathcal{A}} sq(\cdot,a') \right) - f(sq) - sq(x,a) \right\} \\ &= \frac{c(x,a)}{s} + \frac{1}{s}\mathcal{R}_{x,a}\left(s\min_{a'\in\mathcal{A}} q(\cdot,a') \right) - f(q) - q(x,a). \end{aligned}$$

Hence, by Assumption 4.2:

$$H_\infty(q)(x,a) := \lim_{s\to\infty} H_s(q)(x,a) = \mathcal{R}^\infty_{x,a}\left(\min_{a'\in\mathcal{A}} q(\cdot,a') \right) - f(q) - q(x,a).$$

Since $\mathcal{R}^\infty_{x,a}$ exists and the convergence is uniform on all compact subsets of $\mathcal{L}(\mathcal{X})$, the first part follows. Clearly, the origin is an equilibrium point of the ODE (B.8). Following Lemma B.12, $\mathcal{R}^\infty_{x,a}$ satisfies Assumption 2.12, then by Lemma B.11, the origin is also the globally asymptotically stable equilibrium for ODE (B.8). Hence condition (vi) holds. □

B.2.5 Convergence of RVI Q-learning

We are now ready to prove Theorem 4.5.

Proof of Theorem 4.5. The almost sure boundedness condition (iv) is derived from Theorem 2.1 of Borkar and Meyn (2000), which necessitates verifying conditions (i), (vi), (ii) and (iii). These conditions are confirmed using Lemmas B.5, B.13, B.6, and Assumption 4.4. The convergence and optimality of the stochastic approximation then follow from Theorem 2.2 of Borkar and Meyn (2000), where conditions (i), (ii), (iii), and (v) are validated through Lemmas B.5, B.6, B.11, and Assumption 4.4.

Regarding the almost sure convergence of $\pi_n \to \pi^*$, one can first observe that $Q^* : \mathcal{K} \to \mathcal{C} \subset \mathbb{R}$, for some discrete set $\mathcal{C}$ with $|\mathcal{C}| \leq |\mathcal{K}|$. Letting

$$\epsilon = \min_{(x,a),(x',a')\in\mathcal{K}: Q^*(x,a)\neq Q^*(x',a')} |Q^*(x,a) - Q^*(x',a')| > 0,$$

the almost sure convergence of $Q_n \to Q^*$ implies that there is a probability one set of trajectories $\mathcal{Q}$, with each trajectory $\{\bar{Q}_n\} \in \mathcal{Q}$ having the property that there exists an $N \geq 0$ such that $\|\bar{Q}_n - Q^*\|_\infty \leq \epsilon/2$ for all $n \geq N$. This implies that for any $n \geq N$,

$$Q^*(x,a) > Q^*(x',a') \Longrightarrow \bar{Q}_n(x,a) > \bar{Q}_n(x',a'), \quad \forall (x,a),(x',a') \in \mathcal{K}.$$

We can therefore conclude that for all $x \in \mathcal{X}$ and for all $n \geq N$, we have

$$\operatorname*{argmin}_{a\in\mathcal{A}} \bar{Q}_n(x,a) \subseteq \operatorname*{argmin}_{a\in\mathcal{A}} Q^*(x,a), \quad \forall x \in \mathcal{X}.$$

Thus the policy π_n converges to some π^* for all $\{\bar{Q}_n\} \in \mathcal{Q}$ almost surely. □

B.3 Proof of Theorem 4.7

For notation simplicity, we write (4.2) as

$$\hat{\mathcal{H}}(q) = \mathcal{H}_{\hat{P}^1_{N+1}}(q) + \frac{\Delta_N(q)}{p_N},$$

where

$$\Delta_N(q) := \mathcal{H}_{\hat{P}_{N+1}}(q) - \frac{1}{2}\left(\mathcal{H}_{\hat{P}^E_{N+1}}(q) + \mathcal{H}_{\hat{P}^O_{N+1}}(q)\right), \quad \forall q \in \mathcal{L}(\mathcal{K}).$$

To prove Theorem 4.7, we invoke the concentration results under the 1-Wasserstein distance from Fournier and Guillin (2015). The 1-Wasserstein distance between two probability measures μ and ν on $\mathbb{R}$ is defined as

$$d_W(\mu,\nu) := \inf_{\psi\in\Psi(\mu,\nu)} \int |x-y|\psi(dx,dy),$$

where $\Psi(\mu,\nu)$ is the set of all joint probability distributions $\psi(x,y)$ with marginals μ and ν.

Lemma B.14 (Concentration inequalities)**.** *Given any $v \in \mathcal{L}(\mathcal{X})$ and a $\bar{p} \in \mathscr{P}(\mathcal{X})$, let $\hat{p}^k$ be the empirical distribution from k realizations $\{x'_1, x'_2, \ldots, x'_k\}$ drawn i.i.d. from $\bar{p}$. Then,*

$$\mathbb{E}\left[d_W\left(\sum_{x'\in\mathcal{X}} \hat{p}^k(x')\delta_{v(x')}, \sum_{x'\in\mathcal{X}} \bar{p}(x')\delta_{v(x')}\right)\right] \leq \mathfrak{C}_1\|v\|_\infty k^{-1/2}$$

and

$$\mathbb{E}\left[d_W\left(\sum_{x'\in\mathcal{X}} \hat{p}^k(x')\delta_{v(x')}, \sum_{x'\in\mathcal{X}} \bar{p}(x')\delta_{v(x')}\right)^2\right] \leq \mathfrak{C}_2\|v\|^2_\infty k^{-1},$$

for some constant $\mathfrak{C}_1, \mathfrak{C}_2 > 0$ independent of v, $\bar{p}$, k and $\delta_{v(x)}$ is the Dirac measure of $v(x)$.

Proof. To simplify notations, we use $\hat{p}(\cdot)$ to denote $\hat{p}^k(\cdot)$. The first bound follows from Theorem 1 of Fournier and Guillin (2015). Namely, there exists a $\bar{\mathfrak{C}}_1 > 0$ such that:

$$\begin{aligned}\mathbb{E}\left[d_W\left(\sum_{x'\in\mathcal{X}} \hat{p}(x')\delta_{v(x')}, \sum_{x'\in\mathcal{X}} \bar{p}(x')\delta_{v(x')}\right)\right] &\leq 2\bar{\mathfrak{C}}_1\left(\sum_{x'\in\mathcal{X}} v(x')^2\bar{p}(x')\right)^{1/2}/k^{1/2} \\ &\leq 2\bar{\mathfrak{C}}_1\|v\|_\infty k^{-1/2}.\end{aligned}$$

By Lemma 5 and Proposition 10 in Fournier and Guillin (2015), we have that for all $w \in \mathcal{L}(\mathcal{X})$ with $\|w\|_\infty < 1$, there exists constants $\bar{\mathfrak{C}}_2, \bar{\mathfrak{C}}_3 > 0$ such that for all $\lambda \geq 0$:

$$\mathbb{P}\left(d_W\left(\sum_{x'\in\mathcal{X}} \hat{p}(x')\delta_{w(x')}, \sum_{x'\in\mathcal{X}} \bar{p}(x')\delta_{w(x')}\right) \geq \lambda\right) \leq \bar{\mathfrak{C}}_2\exp(-\bar{\mathfrak{C}}_3 k\lambda^2),$$

given the fact that $\bar{p}$ is a distribution on a finite set. We can thus derive that:

$$\begin{aligned}
&\mathbb{E}\left[d_W\left(\sum_{x'\in\mathcal{X}}\hat{p}(x')\delta_{v(x')},\sum_{x'\in\mathcal{X}}\bar{p}(x')\delta_{v(x')}\right)^2\right]\\
&\leq \mathbb{E}\left[4\|v\|_\infty^2 d_W\left(\sum_{x'\in\mathcal{X}}\hat{p}(x')\delta_{w(x')},\sum_{x'\in\mathcal{X}}\bar{p}(x')\delta_{w(x')}\right)^2\right]\\
&= 4\|v\|_\infty^2\int_0^\infty \mathbb{P}\left(d_W\left(\sum_{x'\in\mathcal{X}}\hat{p}(x')\delta_{w(x')},\sum_{x'\in\mathcal{X}}\bar{p}(x')\delta_{w(x')}\right)^2\geq\lambda\right)d\lambda\\
&\leq 4\|v\|_\infty^2\int_0^\infty \bar{\mathfrak{C}}_2\exp(-\bar{\mathfrak{C}}_3 k\lambda)d\lambda = \frac{4\|v\|_\infty^2\bar{\mathfrak{C}}_2}{\bar{\mathfrak{C}}_3 k},
\end{aligned}$$

where $w := (1/2)v/\|v\|_\infty$ is such that $\|w\|_\infty \leq 1/2 < 1$. □

Proof of Theorem 4.7. Assumption 4.6 implies that $\mathcal{R}_{x,a}$ is law invariant, so that $\mathcal{R}_{x,a}(v) = \varrho(\sum_{x'\in\mathcal{X}} P(x'|x,a)\delta_{v(x')})$, with $\varrho_{x,a}$ as the distribution-based risk measure associated to $\mathcal{R}_{x,a}$. Let $\hat{\mathcal{R}}^k_{x,a}$ capture the empirical risk map that employs the same distribution-based risk measure $\varrho_{x,a}$ of $\mathcal{R}_{x,a}$, but on $\sum_{x'\in\mathcal{X}}\hat{P}_k(x'|x,a)\delta_{v(x')}$ instead of $\sum_{x'\in\mathcal{X}}P(x'|x,a)\delta_{v(x')}$, where $\hat{P}_k$ stands for the empirical distribution using 2^k number of samples. We start by establishing two important properties of how $\mathcal{H}_{\hat{P}_k}(q)$ differs from $\mathcal{H}(q)$.

The first property consists of a bound on the expected absolute difference between $\mathcal{H}_{\hat{P}_k}(q)$ and $\mathcal{H}(q)$, where the expectation is taken with respect to the sampling process. Namely, for all $q \in \mathcal{L}(\mathcal{K})$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathcal{H}_{\hat{P}_k}(q)(x,a) - \mathcal{H}(q)(x,a)\right|\right]\\
&= \mathbb{E}\left[\left|\hat{\mathcal{R}}^k_{x,a}\left(c(x,a)+\max_{a'\in\mathcal{A}}q(\cdot,a')\right) - \mathcal{R}_{x,a}\left(c(x,a)+\max_{a'\in\mathcal{A}}q(\cdot,a')\right)\right|\right]\\
&\leq \mathbb{E}\left[\mathfrak{L}d_W\left(\sum_{x'\in\mathcal{X}}\hat{P}_k(x'|x,a)\delta_{v(x')},\sum_{x'\in\mathcal{X}}P(x'|x,a)\delta_{v(x')}\right)\right]\\
&\leq \mathfrak{L}\mathfrak{C}_1\|c(x,a)+\max_{a'\in\mathcal{A}}q(\cdot,a')\|_\infty 2^{-k/2} \leq \mathfrak{L}\mathfrak{C}_1(2\bar{C}+\|q\|_\infty)2^{-k/2},
\end{aligned}$$

where $v(x') := c(x,a)+\max_{a'\in\mathcal{A}}Q(x',a')$. The first inequality follows from Assumption 4.6, and the second one is from Lemma B.14.

Following a similar procedure, we have the second one, which bounds the expected square difference:

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{H}_{\hat{P}_k}(q)(x,a) - \mathcal{H}(q)(x,a)\right)^2\right]\\
&= \mathbb{E}\left[\left(\hat{\mathcal{R}}^k_{x,a}(c(x,a)+\max_{a'\in\mathcal{A}}q(\cdot,a')) - \mathcal{R}_{x,a}(c(x,a)+\max_{a'\in\mathcal{A}}q(\cdot,a'))\right)^2\right]\\
&\leq \mathbb{E}\left[\mathfrak{L}^2 d_W\left(\sum_{x'\in\mathcal{X}}\hat{P}_k(x'|x,a)\delta_{v(x')},\sum_{x'\in\mathcal{X}}P(x'|x,a)\delta_{v(x')}\right)^2\right]\\
&\leq \mathfrak{L}^2\mathfrak{C}_2\|c(x,a)+\max_{a'\in\mathcal{A}}q(\cdot,a')\|_\infty 2^{-k} \leq \mathfrak{L}^2\mathfrak{C}_2(2\bar{C}+\|q\|_\infty)2^{-k},
\end{aligned}$$

where the second inequality is from Lemma B.14.

We are now ready to show that $\mathbb{E}[\hat{\mathcal{H}}(q)] = \mathcal{H}(q)$, for all $q \in \mathcal{L}(\mathcal{K})$, which goes as

$$\begin{aligned}
\mathbb{E}[\hat{\mathcal{H}}(q)] &= \mathbb{E}\left[\mathcal{H}_{\hat{P}^1_{N+1}}(q) + \frac{\Delta_N(q)}{p_N}\right] \\
&= \mathbb{E}[\mathcal{H}_{\hat{P}^1_{N+1}}(q)] + \sum_{k=0}^{\infty} \mathbb{P}(N=k)\mathbb{E}\left[\frac{\Delta_k(q)}{p_k}\big|N=k\right] \\
&= \mathbb{E}[\mathcal{H}_{\hat{P}^1_1}(q)] + \sum_{k=0}^{\infty} \mathbb{E}[\Delta_k(q)] \\
&= \mathbb{E}[\mathcal{H}_{\hat{P}^1_1}(q)] + \sum_{k=0}^{\infty} \mathbb{E}\left[\mathcal{H}_{\hat{P}_{k+1}}(q) - \frac{1}{2}\left(\mathcal{H}_{\hat{P}^E_{k+1}}(q) + \mathcal{H}_{\hat{P}^O_{k+1}}(q)\right)\right] \\
&= \mathbb{E}[\mathcal{H}_{\hat{P}^1_1}(q)] + \sum_{k=0}^{\infty} \left(\mathbb{E}[\mathcal{H}_{\hat{P}_{k+1}}(q)] - \frac{1}{2}\left(\mathbb{E}[\mathcal{H}_{\hat{P}^E_{k+1}}(q)] + \mathbb{E}[\mathcal{H}_{\hat{P}^O_{k+1}}(q)]\right)\right) \\
&= \mathbb{E}[\mathcal{H}_{\hat{P}^1_1}(q)] + \sum_{k=0}^{\infty} \left(\mathbb{E}[\mathcal{H}_{\hat{P}_{k+1}}(q)] - \mathbb{E}[\mathcal{H}_{\hat{P}^E_k}(q)]\right) \\
&= \lim_{k\to\infty} \mathbb{E}[\mathcal{H}_{\hat{P}_k}(q)] = \mathcal{H}(q),
\end{aligned}$$

where the limit is known to exist and identified as $\mathcal{H}(q)$ since for all $(x,a) \in \mathcal{K}$ we have

$$\begin{aligned}
\left|\mathbb{E}[\mathcal{H}_{\hat{P}_k}(q)(x,a)] - \mathcal{H}(q)(x,a)\right| &\le \mathbb{E}[|\mathcal{H}_{\hat{P}_k}(q)(x,a) - \mathcal{H}(q)(x,a)|] \\
&\le \mathfrak{L}\mathfrak{C}_1(2\bar{C} + \|q\|_\infty)2^{-k/2},
\end{aligned}$$

thus implying that $\|\mathbb{E}[\mathcal{H}_{\hat{P}_k}(q)] - \mathcal{H}(q)\|_\infty \to 0$ as $k \to \infty$.

We turn to bounding $\mathrm{Var}[\hat{\mathcal{H}}(q)(x,a)]$. Since for all $(x,a) \in \mathcal{K}$, we have

$$\mathrm{Var}[\hat{\mathcal{H}}(q)(x,a)] = \mathbb{E}[(\hat{\mathcal{H}}(q)(x,a))^2] - (\mathbb{E}[\hat{\mathcal{H}}(q)(x,a)])^2 = \mathbb{E}[(\hat{\mathcal{H}}(q)(x,a))^2] - (\mathcal{H}(q)(x,a))^2,$$

and it is known that $\|(\mathcal{H}(q))^2\|_\infty \le (2\bar{C} + \|q\|_\infty)^2 \le 8\bar{C}^2 + 2\|q\|^2_\infty$. The remaining question is to bound $\mathbb{E}[(\hat{\mathcal{H}}(q)(x,a))^2]$.

We first give a bound on $\mathbb{E}[(\Delta_k(q)(x,a))^2]$. Notice that

$$\begin{aligned}
&\mathbb{E}[\Delta_k(q)(x,a)^2] = \mathbb{E}\left[\left(\mathcal{H}_{\hat{P}_{k+1}}(q)(x,a) - \frac{1}{2}\left(\mathcal{H}_{\hat{P}^E_{k+1}}(q)(x,a) + \mathcal{H}_{\hat{P}^O_{k+1}}(q)(x,a)\right)\right)^2\right] \\
&= \mathbb{E}\left[\left((\mathcal{H}_{\hat{P}_{k+1}}(q)(x,a) - \mathcal{H}(q)(x,a)) - \frac{1}{2}\left(\mathcal{H}_{\hat{P}^E_{k+1}}(q)(x,a) + \mathcal{H}_{\hat{P}^O_{k+1}}(q)(x,a) - 2\mathcal{H}(q)(x,a)\right)\right)^2\right] \\
&\le \mathbb{E}[2(\mathcal{H}_{\hat{P}_{k+1}}(q)(x,a) - \mathcal{H}(q)(x,a))^2 + (\mathcal{H}_{\hat{P}^E_{k+1}}(q)(x,a) - \mathcal{H}(Q_n)(x,a))^2 \\
&\qquad + (\mathcal{H}_{\hat{P}^O_{k+1}}(q)(x,a) - \mathcal{H}(q)(x,a))^2] \\
&= 2\mathbb{E}[(\mathcal{H}_{\hat{P}_{k+1}}(q)(x,a) - \mathcal{H}(q)(x,a))^2] + 2\mathbb{E}[(\mathcal{H}_{\hat{P}_k}(q)(x,a) - \mathcal{H}(q)(x,a))^2] \\
&\le 2\mathfrak{L}^2\mathfrak{C}_2(2\bar{C} + \|q\|_\infty)^2 2^{-k-1} + 2\mathfrak{L}^2\mathfrak{C}_2(2\bar{C} + \|q\|_\infty)^2 2^{-k} \\
&= 3\mathfrak{L}^2\mathfrak{C}_2(2\bar{C} + \|q\|_\infty)^2 2^{-k}.
\end{aligned}$$

Now we are ready to derive the bound for $\mathbb{E}[(\hat{\mathcal{H}}(q)(x,a))^2]$, for any fixed $(x,a) \in \mathcal{K}$. Namely, from definition, we have

$$\begin{aligned}
\mathbb{E}[(\hat{\mathcal{H}}(q)(x,a))^2] &\leq 2\mathbb{E}[(\mathcal{H}_{\hat{P}^1_{N+1}}(q)(x,a))^2] + 2\mathbb{E}\left[\left(\frac{\Delta_N(q)(x,a)}{p_N}\right)^2\right] \\
&= 2(2\bar{C} + \|q\|_\infty)^2 + 2\mathbb{E}\left[\sum_{k=0}^{\infty} p_k \mathbb{E}\left[\left(\frac{\Delta_N(q)(x,a)}{p_N}\right)^2 \middle| N = k\right]\right] \\
&= 2(2\bar{C} + \|Q_n\|_\infty)^2 + 2\sum_{k=0}^{\infty} \frac{1}{p_k}\mathbb{E}[\Delta_k(q)(x,a)^2] \\
&\leq 2(2\bar{C} + \|q\|_\infty)^2 + 2\sum_{k=0}^{\infty} \frac{1}{p_k} 3\mathfrak{L}^2\mathfrak{C}_2(2\bar{C} + \|q\|_\infty)^2 2^{-k} \\
&= 2(2\bar{C} + \|q\|_\infty)^2 + 6\mathfrak{L}^2\mathfrak{C}_2(2\bar{C} + \|q\|_\infty)^2 r^{-1}\sum_{k=0}^{\infty}(2(1-r))^{-k} \\
&= 2(2\bar{C} + \|q\|_\infty)^2 + 6\mathfrak{L}^2\mathfrak{C}_2(2\bar{C} + \|q\|_\infty)^2 r^{-1}(1-(2(1-r))^{-1})^{-1} \\
&= (2 + 6\mathfrak{L}^2\mathfrak{C}_2 r^{-1}(1-(2(1-r))^{-1})^{-1})(2\bar{C} + \|q\|_\infty)^2 \\
&\leq (2 + 6\mathfrak{L}^2\mathfrak{C}_2 r^{-1}(1-(2(1-r))^{-1})^{-1})(8\bar{C}^2 + 2\|q\|_\infty^2),
\end{aligned}$$

where to ensure $\sum_{k=0}^{\infty} 2(1-r)^{-k}$ is finite, we require $r \in (0, 1/2)$. This implies that there exists a uniform bound $C > 0$ such that $\mathrm{Var}[\hat{\mathcal{H}}(q)] = \mathbb{E}[(\hat{\mathcal{H}}(q)(x,a))^2] - (\mathcal{H}(q)(x,a))^2 \leq C(1 + \|q\|_\infty^2)$. This completes the proof. □

B.4 Proof of Theorem 4.10

We impose the following general assumption on the convexity of the loss function ℓ.

Assumption B.15 (Convexity)**.** *The loss function $\ell(x)$ is either convex or concave on $x \geq 0$ and either convex or concave on $x \leq 0$.*

We prove Theorem 4.10 by proving lemmas B.16, B.21 and B.22.

Lemma B.16. *A risk map $\mathcal{R}$ that employs a UBSR measure satisfies assumptions 2.12, 4.2 and 4.6, if the Markov chain satisfies Assumption 4.8 and the loss function satisfies assumptions 4.9 and B.15.*

The following property of UBSR is useful as it establishes a connection between the UBSR measure and the expected utility.

Lemma B.17 (Proposition 4.113, Föllmer and Schied (2016))**.** *Given some random variable v and some $m^* \in \mathbb{R}$, the following statements are equivalent: (i) $\mathrm{SR}(v) = m^*$; (ii) $\mathbb{E}[\ell(v - m^*)] = 0$.*

Lemma B.18. *Under assumptions 4.8, 4.9, the UBSR satisfies Assumption 2.12.*

Proof. By Assumption 4.8, there exists a state $\bar{x} \in \mathcal{X}$ such that $P(\bar{x}|x,a) > 0$ for all $(x,a) \in \mathcal{K}$. Let $\nu(v) := v(\bar{x})$, which trivially satisfies $\nu(\mathbf{0}) = 0$ and is coherent. Meanwhile, choose $0 < \bar{\alpha} < \frac{\epsilon_1}{L_1}\min_{(x,a)\in\mathcal{K}} P(\bar{x}|x,a) \in (0,\, 1)$.

Given any $v \geq v' \in \mathcal{L}(\mathcal{X})$, Lemma B.17 implies that $\sum_{y\in\mathcal{X}} P(y|x,a)\ell(v(y) - \mathrm{SR}_{x,a}(v)) = 0$, and similarly for v'. We therefore have

$$\begin{aligned}
0 &= \sum_{y\in\mathcal{X}} P(y|x,a)\left(\ell(v(y) - \mathrm{SR}_{x,a}(v)) - \ell(v'(y) - \mathrm{SR}_{x,a}(v'))\right) \\
&= \sum_{y\in\mathcal{X}} P(y|x,a)\delta(v,v',x,a,y)((v(y) - \mathrm{SR}_{x,a}(v)) - (v'(y) - \mathrm{SR}_{x,a}(v'))),
\end{aligned}$$

for some $\delta(v, v', x, a, y) \in [\epsilon_1, L_1]$ due to Assumption 4.9. Hence,

$$(\mathrm{SR}_{x,a}(v) - \mathrm{SR}_{x,a}(v')) \sum_{y \in \mathcal{X}} P(y|x,a)\delta(v, v', x, a, y)$$
$$= \sum_{y \in \mathcal{X}} P(y|x,a)\delta(v, v', x, a, y)(v(y) - v'(y)).$$

Therefore, we have

$$\mathrm{SR}_{x,a}(v) - \mathrm{SR}_{x,a}(v') \geq \inf_{\substack{\tilde{\delta} \in \mathcal{L}(\mathcal{K} \times \mathcal{X}) : \tilde{\delta}(x,a,y) \in [\epsilon_1, L_1], \\ (x,a,y) \in \mathcal{K} \times \mathcal{X}}} \frac{\sum_{y \in \mathcal{X}} P(y|x,a)\tilde{\delta}(x,a,y)(v(y) - v'(y))}{\sum_{y \in \mathcal{X}} P(y|x,a)\tilde{\delta}(x,a,y)}$$
$$\geq \frac{\epsilon_1}{L_1} \sum_{y \in \mathcal{X}} P(y|x,a)(v(y) - v'(y)),$$

given that $v \geq v'$. Hence, we have for all $(x, a) \in \mathcal{K}$,

$$\begin{aligned}
&\mathrm{SR}_{x,a}(v) - \bar{\alpha}\nu(v) - \mathrm{SR}_{x,a}(v') + \bar{\alpha}\nu(v') \\
&\geq \left(\frac{\epsilon_1}{L_1} \sum_{y \in \mathcal{X}} P(y|x,a)(v(y) - v'(y)) \right) - \bar{\alpha}\left(\nu(v) - \nu(v')\right) \\
&= \left(\frac{\epsilon_1}{L_1} \sum_{y \in \mathcal{X}} P(y|x,a)(v(y) - v'(y)) \right) - \bar{\alpha}\left(v(\bar{x}) - v'(\bar{x})\right) \\
&\geq \left(\frac{\epsilon_1}{L_1} \min_{(x,a) \in \mathcal{K}} P(\bar{x}|x,a)(v(\bar{x}) - v'(\bar{x})) \right) - \bar{\alpha}\left(v(\bar{x}) - v'(\bar{x})\right) \\
&= \left(\frac{\epsilon_1}{L_1} \min_{(x,a) \in \mathcal{K}} P(\bar{x}|x,a) - \bar{\alpha} \right)\left(v(\bar{x}) - v'(\bar{x})\right) \geq 0.
\end{aligned}$$

This proves Assumption 2.12. □

Lemma B.19. *For any loss function $\ell : \mathbb{R} \to \mathbb{R}$ with $\ell(0) = 0$ satisfying assumptions 4.9 and B.15, define $\ell_s(x) := \frac{1}{s}\ell(sx)$. We have $\ell_s(x) \to \ell_\infty(x)$ uniformly on compact sets as $s \to \infty$, for some $\ell_\infty(x)$ that satisfies assumptions 4.9 and B.15.*

Proof. By Assumption 4.9, $\ell(x)$ is strictly increasing. We consider the case where $\ell(x)$ is convex. In this setting, the right derivative is non-decreasing and bounded above hence the monotone convergence theorem implies that $\ell'_+(x) \to \overline{L} \leq L_1$ as $x \to \infty$, and similarly the left derivative $\ell'_-(x) \to \underline{L} \geq \epsilon$ as $x \to -\infty$, where we slightly abuse the notation and use $\ell'_+(x)$ and $\ell'_-(x)$ to denote the left and right derivative of ℓ at x. Without loss of generality, we assume that $\overline{L} = L_1$ and $\underline{L} = \epsilon_1$. Define $\ell_\infty(x) := L_1 x$ for $x \geq 0$ and $\ell_\infty(x) := \epsilon_1 x$ for $x < 0$. Remember that $\ell_s(x) := \frac{1}{s}\ell(sx)$. We start by demonstrating that $\lim_{s\to\infty} \ell_s(x) = \ell_\infty(x)$ pointwise and will follow with confirming uniform convergence on all compact sets using Ascoli-Arzelà theorem.

Clearly, $\lim_{s\to\infty} \ell_s(0) = 0 = \ell_\infty(0)$. Now consider the case of some $\bar{x} > 0$. Based on Assumption 4.9, for all $s > 0$, we must have $(1/s)\ell(s\bar{x}) \leq (1/s)L_1(s\bar{x} - 0) = L_1\bar{x}$. Hence we have $\lim_{s\to\infty} \ell_s(\bar{x}) \leq L_1\bar{x} = \ell_\infty(\bar{x})$. On the other hand, by convexity of $\ell(x)$ over $x \geq 0$, for any $\varepsilon > 0$, one can identify some $\hat{x} \geq 0$ such that $L_1 - \varepsilon/(2\bar{x}) \in [\ell'_-(\hat{x}), \ell'_+(\hat{x})]$ and therefore for all $x \geq 0$, $\ell(x) \geq \ell(\hat{x}) + (L_1 - \varepsilon/(2\bar{x}))(x - \hat{x})$. Thus we must have that

$$\begin{aligned}
\frac{1}{s}\ell(s\bar{x}) &\geq \frac{1}{s}(\ell(\hat{x}) + (L_1 - \varepsilon/(2\bar{x}))(s\bar{x} - \hat{x})) \\
&= \frac{1}{s}(\ell(\hat{x}) + L_1 s\bar{x} - L_1\hat{x} - (\varepsilon/(2\bar{x}))s\bar{x} + (\varepsilon/(2\bar{x}))\hat{x}) \\
&= L_1\bar{x} - (\varepsilon/(2\bar{x}))\bar{x} + \frac{1}{s}\left(\ell(\hat{x}) - (L_1 - (\varepsilon/(2\bar{x})))\hat{x}\right) \\
&\geq L_1\bar{x} - \varepsilon,
\end{aligned}$$

as long as $s \geq 2|\ell(\hat{x}) - (L_1 - (\varepsilon/(2\bar{x})))\hat{x}|/\varepsilon$. Hence $\lim_{s\to\infty} \ell_s(\bar{x}) \geq \ell_\infty(\bar{x})$. Combining the two results, we conclude that $\lim_{s\to\infty} \ell_s(x) = \ell_\infty(x)$ pointwise for $x \geq 0$.

The case where $\bar{x} < 0$ is treated similarly. Namely, letting $g(x) := -\ell(-x)$, we wish to show that $\lim_{s\to\infty}(1/s)g(s\bar{x}) = \epsilon_1\bar{x}$ for all $\bar{x} > 0$, with $g(x)$ a concave function such that $\epsilon_1 \leq (g(y) - g(x))/(y - x) \leq L_1$ and $g'_+(x) = \ell'_-(-x) \to \epsilon_1$ as $x \to \infty$. We can start with a lower bound argument $(1/s)g(s\bar{x}) \geq (1/s)\epsilon_1(s\bar{x} - 0) = \epsilon_1\bar{x}$. The upper bound is a consequence of the concavity of g, implying the existence of some $\hat{x} > 0$ such that $\epsilon_1 + (\varepsilon/(2\bar{x}))$ belongs to the interval $[g'_-(\hat{x}), g'_+(\hat{x})]$, where $g'_-(x)$ and $g'_+(x)$ denote the left and right derivatives of g at x, respectively. This observation leads to the subsequent argument:

$$\begin{aligned}\frac{1}{s}g(s\bar{x}) &\leq \frac{1}{s}(g(\hat{x}) + (\epsilon_1 + \varepsilon/(2\bar{x}))(s\bar{x} - \hat{x})) \\ &= \epsilon_1\bar{x} + (\varepsilon/(2\bar{x}))\bar{x} + \frac{1}{s}(g(\hat{x}) - (\epsilon_1 + (\varepsilon/(2\bar{x})))\hat{x}) \\ &\leq \epsilon_1\bar{x} + \varepsilon,\end{aligned}$$

for large enough s. This let us conclude that for $x < 0$ it must hold that $\lim_{s\to\infty}(1/s)\ell(sx) = \lim_{s\to\infty} -(1/s)g(-sx) = \epsilon_1 x$.

Concerning the uniform convergence on compacts set, we first observe that $\ell_s(x)$ is uniformly bounded on compact sets. Specifically, for $x \in [x_a, x_b]$, we have $|\ell_s(x)| = |\frac{1}{s}\ell(sx)| \leq \frac{1}{s}L_1(s|x|) \leq L_1 \max\{|x_a|, |x_b|\} < \infty$. Moreover, by Assumption 4.9, both ℓ and ℓ_s are Lipschitz, ensuring equicontinuity. Given the equicontinuity and uniform boundedness of $\ell_s(x)$ on compact sets, along with pointwise convergence, the Ascoli-Arzelà theorem guarantees that $\ell_s(x) \to \ell_\infty(x)$ uniformly on compact sets. To see this, the Ascoli-Arzelà theorem provides subsequential convergence $\ell_{s_i} \to \tilde{\ell}$ uniformly for some function $\tilde{\ell}$, where $s_i \uparrow \infty$ is a subsequence index. Since we also have pointwise convergence $\ell_s \to \ell_\infty$, it follows that $\tilde{\ell} = \ell_\infty$, implying uniform convergence of ℓ_{s_i} to ℓ_∞. Repeating this argument, we show that every subsequence $\{\ell_{s_i}\}$ of $\{\ell_s\}$ has a further subsequence that uniformly converges to ℓ_∞. By the subsequence principle, we conclude that ℓ_s converges to ℓ_∞ uniformly on compact sets.

The cases where $\ell(x)$ is concave or combines convex (or concave) on $x \geq 0$ and concave (or convex) on $x \leq 0$ are derived using a similar approach. Therefore, we conclude that under assumptions 4.9 and B.15, ℓ_s converges to ℓ_∞ uniformly on compact sets. □

Lemma B.20. *Under assumptions 4.9, B.15, the UBSR risk map satisfies Assumption 4.2.*

Proof. We start with establishing that for all $(x, a) \in \mathcal{K}$, all $v \in \mathcal{L}(\mathcal{X})$, and for all $s > 0$, we have

$$\begin{aligned}\frac{1}{s}\mathrm{SR}^{\ell}_{x,a}(sv) &= \frac{1}{s}\inf\{t : \mathbb{E}_{x,a}[\ell(sv - t)] \leq 0\} \\ &= \inf\{t' : \mathbb{E}_{x,a}[\ell(sv - st')] \leq 0\} \\ &= \inf\{t' : (1/s)\mathbb{E}_{x,a}[\ell(s(v - t'))] \leq 0\} \\ &= \inf\{t' : \mathbb{E}_{x,a}[\ell_s(v - t')] \leq 0\} \\ &= \mathrm{SR}^{\ell_s}_{x,a}(v),\end{aligned}$$

where $\ell_s(y) := (1/s)\ell(sy)$. Showing that the UBSR risk map is asymptotically coherent therefore reduces to showing that $\mathrm{SR}^{\ell_s}_{x,a}(v)$ converges uniformly to $\mathrm{SR}^{\ell_\infty}_{x,a}(v)$ on compact sets.

We start with pointwise convergence of $\mathrm{SR}^{\ell_s}_{x,a}(v)$ to $\mathrm{SR}^{\ell_\infty}_{x,a}(v)$ after recalling that by Lemma B.17,

$$\mathbb{E}_{x,a}[\ell_s(v - \mathrm{SR}^{\ell_s}_{x,a}(v))] = 0, \quad \mathbb{E}_{x,a}[\ell_\infty(v - \mathrm{SR}^{\ell_\infty}_{x,a}(v))] = 0, \quad v \in \mathcal{L}(\mathcal{X}), (x, a) \in \mathcal{K}.$$

Specifically, given any $\bar{v} \in \mathcal{L}(\mathcal{X})$, we can define the compact set $\mathcal{V} := [-2\|\bar{v}\|_\infty, 2\|\bar{v}\|_\infty]$. The uniform convergence of ℓ_s to ℓ_∞ on compact sets (see Lemma B.19) implies that for any arbitrarily small $\varepsilon > 0$, there exists a sufficiently large $\bar{s}$ such that we have $|\ell_\infty(y) - \ell_{\bar{s}}(y)| \leq \varepsilon\epsilon_1$, for all

$y \in \mathcal{V}$. Given that $|\mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v})| \leq \|\bar{v}\|_\infty$ for all $(x,a) \in \mathcal{K}$, denoting $\tilde{\varepsilon}(y) := \ell_\infty(y) - \ell_{\bar{s}}(y)$, we have

$$\begin{aligned}
0 = \mathbb{E}_{x,a}[\ell_\infty(\bar{v} - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}))] &= \sum_{x' \in \mathcal{X}} P(x'|x,a)[\ell_{\bar{s}}(\bar{v}(x') - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v})) + \tilde{\varepsilon}(\bar{v}(x') - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}))] \\
&\leq \sum_{x' \in \mathcal{X}} P(x'|x,a)[\ell_{\bar{s}}(\bar{v}(x') - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v})) + \varepsilon\epsilon_1] \\
&\leq \sum_{x' \in \mathcal{X}} P(x'|x,a)[\ell_{\bar{s}}(\bar{v}(x') - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) + \varepsilon)] \\
&= \mathbb{E}_{x,a}[\ell_{\bar{s}}(\bar{v} - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) + \varepsilon))],
\end{aligned}$$

where the last inequality comes from the fact that $\ell_{\bar{s}}(x+\varepsilon) - \ell_{\bar{s}}(x) = \frac{1}{s}\left(\ell(sx + s\varepsilon) - \ell(sx)\right) \geq \frac{1}{s}\epsilon_1 s\varepsilon = \epsilon_1\varepsilon$ due to Assumption 4.9. Similarly, we have

$$0 = \mathbb{E}_{x,a}[\ell_\infty(\bar{v} - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}))] \geq \mathbb{E}_{x,a}[\ell_{\bar{s}}(\bar{v} - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) - \varepsilon)].$$

This implies that

$$\mathbb{E}_{x,a}[\ell_{\bar{s}}(\bar{v} - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) - \varepsilon)] \leq 0 \leq \mathbb{E}_{x,a}[\ell_{\bar{s}}(\bar{v} - \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) + \varepsilon))].$$

By the monotonicity of $\ell_{\bar{s}}$ and Lemma B.17, we conclude that

$$\mathrm{SR}^{\ell_{\bar{s}}}_{x,a}(\bar{v}) \in [\mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) - \varepsilon, \mathrm{SR}^{\ell_\infty}_{x,a}(\bar{v}) + \varepsilon].$$

This implies that $\mathrm{SR}^{\ell_s}_{x,a}(v) \to \mathrm{SR}^{\ell_\infty}_{x,a}(v)$ as $s \to \infty$. Using the same argument as in Lemma B.19, the convergence is also uniform on compact sets.

Therefore, we conclude that the UBSR is asymptotically coherent, i.e. Assumption 4.2 holds. □

Proof of Lemma B.16. The result follows by directly applying lemmas B.18, B.20 and Lemma 15 of Prashanth and Bhat (2022). □

Lemma B.21. *A risk map $\mathcal{R}$ that employs an OCE risk measure satisfies assumptions 2.12 and 4.2, 4.6, if the Markov chain satisfies Assumption 4.8 and the loss function satisfies assumptions 4.9 and B.15.*

Proof. To prove Assumption 2.12 notice that for any $v \geq v' \in \mathcal{L}(\mathcal{X})$, we have

$$\begin{aligned}
\mathrm{OCE}_{x,a}(v) - \mathrm{OCE}_{x,a}(v') &= \inf_{t \in \mathbb{R}}\{\xi + \mathbb{E}_{x,a}[\ell(v-t)]\} - \inf_{t' \in \mathbb{R}}\{t' + \mathbb{E}_{x,a}[\ell(v'-t')]\} \\
&\geq \inf_{t \in \mathbb{R}}\{\mathbb{E}_{x,a}[\ell(v-t) - \ell(v'-t)]\} \\
&= \sum_{y \in \mathcal{X}} P(y|x,a)\delta(v,v',t,y)(v(y) - v'(y)) \\
&\geq \epsilon_1 \sum_{y \in \mathcal{X}} P(y|x,a)(v(y) - v'(y)),
\end{aligned}$$

for some $\delta(v,v',t,y) \in [\epsilon_1, L_1]$ whose existence is guaranteed by Assumption 4.9. We choose the coherent risk measure $\nu(v) := v(\bar{x})$ and $0 < \bar{\alpha} < \epsilon_1 \min_{(x,a)\in\mathcal{K}} P(\bar{x}|x,a) \in (0,1)$, since $\epsilon_1 \leq \ell'_-(0) \leq 1$ from the definition of OCE. By Assumption 4.9, for any $(x,a) \in \mathcal{K}$, we have that for all $(x,a) \in \mathcal{K}$,

$$\begin{aligned}
&\mathrm{OCE}_{x,a}(v) - \bar{\alpha}\nu(v) - \mathrm{OCE}_{x,a}(v') + \bar{\alpha}\nu(v') \\
&\qquad \geq \epsilon_1 \sum_{y \in \mathcal{X}} P(y|x,a)(v(y) - v'(y)) - \bar{\alpha}(v(\bar{x}) - v'(\bar{x})) \\
&\qquad \geq \left(\epsilon_1 \min_{(x,a)\in\mathcal{K}} P(\bar{x}|x,a) - \bar{\alpha}\right)(v(\bar{x}) - v'(\bar{x})) \geq 0,
\end{aligned}$$

This proves Assumption 2.12.

To prove Assumption 4.2, we start with establishing that for all $(x,a) \in \mathcal{K}$, all $v \in \mathcal{L}(\mathcal{X})$ and for all $s > 0$ we have

$$\begin{aligned}\frac{1}{s}\mathrm{OCE}^{\ell}_{x,a}(sv) &= \frac{1}{s}\inf_{\xi\in\mathbb{R}}\{\xi + \mathbb{E}_{x,a}[\ell(sv-\xi)]\}\\ &= \inf_{\xi'\in\mathbb{R}}\{\xi' + (1/s)\mathbb{E}_{x,a}[\ell(sv-s\xi')]\}\\ &= \inf_{\xi'\in\mathbb{R}}\{\xi' + \mathbb{E}_{x,a}[\ell_s(v-\xi')]\}\\ &= \mathrm{OCE}^{\ell_s}_{x,a}(v),\end{aligned}$$

where $\ell_s(y) := (1/s)\ell(sy)$. Showing that the OCE risk map is asymptotically coherent therefore reduces to showing that $\mathrm{OCE}^{\ell_s}_{x,a}(v)$ converges uniformly to $\mathrm{OCE}^{\ell_\infty}_{x,a}(v)$ on compact sets.

Notice that from Proposition 2.1 in Ben-Tal and Teboulle (2008), the infimum in the representation of $\mathrm{OCE}_{x,a}(v)$ can be attained on a member of the bounded interval supporting of the distribution $\sum_{x'} P(x'|x,a)\delta_{v(x')}$. This implies that for any fixed $\bar{v} \in \mathcal{L}(\mathcal{X})$, there exists an optimal $\xi^* \in \mathcal{V} := [\min_{x'\in\mathcal{X}} v(x'), \max_{x'} v(x')]$ such that $\mathrm{OCE}^{\ell}_{x,a}(\bar{v}) = \xi^* + \mathbb{E}_{x,a}[\ell(v-\xi^*)]$. Hence for any fixed $\bar{v} \in \mathcal{L}(\mathcal{X})$, we can let ξ^*_s and $\xi^*_\infty \in \mathcal{V}$ be the optimal ξ for the OCE with loss function ℓ_s and ℓ_∞ such that:

$$\mathrm{OCE}^{\ell_s}_{x,a}(\bar{v}) = \xi^*_s + \mathbb{E}_{x,a}[\ell_s(\bar{v}-\xi^*_s)], \quad \mathrm{OCE}^{\ell_\infty}_{x,a}(\bar{v}) = \xi^*_\infty + \mathbb{E}_{x,a}[\ell_\infty(\bar{v}-\xi^*_\infty)], \quad (x,a) \in \mathcal{K}.$$

Following Lemma B.19, we have $\ell_s(x) \to \ell_\infty(x)$ uniformly on compact sets. Hence for any arbitrarily small $\varepsilon > 0$, there exists a sufficiently large $\bar{s}$ such that $|\ell_\infty(y) - \ell_{\bar{s}}(y)| \le \varepsilon$, for all $y \in \mathcal{V}$. Denoting $\tilde{\varepsilon}(y) := \ell_\infty(y) - \ell_{\bar{s}}(y)$, we have

$$\begin{aligned}\mathrm{OCE}^{\ell_\infty}_{x,a}(v) &= \xi^*_\infty + \mathbb{E}_{x,a}[\ell_\infty(v-\xi^*_\infty)]\\ &= \xi^*_\infty + \sum_{x'\in\mathcal{X}} P(x'|x,a)[\ell_{\bar{s}}(v(x')-\xi^*_\infty) + \tilde{\varepsilon}(\bar{v}(x')-\xi^*_\infty)]\\ &\ge \xi^*_\infty + \sum_{x'\in\mathcal{X}} P(x'|x,a)[\ell_{\bar{s}}(v(x')-\xi^*_\infty) - \varepsilon]\\ &\ge \xi^*_{\bar{s}} + \mathbb{E}_{x,a}[\ell_{\bar{s}}(v-\xi^*_{\bar{s}})] - \varepsilon = \mathrm{OCE}^{\ell_{\bar{s}}}_{x,a}(v) - \varepsilon.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{OCE}^{\ell_{\bar{s}}}_{x,a}(v) &= \xi^*_{\bar{s}} + \mathbb{E}_{x,a}[\ell_{\bar{s}}(v-\xi^*_{\bar{s}})]\\ &= \xi^*_{\bar{s}} + \sum_{x'\in\mathcal{X}} P(x'|x,a)[\ell_\infty(v(x')-\xi^*_{\bar{s}}) - \tilde{\varepsilon}(v(x')-\xi^*_{\bar{s}})]\\ &\ge \xi^*_{\bar{s}} + \sum_{x'\in\mathcal{X}} P(x'|x,a)[\ell_\infty(v(x')-\xi^*_{\bar{s}})] - \varepsilon\\ &\ge \xi^*_\infty + \mathbb{E}_{x,a}[\ell_\infty(v-\xi^*_\infty)] - \varepsilon = \mathrm{OCE}^{\ell_\infty}_{x,a}(v) - \varepsilon.\end{aligned}$$

Hence we have

$$\mathrm{OCE}^{\ell_\infty}_{x,a}(v) - \varepsilon \le \mathrm{OCE}^{\ell_s}_{x,a}(v) \le \mathrm{OCE}^{\ell_\infty}_{x,a}(v) + \varepsilon,$$

which means that $|\mathrm{OCE}^{\ell_s}_{x,a}(v) - \mathrm{OCE}^{\ell_\infty}_{x,a}(v)| < \varepsilon$. This implies $\lim_{s\to\infty} \mathrm{OCE}^{\ell_s}_{x,a}(v) = \mathrm{OCE}^{\ell_\infty}_{x,a}(v)$ pointwise. Since OCE is Lipschitz continuous and uniformly bounded on compact sets, we can apply a similar reasoning as in Lemma B.19 to establish Assumption 4.2.

Assumption 4.6 follows from Lemma 12 of Prashanth and Bhat (2022). □

Lemma B.22. *A risk map $\mathcal{R}$ that employs a spectral risk measure satisfies assumptions 2.12, 4.2, 4.6, if the Markov chain satisfies Assumption 4.8 and the risk spectrum $\phi(\beta) \in [\epsilon_2, L_2]$ for some $\epsilon_2 > 0$ and $L_2 < \infty$ for all $\beta \in [0,1]$.*

Proof. Since spectral risk measures are coherent, Assumption 4.2 holds automatically. Assumption 4.6 follows from Lemma 13 of Prashanth and Bhat (2022). We are left with Assumption 2.12.

Notice that

$$\begin{aligned} M^{\phi}(v) &= \int_0^1 (\phi(\beta) - \epsilon_2) F_v^{-1}(\beta) d\beta + \epsilon_2 \int_0^1 F_v^{-1}(\beta) d\beta \\ &= \epsilon_2 \mathbb{E}[v] + \int_0^1 \tilde{\phi}(\beta) F_v^{-1}(\beta) d\beta = \epsilon_2 \mathbb{E}[v] + M^{\tilde{\phi}}(v), \end{aligned}$$

where $\tilde{\phi}(\beta) := \phi(\beta) - \epsilon_2 \in [0, L_2 - \epsilon_2]$ for all $\beta \in [0,1]$. Therefore, for any $v \geq v' \in \mathcal{L}(\mathcal{X})$, choosing the coherent risk measure $\nu(v) = v(\bar{x})$ and setting $0 < \bar{\alpha} < \epsilon_2 \min_{(x,a)\in\mathcal{K}} P(\bar{x}|x,a) \in (0,1)$, since $\epsilon_2 \leq \int_0^1 \phi(\beta) d\beta = 1$, we have for any $(x,a) \in \mathcal{K}$,

$$\begin{aligned} &M_{x,a}^{\phi}(v) - \bar{\alpha}\nu(v) - M_{x,a}^{\phi}(v') + \bar{\alpha}\nu(v') \\ &\quad = \epsilon_2(\mathbb{E}_{x,a}[v] - \mathbb{E}_{x,a}[v']) + M_{x,a}^{\tilde{\phi}}(v) - M_{x,a}^{\tilde{\phi}}(v') - \bar{\alpha}(\nu(v) - \nu(v')) \\ &\quad \geq \epsilon_2 \sum_{y\in\mathcal{X}} P(y|x,a)(v(y) - v'(y)) - \bar{\alpha}(v(\bar{x}) - v'(\bar{x})) \\ &\quad \geq \min_{(x,a)\in\mathcal{K}} P(\bar{x}|x,a)(\epsilon_2 - \epsilon_2)(v(\bar{x}) - v'(\bar{x})) \geq 0, \end{aligned}$$

where the first inequality follows from the fact that spectral risk measures are monotone. Therefore, Assumption 2.12 holds. □

It is worth noting that for the widely used OCE measure CVaR, the corresponding loss function, given by $\ell(x) = (1-\alpha)^{-1}(x)^+$, has a minimum slope of 0; its risk spectrum, defined as $\phi(\beta) = (1-\alpha)^{-1}\mathbf{1}\{\beta \geq \alpha\}$, attains a minimal value of 0. As a result, CVaR does not satisfy the condition required in Theorem 4.10. However, when mixed with the expectation, the mean-CVaR risk measure with $\eta > 0$ fulfills the necessary condition for spectral risk measure in Theorem 4.10 and hence satisfies Assumption 2.12.

Proof of Theorem 4.10. The result follows directly by applying lemmas B.16, B.21 and B.22. □

C Additional Details and Results of Experiments

This section presents further experiments on the convergence of the MLMC Q-learning algorithm (4.1), along with statistical experiments on its sample efficiency. In addition, we conduct further experiments regarding the convergence of synchronous and asynchronous UBSR Q-learning with different loss functions, along with statistical results comparing this algorithm to the MLMC Q-learning algorithm (4.1). Furthermore, we provide details on the application setups used in the main text. We also include a risk analysis based on the expectile parameters across different application scenarios. All the experiments were carried out using Python 3.9 on a Linux server equipped with a 64-core AMD EPYC 7763 processor.

C.1 Statistical Experiments on MLMC

Although Theorem 4.7 ensures controllable variance for $r \in (0, 1/2)$, it requires an infinite number of samples in expectation per iteration to achieve this. However, our experiments indicate that controllable variance can still be attained for some $r \in (1/2, 3/4)$, as demonstrated in Wang et al. (2023a) for a special case of distributionally robust discounted MDP.

Table C.1 shows the statistical results (average number of samples, average estimated optimal risk and standard deviation of estimated optimal risk) from 100 simulations, each consisting of 1,000 iterations of the MLMC Q-learning algorithm (4.1) based on MLMC, for different values of the geometric parameter r changing from 0.49 to 0.9, under a randomly generated MDP with 3 states and 3 actions, following the generation procedure outlined in Section 5.1. We observe that for small values of r, the number of samples required to estimate the risk measure is quite large, but it decreases as r increases. Additionally, the final estimated optimal average risk closely approximates the true average risk, which is 0.2968, computed via risk-aware RVI, suggesting that the MLMC Q-learning algorithm indeed converges to the right value. The standard deviation of estimated optimal risk initially decreases starting at $r = 0.49$, but begins to rise again at $r = 0.70$. This supports the findings in Wang et al. (2023a), indicating that MLMC could offer finite sample guarantee and controllable variance for some $r \in (1/2, 3/4)$.

Table C.1: Statistical properties of MLMC Q-learning algorithm for different r

r	**Average Number of Samples**	**Average Estimated Optimal Risk**	**Standard Deviation of Est. Opt. Risk**
0.49	202615.68	0.2956	0.0159
0.50	174297.02	0.2973	0.0163
0.55	92635.72	0.2976	0.0153
0.60	52500.64	0.2974	0.0164
0.65	39152.40	0.2963	0.0157
0.70	31476.86	0.2983	0.0188
0.75	26955.32	0.3002	0.0210
0.80	23967.84	0.2966	0.0221
0.90	20257.32	0.2902	0.0326

C.2 Additional Experiments for UBSR Q-learning Algorithm

For completeness, we present the synchronous version of UBSR Q-learning algorithm as follows: for all $(x, a) \in \mathcal{K}$,

$$Q_{n+1}(x,a) = Q_n(x,a) + \gamma(n)\ell\Big(c(x,a) + \min_{a' \in \mathcal{A}} Q_n(x',a') - f(Q_n) - Q_n(x,a)\Big), \tag{C.1}$$

where $x' \sim P(\cdot|x,a)$, $\gamma(n)$ is the step size satisfying Assumption 4.4 and $f(Q_n)$ serves as the relative value satisfying Assumption 3.3 and 4.3.

In addition to the expectile experiment presented in Section 4.3, we also provide the convergence results for the special case of the synchronous UBSR Q-learning algorithms with polynomial mixed utility (also referred to as the S-shape utility, as discussed in Shen et al. (2014)) and the soft quantile (as discussed in Hau et al. (2025)). The polynomial mixed utility function, derived from prospect

theory, is defined as follows:

$$\ell_{\mathrm{PM}}(x) := \begin{cases} k_1 x^{b_1}, & x \geq 0, \\ -k_2(-x)^{b_2}, & x < 0, \end{cases}$$

where $k_1, k_2 > 0, b_1, b_2 \geq 0$. The soft quantile, used as an approximation for the quantile, is defined as follows:

$$\ell_{\mathrm{SQ}}(x) := \begin{cases} (1-\alpha)(\kappa x + \kappa^2 - 1), & x < -\kappa, \\ \frac{1-\alpha}{\kappa} x, & -\kappa \leq x < 0, \\ \frac{\alpha}{\kappa} x, & 0 \leq x < \kappa, \\ \alpha(\kappa x - \kappa^2 + 1), & x \geq \kappa, \end{cases}$$

with $\alpha \in [0, 1]$ and $\kappa > 0$. We choose $k_1 = 1 - k_2 = 0.3$, $b_1 = b_2 = 0.5$ for the polynomial mixed utility and $\alpha = 0.2$, $\kappa = 2$ for the soft quantile. It is worth noticing that both two loss functions are generally neither convex nor concave on the whole domain.

Under the same MDP and step size settings as in Section 5.1, we run our algorithm 100 times independently and plot the trajectory obtained from the value iteration, risk-aware RVI algorithm (3.2), mean value of $f(Q)$ across all 100 trajectories from the synchronous UBSR Q-learning algorithm (C.1) and 95th and 5th percentiles as the upper and lower bound of the 100 trajectories as the confidence interval. The results are presented in Figure C.1. It appears that our synchronous UBSR Q-learning algorithm successfully converges to the true optimal average risk with high probability under both instances of the loss functions.

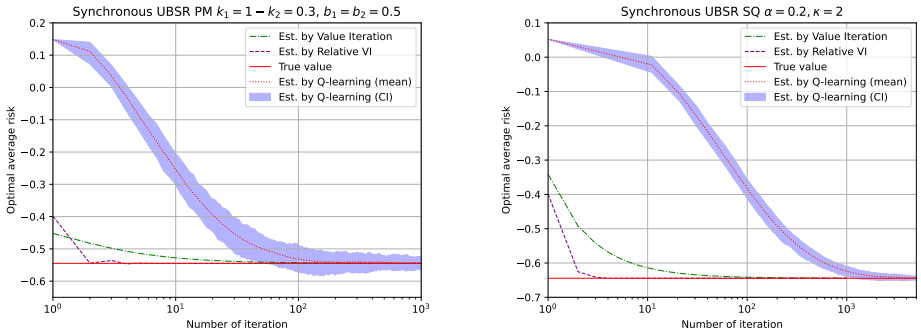

Figure C.1: Convergence of the synchronous UBSR Q-learning algorithm (C.1) for polynomial mixed utility and soft quantile.

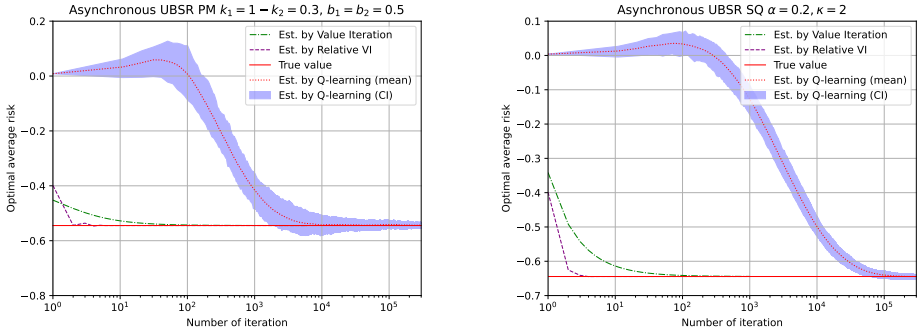

Figure C.2: Convergence of the asynchronous UBSR Q-learning algorithm (4.3) for polynomial mixed utility and soft quantile.

Figure C.2 presents the convergence experiments for the asynchronous UBSR Q-learning algorithm (4.3) under the same settings as in Section 5.1, with polynomial mixed utility parameters $k_1 =$

$1 - k_2 = 0.3$, $b_1 = b_2 = 0.5$ and soft quantile parameters $\alpha = 0.2$, $\kappa = 2$, under 300,000 iterations. The results provide evidence that the asynchronous algorithm also converges with high probability, confirming the applicability of the off-policy UBSR Q-learning algorithm (4.3).

Table C.2 shows the statistical properties (average number of iterations, average estimated optimal risk, standard deviation of estimated optimal risk and average risk of estimated policy) of 100 simulations comparing the MLMC Q-learning algorithm (4.1) with the synchronous UBSR (S-UBSR) Q-learning algorithm (C.1) and asynchronous UBSR (A-UBSR) Q-learning algorithm (4.3) under the same settings as in Section 5.1. All the algorithms are using the equivalent number of samples. The total sample size for the MLMC-based and A-UBSR Q-learning algorithms is 300,000, equivalent to 1,000 iterations for the S-UBSR Q-learning algorithm since the S-UBSR Q-learning generates one sample per state-action pair during each iteration. For the MLMC Q-learning algorithm, we select $r > 0.5$ to ensure a finite average number of samples per iteration. The optimal average risk, computed through the risk-aware RVI (3.2), is -0.1076. Additionally, we compare the mean average risk derived from the policies produced by the algorithms to assess whether the algorithms provide the optimal policy.

Table C.2: Statistics of solutions from risk-aware RVI Q-learning algorithms after 300,000 (x, a, x') observations in a setting where the true optimal average risk is -0.1076.

Algorithm	r	**Average Number of Iterations**	**Average Estimated Optimal Risk**	**Standard Deviation of Est. Opt. Risk**	**Average Risk of Estimated Policy**
MLMC	0.55	729.16	-0.1086	0.0137	-0.1052
	0.60	1039.22	-0.1099	0.0103	-0.1060
	0.65	1396.86	-0.1067	0.0104	-0.1060
	0.70	1731.77	-0.1096	0.0099	-0.1065
	0.75	2000.43	-0.1097	0.0089	-0.1053
	0.80	2250.29	-0.1113	0.0114	-0.1059
	0.90	2667.43	-0.1165	0.0206	-0.1050
S-UBSR	–	6000	-0.1076	0.0030	-0.1076
A-UBSR	–	300000	-0.1074	0.0029	-0.1076

Figure C.3 illustrates the convergence rate of the synchronous UBSR Q-learning algorithm under the expectile risk measure for different values of τ. The y-axis denotes the absolute error between the current value and the optimal average risk. Recall that under the expectile risk measure, $\tau = 0.5$ represents a risk-neutral agent, $\tau > 0.5$ corresponds to a risk-averse agent, and $\tau < 0.5$ to a risk-seeking agent. The results show that convergence occurs faster for a risk-neutral agent and more slowly as the agent becomes more risk-aware. This suggests that incorporating risk-awareness in a way that is more sensitive to the tail events (i.e. τ going towards 0 or 1) increases the computational effort required.

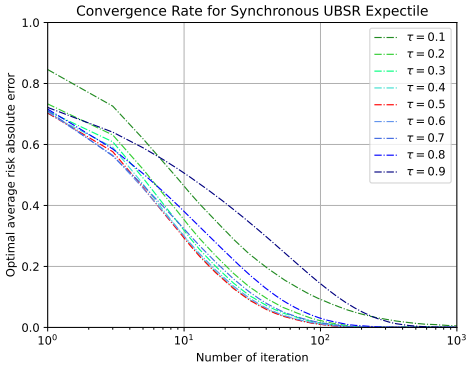

Figure C.3: Convergence rate of synchronous UBSR Q-learning for expectile under different τ.

From the experiments, we conclude that although constrained to the UBSR class of risk measures, the UBSR Q-learning algorithm shows significantly higher computational efficiency. It achieves notably lower standard error, faster convergence, and greater ease of implementation.

Although the convergence experiment results are promising, analyzing the almost sure convergence of the UBSR Q-learning algorithms (C.1) remains a greater challenge. The standard approach for proving the almost sure convergence of the average reward Q-learning algorithm relies on the ODE analysis of stochastic approximation (Abounadi et al., 2001; Borkar and Meyn, 2000). In this case, the analysis leads to a high-dimensional nonlinear ODE system, which lacks the desirable property observed in the risk-neutral setting and our MLMC-based approach, where the difference between the reference ODE q and the target ODE p, when starting from the same initial point, remains a scalar function over time. Consequently, Lemma B.10 does not hold.

C.3 Application Setups

The *machine replacement* problem (e.g. Section 6.10.4, Puterman (1994)) involves managing a machine that deteriorates over time, with the goal of minimizing the long-term average cost. The machine can be in various states representing its condition, ranging from new to totally break down. At each time step, the agent chooses between two actions: keep operating the machine or replace it with a new one. Operating the machine incurs maintenance and operational costs, which increase as the machine deteriorates, while replacing it incurs a significant one-time cost but resets the machine to its best condition.

For the parameters, we define a scenario with 30 degradation states, where state 0 represents a fully new machine and state 29 corresponds to a failure. The degradation probabilities are generated randomly, with a decreasing probability of transitioning to higher degradation states as the machine's condition worsens. Additionally, there is always a positive probability of transitioning to the failure state. The replacement cost is set to $30^{1.5}$, the operating cost is $1 \times s$, and the maintenance cost is $0.5 \times s^{1.5}$, where s denotes the current state level. Additionally, the failure cost is twice the replacement cost, ensuring significant penalties for machine failure.

The *water reservoir management* problem (e.g. Section 1.3, Hernández-Lerma (1989)) involves managing a reservoir to balance water supply, demand, and the risk of overflow or shortage. The reservoir has discrete states representing water levels, and at each time step, the decision-maker chooses how much water to release. The goal is to minimize the long-term average cost, which includes penalties for water shortages, overflows, and operational costs.

For the parameters, we define the maximum water level as 19 and the maximum release as 5. The demand is set to 4, with a shortage cost of 15 per level shortage, an overflow cost of 20 per level overflow, and an operational cost of 2 per unit of water released. The probability of the incoming water level is randomly generated, with a decreasing probability of transitioning to higher water levels, reflecting the natural variability of inflows. However, there is always a positive probability of reaching the maximum water level, ensuring that the risk of overflow is accounted for in every state.

The *inventory management* problem (e.g. Section 1.3, Hernández-Lerma (1989)) involves managing stock levels to meet stochastic demand while minimizing long-term average costs. The system has discrete states representing inventory levels, and at each time step, the agent chooses how much to order to replenish stock. Costs include holding costs for inventory, ordering costs for placing orders, and shortage costs for unmet demand.

For the parameters, we set the maximum inventory level to 9 and the maximum demand to 9. The probability of the incoming demand is generated randomly with lower probability for higher demand. The holding cost per unit of inventory is 1, the ordering cost per unit is 5, and the shortage cost per unit of unmet demand is 10.

C.4 Risk Analysis Based on Parameter of Expectile

We visualize the results of the machine replacement and water reservoir management problems under different τ parameters of the expectile in Figure C.4. This figure illustrates the difference between the τ-optimal average risk and the average risk evaluated under the risk-neutral policy. The findings confirm that for $\tau < 0.5$, the agent exhibits risk-seeking behavior, whereas for $\tau > 0.5$, the agent becomes risk-averse. Notably, when $\tau > 0.5$, the τ-optimal policy achieves a lower average risk than the risk-neutral policy, reaching the minimum average risk at the corresponding τ.

In Figure C.5, we present 30 simulation trajectories for both the risk-neutral policy and the risk-averse expectile policy (with $\tau = 0.9$) across 1,000 iterations for the inventory management prob-

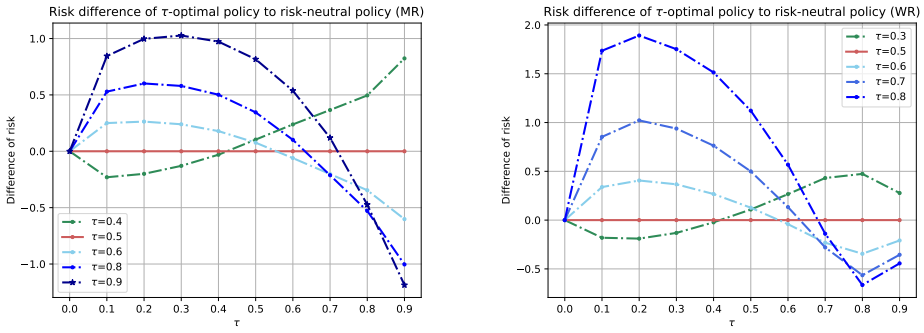

Figure C.4: Risk difference between τ-optimal policy and risk-neutral policy under different τ-values for MR and WR.

lem. The risk-neutral policy results in an optimal policy of (5,4,3,2,1,0,0,0,0,0), while the risk-averse policy yields an optimal policy of (2,1,0,0,0,0,0,0,0,0), with the first element of the vector representing zero inventory. It is evident that the risk-averse policy produces trajectories with lower variance, suggesting that it could offer greater stability when observed over a shorter time frame.

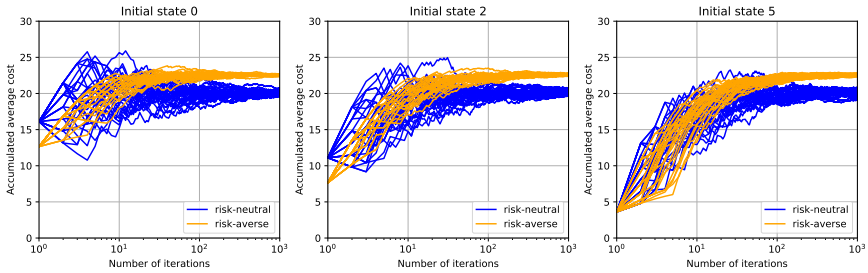

Figure C.5: Simulation trajectories of risk-neutral policy and risk-averse policy under the risk-neutral setting for IM.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: Theoretical results are provided in Sections 3 and 4 with complete proofs in the Appendix B. These results are verified numerically in Section 5 and Appendix C.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The limitations are discussed in Section 5.1 and further experiments addressing these issues are provided in Section C. Future research directions are discussed in Section 6.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: Assumptions are stated and justified in Sections 2, 3, 4. Proofs are in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Experiment details are well documented in the paper. We provide the pseudo code for the algorithms in Appendix A and the setup of the application environments in Appendix C.3.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]

Justification: We provide the code for implementation in the anonymous repository: https://anonymous.4open.science/r/P-L_ARMDP-NeurIPS2025-3471.
Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: The details are described in Section 5.1 and Appendix C.3.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: We report standard errors for the experimental results in Appendices C.1 and C.2.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: The information is provided in Appendix C.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
   Answer: [Yes]
   Justification: We have reviewed the NeurIPS Code and Ethics and confirm that our paper conform with it.
   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [NA]
    Justification: This is a theoretical paper that presents foundational research in reinforcement learning algorithms. We do not see a direct path to any negative societal impact.
    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is a theoretical paper that presents foundational research in reinforcement learning algorithms. It does not involve any high-risk data or model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the papers on which our research is built on.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: : The paper does not present new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [NA]
    Justification: Our paper does not involve crowdsourcing nor research with human subjects.
    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [NA]
    Justification: Our paper does not involve crowdsourcing nor research with human subjects.
    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
    Answer: [NA]
    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.
    Guidelines:
    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.