Q-learning for Quantile MDPs: A Decomposition, Performance, and Convergence Analysis

Jia Lin Hau
University of New Hampshire
Durham, NH

Erick Delage
GERAD, HEC Montréal, and
MILA - Quebec AI Institute

Esther Derman
MILA - Quebec AI Institute and
Université de Montréal

Mohammad Ghavamzadeh
Amazon AGI
Sunnyvale, CA

Marek Petrik
University of New Hampshire
Durham, NH

Abstract

In Markov decision processes (MDPs), quantile risk measures such as Value-at-Risk are a standard metric for modeling RL agents' preferences for certain outcomes. This paper proposes a new Q-learning algorithm for quantile optimization in MDPs with strong convergence and performance guarantees. The algorithm leverages a new, simple dynamic program (DP) decomposition for quantile MDPs. Compared with prior work, our DP decomposition requires neither known transition probabilities nor solving complex saddle point equations and serves as a suitable foundation for other model-free RL algorithms. Our numerical results in tabular domains show that our Q-learning algorithm converges to its DP variant and outperforms earlier algorithms.

1 INTRODUCTION

The practicality of reinforcement learning (RL) models has led to their widespread integration in autonomous decision-making (Kiran et al., 2021; Kiumarsi et al., 2017). Traditional metrics focusing on expected value fail to address the uncertainties of random returns, thus acknowledging that a one-size-fits-all model is insufficient (Howard and Matheson, 1972; Tamar et al., 2012). This observation is blatant in various applications such as motion control (Ahmadi et al., 2021; Braun et al., 2011; Hakobyan and Yang, 2021), autonomous systems (Jin et al., 2019; Wang and Chapman, 2022), healthcare (Köse, 2016; Singh et al., 2020), or capital investment (Min et al., 2022), where each necessitates a specific risk-sensitive objective. Risk-averse RL (RARL) endeavors to align with the decision-maker's preferences by tailoring the objective function according to their risk interests (Yoo et al., 2024). Common measures in RARL include exponential utility (Borkar, 2002; Hau et al., 2023b), target value (Lin et al., 2003; Wu and Lin, 1999), value-at-risk (VaR) (Chow et al., 2018; Hau et al., 2023a; Li et al., 2022), conditional VaR (CVaR) (Bäuerle and Ott, 2011; Chow and Ghavamzadeh, 2014; Lim and Malik, 2022), or mean-variance objectives (Luo et al., 2024; Tamar et al., 2012), among others.

Quantile measures such as VaR are used in data centers to ensure the reliability of cloud computing services (DeCandia et al., 2007), in epidemiology to understand how exposure to disease differs across continuous health outcomes distributions, e.g., BMI (Wei et al., 2019), or more standardly in financial markets to assess counterparty risk (Alexander and Sarabia, 2012; Emmer et al., 2015). Although VaR does not enjoy the mathematical properties of CVaR and thus is not a coherent risk, its numerical advantages for backtesting have led the Basel Committee to retain it for credit value adjustment (Embrechts et al., 2018; on Banking Supervision, 2023). Unlike expected value, quantile risk measures account for the return distribution. For example, the expectation may favor low probability outcomes yielding a high return, whereas the α-level-VaR guarantees the gain is greater or equal to that value with high probability $1-\alpha$. In RARL, recent years have witnessed an increasing interest in quantile measures, especially since the emergence of distributional RL (Bellemare et al., 2017; Dabney et al., 2018b) and its successful

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

application in robotics (Majumdar and Pavone, 2017).

A major challenge in deriving practical RL solutions is that a full knowledge of the environment is often inaccessible. This can be even more problematic in RARL where risk-averse strategies must involve the full return distribution rather than just its expectation. Previous works have presented model-free methods to find optimal CVaR policies (Dabney et al., 2018a; Keramati et al., 2020; Lim and Malik, 2022), but all assumed the existence of an optimal Markov policy. Such assumption is valid for mean and entropic risk objectives but generally not for quantile-based objectives such as CVaR or VaR (Ben-Tal and Teboulle, 2007; Hau et al., 2023a). Indeed, optimal-VaR policies can potentially all be history-dependent, so restricting the search to Markov policies can produce suboptimal return (Hau et al., 2023a; Li et al., 2022).

In this study, we propose a new dynamic programming (DP) formulation for quantile MDPs (see Li et al. 2022) that we refer to as VaR-MDP since it seeks a policy that is optimal with respect to the VaR of the total discounted reward. Compared with prior work, our DP decomposition requires neither known transition probabilities nor solving complex saddle point equations. As a second contribution, we introduce VaR-Q-learning, the first model-free method that provably optimizes the return distribution's VaR. We develop a rigorous proof of convergence to an optimal policy, thus ensuring the validity of our approach. Although our VaR-Q-learning can be seen as a simple variant of the risk-sensitive Q-learning methods developed in distributional RL (Bellemare et al., 2017; Dabney et al., 2018a,b), we show that our slight modification makes a meaningful difference in the quality of the computed policy.

We first delineate our research setting and preliminary concepts in Section 2. Then, we introduce new DP equations in Section 3 to solve VaR-MDPs while being simpler than existing methods. This enables us to propose a novel quantile Q-learning algorithm in Section 4, which optimizes the VaR-MDP from sampled trajectories, and establish its convergence property. Finally, numerical experiments presented in Section 5 illustrate the effectiveness of our algorithm.

2 PRELIMINARIES AND FORMAL MODEL

We first introduce our notations and overview relevant properties for quantile and VaR risk measures. We then formalize the MDP framework with VaR objective, which we name VaR-MDP in short. We prove the claims of this section in Appendix A.

Notation. The augmented reals are $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ and we denote by $\mathcal{I}$ the class of closed intervals in $\mathbb{R}$. Given a measurable space $\mathcal{E}$, we abuse notation and denote by $\mathbb{R}^{\mathcal{E}}$ the set of all measurable functions from $\mathcal{E}$ to $\mathbb{R}$, with $\bar{\mathbb{R}}^{\mathcal{E}}$ the measurable functions to $\bar{\mathbb{R}}$, and finally with $\mathcal{I}^{\mathcal{E}}$ the functions to $\mathcal{I}$. Given a finite set $\mathcal{Z} = \{1, \ldots, Z\}$, the probability simplex is $\Delta_{\mathcal{Z}} := \{y \in \mathbb{R}_+^{Z} \mid \mathbf{1}^\top y = 1\}$. For conciseness, we denote by $[n]$ the sequence of integers from 0 to n. The set of discrete real-valued random variables with finite support is denoted by $\mathbb{X}$. Random variables are marked with a tilde, e.g., $\tilde{x} \in \mathbb{X}$.

Quantiles and Value-at-Risk The quantile of a random variable $\tilde{x} \in \mathbb{X}$ at level $\alpha \in [0,1]$ is any $\tau \in \bar{\mathbb{R}}$ such that $\mathbb{P}[\tilde{x} \le \tau] \ge \alpha$ and $\mathbb{P}[\tilde{x} \ge \tau] \ge 1-\alpha$. It may not be unique but lies in the interval $[\mathfrak{q}_\alpha^-(\tilde{x}), \mathfrak{q}_\alpha^+(\tilde{x})]$, where

$$\begin{aligned}\mathfrak{q}_\alpha^-(\tilde{x}) &:= \min\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}[\tilde{x} \le \tau] \ge \alpha\right\},\\ \mathfrak{q}_\alpha^+(\tilde{x}) &:= \max\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}[\tilde{x} < \tau] \le \alpha\right\}.\end{aligned} \tag{1}$$

The maximum in Eq. (1) exists because the mapping $\tau \mapsto \mathbb{P}[\tilde{x} < \tau]$ is lower semi-continuous. So does the minimum, since $\tau \mapsto \mathbb{P}[\tilde{x} \le \tau]$ is upper semi-continuous. Also, $\mathfrak{q}_0^-(\tilde{x}) = -\infty$ and $\mathfrak{q}_1^+(\tilde{x}) = \infty$, while $\mathfrak{q}_\alpha^-(\tilde{x}) \in \mathbb{R}$ for all $\alpha \in (0,1]$ and $\mathfrak{q}_\alpha^+(\tilde{x}) \in \mathbb{R}$ for all $\alpha \in [0,1)$.

Monetary risk measures generalize the average criterion to account for uncertain outcomes. Among them, quantile-based measures like VaR are the most common (Follmer and Schied, 2016; Shapiro et al., 2014). Given a risk level $\alpha \in [0,1]$, the VaR function $\mathrm{VaR}_\alpha \colon \mathbb{X} \to \bar{\mathbb{R}}$ is defined as the largest $1-\alpha$ confidence lower bound on the value of $\tilde{x}$, i.e.,

$$\mathrm{VaR}_\alpha[\tilde{x}] \;:=\; \mathfrak{q}_\alpha^+(\tilde{x}). \tag{2}$$

By convention, $\mathrm{VaR}_\alpha[0] = 0$ if $\alpha \in [0,1)$ and ∞ otherwise.

Elicitability. Based on the works of Gneiting (2011) and Bellini and Bignozzi (2015), a risk measure is *elicitable* if it is the solution of an empirical risk minimization problem. In particular, a quantile can be estimated via quantile regression with the loss (Koenker and Bassett, 1978):

$$\ell_\alpha(\delta) := \max\{\alpha\delta, -(1-\alpha)\delta\}\ , \tag{3}$$

where δ represents the residual error between the prediction and the noisy observation. Thus, quantile measures are elicitable as stated below.

Lemma 2.1. *For any $\tilde{x} \in \mathbb{X}$ and $\alpha \in [0,1]$, it holds that*

$$\operatorname*{argmin}_{y \in \mathbb{R}} \mathbb{E}[\ell_\alpha(\tilde{x} - y)] \;=\; [\mathfrak{q}_\alpha^-(\tilde{x}), \mathfrak{q}_\alpha^+(\tilde{x})] \cap \mathbb{R}.$$

As explained in (Bellini and Bignozzi, 2015, Ex. 3.8), VaR is not elicitable unless $\tilde{x}$ is continuous, i.e. $\mathfrak{q}^-_\alpha(\tilde{x}) = \mathfrak{q}^+_\alpha(\tilde{x})$. Hence, methods that rely on statistical estimation of conditional VaR using Lemma 2.1 must contend with potential underestimation.

MDPs with Value-at-Risk Objective We formulate the decision process as an MDP $(\mathcal{S}, \mathcal{A}, r, p, \gamma, T)$ that comprises a set of states $\mathcal{S} = \{1, \dots, S\}$, a set of actions $\mathcal{A} = \{1, \dots, A\}$, a reward function $r \colon \mathcal{S} \times \mathcal{A} \to [\underline{R}, \bar{R}]$, and a transition probability $p \colon \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$, where $p(s, a, s')$ denotes the probability to transit from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$. The coefficient $\gamma \in [0, 1]$ is a discount factor and $T \in \mathbb{N}$ is the decision horizon. We allow for a discount factor $\gamma \in [0, 1]$ to study a more general framework, which can be extended to discounted-infinite horizon MDPs when $\gamma < 1$ (Li et al., 2022, Theorem 6). The agent aims to find a policy π that optimizes the *static* VaR of the discounted sum of returns,

$$\rho(\pi) := \operatorname{VaR}^{\pi, s_0}_{\alpha_0}\left[\sum_{k=0}^{T-1} \gamma^k r(\tilde{s}_k, \tilde{a}_k)\right], \tag{4}$$

for some initial state $s_0 \in \mathcal{S}$ and reference risk level $\alpha_0 \in (0, 1)$. In Eq. (4), the distribution of $\tilde{s}_k$ for $k \geq 1$ is implicitly governed by the transition model p while the superscript s_0 fixes the initial state. The policy π governs the realization of actions $\tilde{a}_k$ at all steps k, which we formalize next.

Defining a history at time $k \in [T-1]$ as $h_k := (s_0, a_0, s_1, a_1, \dots, s_k) \in \mathcal{H}_k := (\mathcal{S} \times \mathcal{A})^k \times \mathcal{S}$, its appending to $a \in \mathcal{A}$ and $s' \in \mathcal{S}$ is denoted by $\langle h_k, a, s' \rangle \in \mathcal{H}_{k+1}$. Given a time horizon $t \in 1{:}T$, a *history-dependent policy* $\pi := (\pi_k)_{k=0}^{t-1}$ is a sequence of decision rules $\pi_k \colon \mathcal{H}_k \to \mathcal{A}$ from histories to actions. Focusing on the class of Markov or even stationary policies is standard for risk-neutral objectives because they are optimal (Puterman, 2014). For risk-averse objectives, they are generally not, so we must optimize over history-dependent policies. We note that Hau et al. (2023a) established the existence of optimal deterministic policies in Eq. (4), so we can ignore stochastic policies without impairing optimality. Let Π^t_{HD} be the set of all history-dependent deterministic policies over horizon t. All in all, given a quantile level α_0 and an initial state s_0, we aim to find $\max_{\pi \in \Pi^T_{\mathrm{HD}}} \rho(\pi)$.

Although the optimal policy for Eq. (4) is history-dependent, it can still be computed using DP and value iteration (Hau et al., 2023a; Li et al., 2022). As we will see, the difference with standard DP lies in the fact that the optimal state-action value function must also adapt the quantile level $\alpha \in [0, 1]$ at each state $s \in \mathcal{S}$ and $t \in [T]$. Let thus $q^\star_t \colon \mathcal{S} \times [0, 1] \times \mathcal{A} \to \bar{\mathbb{R}}$ be the optimal state-action value function for $t \in [T]$:

$$q^\star_t(s, \alpha, a) := \max_{\substack{\pi \in \Pi^t_{\mathrm{HD}}: \\ \pi_0(s) = a}} \operatorname{VaR}^{\pi, s}_\alpha\left[\sum_{k=0}^{t-1} \gamma^k r(\tilde{s}_k, \tilde{a}_k)\right]. \tag{5}$$

It is also convenient to define the optimal state value function $v^\star_t \colon \mathcal{S} \times [0, 1] \to \bar{\mathbb{R}}$ for horizon $t \in [T]$ as

$$v^\star_t(s, \alpha) := \max_{\pi \in \Pi^t_{\mathrm{HD}}} \operatorname{VaR}^{\pi, s}_\alpha\left[\sum_{k=0}^{t-1} \gamma^k \cdot r(\tilde{s}_k, \tilde{a}_k)\right].$$

Similar to risk-neutral MDPs, the state value function is related to the state-action value function through $v^\star_t(s, \alpha) = \max_{a \in \mathcal{A}} q^\star_t(s, \alpha, a), \forall t \in [T]$ (see Appendix A.2).

3 VAR DYNAMIC PROGRAMMING

In this section, we devise a DP method to compute an optimal policy for the static VaR objective in (4). This section assumes that the model is known and builds on the analysis in Hau et al. (2023a); Li et al. (2022). Section 4 extends the approach to the model-free setting. Regardless of the methodology, the key idea of VaR-DP is to augment the state space with a risk-level input and to perform Bellman recursions on the augmented state-action value function. This augmentation should not be arbitrary, as the 'risk-level state' must evolve in a specific way to yield an optimal policy. We report the proofs of this section in Appendix B.

3.1 Model-based VaR-DP

We present the VaR-DP introduced by Li et al. (2022) and revised in Hau et al. (2023a). Let $B_{\max} \colon \bar{\mathbb{R}}^{\mathcal{S} \times [0,1] \times \mathcal{A}} \to \bar{\mathbb{R}}^{\mathcal{S} \times [0,1] \times \mathcal{A}}$ be the following Bellman operator. For all $s \in \mathcal{S}$, $\alpha \in [0, 1]$, and $a \in \mathcal{A}$:

$$(B_{\max} q)(s, \alpha, a) := r(s, a) + \gamma \cdot \max_{o \in \mathcal{O}_{sa}(\alpha)} \min_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} q(s', o_{s'}, a'), \tag{6}$$

$$\mathcal{O}_{sa}(\alpha) := \left\{ o \in [0, 1]^S \mid \sum_{s' \in \mathcal{S}} o_{s'} \cdot p(s, a, s') \leq \alpha \right\}.$$

Further consider the sequence $q_0(s, \alpha, a) := \operatorname{VaR}_\alpha[0]$, and $q_{t+1} := B_{\max} q_t$ for all $t \in [T-1]$. The constraint set above depends on the transition model p so Bellman recursions are model-based. Correspondingly, a proper recursion on risk levels leads to an optimal policy in terms of VaR return. Namely, for $k \in [T-1]$, let $\alpha_k : \mathcal{H}_k \to [0, 1]$ such that $\alpha_0(s) = \alpha_0, \quad \forall s \in \mathcal{S}$, while $\alpha_{k+1}(h_{k+1})$ satisfies both

$\alpha_{k+1}(\langle h_k, a, \cdot\rangle) \in \mathcal{O}_{sa}(\alpha_k(h_k))$ and

$$q_{T-k}(s, \alpha_k(h_k), a) = r(s,a) \\ + \gamma \min_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} q_{T-k-1}(s', \alpha_{k+1}(\langle h_k, a, s'\rangle), a').$$

Then, one can derive a greedy policy $\pi := (\pi_k)_{k=0}^{T-1}$ from the constructed risk level mappings $(\alpha_k)_{k=0}^{T-1}$, i.e., at each step $k \in [T-1]$ as

$$\pi_k(h_k) \in \operatorname*{argmax}_{a \in \mathcal{A}} q_{T-k}(s, \alpha_k(h_k), a), \quad \forall h_k \in \mathcal{H}_k. \tag{7}$$

The following theorem shows that the above Bellman operator is optimal, in the sense that the sequence $q := (q_t)_{t=0}^{T-1}$ resulting from these recursions yields Eq. (5). Accordingly, a policy π that is greedy for that sequence q of state-action value functions at the constructed risk-level sequence is optimal for the VaR objective (4).

Theorem 3.1. *Let a sequence $q = (q_t)_{t=0}^{T}$ be such that $q_0(s, \alpha, a) = \mathrm{VaR}_\alpha[0]$ and $q_{t+1} := B_{\max} q_t$ for $t \in [T-1]$. Then, $q_t = q_t^\star$ for all $t \in [T]$, where $q_t^\star$ is defined in (5). Moreover, if a policy $\pi = (\pi_k)_{k=0}^{T-1}$ is greedy w.r.t. q as in (7), then it maximizes the VaR objective (4).*

Theorem 3.1 enables solving MDPs with static VaR objectives through DP equations. An important limitation of this representation is that it requires access to the underlying transition model. This implicitly appears in the constraint set $\mathcal{O}_{sa}(\alpha)$ from Eq. (6), which is required to find an optimal greedy policy according to the risk level mappings $(\alpha_k)_{k=0}^{T-1}$. Additionally, each Bellman update requires solving a constrained optimization problem, which slows down the learning process.

This paper proposes a model-free learning algorithm with theoretical convergence guarantees. To assess the optimality of our value function updates, we reformulate VaR-DP equations in terms of Bellman operators in Section 3.2. Our VaR-DP formulation can be seamlessly used for known or unknown transition probability models, as it allows for statistical estimation from sampled trajectories.

3.2 Nested VaR-DP

Let $B_{\mathrm{u}} \colon \bar{\mathbb{R}}^{\mathcal{S}\times[0,1]\times\mathcal{A}} \to \bar{\mathbb{R}}^{\mathcal{S}\times[0,1]\times\mathcal{A}}$ be the following VaR Bellman operator:

$$(B_{\mathrm{u}} q)(s, \alpha, a) := \mathrm{VaR}_\alpha^{a,s}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q(\tilde{s}_1, \tilde{u}, a')],$$

where the VaR is based on the joint distribution of $(\tilde{s}_1, \tilde{u})$ with $\tilde{s}_1 \sim p(s, a, \cdot)$ and an independent $\tilde{u}$ uniformly distributed on $[0,1]$ ($\tilde{u} \sim U([0,1])$). Correspondingly, consider the following DP equations:

$$q_0^{\mathrm{u}}(s, \alpha, a) := \mathrm{VaR}_\alpha[0], \quad \forall s \in \mathcal{S}, \alpha \in [0,1], a \in \mathcal{A}, \\ q_{t+1}^{\mathrm{u}} := B_{\mathrm{u}} q_t^{\mathrm{u}}, \quad \forall t \in [T-1]. \tag{8}$$

We must also adapt the risk level mappings to this new Bellman recursion. Let $\hat{\alpha}_0^{\mathrm{u}}(s_0) := \alpha_0$ and for $k \in [T-1]$,

$$\hat{\alpha}_{k+1}^{\mathrm{u}}(h_{k+1}) := \min\left\{ o \in [0,1] \mid \max_{a \in \mathcal{A}} q_{T-k-1}^{\mathrm{u}}(s_{k+1}, o, a) \\ \geq \frac{q_{T-k}^{\mathrm{u}}(s_k, \hat{\alpha}_k(h_k), a_k) - r(s_k, a_k)}{\gamma} \right\}. \tag{9}$$

The greediness criterion remains unchanged: at each step $k \in [T-1]$, construct

$$\pi_k^{\mathrm{u}}(h_k) \in \operatorname*{argmax}_{a \in \mathcal{A}} q_{T-k}^{\mathrm{u}}(s, \hat{\alpha}_k^{\mathrm{u}}(h_k), a), \quad h_k \in \mathcal{H}_k. \tag{10}$$

The following theorem states that this alternative method is still valid for finding an optimal solution.

Theorem 3.2. *Let a sequence $q^{\mathrm{u}} = (q_t^{\mathrm{u}})_{t=0}^{T}$ be such that $q_0^{\mathrm{u}}(s, \alpha, a) := \mathrm{VaR}_\alpha[0]$ and $q_{t+1}^{\mathrm{u}} := B_{\mathrm{u}} q_t^{\mathrm{u}}$ for $t \in [T-1]$. Then, $q_t^{\mathrm{u}} = q_t^\star$ for all $t \in [T]$, where $q_t^\star$ is defined in (5). Moreover, if a policy $\pi^{\mathrm{u}} = (\pi_k^{\mathrm{u}})_{k=0}^{T-1}$ is greedy w.r.t. q^{u} as in (10), then it maximizes the VaR objective in (4).*

In contrast to Hau et al. (2023a); Li et al. (2022), Theorem 3.2 does not require knowing the transition model. More importantly, it reduces VaR-MDPs to a nested VaR conditional mapping:

$$v_T^\star(s_0, \alpha_0) = \max_{a_0 \in \mathcal{A}} \mathrm{VaR}_{\alpha_0}^{a_0, s_0}\Big[r(s_0, a_0) + \gamma \cdot \\ \max_{a_1 \in \mathcal{A}} \mathrm{VaR}_{\tilde{u}_1}\big[r(\tilde{s}_1, a_1) + \cdots + \gamma \cdot \\ \max_{a_{T-2} \in \mathcal{A}} \mathrm{VaR}_{\tilde{u}_{T-2}}[r(\tilde{s}_{T-2}, a_{T-2}) + \gamma \cdot \\ \max_{a_{T-1} \in \mathcal{A}} r(\tilde{s}_{T-1}, a_{T-1}) | \tilde{s}_{1:T-2}, \tilde{u}_{1:T-2}] \ldots | \tilde{s}_1, \tilde{u}_1\big]\Big].$$

The value is still over an augmented state-space, but this time, the risk tolerance is independently and uniformly drawn from $[0,1]$ at each step. This is particularly suitable for sample-based RL and more amenable to deep settings such as Dabney et al. (2018a,b); Lim and Malik (2022). Yet, two issues remain to produce a Q-learning procedure. First, the state space is infinite because the risk level α is continuous. Second, the elicitation procedure described in Lemma 2.1 underestimates the VaR when the risk-to-go variable is discrete, so we cannot directly employ quantile regression. To address these issues, in the next section, we propose an approximation scheme that replaces the Bellman operator of Eq. (8) with either lower or upper bounds of the appropriate quantiles.

4 Q-LEARNING ALGORITHM AND ANALYSIS

This section builds on the DP of Section 3 to derive a new Q-learning algorithm for the static VaR objective. Section 4.1 introduces an approximate Bellman operator that can be used to compute Eq. (8) in a tractable way. Then, Section 4.2 proposes the VaR-Q-learning algorithm and shows its convergence guarantees.

4.1 Discretized Quantile Q-functions

The challenge in computing the value function in Eq. (8) stems from the fact that it is defined over a continuous $\alpha \in [0,1]$. To make the computation tractable, we propose to approximate the risk level α with properly defined functions that yield lower and upper bounds on the quantile value function.

Definition 4.1. Let $\bar{f}, \underline{f} \colon [0,1] \to [0,1]$ be two non-decreasing right-continuous functions such that $\bar{f}(\alpha) > \alpha \geq \underline{f}(\alpha)$ for all $\alpha \in [0,1)$, while $\bar{f}(1) = 1 \geq \underline{f}(1)$.

The following result shows how the functions $\underline{f}, \bar{f}$ can yield upper and lower bounds on VaR, and thus, on the state-action value function q. This development exploits that $\alpha \mapsto q(s,\alpha,a)$ is non-decreasing for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Lemma 4.2. *For any $\tilde{x} \in \mathbb{X}$ and $\alpha \in (0,1)$, we have*

$$\mathrm{VaR}_\alpha(\tilde{x}) \geq \mathfrak{q}^+_{\underline{f}(\alpha)}(\tilde{x}) = \max \operatorname*{argmin}_{q \in \mathbb{R}} \mathbb{E}[\ell_{\underline{f}(\alpha)}(\tilde{x} - q)],$$
$$\mathrm{VaR}_\alpha(\tilde{x}) \leq \mathfrak{q}^-_{\bar{f}(\alpha)}(\tilde{x}) = \min \operatorname*{argmin}_{q \in \mathbb{R}} \mathbb{E}[\ell_{\bar{f}(\alpha)}(\tilde{x} - q)],$$

where ℓ_α is defined as in Eq. (3) and $\underline{f}, \bar{f}$ as in Definition 4.1.

We now derive Bellman operators that facilitate the construction of the Q-learning algorithm. First, the set-valued operator $\mathcal{B}_{\mathrm{u}} \colon \mathbb{R}^{\mathcal{S}\times[0,1]\times\mathcal{A}} \to \mathcal{I}^{\mathcal{S}\times[0,1]\times\mathcal{A}}$ generalizes B_{u} as an empirical risk minimizer, i.e.,

$$(\mathcal{B}_{\mathrm{u}} q)(s,\alpha,a) := \tag{11}$$
$$\operatorname*{argmin}_{x \in \mathbb{R}} \mathbb{E}^{a,s}\left[\ell_\alpha\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q(\tilde{s}_1, \tilde{u}, a') - x\right)\right],$$

where $\tilde{u} \sim U([0,1])$. Second, we define the upper and lower bounding Bellman operators $\mathcal{B}^{\bar{f}}_{\mathrm{u}}$ and $\mathcal{B}^{\underline{f}}_{\mathrm{u}}$ for each $b = (s,\alpha,a)$, $s \in \mathcal{S}$, $\alpha \in [0,1]$, and $a \in \mathcal{A}$ as

$$(\mathcal{B}^{\bar{f}}_{\mathrm{u}} q)(b) := \begin{cases} (\mathcal{B}_{\mathrm{u}} q)(s, \bar{f}(\alpha), a) & \text{if } \bar{f}(\alpha) < 1, \\ \bar{R} + \max_{s' \in \mathcal{S}, a' \in \mathcal{A}} q(s', 1, a') & \text{if } \bar{f}(\alpha) = 1, \end{cases}$$

$$(\mathcal{B}^{\underline{f}}_{\mathrm{u}} q)(b) := \begin{cases} (\mathcal{B}_{\mathrm{u}} q)(s, \underline{f}(\alpha), a) & \text{if } \underline{f}(\alpha) > 0, \\ \underline{R} + \min_{s' \in \mathcal{S}, a' \in \mathcal{A}} q(s', 0, a') & \text{if } \underline{f}(\alpha) = 0. \end{cases}$$

The following theorem shows how to use these operators to bound the value functions and the performance of the computed policy.

Theorem 4.3. *Suppose that $\bar{q}^{\mathrm{u}}_0 = \underline{q}^{\mathrm{u}}_0 = 0$, and that $\bar{q}^{\mathrm{u}}_{t+1}$ and $\underline{q}^{\mathrm{u}}_{t+1}$ are right-continuous non-decreasing functions in α satisfying $\bar{q}^{\mathrm{u}}_{t+1} \in (\mathcal{B}^{\bar{f}}_{\mathrm{u}} \bar{q}^{\mathrm{u}}_t)$ and $\underline{q}^{\mathrm{u}}_{t+1} \in (\mathcal{B}^{\underline{f}}_{\mathrm{u}} \underline{q}^{\mathrm{u}}_t)$ for all $t \in [T-1]$.*[1] *Then, $\bar{q}^{\mathrm{u}}_t \geq q^\star_t \geq \underline{q}^{\mathrm{u}}_t$ for all $t \in [T]$, where $q^\star_t$ is defined in (5). Moreover, if a policy $\underline{\pi} := (\underline{\pi}_k)_{k=0}^{T-1}$ is greedy for $\underline{q}^{\mathrm{u}}$, in the sense that*

$$\underline{\pi}_k(h_k) \in \operatorname*{argmax}_{a \in \mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k}(s, \underline{\alpha}^{\mathrm{u}}_k(h_k), a),$$

with $\underline{\alpha}^{\mathrm{u}}$ as in (9), then it satisfies $\max_{a \in \mathcal{A}} \underline{q}^{\mathrm{u}}_T(s_0, \alpha_0, a) \leq \rho(\underline{\pi})$.

To simplify the exposition, we focus on the simple approximation scheme that discretizes the risk-level α using a uniform grid as below.

Example 4.4 (J-uniform discretization). Define $\underline{f}, \bar{f} \colon [0,1] \to [0,1]$ as

$$\underline{f}(\alpha) := \max\left\{ {}^{j}\!/\!{}_{J} \mid {}^{j}\!/\!{}_{J} \leq \alpha, j \in [J-1] \right\},$$
$$\bar{f}(\alpha) := \max\left\{ {}^{j+1}\!/\!{}_{J} \mid {}^{j}\!/\!{}_{J} \leq \alpha, j \in [J-1] \right\},$$

for $J \geq 2$. These functions satisfy the conditions of Definition 4.1.

Under this uniform discretization scheme, $\mathcal{B}_{\mathrm{u}}$ becomes $\mathcal{B}^{\mathrm{d}}_{\mathrm{u}} \colon \mathbb{R}^{\mathcal{S}\times[J-1]\times\mathcal{A}} \to \mathcal{I}^{\mathcal{S}\times[J-1]\times\mathcal{A}}$, defined for $s \in \mathcal{S}$, $j \in [J-1]$, and $a \in \mathcal{A}$ as

$$(\mathcal{B}^{\mathrm{d}}_{\mathrm{u}} q)(s,j,a) \; := \; \operatorname*{argmin}_{x \in \mathbb{R}}$$
$$\frac{1}{J} \sum_{j'=0}^{J-1} \mathbb{E}^{a,s}\left[\ell_{\frac{j}{J}}\left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q(\tilde{s}_1, j', a') - x\right)\right].$$

Similarly, the Bellman operator $\mathcal{B}^{\underline{f}}_{\mathrm{u}}$ becomes $\underline{\mathcal{B}}^{\mathrm{d}}_{\mathrm{u}} \colon \mathbb{R}^{\mathcal{S}\times[J-1]\times\mathcal{A}} \to \mathcal{I}^{\mathcal{S}\times[J-1]\times\mathcal{A}}$, which is defined for $b = (s,j,a)$, $s \in \mathcal{S}$, $j \in [J-1]$, and $a \in \mathcal{A}$ as

$$(\underline{\mathcal{B}}^{\mathrm{d}}_{\mathrm{u}} q)(b) := \begin{cases} (\mathcal{B}^{\mathrm{d}}_{\mathrm{u}} q)(s,j,a) & \text{if } j \geq 1, \\ \underline{R} + \min_{s \in \mathcal{S}, a \in \mathcal{A}} q(s,0,a) & \text{if } j = 0. \end{cases} \tag{12}$$

The following proposition states the correctness of the discretized Bellman operators.

Proposition 4.5. *Given the J-uniform discretization of Example 4.4, let $\underline{q}^{\mathrm{d}}_0 := 0$ and $\underline{q}^{\mathrm{d}}_{t+1} \in \underline{\mathcal{B}}^{\mathrm{d}}_{\mathrm{u}} \underline{q}^{\mathrm{d}}_t, t \in [T-1]$. Then, the sequence $\underline{q} := (\underline{q}_t)_{t=0}^T$ defined for $t \in [T]$ as*

$$\underline{q}_t(s,\alpha,a) := \underline{q}^{\mathrm{d}}_t(s, J \cdot \underline{f}(\alpha), a), \forall s \in \mathcal{S}, \alpha \in [0,1], a \in \mathcal{A},$$

provides a lower-bound for the optimal value function $q^\star$ defined in Eq. (5). Moreover, it can be used to build a policy $\underline{\pi} := (\underline{\pi}_t)_{t=0}^T$ that achieves this lower-bound.

[1] Right-continuity and non-decreasingness can be obtained by ensuring $\bar{q}_t(s,\alpha,a) = \bar{q}_t(s,\alpha',a)$ if $\bar{f}(\alpha) = \bar{f}(\alpha')$.

Algorithm 1 presents a procedure that constructs the policy $\underline{\pi}$ described in Proposition 4.5.

Algorithm 1: Static VaR Policy Execution

```
Input: $s_0 \in \mathcal{S}, \alpha_0 \in (0,1), T, J \in \mathbb{N}$,
       $\underline{q}^{\mathrm{d}} : [T] \times \mathcal{S} \times [J-1] \times \mathcal{A} \to \bar{\mathbb{R}}$
$(s, j) \leftarrow (s_0, \lfloor J \cdot \alpha_0 \rfloor)$
for $t = T, \ldots, 1$ do
    $a^\star \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \underline{q}^{\mathrm{d}}_t(s, j, a)$
    Execute $a^\star$ and observe $r$ and $s'$
    $\tau \leftarrow \gamma^{-1}(\underline{q}^{\mathrm{d}}_t(s, j, a^\star) - r)$
    $\mathcal{J} \leftarrow \left\{ j' \in [J-1] \mid \max_{a' \in \mathcal{A}} \underline{q}^{\mathrm{d}}_{t-1}(s', j', a') \geq \tau \right\}$
    $j \leftarrow J - 1$ ; // arbitrary initialization
    if $\mathcal{J}$ is not empty then
        $j \leftarrow \min \mathcal{J}$
    $s \leftarrow s'$
```

Note that we limit our analysis to schemes that calculate a posteriori error bounds due to the discretization of α. It is also important to study a priori structural error bounds, which can guide the choice of the discretization. However, such analysis is beyond the scope of this work.

4.2 VaR-Q-learning Algorithm

We now use the Bellman operators from Section 4.1 to develop and analyze a new Q-learning algorithm that solves the VaR objective in a model-free way.

An important obstacle in developing the Q-learning algorithm is that the Bellman operators in Section 4.1 are set-valued and lack a unique solution. The operators are set-valued because the quantile is not unique. As a result, even a well-designed iterative algorithm may oscillate among multiple possible solutions. It is common in RARL to replace the quantile loss function with Huber's loss to guarantee differentiability (Dabney et al., 2018a); however, as we show in Appendix C.5, Huber's loss is insufficient to guarantee the uniqueness of value function. Instead, we replace the loss ℓ_α of Eq. (3) with the *soft-quantile loss* $\ell^\kappa_\alpha \colon \mathbb{R} \to \mathbb{R}$ defined for $\kappa \in (0,1]$ and $\alpha \in (0,1)$ as

$$\ell^\kappa_\alpha(\delta) := \begin{cases} \frac{(1-\alpha)\kappa}{2} \left((\delta + \kappa)^2 - \frac{2\delta}{\kappa} - 1 \right) & \text{if } \delta < -\kappa, \\ (1-\alpha) \left(\frac{\delta^2}{2\kappa} \right) & \text{if } \delta \in [-\kappa, 0), \\ \alpha \left(\frac{\delta^2}{2\kappa} \right) & \text{if } \delta \in [0, \kappa), \\ \frac{\alpha\kappa}{2} \left((\delta - \kappa)^2 + \frac{2\delta}{\kappa} - 1 \right) & \text{if } \delta \geq \kappa. \end{cases} \tag{13}$$

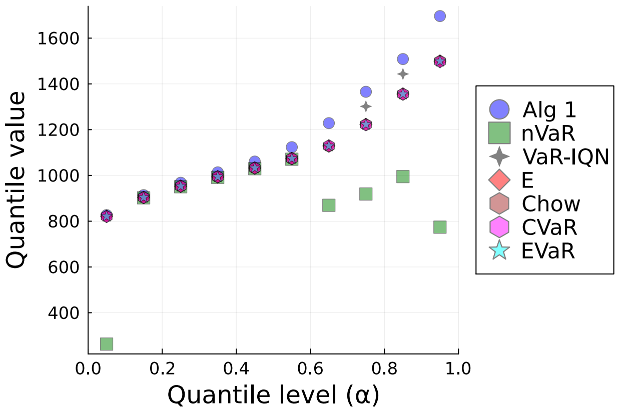

Figure 1: Policy performance $\rho(\pi)$ on INV2.

We also need the derivative $\partial \ell^\kappa_\alpha$ of ℓ^κ_α, which is

$$\partial \ell^\kappa_\alpha(\delta) := \begin{cases} (1-\alpha) \left(\kappa\delta + \kappa^2 - 1 \right) & \text{if } \delta < -\kappa, \\ \frac{1-\alpha}{\kappa} \delta & \text{if } \delta \in [-\kappa, 0), \\ \frac{\alpha}{\kappa} \delta & \text{if } \delta \in [0, \kappa), \\ \alpha \left(\kappa\delta - \kappa^2 + 1 \right) & \text{if } \delta \geq \kappa. \end{cases} \tag{14}$$

As the following lemma states, the function ℓ^κ_α is strongly convex and has a Lipschitz-continuous gradient. These properties are instrumental in showing the value function's uniqueness and analyzing our Q-learning algorithm.

Lemma 4.6. *The function* $m \mapsto \mathbb{E}[\ell^\kappa_\alpha(\tilde{x} - m)]$ *with* $\tilde{x}$ *discrete is* μ-strongly convex *and has an* L-Lipschitz continuous derivative *for each* $\alpha \in (0,1), \kappa \in (0,1]$ *with*

$$\mu = \min\{\alpha, 1-\alpha\}\, \kappa, \qquad L = \max\{\alpha, 1-\alpha\}\, \kappa^{-1}.$$

Algorithm 2: VaR-Q-learning Algorithm

```
Input: Step sizes $\beta_i$, stream of sampled
       transitions $(t_i, s_i, j_i, a_i, s_i')$, for all $i \in \mathbb{N}$
Init: $\underline{q}^{\mathrm{d}}_0(b) \leftarrow t\underline{R},\ \forall b \in [T] \times \mathcal{S} \times [J-1] \times \mathcal{A}$
for $i = 0, 1, 2, \ldots$ do
    $b_i \leftarrow (t_i, s_i, j_i, a_i)$
    if $j_i > 0$ and $t_i > 0$ then
        $\underline{q}^{\mathrm{d}}_{i+1}(b_i) \leftarrow \underline{q}^{\mathrm{d}}_i(b_i) + \frac{\beta_i}{J} \sum_{j'=0}^{J-1} \partial \ell^\kappa_{\frac{j_i}{J}} \big( r(s_i, a_i)$
            $+ \gamma \cdot \max_{a' \in \mathcal{A}} \underline{q}^{\mathrm{d}}_i(t_i - 1, s_i', j', a') - \underline{q}^{\mathrm{d}}_i(b_i) \big)$
    else
        // Effectively do nothing
        $\underline{q}^{\mathrm{d}}_{i+1}(b_i) \leftarrow \underline{q}^{\mathrm{d}}_i(b_i) + \beta_i(t_i\underline{R} - \underline{q}^{\mathrm{d}}_i(b_i))$
```

Having introduced the soft-quantile function in (13), we are now ready to adapt the Bellman operators to ensure that the Q-learning algorithm converges to a unique solution. In particular, we replace $\mathcal{B}^{\mathrm{d}}_{\mathrm{u}}$ with B^{d}_κ

which is defined for $b = (t, s, j, a)$, $t \in 1{:}T$, $s \in \mathcal{S}$, $j \in 1{:}J-1$, and $a \in \mathcal{A}$ as

$$(B^{\mathrm{d}}_{\kappa} q)(b) \;:=\; \operatorname*{argmin}_{x\in\mathbb{R}} \\ \mathbb{E}^{a,s}\left[\ell^{\kappa}_{\frac{j}{J}}\left(r(s,a) + \gamma \max_{a'\in\mathcal{A}} q(t-1, \tilde{s}_1, \tilde{j}', a') - x\right)\right],$$

and we replace the lower-bound operator $\underline{\mathcal{B}}^{\mathrm{d}}_{\mathrm{u}}$ by $\underline{B}^{\mathrm{d}}_{\kappa}$:

$$(\underline{B}^{\mathrm{d}}_{\kappa} q)(b) := \begin{cases} \underline{R}\cdot t & \text{if } j = 0 \vee t = 0, \\ (B^{\mathrm{d}}_{\kappa} q)(b) & \text{otherwise.} \end{cases}$$

Note that the operators B^{d}_{κ} and $\underline{B}^{\mathrm{d}}_{\kappa}$ are not calligraphic because their objective functions possess unique minimizers. In addition, B^{d}_{κ} and $\underline{B}^{\mathrm{d}}_{\kappa}$ are applied to value functions defined across all time steps simultaneously. This representation is convenient because we focus on the finite-horizon objective and must separate the time step from the Q-learning iteration. That is, $\underline{q}^{\mathrm{d}}_i(t, s, j, a)$ represents the value function in the i-th iteration evaluated at time t, state s, risk level j, and action a.

Equipped with the above definitions, we now introduce the *VaR-Q-learning* algorithm in Algorithm 2. The algorithm seeks to identify the fixed-point $\underline{q}^{\mathrm{d}} = \underline{B}^{\mathrm{d}}_{\kappa}\underline{q}^{\mathrm{d}}$, which is unique, as we show in Appendix C. The algorithm adapts the standard Q-learning approach to the risk-averse setting. It is well known that the standard Q-learning algorithm can be seen as a stochastic gradient descent on the quadratic loss function (e.g., Asadi et al. (2023)). Algorithm 2 also follows a sequence of stochastic gradient steps, but it replaces the quadratic loss function with the soft-quantile loss function ℓ^{α}_{κ}.

Algorithm 2 takes a stream of samples as input, and thus, implies that it is an offline algorithm. However, the algorithm and its analysis also apply to the online setting in which the sample $(t_i, s_i, j_i, a_i, s'_i)$ and step size β_i can depend on the values $\underline{q}^{\mathrm{d}}_0, \ldots, \underline{q}^{\mathrm{d}}_i$ and are generated during the execution of the algorithm.

In an actual implementation, Line 6 in Algorithm 2 may be omitted since it does not modify the value function. We include this step to simplify our convergence analysis, which considers each update as a stochastic gradient step towards a contractive minimizer.

We require the following standard assumption to prove the convergence of Algorithm 2.

Assumption 4.7. The input to Algorithm 2 satisfies $\forall i \in \mathbb{N}$:

$$\mathbb{P}\left[\tilde{s}'_i = s' \mid \mathcal{G}_{i-1}, \tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i, \tilde{\beta}_i\right] = p(\tilde{s}_i, \tilde{a}_i, s'),\ \forall s' \in \mathcal{S},$$

almost surely, where $\mathcal{G}_{i-1} := (\tilde{\beta}_l, (\tilde{t}_l, \tilde{s}_l, \tilde{j}_l, \tilde{a}_l, \tilde{s}'_l))_{l=0}^{i-1}$.

The following theorem shows that Algorithm 2 enjoys convergence guarantees that are comparable to those in standard Q-learning.

Theorem 4.8. *Let $\kappa \in (0, 1]$. Assume that the sequences $\{\tilde{\beta}_i\}_{i=0}^{\infty}$ and $\{(\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i, \tilde{s}'_i)\}_{i=0}^{\infty}$ used in Algorithm 2 satisfy Assumption 4.7 and the step-size conditions*

$$\sum_{i\in\tilde{\mathfrak{I}}(t,s,j,a)}^{\infty} \tilde{\beta}_i = \infty, \qquad \sum_{i\in\tilde{\mathfrak{I}}(t,s,j,a)} \tilde{\beta}_i^2 < \infty, \quad a.s.\ ,$$

where $\tilde{\mathfrak{I}}(t, s, j, a) := \{i \in \mathbb{N} \mid (\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i) = (t, s, j, a)\}$. Then, the sequence $(\tilde{\underline{q}}^{\mathrm{d}}_i)_{i=0}^{\infty}$ produced by Algorithm 2 converges almost surely to $\underline{q}^{\mathrm{d}}_{\infty}$ such that $\underline{q}^{\mathrm{d}}_{\infty} = \underline{B}^{\mathrm{d}}_{\kappa}\underline{q}^{\mathrm{d}}_{\infty}$.

The proof of Theorem 4.8 follows an approach similar to that in the proofs of standard Q-learning (Bertsekas and Tsitsiklis, 1996) with two main differences. First, the algorithm converges even when $\gamma = 1$ and $\underline{B}^{\mathrm{d}}_{\kappa}$ is not an L_∞ contraction. Instead, we show that $\underline{B}^{\mathrm{d}}_{\kappa}$ is a contraction w.r.t. a particular weighted norm. Second, the use of a non-quadratic function ℓ^{κ}_{α} requires a more careful choice of the step-sizes than the standard analysis. Moreover, our analysis of the non-quadratic function extends the Q-learning analysis for risk-sensitive RL with nested risk measures in Shen et al. (2014).

5 NUMERICAL EXPERIMENTS

In this section, we empirically test our theoretical results and algorithms on 7 tabular domains: machine replacement (MR) (Delage and Mannor, 2010), gamblers ruin (GR) (Bäuerle and Ott, 2011; Li et al., 2022), two inventory management problems (INV1 and INV2) (Ho et al., 2021), riverswim (RS) (Strehl and Littman, 2008), population (POP) (Kéry and Schaub, 2011; Tirinzoni et al., 2018), and cliffwalk (CW) (Sutton and Barto, 2018). We set the horizon to $T = 100$ with $\gamma = 0.9$ to evaluate the risk of the random discounted return. More details on each experiment can be found in Appendix D and our code can be found at `https://github.com/MonkieDein/DRA-Q-LA`.

Policy execution. We first validate the discretization scheme presented in Algorithm 1 for model-free policy execution. To this aim, we compare the performance of Algorithm 1 (Alg 1) with other risk-averse algorithms. As a standard baseline, we include the risk-neutral objective (E). Other baselines are nested VaR (nVaR) (Ruszczyński, 2010), conditional VaR (CVaR) (Bäuerle and Ott, 2011), Chow (Chow et al., 2015), distributional VaR (VaR-IQN) (Dabney et al., 2018a,b), and entropic VaR (EVaR) (Hau et al., 2023b). We take a quantile discretization level of $J = 4096$ to train Algorithm 1, Chow, and VaR-IQN. Detail of all algorithms can be found in Appendix D.2.

π	CW	INV1	INV2	MR	POP	RS	GR
$\bar{q}^{\mathrm{d}}$	-9.11	237.19	970.08	-2.79	-14348.60	50.0	4.78
Algorithm 1	**-9.11**	**237.02**	**968.01**	**-2.84**	**-14348.60**	**50.0**	**4.78**
$\underline{q}^{\mathrm{d}}$	-9.11	236.88	967.60	-2.85	-14348.60	50.0	4.78
nVaR	-87.20	202.46	950.60	-20.00	**-14348.60**	**50.0**	0.00
VaR-IQN	-9.20	234.61	950.60	-18.21	**-14348.60**	**50.0**	0.00
E	-9.72	234.43	953.00	-2.96	-15101.04	33.3	3.14
Chow	-9.72	232.18	952.72	-4.13	**-14348.60**	**50.0**	3.14
CVaR	-9.72	235.27	953.07	-3.00	**-14348.60**	**50.0**	3.14
EVaR	-9.72	234.90	953.00	-2.96	**-14348.60**	**50.0**	2.82

Table 1: 25%-quantile of policy return: $\rho(\pi)$ for $\mathrm{VaR}_{0.25}$.

Table 1 shows the 25%-quantile value obtained after training, where each entry is the performance obtained from 100,000 episodes generated from the final policy. As we can see, our algorithm consistently outperforms all other algorithms across all tested domains. We also test our algorithm on a range of quantile levels $\alpha_0 \in \{0.05, 0.15, \dots, 0.85, 0.95\}$. Fig. 1 shows the quantile value obtained on INV2 after training each baseline. Our method shows an insensitive behavior to risk levels. All other domains exhibit a similar trend across quantile levels (see Appendix D), thus illustrating the robustness of Algorithm 1 to different environments and risk levels.

We perform an ablative study to understand how the discretization and selection of $\underline{q}^{\mathrm{d}}$ in Algorithm 1 contribute to the solution quality. We compare the performance of $\underline{\pi}$ with that of $\bar{\pi}$ (defined analogously to Eq. (10)) by confronting them to the bounds $\underline{q}^{\mathrm{d}}$ and $\bar{q}^{\mathrm{d}}$ from Example 4.4. We take $J \in \{16, 256, 4096\}$. Fig. 2 demonstrates that the performance of $\underline{\pi}$ lies within $[\underline{q}^{\mathrm{d}}, \bar{q}^{\mathrm{d}}]$, whereas $\bar{\pi}$ sometimes performs worse than $\underline{q}$ on INV2. Furthermore, as the discretization level increases, the bounding gap $\bar{q}^{\mathrm{d}} - \underline{q}^{\mathrm{d}}$ shrinks, suggesting that $\underline{\pi}$ converges to $\pi^\star$.

Quantile Q-learning. We now check the convergence and performance of Algorithm 2, which approximates the VaR computed from DP. In Section 4.2, we introduced a general VaR-Q-learning handling sampled time horizons. However, in practice, we remove the time index to reduce the computation overhead associated with updating time-indexed value functions. We take $\kappa \in \{10^{-4}, 10^{-8}, 10^{-12}, 0\}$ for the κ-soft quantile loss with a uniform discretization of $J = 256$. For $\kappa = 0$, the loss is that of Eq. (3) while for positive values, it is that of Eq. (13). Fig. 3 displays the 1-Wasserstein distance between the quantile value estimated from VaR-Q-learning and the quantile value $\underline{q}^{\mathrm{d}}$ computed via DP (Eq. (12)). For all κ's, we see that the distance converges to zero as the number of samples increases. Furthermore, the VaR-Q-learning policy performs similarly to DP (see also Fig. 8 in Appendix D).

To summarize, our experiments illustrate that the policy returned by the VaR-Q-learning algorithm: (1) Outperforms other baselines across both domains and quantile levels; (2) Lies in $[\underline{q}, \bar{q}]$; and (3) Performs similarly as the DP optimal policy $\underline{\pi}$.

6 RELATED WORK AND DISCUSSION

Several works propose model-free methods for RARL. Mihatsch and Neuneier (2002) introduce a temporal difference scheme for prediction and control with convergence guarantees, but focus on a specific utility-based shortfall risk. Converging Q-learning algorithms are further extended in Borkar and Chandak (2021); Shen et al. (2014) to a larger class of utility functions. However, all these works focus on a *nested* risk measure, which provides Bellman equations at the expense of interpretable policies. Differently, Stanko and Macek (2021) study Q-learning for *static* CVaR, but their analysis relies on the DP equations of Chow et al. (2015), which were shown to be incorrect by Hau et al. (2023a). When considering static risk measures, one must use a proper state augmentation to guarantee that an optimal policy is identified (Bäuerle and Ott, 2011; Hau et al., 2023a). Otherwise, limiting assumptions such as the existence of an optimal Markov policy are required (Lim and Malik, 2022).

Dabney et al. (2018a) present a similar goal as our study and propose to train risk-sensitive policies using a quantile representation of the return distribution. By leveraging Q-learning for distributional RL (Bellemare et al., 2017), they introduce IQN, an algorithm capable of achieving risk-sensitive behavior. Yet, as first pointed out in Lim and Malik (2022), the greedy step employed in IQN lacks a clear criterion of optimality for the trained policy. In this regard, our Q-learning algorithm modifies the optimal action selected at each

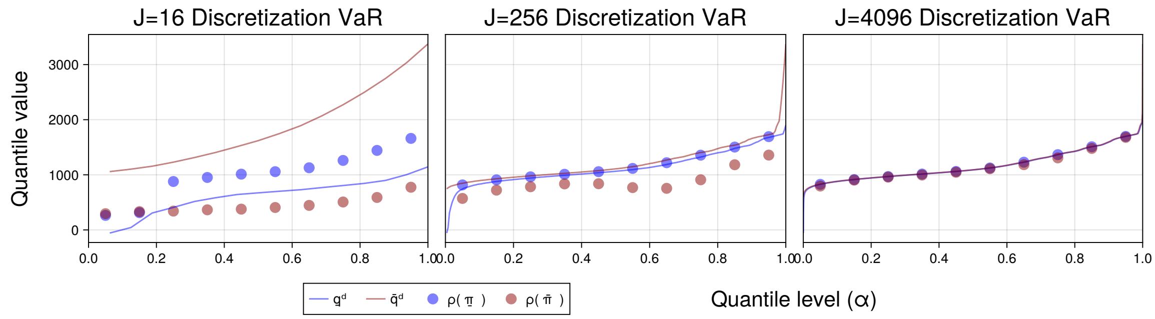

Figure 2: Impact of discretization level J on VaR-MDP performance and Q-functions.

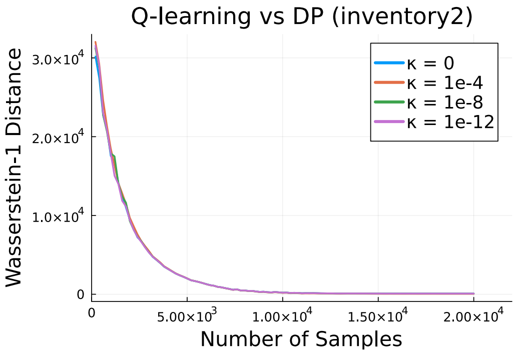

Figure 3: Distance between Q-values of Algorithm 2 and DP value function for varying κ.

state-risk level pair, thus fixing IQN's deficiency in policy optimization. For policy evaluation, our algorithm reduces to a variant of IQN ensuring that the resulting approximated distribution functions under-estimate (in terms of stochastic ordering) the return distribution. A detailed discussion on the differences between IQN and our VaR-Q-learning algorithm can be found in Appendix E.

Most related to our work is the one by Gilbert and Weng (2016) in which the authors propose a Q-learning algorithm to identify a VaR-optimal policy on a special class of MDPs with end states. There, preferences are expressed using an ordering over end states. Following Borkar (1997), their algorithm is based on stochastic approximation with two time-scales and its convergence is only empirically demonstrated. They leave open the question of how to generalize the approach to other forms of MDPs and raise the question of whether quantile regression methods could be used, which we address it in this work.

Looking forward, the question of extending our results to an infinite horizon setting with continuous state and/or action spaces is definitely interesting. One might also be able to adapt the convergence analysis of policy evaluation for distributional RL in Rowland et al. (2024) to formally establish the convergence properties of Algorithm 2 under the non-strongly convex objective ℓ_α.

Acknowledgements

We would like to thank Marc G. Bellemare for valuable discussions on the topic of distributional RL and the IQN algorithm. Esther Derman was partially funded by IVADO. 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]. Jia Lin Hau was partially funded by the University of New Hampshire dissertation year fellowship 2024-25. Jia Lin Hau and Marek Petrik were partially supported by NSF grants 2144601 and 2218063.

References

Ahmadi, M., Xiong, X., and Ames, A. D. (2021). Risk-averse control via CVaR barrier functions: Application to bipedal robot locomotion. *IEEE Control Systems Letters*, 6:878–883.

Alexander, C. and Sarabia, J. M. (2012). Quantile uncertainty and value-at-risk model risk. *Risk Analysis: An International Journal*, 32(8):1293–1308.

Asadi, K., Sabach, S., Liu, Y., Gottesman, O., and Fakoor, R. (2023). TD convergence: An optimization perspective. *Advances in Neural Information Processing Systems*, 37.

Bäuerle, N. and Ott, J. (2011). Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379.

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458.

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

Ben-Tal, A. and Teboulle, M. (2007). An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294.

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

Borkar, V. S. and Chandak, S. (2021). Prospect-theoretic q-learning. *Systems & Control Letters*, 156:105009.

Braun, D. A., Nagengast, A. J., and Wolpert, D. M. (2011). Risk-sensitivity in sensorimotor control. *Frontiers in Human Neuroscience*, 5:1.

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

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

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2015). Risk-sensitive and robust decision-making: a CVaR optimization approach. *Advances in Neural Information Processing Systems*, 28.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. (2018a). Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, pages 1096–1105.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. (2018b). Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, volume 32.

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. *ACM SIGOPS Operating Systems Review*, 41(6):205–220.

Delage, E. and Mannor, S. (2010). Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213.

Embrechts, P., Liu, H., and Wang, R. (2018). Quantile-based risk sharing. *Operations Research*, 66(4):936–949.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk*, 18(2):31–60.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447.

Follmer, H. and Schied, A. (2016). *Stochastic Finance: Introduction in Discrete Time*. De Gruyter Graduate, fourth edition.

Gilbert, H. and Weng, P. (2016). Quantile reinforcement learning. *arXiv preprint arXiv:1611.00862*.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Hakobyan, A. and Yang, I. (2021). Wasserstein distributionally robust motion control for collision avoidance using conditional value-at-risk. *IEEE Transactions on Robotics*, 38(2):939–957.

Hau, J. L., Delage, E., Ghavamzadeh, M., and Petrik, M. (2023a). On dynamic programming decompositions of static risk measures in Markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hau, J. L., Petrik, M., and Ghavamzadeh, M. (2023b). Entropic risk optimization in discounted MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 47–76.

Ho, C. P., Petrik, M., and Wiesemann, W. (2021). Partial policy iteration for L1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46.

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

Jin, I. G., Schürmann, B., Murray, R. M., and Althoff, M. (2019). Risk-aware motion planning for automated vehicle among human-driven cars. In *2019 American Control Conference (ACC)*, pages 3987–3993. IEEE.

Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. (2020). Being optimistic to be conservative: Quickly learning a CVaR policy. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic press.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926.

Kiumarsi, B., Vamvoudakis, K. G., Modares, H., and Lewis, F. L. (2017). Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6):2042–2062.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33.

Köse, Ü. E. (2016). *Optimal timing of living-donor liver transplantation under risk-aversion*. PhD thesis, Bilkent Universitesi (Turkey).

Li, X., Zhong, H., and Brandeau, M. L. (2022). Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447.

Lim, S. H. and Malik, I. (2022). Distributional reinforcement learning for risk-sensitive policies. *Advances in Neural Information Processing Systems*, 35:30977–30989.

Lin, Y., Wu, C., and Kang, B. (2003). Optimal models with maximizing probability of first achieving target value in the preceding stages. *Science in China Series A: Mathematics*, 46:396–414.

Luo, Y., Liu, G., Poupart, P., and Pan, Y. (2024). An alternative to variance: Gini deviation for risk-averse policy gradient. *Advances in Neural Information Processing Systems*, 36.

Majumdar, A. and Pavone, M. (2017). How should a robot assess risk. *Towards an Axiomatic Theory of Risk in Robotics*, pages 75–84.

Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–290.

Min, S., Moallemi, C. C., and Maglaras, C. (2022). Risk-sensitive optimal execution via a conditional value-at-risk objective. *arXiv preprint arXiv:2201.11962*.

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer, 2nd edition.

on Banking Supervision, B. C. (2023). The Basel framework. *Basel III*.

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

Rockafellar, R. T. and Wets, R. J. (2009). *Variational Analysis*. Springer.

Rowland, M., Munos, R., Azar, M. G., Tang, Y., Ostrovski, G., Harutyunyan, A., Tuyls, K., Bellemare, M. G., and Dabney, W. (2024). An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25(163):1–47.

Rudin, W. et al. (1964). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.

Ruszczyński, A. (2010). Risk-averse dynamic programming for Markov decision processes. *mathematical Programming*, 125:235–261.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

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

Singh, A., Halpern, Y., Thain, N., Christakopoulou, K., Chi, E., Chen, J., and Beutel, A. (2020). Building healthy recommendation sequences for everyone: A safe reinforcement learning approach. In *FAccTRec Workshop*.

Stanko, S. and Macek, K. (2021). CVaR Q-learning. In *Computational Intelligence: 11th International Joint Conference, IJCCI 2019, Vienna, Austria, September 17–19, 2019, Revised Selected Papers*, pages 333–358. Springer.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for Markov decision processes. *Elsevier*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tamar, A., Di Castro, D., and Mannor, S. (2012). Policy gradients with variance related risk criteria. In *international Conference on Machine Learning*, pages 387–396.

Tirinzoni, A., Petrik, M., Chen, X., and Ziebart, B. (2018). Policy-conditioned uncertainty sets for robust markov decision processes. *Advances in neural information processing systems*, 31.

Wang, Y. and Chapman, M. P. (2022). Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control. *Artificial Intelligence*, 311:103743.

Weber, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16(2):419–441.

Wei, Y., Kehm, R. D., Goldberg, M., and Terry, M. B. (2019). Applications for quantile regression in epidemiology. *Current Epidemiology Reports*, 6:191–199.

Wu, C. and Lin, Y. (1999). Minimizing risk models in Markov decision processes with policies depending on target values. *Journal of Mathematical Analysis and Applications*, 231(1):47–67.

Yoo, G., Park, J., and Woo, H. (2024). Risk-conditioned reinforcement learning: A generalized approach for adapting to varying risk measures. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 16513–16521.

Q-learning for Quantile MDPs: A Decomposition, Performance, and Convergence Analysis Supplementary Materials

A PROOFS OF SECTION 2

A.1 Proof of Lemma 2.1

Proof. We aim to show that for $\tilde{x} \in \mathbb{X}, \alpha \in [0,1]$,

$$\operatorname*{argmin}_{y\in\mathbb{R}} \mathbb{E}[\max(\alpha(\tilde{x}-y), -(1-\alpha)(\tilde{x}-y))] = [\mathfrak{q}^-_\alpha(\tilde{x}), \mathfrak{q}^+_\alpha(\tilde{x})] \cap \mathbb{R},$$

where $[\mathfrak{q}^-_0(\tilde{x}), \mathfrak{q}^+_0(\tilde{x})] \cap \mathbb{R} = (-\infty, \mathfrak{q}^+_0(\tilde{x})]$ and $[\mathfrak{q}^-_1(\tilde{x}), \mathfrak{q}^+_1(\tilde{x})] = [\mathfrak{q}^-_0(\tilde{x}), \infty)$. For the case $\alpha \in (0,1)$, we refer the reader to (Gneiting, 2011, Thm. 9). We study the case $\alpha = 0$ as a similar set of arguments holds when $\alpha = 1$. By definition of $\mathfrak{q}^+_0(\tilde{x})$, for all $\bar{y} \in \mathbb{R}$ such that $\bar{y} \le \mathfrak{q}^+_0(\tilde{x})$, $\mathbb{P}[\tilde{x} < \bar{y}] \le \mathbb{P}\left[\tilde{x} < \mathfrak{q}^+_0(\tilde{x})\right] \le 0$ (i.e., equals zero) and therefore, by the law of total probability:

$$0 \le \mathbb{E}[\ell_0(\tilde{x}-\bar{y})] = \mathbb{E}[0|\tilde{x} \ge \bar{y}]\mathbb{P}[\tilde{x} \ge \bar{y}] + \mathbb{E}[-(\tilde{x}-\bar{y})|\tilde{x} < \bar{y}]\mathbb{P}[\tilde{x} < \bar{y}] = 0.$$

Hence, $\operatorname{argmin}_{y\in\mathbb{R}} \mathbb{E}[\ell_0(\tilde{x}-y)] \supseteq (-\infty, \mathfrak{q}^+_0(\tilde{x})]$.

By definition of $\mathfrak{q}^+_0(\tilde{x})$, for all $\bar{y} \in \mathbb{R}$ such that $\bar{y} > \mathfrak{q}^+_0(\tilde{x})$, $\mathbb{P}[\tilde{x} < \bar{y}] > 0$ and:

$$\mathbb{E}[\ell_0(\tilde{x}-\bar{y})] = \mathbb{E}[0|\tilde{x} \ge \bar{y}]\mathbb{P}[\tilde{x} \ge \bar{y}] + \mathbb{E}[-(\tilde{x}-\bar{y})|\tilde{x} < \bar{y}]\mathbb{P}[\tilde{x} < \bar{y}] = \mathbb{E}[\bar{y}-\tilde{x}|\tilde{x} < \bar{y}]\mathbb{P}[\tilde{x} < \bar{y}] > 0.$$

The last inequality exploits two facts. First, by left continuity of $h(z) := \mathbb{P}[\tilde{x} < \bar{y} + z]$, $\mathbb{P}[\tilde{x} < \bar{y}] > 0$ implies that there must be some $\epsilon > 0$ for which $\mathbb{P}[\tilde{x} \le \bar{y} - \epsilon] > 0$. Second, by the Markov inequality $\mathbb{E}[\bar{y}-\tilde{x}|\bar{y}-\tilde{x} > 0] \ge \epsilon\mathbb{P}[\bar{y}-\tilde{x} \ge \epsilon|\bar{y}-\tilde{x} > 0] = \epsilon\mathbb{P}[\bar{y}-\tilde{x} \ge \epsilon]/\mathbb{P}[\bar{y}-\tilde{x} > 0] > 0$. Hence, $\operatorname{argmin}_{y\in\mathbb{R}} \mathbb{E}[\ell_0(\tilde{x}-y)] \subseteq (-\infty, \mathfrak{q}^+_0(\tilde{x})]$. We have shown $\operatorname{argmin}_{y\in\mathbb{R}} \mathbb{E}[\ell_0(\tilde{x}-y)] = (-\infty, \mathfrak{q}^+_0(\tilde{x})]$, which ends the proof for $\alpha = 0$ since $\mathfrak{q}^-_0(\tilde{x}) = -\infty$. □

A.2 From VaR state-action to VaR state value function

Proposition A.1. *The optimal value functions satisfy for each $s \in \mathcal{S}$, $\alpha \in [0,1]$, and $t \in [T]$ that*

$$v^\star_t(s,\alpha) \;=\; \max_{a\in\mathcal{A}} q^\star_t(s,\alpha,a).$$

Proof. The result follows straightforwardly from the definition of $v^\star_t$ and $q^\star_t$. Namely,

$$v^\star_t(s,\alpha) = \max_{\pi\in\Pi^t_{\mathrm{HD}}} \operatorname{VaR}^{\pi,s}_\alpha\left[\sum_{k=0}^{t-1}\gamma^k r(\tilde{s}_k,\tilde{a}_k)\right] = \max_{a\in\mathcal{A},\pi\in\Pi^t_{\mathrm{HD}}:\pi_0(s)=a} \operatorname{VaR}^{\pi,s}_\alpha\left[\sum_{k=0}^{t-1}\gamma^k r(\tilde{s}_k,\tilde{a}_k)\right] = \max_{a\in\mathcal{A}} q^\star_t(s,\alpha,a).$$

□

B PROOFS OF SECTION 3

B.1 Proof of Theorem 3.1

Proof. This proof extends the results obtained in (Hau et al., 2023a, Appx. C) to the case where there exist (s,a,s') tuples for which $p(s,a,s')=0$, and reparameterizes the representation. The result on the definition and optimality of $\pi^\star$ comes directly from (Hau et al., 2023a). Specifically, by (Hau et al., 2023a, Appx. C), it holds that $\ddot{q}_t = q_t^\star, \forall t \in [T]$, for all sequences $\ddot{q} := (\ddot{q}_t)_{t=0}^T$ such that $\ddot{q}_0(s,\alpha,a) = \mathrm{VaR}_\alpha[0]$ and

$$\ddot{q}_{t+1}(s,\alpha,a) \;=\; r(s,a) + \gamma \max_{\zeta\in\Xi_{sa}(\alpha)} \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} \ddot{q}_t\left(s', \frac{\alpha\zeta_{s'}}{p(s,a,s')}, a'\right), \quad \forall t\in[T-1],$$

where

$$\Xi_{sa}(\alpha) := \{\zeta\in[0,1]^S \mid \sum_{s'\in\mathcal{S}} \zeta_{s'} = 1, \alpha\zeta_{s'} \le p(s,a,s'), \quad \forall s'\in\mathcal{S}\}.$$

By construction of $q = (q_t)_{t=0}^T$ from the theorem statement, we have $q_0 = \ddot{q}_0$, so it remains to establish $B_{\max} q_t = \ddot{q}_{t+1}$ for all $t\in[T-1]$. By mathematical induction, assuming that $q_t = \ddot{q}_t$ for some $t\in[T-1]$, we show that $q_{t+1} := B_{\max} q_t = \ddot{q}_{t+1}$ in two steps : (1) $B_{\max} q_t \le \ddot{q}_{t+1}$ and (2) $B_{\max} q_t \ge \ddot{q}_{t+1}$.

Step 1: Establishing $B_{\max} q_t \le \ddot{q}_{t+1}$.

Let $o^\star \in \mathcal{O}_{sa}(\alpha)$ (which is non-empty given that $0\in\mathcal{O}_{sa}(\alpha)$) be an optimal point for operator $B_{\max}$ in Eq. (6) and consider the function:

$$\begin{aligned} g(o) :&= \min_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} q_t(s', o_{s'}, a') \\ &= \min_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} q_t^\star(s', o_{s'}, a') \\ &= \min_{s'\in\mathcal{S}} \max_{\pi\in\Pi_{\mathrm{HD}}^t} \mathrm{VaR}_{o_{s'}}^{\pi,s'}\left[\sum_{k=0}^{t-1} \gamma^k \cdot r(\tilde{s}_k,\tilde{a}_k)\right]. \end{aligned}$$

By properties of VaR and min operators, g is non-decreasing in o. Therefore, $\max_{o\in\mathcal{O}_{sa}(\alpha)} g(o)$ is achieved inside $\{o\in[0,1]^S \mid \sum_{s'\in\mathcal{S}} o_{s'} p(s,a,s') = \alpha\}$ and necessarily, $\sum_{s'\in\mathcal{S}} o^\star_{s'} p(s,a,s') = \alpha$.

Now let $\zeta\in\mathbb{R}^S$ be defined as:

$$\zeta_{s'} := \begin{cases} \frac{o^\star_{s'} p(s,a,s')}{\alpha} & \text{if } \alpha > 0 \\ p(s,a,s') & \text{otherwise.} \end{cases}$$

for all $s'\in\mathcal{S}$. We aim to show that $\zeta\in\Xi_{sa}(\alpha)$. When $\alpha=0$, the claim follows by definition of a transition kernel p. When $\alpha>0$, we get

- $\sum_{s'\in\mathcal{S}} \zeta_{s'} = \frac{\sum_{s'\in\mathcal{S}} o^\star_{s'} p(s,a,s')}{\alpha} = 1$
- $p(s,a,s') \ge 0, o^\star_{s'} \ge 0 \implies \zeta_{s'} = \frac{o^\star_{s'} p(s,a,s')}{\alpha} \ge 0 \;, \forall s'\in\mathcal{S}$
- $(\zeta_{s'} \ge 0 \;\forall s'\in\mathcal{S}\;,\; \sum_{s'\in\mathcal{S}} \zeta_{s'} = 1) \implies (\zeta_{s'} \le 1 \;\forall s'\in\mathcal{S})$
- $o^\star_{s'} \le 1 \implies \alpha\zeta_{s'} = o^\star_{s'} p(s,a,s') \le p(s,a,s') \;, \forall s'\in\mathcal{S}$

In both cases, $\alpha\zeta_{s'} = o^\star_{s'} p(s,a,s'), \forall s'\in\mathcal{S}$. Indeed, for $\alpha=0$, the condition $\sum_{s'\in\mathcal{S}} o^\star_{s'} p(s,a,s') = \alpha$ implies

$o^\star_{s'} = 0$ whenever $p(s,a,s') > 0$, so $o^\star_{s'} p(s,a,s') = 0 = \alpha\zeta_{s'}$ for all $s' \in \mathcal{S}$. We can thus deduce:

$$\begin{aligned}
B_{\max} q_t(s,\alpha,a) &= r(s,a) + \gamma \max_{o\in\mathcal{O}_{sa}(\alpha)} \min_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} q_t(s', o_{s'}, a') \\
&= r(s,a) + \gamma \min_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} q_t(s', o^\star_{s'}, a') \\
&\le r(s,a) + \gamma \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t(s', o^\star_{s'}, a') \\
&= r(s,a) + \gamma \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t\left(s', \frac{\alpha\zeta_{s'}}{p(s,a,s')}, a'\right) \\
&\le r(s,a) + \gamma \max_{\zeta\in\Xi_{sa}(\alpha)} \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t\left(s', \frac{\alpha\zeta_{s'}}{p(s,a,s')}, a'\right) = \ddot{q}_{t+1}(s,\alpha,a).
\end{aligned}$$

Step 2: Establishing $B_{\max} q_t \ge \ddot{q}_{t+1}$. We proceed similarly. Let an optimal $\zeta^\star \in \Xi_{sa}(\alpha)$ (it exists since $\zeta := p(s,a,\cdot)$ is always feasible) satisfying:

$$\ddot{q}_{t+1}(s,\alpha,a) = r(s,a) + \gamma \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t\left(s', \frac{\alpha\zeta^\star_{s'}}{p(s,a,s')}, a'\right),$$

and define $o \in \mathbb{R}^S$ as

$$o_{s'} := \begin{cases} \frac{\alpha\zeta^\star_{s'}}{p(s,a,s')} & \text{if } p(s,a,s') > 0, \\ 1 & \text{otherwise.} \end{cases}$$

To check if $o \in \mathcal{O}_{sa}(\alpha)$, we remark that for any $s' \in \mathcal{S}$ with $p(s,a,s') > 0$,

$$\begin{aligned}
\frac{\alpha\zeta^\star_{s'}}{p(s,a,s')} &\ge 0 && [\alpha \ge 0, \zeta^\star \ge 0, p(s,a,s') > 0] \\
\frac{\alpha\zeta^\star_{s'}}{p(s,a,s')} &\le 1, && [\alpha\zeta^\star_{s'} \le p(s,a,s')]
\end{aligned}$$

so $o_{s'} \in [0,1]$ when $p(s,a,s') > 0$. Otherwise, $o_{s'} = 1 \in [0,1]$. Additionally,

$$\sum_{s'\in\mathcal{S}} o_{s'} p(s,a,s') = \sum_{s'\in\mathcal{S}:p(s,a,s')>0} \frac{\alpha\zeta^\star_{s'}}{p(s,a,s')} \cdot p(s,a,s') = \sum_{s'\in\mathcal{S}:p(s,a,s')>0} \alpha\zeta^\star_{s'} = \alpha,$$

so $o \in \mathcal{O}_{sa}(\alpha)$. We can thus establish:

$$\begin{aligned}
B_{\max} q_t(s,\alpha,a) &= r(s,a) + \gamma \max_{o\in\mathcal{O}_{sa}(\alpha)} \min_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} q_t(s', o_{s'}, a') \\
&\ge r(s,a) + \gamma \min_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} q_t(s', o_{s'}, a') \\
&\overset{\text{(a)}}{=} r(s,a) + \gamma \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t(s', o_{s'}, a') \\
&= r(s,a) + \gamma \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t\left(s', \frac{\alpha\zeta^\star_{s'}}{p(s,a,s')}, a'\right) \\
&= r(s,a) + \gamma \max_{\zeta\in\Xi_{sa}(\alpha)} \min_{s'\in\mathcal{S}:p(s,a,s')>0} \max_{a'\in\mathcal{A}} q_t\left(s', \frac{\alpha\zeta_{s'}}{p(s,a,s')}, a'\right) \\
&= \ddot{q}_{t+1}(s,\alpha,a).
\end{aligned}$$

Equality (a) stems from the implications: $p(s,a,s') = 0 \implies o_{s'} = 1 \implies q_t(s', o_{s'}, a') = \infty$ and for any $c \in \bar{\mathbb{R}}$, $\min(\infty, c) = c$. We now conclude the proof by induction, as $q_{t+1} := B_{\max} q_t = \ddot{q}_{t+1} \forall t \in [T-1]$. □

B.2 Proof of Theorem 3.2

Before diving into the proof of the theorem, we show the following general lemmas which will eventually be used for establishing Theorem 3.2.

Lemma B.1. *For all $t \in [T]$ and $s \in \mathcal{S}$, $\alpha \mapsto v_t^\star(s, \alpha)$ is non-decreasing, right-continuous on $[0, 1]$.*

Proof. Proposition A.1 indicates that for any $t \in [T]$,

$$v_t^\star(s, \alpha) = \mathrm{VaR}_\alpha^{\pi^\star, s}\left[\sum_{k=0}^{t-1} \gamma^k r(\tilde{s}_k, \tilde{a}_k)\right].$$

By (Follmer and Schied, 2016, Lem. A.19), $\alpha \mapsto \mathrm{VaR}_\alpha[\tilde{x}]$ is non-decreasing and right-continuous for any $\tilde{x} \in \mathbb{X}$, so the result follows. □

Lemma B.2. *Let $f : [0, 1] \to \bar{\mathbb{R}}$ be non-decreasing and $\tilde{u}$ be a uniform random variable over $[0, 1]$. Then, we have that $\mathrm{VaR}_\alpha[f(\tilde{u})] = f(\alpha)$ for all $\alpha \in [0, 1)$ where $f(\alpha)$ is right-continuous. Moreover, if $f(1) = \infty$, the equality also holds at $\alpha = 1$.*

Proof. By definition of VaR (see Eq. (2)), we have: $\mathrm{VaR}_\alpha\left[f(\tilde{u})\right] = \max\{z \in \bar{\mathbb{R}} | \mathbb{P}\left[f(\tilde{u}) < z\right] \leq \alpha\}$. When $\alpha = 1$, $\mathbb{P}\left[f(\tilde{u}) < z\right] \leq \alpha$ for all $z \in \bar{\mathbb{R}}$ so that $\mathrm{VaR}_\alpha[f(\tilde{u})] = \infty$ and the second part of the statement holds. Let thus $\alpha \in [0,\ 1)$ be such that $f(\alpha)$ is right-continuous at α. By assumption on f being non-decreasing, $\tilde{u} \geq \alpha$ implies $f(\tilde{u}) \geq f(\alpha)$ and we can establish:

$$\begin{aligned}
& \mathbb{P}\left[\tilde{u} \geq \alpha\right] \leq \mathbb{P}\left[f(\tilde{u}) \geq f(\alpha)\right] \\
\iff & 1 - \mathbb{P}\left[\tilde{u} < \alpha\right] \leq 1 - \mathbb{P}\left[f(\tilde{u}) < f(\alpha)\right] \\
\iff & 1 - \alpha \leq 1 - \mathbb{P}\left[f(\tilde{u}) < f(\alpha)\right] && \left[\mathbb{P}\left[\tilde{u} < \alpha\right] = \mathbb{P}\left[\tilde{u} \leq \alpha\right] = \alpha\right] \\
\iff & \alpha \geq \mathbb{P}\left[f(\tilde{u}) < f(\alpha)\right].
\end{aligned}$$

As a result, $f(\alpha) \leq \mathrm{VaR}_\alpha\left[f(\tilde{u})\right]$.

On the other hand, the right-continuity of f at α ensures that for all $\epsilon > 0$ there exists a $\delta > 0$ such that $f(\alpha + \delta) < f(\alpha) + \epsilon$. Thus

$$\begin{aligned}
\mathbb{P}\left[f(\tilde{u}) < f(\alpha) + \epsilon\right] & \geq \mathbb{P}\left[f(\tilde{u}) \leq f(\alpha + \delta)\right] && \left[\text{By construction: } f(\alpha + \delta) < f(\alpha) + \epsilon\right] \\
& \geq \mathbb{P}\left[\tilde{u} \leq \alpha + \delta\right] && \left[\tilde{u} \leq \alpha + \delta \implies f(\tilde{u}) \leq f(\alpha + \delta)\right] \\
& = \alpha + \delta \\
& > \alpha.
\end{aligned}$$

Hence, $\mathrm{VaR}_\alpha\left[f(\tilde{u})\right] \leq f(\alpha) + \epsilon$ for all $\epsilon > 0$. Setting $\epsilon \to 0$, $\mathrm{VaR}_\alpha\left[f(\tilde{u})\right] \leq f(\alpha)$.

We conclude that $\mathrm{VaR}_\alpha\left[f(\tilde{u})\right] = f(\alpha)$ for all $\alpha \in [0,\ 1)$ at which $f(\alpha)$ is right-continuous. □

The result below directly follows from Lemmas B.1 and B.2.

Corollary B.3. *For all $t \in [T]$ and $s \in \mathcal{S}$, we have $\mathrm{VaR}_\alpha[v_t^\star(s, \tilde{u})] = v_t^\star(s, \alpha)$, where $\tilde{u}$ is a uniform random variable on $[0, 1]$.*

Lemma B.4. *Let $S \in \mathbb{N}$ non-decreasing functions $f_i \colon [0, 1] \to \bar{\mathbb{R}}, i \in 1{:}S$ with each $f_i(\alpha) \in \mathbb{R}$ for all $\alpha \in (0, 1)$. Let also $\tilde{y}$ be a discrete random variable on $1{:}S$ with probability mass function $\hat{p}_i := \mathbb{P}[\tilde{y} = i], \quad \forall i \in 1{:}S$, and $\tilde{u}$ an independent random variable with uniform distribution on $[0, 1]$. Then, we have*

$$\mathrm{VaR}_\alpha\left[f_{\tilde{y}}(\tilde{u})\right] \quad = \max_{\boldsymbol{o} \in [0,1]^S} \left\{ \min_{i \in 1:S} \mathrm{VaR}_{o_i}\left[f_i(\tilde{u})\right] \mid \sum_{j=1}^{S} o_j \hat{p}_j \leq \alpha \right\}. \tag{15}$$

Proof. This proof closely follows that of (Hau et al., 2023a, Thm. 5.1) but relaxes the assumption $\hat{p}_i > 0$ and simplifies the notation. Let $\mathfrak{I} := \{i \in 1{:}S \mid \hat{p}_i > 0\}$. We decompose VaR based on its definition in Eqs. (1) and (2):

$$\begin{aligned}
\operatorname{VaR}_\alpha\left[f_{\tilde{y}}(\tilde{u})\right] &= \max\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[f_{\tilde{y}}(\tilde{u}) < \tau\right] \le \alpha\right\} \\
&\overset{\text{(b)}}{=} \max\left\{\tau \in \bar{\mathbb{R}} \mid \sum_{i\in\mathfrak{I}} \mathbb{P}\left[f_{\tilde{y}}(\tilde{u}) < \tau \mid \tilde{y} = i\right] \cdot \hat{p}_i \le \alpha\right\} \\
&= \max\left\{\tau \in \bar{\mathbb{R}} \mid \sum_{i\in\mathfrak{I}} \mathbb{P}\left[f_i(\tilde{u}) < \tau\right] \cdot \hat{p}_i \le \alpha\right\} \\
&\overset{\text{(c)}}{=} \max\left\{\tau \in \bar{\mathbb{R}} \mid \exists \boldsymbol{o} \in [0,1]^S,\ \mathbb{P}\left[f_i(\tilde{u}) < \tau\right] \le o_i,\ \forall i \in \mathfrak{I},\ \sum_{j\in\mathfrak{I}} o_j \hat{p}_j \le \alpha\right\} \\
&= \max\left\{\tau \in \bar{\mathbb{R}} \mid \exists \boldsymbol{o} \in [0,1]^S,\ \mathbb{P}\left[f_i(\tilde{u}) < \tau\right] \le o_i,\ \forall i \in \mathfrak{I},\ \sum_{j=1}^{S} o_j \hat{p}_j \le \alpha\right\} \\
&\overset{\text{(d)}}{=} \max_{\boldsymbol{o}\in[0,1]^S}\left\{\max\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[f_i(\tilde{u}) < \tau\right] \le o_i,\ \forall i \in \mathfrak{I}\right\} \mid \sum_{j=1}^{S} o_j \hat{p}_j \le \alpha\right\} \\
&= \max_{\boldsymbol{o}\in[0,1]^S}\left\{\max \bigcap_{i\in\mathfrak{I}} \left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[f_i(\tilde{u}) < \tau\right] \le o_i\right\} \mid \sum_{j=1}^{S} o_j \hat{p}_j \le \alpha\right\} \\
&\overset{\text{(e)}}{=} \max_{\boldsymbol{o}\in[0,1]^S}\left\{\min_{i\in\mathfrak{I}} \max\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[f_i(\tilde{u}) < \tau\right] \le o_i\right\} \mid \sum_{j=1}^{S} o_j \hat{p}_j \le \alpha\right\} \\
&\overset{\text{(f)}}{=} \max_{\boldsymbol{o}\in[0,1]^S}\left\{\min_{i\in\mathfrak{I}}\ \operatorname{VaR}_{o_i}\left[f_i(\tilde{u})\right] \mid \sum_{j=1}^{S} o_j \hat{p}_j \le \alpha\right\}.
\end{aligned} \tag{16}$$

In the derivation above, step (b) follows from the law of total probability and omitting zero probability events. Then we lower-bound them by an auxiliary variable o_i in step (c). In step (d) we replace the joint maximum over τ and $\boldsymbol{o}$ by sequential max, and then we replace the max of an intersection by the minimum of the maxima of sets in (e). The equality in (e) holds because $\tau \mapsto \mathbb{P}\left[f_i(\tilde{u}) < \tau\right]$ is monotone and, therefore, the sets $\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[f_i(\tilde{u}) < \tau \mid \tilde{y} = i\right] \le o_i\right\}$ are nested. Step (f) holds by definition of VaR.

It remains to show that (16) equals (15). Suppose that $o^\star \in [0,1]^S$ is optimal in (16) and construct $\bar{o} \in [0,1]^S$ as

$$\bar{o}_i = \begin{cases} o_i^\star & \text{if } i \in \mathfrak{I}, \\ 1 & \text{otherwise,} \end{cases} \qquad \forall i \in 1{:}S.$$

Since $\bar{o}$ is feasible in (15) with the same objective, (16) $\le$ (15). To show that (16) $\ge$ (15) suppose that $o^\star \in [0,1]^S$ is optimal in (15). The inequality then holds because $o^\star$ is feasible in (16) and because

$$\min_{i\in\mathfrak{I}}\ \operatorname{VaR}_{o_i^\star}\left[f_i(\tilde{u})\right] \ge \min_{i\in 1:S}\ \operatorname{VaR}_{o_i^\star}\left[f_i(\tilde{u})\right].$$

□

Lemma B.5. *Assume non-decreasing functions $f_i\colon [0,1] \to \bar{\mathbb{R}}, i \in 1{:}S$ with each $f_i(\alpha) \in \mathbb{R}$ for all $\alpha \in (0,1)$. Then, it holds that $\max_{i\in 1:S} \operatorname{VaR}_{\tilde{u}_1}[f_i(\tilde{u}_2)] = \max_{i\in 1:S} f_i(\tilde{u}_1)$ almost surely, where $\tilde{u}_1$ and $\tilde{u}_2$ are two independent uniform random variables on $[0,1]$.*

Proof. By Lemma B.2, for any $j \in 1{:}S$, $\operatorname{VaR}_{\bar{\alpha}}[f_j(\tilde{u})] = f_j(\bar{\alpha})$ at any value of $\bar{\alpha} \in [0,1)$ where $f_j(\cdot)$ is right-continuous. This implies that for all $\alpha \in [0,1)$ where all f_i's are right-continuous, $\max_{i\in 1:S} \operatorname{VaR}_\alpha[f_i(\tilde{u})] = \max_{i\in 1:S} f_i(\alpha)$ since the maximum is necessarily right-continuous then. For $i \in 1{:}S$, f_i is monotone on the interval $(0,1)$, so by Froda's theorem (Rudin et al., 1964, Thm. 4.30), the number of discontinuities of f_i must be at most countable on $(0,1)$ so therefore also on $[0,1]$. This implies that the number of points α at which some f_i from

$i \in 1{:}S$ is discontinuous is at most countable. We thus conclude that $\max_{i\in 1:S} f_i(\tilde{u}_1) = \max_{i \in 1:S} \mathrm{VaR}_{\tilde{u}_1}[f_i(\tilde{u}_2)]$ with probability one. □

Lemma B.6. *Let S non-decreasing functions $f_i\colon [0,1] \to \bar{\mathbb{R}}, i \in 1{:}S$ with each $f_i(\alpha) \in \mathbb{R}$ for all $\alpha \in (0,1)$. Then $\mathrm{VaR}_\alpha[\max_{i\in 1:S} f_i(\tilde{u})] = \max_{i\in 1:S} \mathrm{VaR}_\alpha[f_i(\tilde{u})]$, where $\tilde{u}$ is a uniform random variable on $[0,1]$*

Proof. First, the claim trivially applies to $\alpha = 1$ since $\mathrm{VaR}_1[\tilde{x}] = \infty$ for all random variables $\tilde{x} \in \mathbb{X}$. Thus, we focus on the case $\alpha \in [0,1)$. Since $\alpha \mapsto \mathrm{VaR}_\alpha[\cdot]$ is non-decreasing (see (Follmer and Schied, 2016, Lem. A.19)), $\mathrm{VaR}_\alpha[\max_{i\in 1:S} f_i(\tilde{u})] \geq \mathrm{VaR}_\alpha[f_j(\tilde{u})]$ for all $j \in 1{:}S$, hence $\mathrm{VaR}_\alpha[\max_{i\in 1:S} f_i(\tilde{u})] \geq \max_{i\in 1:S} \mathrm{VaR}_\alpha[f_i(\tilde{u})]$. We are therefore left with showing that $\mathrm{VaR}_\alpha[\max_{i\in 1:S} f_i(\tilde{u})] \leq \max_{i\in 1:S} \mathrm{VaR}_\alpha[f_i(\tilde{u})]$. We do so by contradiction. Assume that

$$\mathrm{VaR}_\alpha[\max_{i\in 1:S} f_i(\tilde{u})] > \max_{i\in 1:S} \mathrm{VaR}_\alpha[f_i(\tilde{u})] =: \nu^\star.$$

Applying Lemma B.5, $\max_{i\in 1:S} f_i(\tilde{u}) = \max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)]$ almost surely, so we must have $\nu^\star < \mathrm{VaR}_\alpha[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)]]$ where $\tilde{u}_2$ is uniformly distributed on $[0,1]$. By definition of $\mathrm{VaR}_\alpha[\cdot]$ (Eq. (2)), this implies that there exists $\epsilon > 0$ such that:

$$\mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon\right] \leq \alpha.$$

Since $\nu^\star = \max_{i\in 1:S} \mathrm{VaR}_\alpha[f_i(\tilde{u})]$ and $\alpha \mapsto \mathrm{VaR}_\alpha[f_i(\tilde{u})]$ is non-decreasing, we must have $\max_{i\in 1:S} \mathrm{VaR}_{\alpha'}[f_i(\tilde{u})] \leq \nu^\star < \nu^\star + \epsilon$ for all $\alpha' \leq \alpha$. In addition, by the law of total probability:

$$\begin{aligned}
\alpha &\geq \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon\right] \\
&= \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} > \alpha\right] \mathbb{P}\left[\tilde{u} > \alpha\right] \\
&\quad + \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} \leq \alpha\right] \mathbb{P}\left[\tilde{u} \leq \alpha\right] \\
&= (1-\alpha)\mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} > \alpha\right] + \alpha\mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} \leq \alpha\right] \\
&= (1-\alpha)\mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} > \alpha\right] + \alpha,
\end{aligned}$$

so necessarily:

$$\mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} > \alpha\right] = 0.$$

Yet, since $\max_{i\in 1:S} \mathrm{VaR}_{\alpha'}[f_i(\tilde{u}_2)]$ is right-continuous, non-decreasing in α' and evaluates at $\nu^\star$ for $\alpha' = \alpha$, there must be a $\delta > 0$ such that $\max_{i\in 1:S} \mathrm{VaR}_{\alpha'}[f_i(\tilde{u}_2)] \leq \nu^\star + \epsilon$ for all $\alpha' \leq \alpha + \delta$. This leads to a contradiction since:

$$\begin{aligned}
0 < \delta &= \mathbb{P}\left[\tilde{u} \in (\alpha, \alpha+\delta]\right] \\
&= \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} \in (\alpha, \alpha+\delta]\right] \mathbb{P}\left[\tilde{u} \in (\alpha, \alpha+\delta]\right] \\
&= \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon, \tilde{u} \in (\alpha, \alpha+\delta]\right] \\
&\leq \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon, \tilde{u} > \alpha\right] \\
&= \mathbb{P}\left[\max_{i\in 1:S} \mathrm{VaR}_{\tilde{u}}[f_i(\tilde{u}_2)] < \nu^\star + \epsilon | \tilde{u} > \alpha\right] (1-\alpha) = 0.
\end{aligned}$$

□

We are now ready to prove Theorem 3.2, whose statement is recalled below.

Theorem. *Let a sequence $q^{\mathrm{u}} = (q_t^{\mathrm{u}})_{t=0}^{T}$ be such that $q_0^{\mathrm{u}}(s,\alpha,a) = \mathrm{VaR}_\alpha[0]$ and $q_{t+1}^{\mathrm{u}}(s,\alpha,a) := B_{\mathrm{u}} q_t^{\mathrm{u}}(s,\alpha,a)$ for $t \in [T-1]$. Then, $q_t^{\mathrm{u}} = q_t^\star$ for all $t \in [T]$, where $q_t^\star$ is defined in Eq. (5). Moreover, if a policy $\pi^{\mathrm{u}} = (\pi_k^{\mathrm{u}})_{k=0}^{T-1}$ is greedy for q^{u} as in Eq. (10), then it maximizes the value-at-risk objective (4).*

Proof. We start with demonstrating that $q_t^\star = q_t^{\mathrm{u}}$ $,\forall t \in [T]$, recursively with mathematical induction. Assuming that $q_t^{\mathrm{u}} = q_t^\star$ we want to show that $q_{t+1}^{\mathrm{u}} := B_{\mathrm{u}} q_t^{\mathrm{u}} = q_{t+1}^\star$ as written in Eq. (8). We then prove the optimality of $\pi^\star$ constructed using $\hat{\alpha}^{\mathrm{u}}$.

Step 1: For all $s \in \mathcal{S}, \alpha \in [0,1]$ and $a \in \mathcal{A}$, $q_0^\star(s,\alpha,a) = q_0^{\mathrm{u}}(s,\alpha,a) = \mathrm{VaR}_\alpha[0]$ so the base case holds. Assume that $q_t^{\mathrm{u}} = q_t^\star$ for some $t \in [T-1]$. Then, one can derive:

$$\begin{aligned}
q_{t+1}^\star(s,\alpha,a) &= r(s,a) + \gamma \cdot \max_{o \in \mathcal{O}_{sa}(\alpha)} \min_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} q_t^\star(s', o_{s'}, a') && \text{[By Theorem 3.1]} \\
&= r(s,a) + \gamma \cdot \max_{o \in \mathcal{O}_{sa}(\alpha)} \min_{s' \in \mathcal{S}} v_t^\star(s', o_{s'}) && \text{[By Proposition A.1]} \\
&= r(s,a) + \gamma \cdot \max_{o \in \mathcal{O}_{sa}(\alpha)} \min_{s' \in \mathcal{S}} \mathrm{VaR}_{o_{s'}}[v_t^\star(s', \tilde{u})] && \text{[By Corollary B.3]} \\
&= r(s,a) + \gamma \cdot \mathrm{VaR}_\alpha^{a,s}[v_t^\star(\tilde{s}_1, \tilde{u})] && \text{[By Lemma B.4]} \\
&= \mathrm{VaR}_\alpha^{a,s}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q_t^\star(\tilde{s}_1, \tilde{u}, a')] && \text{[By Proposition A.1]} \\
&= \mathrm{VaR}_\alpha^{a,s}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q_t^{\mathrm{u}}(\tilde{s}_1, \tilde{u}, a')] && \text{[By inductive assumption]} \\
&= B_{\mathrm{u}} q_t^{\mathrm{u}}(s,\alpha,a) = q_{t+1}^{\mathrm{u}}(s,\alpha,a) && \text{[By Eq. (8)].}
\end{aligned}$$

This confirms that $q^\star$ satisfies $q_{t+1}^\star = B_{\mathrm{u}} q_t^\star = B_{\mathrm{u}} q_t^{\mathrm{u}} = q_{t+1}^{\mathrm{u}}$ for all $t \in [T-1]$, so that $q^\star = q^{\mathrm{u}}$.

Step 2: We now show that $\hat{\alpha}_k^{\mathrm{u}}(\cdot)$ constructed according to:

$$\hat{\alpha}_{k+1}^{\mathrm{u}}(h_{k+1}) := \min \left\{ o \in [0,1] \mid \max_{a \in \mathcal{A}} q_{T-k-1}^{\mathrm{u}}(s_{k+1}, o, a) \geq \frac{q_{T-k}^{\mathrm{u}}(s_k, \hat{\alpha}_k(h_k), a_k) - r(s_k, a_k)}{\gamma} \right\}$$

defines an optimal policy. Namely,

$$\hat{\alpha}_{k+1}^{\mathrm{u}}(h_{k+1}) := \min \left\{ o \in [0,1] \mid \max_{a \in \mathcal{A}} q_{T-k-1}^{\mathrm{u}}(s_{k+1}, o, a) \geq \frac{q_{T-k}^{\mathrm{u}}(s_k, \hat{\alpha}_k^{\mathrm{u}}(h_k), a_k) - r(s_k, a_k)}{\gamma} \right\}.$$

Based on what has been shown in Step 1, we can interchangeably write q_{T-k-1}^{u} or $q_{T-k-1}^\star$ in the construction of $\hat{\alpha}_k^{\mathrm{u}}, k \in [T-1]$, so that

$$\hat{\alpha}_{k+1}^{\mathrm{u}}(h_{k+1}) := \min \left\{ o \in [0,1] \mid \max_{a \in \mathcal{A}} q_{T-k-1}^\star(s_{k+1}, o, a) \geq \frac{q_{T-k}^\star(s_k, \hat{\alpha}_k^{\mathrm{u}}(h_k), a_k) - r(s_k, a_k)}{\gamma} \right\}.$$

By Lemma B.1, $\alpha \mapsto v_t^\star(s,\alpha)$ is right-continuous and non-decreasing so the minimum above is well-defined. We are left to check that $\hat{\alpha}_k^{\mathrm{u}}(\cdot)$ leads to an associated $\hat{o}^{ka}(h_k) \in \mathcal{O}_{sa}(\hat{\alpha}_k^{\mathrm{u}}(h_k))$ satisfying:

$$q_{T-k}(s, \alpha_k(h_k), a) = r(s,a) + \gamma \cdot \min_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} q_{T-k-1}(s', \alpha_{k+1}(\langle h_k, a, s' \rangle), a') \tag{17}$$

This can be done in two steps.

We can first show that for all $h_k \in \mathcal{H}_k, a \in \mathcal{A}$, the vector $\hat{o}^k \in \mathbb{R}^{\mathcal{S}}$ composed of $\hat{o}_{s'}^k := \hat{\alpha}_{k+1}^{\mathrm{u}}(\langle h_k, a, s' \rangle)$ for all $s' \in \mathcal{S}$, is in $\mathcal{O}_{sa}(\hat{\alpha}_k^{\mathrm{u}}(h_k))$. To do so, we make use of the fact that for all t the maximum over $\max_{o \in \mathcal{O}_{sa}}$ in (6) is achieved thus implying that it is achieved at $T-k$ by some $\hat{o}^{k\star}$ when $s = s_k$ and $\alpha = \hat{\alpha}_k^{\mathrm{u}}(h_k)$. Namely, $\hat{o}^{k\star}$ satisfies:

$$q_{T-k}^\star(s_k, \hat{\alpha}_k^{\mathrm{u}}(h_k), a) = \min_{s' \in \mathcal{S}} r(s_k, a) + \gamma v_{T-k-1}^\star(s', \hat{o}_{s'}^{k\star}) \leq r(s_k, a) + \gamma v_{T-k-1}^\star(s', \hat{o}_{s'}^{k\star})\,,\ \forall s' \in \mathcal{S}$$

where we replaced $\max_{a' \in \mathcal{A}} q_{T-k-1}^\star(s', \hat{o}_{s'}^{k\star}, a')$ with $v_{T-k-1}^\star(s', \hat{o}_{s'}^{k\star})$. We can therefore easily conclude that:

$$\sum_{s' \in \mathcal{S}} \hat{o}_{s'}^k p(s_k, a, s') = \sum_{s' \in \mathcal{S}} \hat{\alpha}_{k+1}^{\mathrm{u}}(\langle h_k, a, s' \rangle) p(s_k, a, s') \sum_{s' \in \mathcal{S}} \hat{o}_{s'}^{k\star} p(s_k, a, s') \leq \hat{\alpha}_k^{\mathrm{u}}(h_k),$$

which establishes that $\hat{o}^k \in \mathcal{O}_{sa}(\hat{\alpha}^{\mathrm{u}}_k(h_k))$.

We can then show that equation (17), with q_{T-k} and q_{T-k-1} respectively replaced by their equivalent $q^\star_{T-k}$ and $q^\star_{T-k-1}$ (based on Theorem 3.1), is satisfied based on:

$$\begin{aligned}
q^\star_{T-k}(s, \hat{\alpha}^{\mathrm{u}}_k(h_k), a) &= \max_{o \in \mathcal{O}_{sa}(\hat{\alpha}^{\mathrm{u}}_k(h_k))} \min_{s' \in \mathcal{S}} \left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q^\star_{T-k-1}(s', o_{s'}, a') \right) && \text{[By Eq. (6)]} \\
&\geq \min_{s' \in \mathcal{S}} \left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q^\star_{T-k-1}(s', \hat{\alpha}^{\mathrm{u}}_{k+1}(\langle h_k, a, s' \rangle), a') \right) && [\hat{o}^k \in \mathcal{O}_{sa}(\hat{\alpha}^{\mathrm{u}}_k(h_k))] \\
&= \min_{s' \in \mathcal{S}} \left(r(s,a) + \gamma \cdot v^\star_{T-k-1}(s', \hat{\alpha}^{\mathrm{u}}_{k+1}(\langle h_k, a, s' \rangle)) \right) && \text{[By Proposition A.1]} \\
&\geq \min_{s' \in \mathcal{S}} \left(r(s,a) + \gamma \cdot \frac{q^\star_{T-k}(s, \hat{\alpha}^{\mathrm{u}}_k(h_k), a) - r(s,a)}{\gamma} \right) && [\text{Definition of } \hat{\alpha}^{\mathrm{u}}_k] \\
&= \min_{s' \in \mathcal{S}} q^\star_{T-k}(s, \hat{\alpha}^{\mathrm{u}}_k(h_k), a) = q^\star_{T-k}(s, \hat{\alpha}^{\mathrm{u}}_k(h_k), a),
\end{aligned}$$

The above inequalities must therefore be equalities, so equation (17) holds for $\hat{\alpha}^{\mathrm{u}}_k(h_k)$. As a result, the policy $\pi^\star$ constructed using $\hat{\alpha}^{\mathrm{u}}$ is optimal. □

C PROOFS OF SECTION 4

C.1 Risk Measures

Definition C.1 (Monetary risk measure)**.** A monetary risk measure is a mapping $\varrho : \mathbb{X} \to \bar{\mathbb{R}}$ satisfying the following properties:

1. *Translation invariance:* For all $\tilde{x} \in \mathbb{X}, c \in \mathbb{R}, \varrho(\tilde{x} + c) = \varrho(x) + c$
2. *Monotonicity:* For all $\tilde{x}, \tilde{y} \in \mathbb{X}, \tilde{x} \leq \tilde{y} \implies \varrho(\tilde{x}) \leq \varrho(\tilde{y})$.

Lemma C.2. *For any monetary risk measure $\varrho \colon \mathbb{X} \to \mathbb{R}$ where $\mathbb{X}$ is defined in a finite outcome space $\omega \in \Omega$, it holds that:*

$$|\varrho(\tilde{x}) - \varrho(\tilde{y})| \leq \max_{\omega \in \Omega} |\tilde{x}(\omega) - \tilde{y}(\omega)|.$$

Proof. Define $\epsilon := \max_{\omega \in \Omega} |\tilde{x}(\omega) - \tilde{y}(\omega)| \geq 0$. We prove that $\varrho(\tilde{x}) - \varrho(\tilde{y}) \leq \epsilon$. The second inequality $\varrho(\tilde{y}) - \varrho(\tilde{x}) \leq \epsilon$ follows analogously. Let $\tilde{z} := \max\{\tilde{x}, \tilde{y}\}$. Then, for all $\omega \in \Omega$:

$$\begin{aligned}
\tilde{x}(\omega) &\leq \max\{\tilde{x}(\omega), \tilde{y}(\omega)\} \\
&= \tilde{z}(\omega) \\
&= \tilde{y}(\omega) + \max\{\tilde{x}(\omega) - \tilde{y}(\omega), 0\} \\
&\leq \tilde{y}(\omega) + |\tilde{x}(\omega) - \tilde{y}(\omega)| && [\tilde{x}(\omega) - \tilde{y}(\omega) \leq |\tilde{x}(\omega) - \tilde{y}(\omega)|, 0 \leq |\tilde{x}(\omega) - \tilde{y}(\omega)|] \\
&\leq \tilde{y}(\omega) + \max_{\omega' \in \Omega} |\tilde{x}(\omega') - \tilde{y}(\omega')| \\
&= \tilde{y}(\omega) + \epsilon.
\end{aligned}$$

As a result, $\tilde{x} \leq \tilde{z} \leq \tilde{y} + \epsilon$. Since ϱ is a monetary risk measure, it is monotonous and translation invariant, so that

$$\varrho(\tilde{x}) \leq \ \varrho(\tilde{z}) \leq \ \varrho(\tilde{y} + \epsilon) = \varrho(\tilde{y}) + \epsilon,$$

and $\varrho(\tilde{x}) - \varrho(\tilde{y}) \leq \epsilon$. □

C.2 Proof of Lemma 4.2

Proof. The inner inequalites follow from

$$\begin{aligned}
\mathfrak{q}^{+}_{\underline{f}(\alpha)}(\tilde{x}) &\le \mathfrak{q}^{+}_{\alpha}(\tilde{x}) && \text{[By monotonicity]}\\
&= \mathrm{VaR}_{\alpha}[\tilde{x}]\\
&= \max\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[\tilde{x} < \tau\right] \le \alpha\right\}\\
&\le \sup\left\{\tau \in \bar{\mathbb{R}} \mid \mathbb{P}\left[\tilde{x} < \tau\right] < \bar{f}(\alpha)\right\}\\
&= \mathfrak{q}^{-}_{\bar{f}(\alpha)}(\tilde{x}). && \text{[(Follmer and Schied, 2016, Appx. A.3, Def. A.24)]}
\end{aligned}$$

The outer inequalities follow from Lemma 2.1. □

C.3 Proof of Theorem 4.3

Proof. The proof is broken down into two parts. First, we address the bounding on $q^\star$, then we demonstrate the stated properties of our constructed policy.

Step 1: Upper and lower bound on $q^\star$. We prove by induction on $t \in [T]$ that for all $(s, \alpha, a) \in \mathcal{S} \times (0,1) \times \mathcal{A}$:

$$t\underline{R} \le \underline{q}^{\mathrm{u}}_t(s,\alpha,a) \le q^\star_t(s,\alpha,a) \le \bar{q}^{\mathrm{u}}_t(s,\alpha,a) \le t\bar{R},$$

or more succinctly, that $t\underline{R} \le \underline{q}^{\mathrm{u}}_t \le q^\star_t \le \bar{q}^{\mathrm{u}}_t \le t\bar{R}$. At $t = 0$, the bounds are obtained by definition given that $\bar{q}^{\mathrm{u}}_0 = \underline{q}^{\mathrm{u}}_0 = q^\star_0 = 0$. Assume that the statement holds for some $t \in [T-1]$ and let's check if the proposition is preserved at $t+1$.

Case 1: $\bar{f}(\alpha) < 1$. For all $\alpha \in (0,1)$ such that $\bar{f}(\alpha) < 1$, we have:

$$\begin{aligned}
q^\star_{t+1}(s,\alpha,a) &= \mathrm{VaR}^{a,s}_{\alpha}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q^\star_t(\tilde{s}_1, \tilde{u}, a')] && \text{[By Theorem 3.2]}\\
&= \mathfrak{q}^{+\,a,s}_{\alpha}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q^\star_t(\tilde{s}_1, \tilde{u}, a')] && \text{[By definition of } \mathrm{VaR}_\alpha\text{]}\\
&\le \mathfrak{q}^{+\,a,s}_{\alpha}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(\tilde{s}_1, \tilde{u}, a')] && \text{[By inductive assumption]}\\
&\le \mathfrak{q}^{-\,a,s}_{\bar{f}(\alpha)}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(\tilde{s}_1, \tilde{u}, a')] && \text{[By Lemma 4.2]}
\end{aligned}$$

Finally, we exploit the elicitability of $\mathfrak{q}_{\bar{f}(\alpha)}$ (see Lemma 2.1), due to the random variable $r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(s', \tilde{u}, a')$ being supported on the interval $[(t+1)\underline{R}, (t+1)\bar{R}]$, to obtain:

$$\begin{aligned}
q^\star_{t+1}(s,\alpha,a) &\le \mathfrak{q}^{-\,a,s}_{\bar{f}(\alpha)}[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(\tilde{s}_1, \tilde{u}, a')]\\
&= \min \operatorname*{argmin}_{q} \mathbb{E}^{a,s}\left[\ell_{\bar{f}(\alpha)}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(\tilde{s}_1, \tilde{u}, a') - q\right)\right]\\
&= \min\left\{(\mathcal{B}^{\bar{f}}_{\mathrm{u}} \bar{q}^{\mathrm{u}}_t)(s,\alpha,a)\right\}\\
&\le \bar{q}^{\mathrm{u}}_{t+1}(s,\alpha,a),
\end{aligned}$$

where the last inequality is due to the fact that $\bar{q}^{\mathrm{u}}_{t+1} \in (\mathcal{B}^{\bar{f}}_{\mathrm{u}} \bar{q}^{\mathrm{u}}_t)$ by construction:

$$\bar{q}^{\mathrm{u}}_{t+1} \in (\mathcal{B}^{\bar{f}}_{\mathrm{u}} \bar{q}^{\mathrm{u}}_t), \quad \underline{q}^{\mathrm{u}}_{t+1} \in (\mathcal{B}^{\underline{f}}_{\mathrm{u}} \underline{q}^{\mathrm{u}}_t), \; \forall\, t \in [T-1]. \tag{18}$$

Moreover, we have that:

$$\operatorname*{argmin}_{q} \mathbb{E}^{a,s}\left[\ell_{\bar{f}(\alpha)}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(\tilde{s}_1, \tilde{u}, a') - q\right)\right] \subset \left(-\infty, (t+1)\bar{R}\right]$$

since it is a quantile of $r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \bar{q}^{\mathrm{u}}_t(s', \tilde{u}, a') \le (t+1)\bar{R}$. This confirms the statement for $\bar{f}(\alpha) < 1$.

Case 2: $\bar{f}(\alpha) = 1$. We instead rely on the following inequalities:

$$\begin{aligned} q^\star_{t+1}(s,\alpha,a) &= \max_{\pi\in\tilde{\Pi}_{\mathrm{HR}}} \mathrm{VaR}_\alpha^{(a,\pi_{1:t}),s}\left[\sum_{k=0}^{t}\gamma^k r(\tilde{s}_k,\tilde{a}_k)\right] \\ &\le \sum_{k=0}^{t}\gamma^k \bar{R} \le (t+1)\bar{R} = \bar{R} + \max_{s\in\mathcal{S},a'\in\mathcal{A}} \bar{q}^{\mathrm{u}}_t(s,1,a') = \bar{q}^{\mathrm{u}}_{t+1}(s,\alpha,a). \end{aligned}$$

A similar set of arguments applies for the lower bound. Namely,

$$\begin{aligned} q^\star_t(s,\alpha,a) &= \mathfrak{q}^{+\,a,s}_{\alpha}[r(s,a) + \gamma\cdot\max_{a'\in\mathcal{A}} q^\star_{t-1}(\tilde{s}_1,\tilde{u},a')] && \text{[By Theorem 3.2]} \\ &\ge \mathfrak{q}^{+\,a,s}_{\alpha}[r(s,a) + \gamma\cdot\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{t-1}(\tilde{s}_1,\tilde{u},a')] && \text{[By inductive assumption]} \\ &\ge \mathfrak{q}^{+\,a,s}_{\underline{f}(\alpha)}[r(s,a) + \gamma\cdot\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{t-1}(\tilde{s}_1,\tilde{u},a')]. && \text{[By Lemma 4.2]} \end{aligned}$$

Similarly, we exploit the elicitability and the fact that $\underline{q}^{\mathrm{u}}_{t+1} \in (\mathcal{B}^{\underline{f}}_{\mathrm{u}}\underline{q}^{\mathrm{u}}_t)$ to write

$$\begin{aligned} q^\star_{t+1}(s,\alpha,a) &\ge \mathfrak{q}^{+\,a,s}_{\underline{f}(\alpha)}[r(s,a) + \gamma\cdot\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{t}(\tilde{s}_1,\tilde{u},a')] \\ &= \max \operatorname*{argmin}_q \mathbb{E}^{a,s}\left[\ell_{\underline{f}(\alpha)}\left(r(s,a) + \gamma\cdot\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_t(\tilde{s}_1,\tilde{u},a') - q\right)\right] \\ &= \max\left\{(\mathcal{B}^{\underline{f}}_{\mathrm{u}}\underline{q}^{\mathrm{u}}_t)(s,\alpha,a)\right\} \\ &\ge \underline{q}^{\mathrm{u}}_{t+1}(s,\alpha,a), \end{aligned}$$

where the difference lies in the second inequality, i.e. one needs to exploit the monotonicity of $\alpha\mapsto\mathfrak{q}^+_\alpha(\tilde{x})$. The case where $\underline{f}(\alpha)=0$ also follows naturally:

$$q^\star_{t+1}(s,\alpha,a) = \max_{\pi\in\tilde{\Pi}_{\mathrm{HR}}} \mathrm{VaR}_\alpha^{(a,\pi_{1:t}),s}\left[\sum_{k=0}^{t}\gamma^k r(\tilde{s}_k,\tilde{a}_k)\right] \ge \sum_{k=0}^{t}\gamma^k \underline{R} \ge (t+1)\underline{R} = \underline{q}^{\mathrm{u}}_{t+1}(s,\alpha,a).$$

This completes our inductive proposition: $t\underline{R} \le \underline{q}^{\mathrm{u}}_t \le q^\star_t \le \bar{q}^{\mathrm{u}}_t \le t\bar{R}$.

Step 2: Bounds on the performance of $\underline{\pi}_k(h_k)$. We start by proving that

$$\max_{a\in\mathcal{A}} \underline{q}^{\mathrm{u}}_T(s_0,\alpha_0,a) \le \mathrm{VaR}^{\underline{\pi},s_0}_{\alpha_0}\left[\sum_{k=0}^{T-1}\gamma^k r(\tilde{s}_k,\tilde{a}_k)\right].$$

The rest follows based on the bounding properties of $\bar{q}^{\mathrm{u}}_T$ and $\underline{q}^{\mathrm{u}}_T$.

First, we can confirm that $\underline{o}^k$ composed using $\underline{o}^k_{s'} := \underline{\alpha}^{\mathrm{u}}_{k+1}(\langle h_k,a,s'\rangle)$ is in $\mathcal{O}_{sa}(\underline{\alpha}^{\mathrm{u}}_k(h_k))$. To do so, we exploit the fact that:

$$\begin{aligned}
\underline{q}^{\mathrm{u}}_{T-k}(s_k, \underline{\alpha}^{\mathrm{u}}_k(h_k), a) &\overset{(18)}{\in} \operatorname*{argmin}_{q\in\mathbb{R}} \mathbb{E}^{a,s_k}\left[\ell_{\underline{f}(\underline{\alpha}^{\mathrm{u}}_k(h_k))}\left(r(s_k,a) + \gamma\cdot\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(\tilde{s}_1, \tilde{u}, a') - q\right)\right] \\
&\overset{4.2}{\leq} \mathrm{VaR}^{a,s_k}_{\underline{f}(\underline{\alpha}^{\mathrm{u}}_k(h_k))}[r(s_k,a) + \gamma \max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(\tilde{s}_1, \tilde{u}_1, a')] \\
&\overset{4.2}{\leq} \mathrm{VaR}^{a,s_k}_{\underline{\alpha}^{\mathrm{u}}_k(h_k)}[r(s_k,a) + \gamma \max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(\tilde{s}_1, \tilde{u}_1, a')] \\
&\overset{B.4}{=} \max_{o\in\mathcal{O}_{s_k a}(\underline{\alpha}^{\mathrm{u}}_k(h_k))} \min_{s'\in\mathcal{S}} r(s_k,a) + \gamma\,\mathrm{VaR}_{o_{s'}}[\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(s', \tilde{u}_1, a')] \\
&= \min_{s'\in\mathcal{S}} r(s_k,a) + \gamma\,\mathrm{VaR}_{o^\star_{s'}}[\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(s', \tilde{u}_1, a')] \\
&\overset{B.6}{=} \min_{s'\in\mathcal{S}} r(s_k,a) + \gamma \max_{a'\in\mathcal{A}} \mathrm{VaR}_{o^\star_{s'}}[\underline{q}^{\mathrm{u}}_{T-k-1}(s', \tilde{u}_1, a')] \\
&\overset{B.2}{=} \min_{s'\in\mathcal{S}} r(s_k,a) + \gamma \max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(s', o^\star_{s'}, a') \\
&\leq r(s_k,a) + \gamma \max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(s', o^\star_{s'}, a') \quad ,\forall s'\in\mathcal{S}
\end{aligned}$$

for optimal $o^\star \in \mathcal{O}_{s_k a}(\underline{\alpha}^{\mathrm{u}}_k(h_k))$. This implies that

$$\max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{T-k-1}(s', o^\star_{s'}, a') \geq \frac{\underline{q}^{\mathrm{u}}_{T-k}(s_k, \underline{\alpha}^{\mathrm{u}}_k(h_k), a) - r(s_k,a)}{\gamma}$$

By construction of $\underline{o}^k$, we have that $\underline{o}^k_{s'} \leq o^\star_{s'}$, hence

$$\sum_{s'\in\mathcal{S}} \underline{o}^k_{s'} p(s_k,a,s') \leq \sum_{s'\in\mathcal{S}} o^\star_{s'} p(s_k,a,s') \leq \underline{\alpha}^{\mathrm{u}}_k(h_k).$$

Let us now construct a series of policies $\underline{\pi}^t_k$ defined as:

$$\underline{\pi}^t_k(h_k) \in \operatorname*{argmax}_{a\in\mathcal{A}} \underline{q}^{\mathrm{u}}_{t-k}(s_k, \underline{\alpha}^{\mathrm{u},t}_k(h_k), a),$$

where $\underline{\alpha}^{\mathrm{u},t-1}_0(s') := \underline{\alpha}^{\mathrm{u},t}_1(\langle s,a,s'\rangle)$ and $\underline{\pi}^{t-1}_{k-1}(h_{1:k}) := \underline{\pi}^t_k(h_k)$ are defined without loss of generality for all k, α, t, h_k. The superscript t represents the horizon of the sub-problem, we may omit the superscript when it represent the final objective decision horizon T as $\underline{\pi} = \underline{\pi}^T$ and $\underline{\alpha}^{\mathrm{u}} = \underline{\alpha}^{\mathrm{u},T}$.

Letting

$$v^\pi_t(s,\alpha) := \mathrm{VaR}^{\pi,s}_\alpha[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k)] \tag{19}$$

and $\underline{v}^{\mathrm{u}}_t(s,\alpha) := \max_{a\in\mathcal{A}} \underline{q}^{\mathrm{u}}_t(s,\alpha,a)$, we now prove by induction on t that $v^{\underline{\pi}^t}_t(s,\alpha) \geq \underline{v}^{\mathrm{u}}_t(s,\alpha)\ \forall t\in[T]$. This is obviously the case at $t=0$ since $v^\pi_0(s,\alpha) = 0 = \underline{q}^{\mathrm{u}}_0(s,\alpha,a)$ for all s, a, $\alpha\in(0,1)$, and π. Now assuming that for some $t\in 1:T$, $v^{\underline{\pi}^{t-1}}_{t-1}(s,\alpha) \geq \underline{v}^{\mathrm{u}}_{t-1}(s,\alpha)$, letting $\underline{\alpha}^{\mathrm{u},T}_0(s) = \alpha$ and $a = \underline{\pi}^t_0(s)$ for $v^{\underline{\pi}^t}_t(s,\alpha)$ we obtain:

$$\begin{aligned}
v^{\underline{\pi}^t}_t(s,\alpha) &\overset{(19)}{=} \mathrm{VaR}^{\underline{\pi}^t,s}_\alpha(\sum_{k=0}^{t-1} \gamma^k r(\tilde{s}_k, \tilde{a}_k)) \\
&\overset{B.4}{=} \max_{o\in\mathcal{O}_{sa}(\alpha)} \min_{s'\in\mathcal{S}} r(s,a) + \gamma\,\mathrm{VaR}^s_{o_{s'}}[\sum_{k=1}^{t-1} \gamma^{k-1} r(\tilde{s}_k, \pi^t_k(\tilde{h}_k)|\tilde{a}_0 = a, \tilde{s}_1 = s']
\end{aligned}$$

Let rewrite the second term as follows

$$\begin{aligned}
&\mathrm{VaR}^{s}_{o_{s'}}\left[\sum_{k=1}^{t-1}\gamma^{k-1} r(\tilde{s}_k, \underline{\pi}^t_k(\tilde{h}_k)) | \tilde{a}_0 = a, \tilde{s}_1 = s'\right]\\
&= \mathrm{VaR}^{s}_{o_{s'}}\left[\sum_{k=1}^{t-1}\gamma^{k-1} r(\tilde{s}_k, \underline{\pi}^{t-1}_{k-1}(h_{1:k} \mid \underline{\alpha}_0^{\mathrm{u},t-1}(\tilde{s}_1) = \underline{\alpha}_1^{\mathrm{u},t}(\langle s, \tilde{a}_0, \tilde{s}_1\rangle))) \mid \tilde{a}_0 = a, \tilde{s}_1 = s'\right]\\
&= \mathrm{VaR}^{s'}_{o_{s'}}\left[\sum_{k'=0}^{t-2}\gamma^{k'} r(\tilde{s}_{k'}, \underline{\pi}^{t-1}_{k'}(h_{k'}))\right]\\
&= \mathrm{VaR}^{\underline{\pi}^{t-1},s'}_{o_{s'}}\left[\sum_{k'=0}^{t-2}\gamma^{k'} r(\tilde{s}_{k'}, \tilde{a}_{k'})\right]
\end{aligned}$$

Now we have

$$\begin{aligned}
v_t^{\underline{\pi}^t}(s, \alpha) &= \max_{o \in \mathcal{O}_{sa}(\alpha)} \min_{s' \in \mathcal{S}} r(s,a) + \mathrm{VaR}^{\underline{\pi}^{t-1},s'}_{o_{s'}}\left[\sum_{k'=0}^{t-2}\gamma^{k'} r(\tilde{s}_{k'}, \tilde{a}_{k'})\right] && \text{[From derivation above]}\\
&\geq \min_{s' \in \mathcal{S}} r(s,a) + \mathrm{VaR}^{\underline{\pi}^{t-1},s'}_{\underline{\alpha}_1^{\mathrm{u},t}(\langle s,a,s'\rangle)}\left[\sum_{k'=0}^{t-2}\gamma^{k'} r(\tilde{s}_{k'}, \tilde{a}_{k'})\right] && \text{[Property of max]}\\
&= \min_{s' \in \mathcal{S}} r(s,a) + \gamma \underline{v}^{\underline{\pi}^{t-1}}_{t-1}(s', \underline{\alpha}_1^{\mathrm{u},t}(\langle s,a,s'\rangle)) && \text{[By Eq. (19)]}\\
&\geq \min_{s' \in \mathcal{S}} r(s,a) + \gamma \underline{v}^{\mathrm{u}}_{t-1}(s', \underline{\alpha}_1^{\mathrm{u},t}(\langle s,a,s'\rangle)) && \text{[By Inductive assumption]}\\
&\geq \min_{s' \in \mathcal{S}} \underline{v}^{\mathrm{u}}_{t}(s, \alpha) = \underline{v}^{\mathrm{u}}_{t}(s, \alpha) && \text{[By Eq. (18)].}
\end{aligned}$$

We could conclude with induction that:

$$\mathrm{VaR}^{\underline{\pi},s_0}_{\alpha_0}\left[\sum_{k=0}^{T-1}\gamma^{k} r(\tilde{s}_k, \tilde{a}_k)\right] = v_T^{\underline{\pi}^T}(s_0, \alpha_0) \geq \underline{v}^{\mathrm{u}}_{T}(s_0, \alpha_0) = \max_{a \in \mathcal{A}} \underline{q}^{\mathrm{u}}_{T}(s_0, \alpha_0, a).$$

□

C.4 Proof of Proposition 4.5

Proof. This result follows by induction from showing that the constructed $\underline{q}$ satisfies the conditions identified in Theorem 4.3 under $\underline{f}$ as defined in Example 4.4. Naturally, the initial condition is satisfied:

$$\underline{q}_0(s, \alpha, a) = \underline{q}^{\mathrm{d}}_0(s, J \cdot \underline{f}(\alpha), a) = 0.$$

To prove the inductive step tor any $t \in 1{:}T-1$, we have that if $\underline{f}(\alpha) = 0$ then :

$$\underline{q}_{t+1}(s, \alpha, a) = \underline{q}^{\mathrm{d}}_{t+1}(s, 0, a) = \underline{R} + \min_{s \in \mathcal{S}, a \in \mathcal{A}} \underline{q}^{\mathrm{d}}_t(s, 0, a) = \underline{R} + \min_{s \in \mathcal{S}, a \in \mathcal{A}} \underline{q}_t(s, 0, a)$$

since $\underline{f}(0) = 0$. If $\underline{f}(\alpha) = j/J$ with $j \geq 1$, then

$$\begin{aligned}
\underline{q}_{t+1}(s, \alpha, a) &= \underline{q}^{\mathrm{d}}_{t+1}(s, J \cdot \underline{f}(\alpha), a)\\
&\in \operatorname*{argmin}_{x \in \mathbb{R}} \frac{1}{J}\sum_{j'=0}^{J-1} \mathbb{E}^{a,s}\left[\ell_{j/J}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \underline{q}^{\mathrm{d}}_t(\tilde{s}_1, j', a') - x\right)\right]\\
&= \operatorname*{argmin}_{x \in \mathbb{R}} \frac{1}{J}\sum_{j'=0}^{J-1} \mathbb{E}^{a,s}\left[\ell_{j/J}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \underline{q}^{\mathrm{u}}_t(\tilde{s}_1, j'/J, a') - x\right)\right]\\
&= \operatorname*{argmin}_{x \in \mathbb{R}} \mathbb{E}^{a,s}\left[\ell_{j/J}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} \underline{q}^{\mathrm{u}}_t(\tilde{s}_1, \underline{f}(\tilde{u}), a') - x\right)\right]\\
&= \mathcal{B}^{\underline{f}}_{\mathrm{u}} \underline{q}_t.
\end{aligned}$$

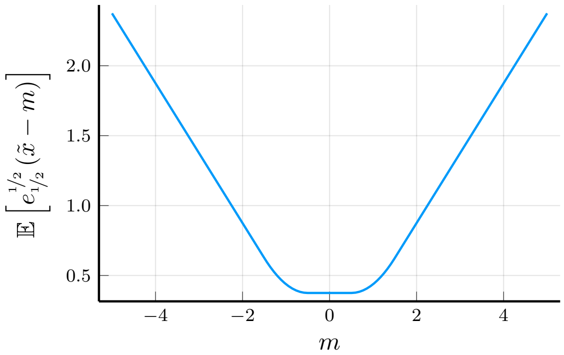

Figure 4: A example used to prove the non-uniqueness of an optimal solution in Proposition C.3.

Moreover, since $\underline{f}$ is right-continuous and non-decreasing $\underline{q}$ is right-continuous and non-decreasing by construction. □

C.5 Why not Huber's Loss

A common differentiable function used in quantile regression is the *Huber's* quantile regression loss function and is commonly defined as (and in general differs from the Moreau envelope of the quantile loss function):

$$e_\alpha^\kappa(\delta) \ := \begin{cases} -(1-\alpha)(\delta+\kappa)+\frac{1}{2}(1-\alpha)\kappa & \text{if } \delta < -\kappa \\ (1-\alpha)\left(\frac{\delta^2}{2\kappa}\right) & \text{if } \delta \in [-\kappa, 0] \\ \alpha\left(\frac{\delta^2}{2\kappa}\right) & \text{if } \delta \in (0,\kappa) \\ \alpha(\delta-\kappa)+\frac{1}{2}\alpha\kappa & \text{if } \delta > \kappa \end{cases} \tag{20}$$

Although some definitions of Huber's loss are scaled by a positive constant compared with (20), such scaling does not affect the minimizer.

The next proposition demonstrates that it is not sufficient to consider the popular Huber's loss function in place of the quantile loss function. Suppose that one defines a risk measure as

$$\xi(\tilde{x}) := \operatorname*{argmin}_{m\in\mathbb{R}} \mathbb{E}[e_\alpha^\kappa(\tilde{x}-m)] \tag{21}$$

where Huber's loss e_α^κ is defined in (20).

Proposition C.3. *The minimization problem in* (21) *may not have a unique solution.*

Proof. Consider a risk level of $\alpha = 0.5$, $\kappa = 0.5$, and random variable $\tilde{x}$ such that $\mathbb{P}[\tilde{x} = -1] = \mathbb{P}[\tilde{x} = 1] = 1/2$. Then the risk measure defined in (21) becomes

$$\xi(\tilde{x}) \ := \ \operatorname*{argmin}_{m\in\mathbb{R}} \mathbb{E}[e_\alpha^\kappa(\tilde{x}-m)] \ = \ \operatorname*{argmin}_{m\in\mathbb{R}} \mathbb{E}\left[e_{1/2}^{1/2}(\tilde{x}-m)\right].$$

Simple algebraic manipulation shows that the minimization above does not have a unique optimal solution and instead $m \in [-1/2, 1/2]$ is optimal, as Fig. 4 illustrates. □

C.6 Proof of Lemma 4.6

Before proving Lemma 4.6, first define the notions it invokes.

Definition C.4. Given $L, \mu \in \mathbb{R}_{++}$, a differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ is *μ-strongly convex* if

$$f(x) \ \geq \ f(y) + f'(y)(x-y) + \frac{1}{2}\mu(x-y)^2, \qquad \forall x, y \in \mathbb{R}.$$

It has an *L-Lipschitz continuous derivative* if

$$|f'(x) - f'(y)| \;\le\; L|x-y|, \qquad \forall x, y \in \mathbb{R}.$$

We always have $L \ge \mu$ (Nesterov, 2018, Thm. 2.1.10).

The next lemma focuses on the properties of the loss function, after which we can prove Lemma 4.6.

Lemma C.5. *Suppose that $\kappa \in (0,1]$. Then the function ℓ_α^κ is continuously differentiable and satisfies for all $\delta_1, \delta_2 \in \mathbb{R}$ that*

$$\min\{\alpha, 1-\alpha\}\;\kappa \;\le\; \frac{|\partial\ell_\alpha^\kappa(\delta_2) - \partial\ell_\alpha^\kappa(\delta_1)|}{|\delta_2 - \delta_1|} \;\le\; \frac{\max\{\alpha, 1-\alpha\}}{\kappa}. \tag{22}$$

Moreover, the function ℓ_α^κ is strongly convex *with constant* $\min\{\alpha, 1-\alpha\}\,\kappa$ *and has a Lipschitz-continuous derivative with constant* $\max\{\alpha, 1-\alpha\}\,\kappa^{-1}$.

Proof. The inequality in (22) follows from the fact that $\partial\ell_\alpha^\kappa$, derived in (14), is a piecewise linear function and $\kappa \le 1$. Then, the function ℓ_α^κ is strongly convex from the strong monotonicity of its derivative in (14) by (Rockafellar and Wets, 2009, Exercise 12.59). The function ℓ_α^κ has a Lipschitz continous derivative immediately by the upper bound in (22). □

Proof of Lemma 4.6. By Lemma C.5 and (Nesterov, 2018, Lem. 2.1.6), we have that the objective in (24) is also strongly convex with constant μ given in the lemma. By Lemma 4.6 and the linearity of the expectation operator, the derivative of (24) is also Lipschitz continuous with constant L given in the lemma. □

C.7 Proof of Theorem 4.8

C.7.1 Properties of Soft-Quantile Measure

Definition C.6. A risk measure $\hat{\mathfrak{q}}_\alpha^\kappa$ is a *shortfall risk measure* (see Föllmer and Schied (2002); Weber (2006)) defined as:

$$\hat{\mathfrak{q}}_\alpha^\kappa(\tilde{x}) := \sup\{m \in \mathbb{R} : \mathbb{E}[\partial\ell_\alpha^\kappa(\tilde{x} - m)] \ge 0\}, \tag{23}$$

where $\partial\ell_\alpha^\kappa(\cdot)$ is defined in (14).

Lemma C.7. *The measure $\hat{\mathfrak{q}}_\alpha^\kappa$ satisfies monotonicity, translation invariance, and is elicitable as:*

$$\hat{\mathfrak{q}}_\alpha^\kappa(\tilde{x}) := \operatorname*{argmin}_{m \in \mathbb{R}} \mathbb{E}[\ell_\alpha^\kappa(\tilde{x} - m)] \tag{24}$$

with ℓ_α^κ as defined in (13).

Proof. Monotonicity and translation invariance follow naturally from the properties of shortfall risk measures as defined in Föllmer and Schied (2002); Weber (2006), after confirming that $\partial\ell_\alpha^\kappa(\delta)$ is increasing and non-constant. Elicitability follows from Theorem 4.3 in Bellini and Bignozzi (2015) after confirming that $\partial\ell_\alpha^\kappa(\delta)$ is strictly increasing and left continuous. □

C.7.2 Standard Operator Convergence Results

Our convergence analysis follows the framework presented in Section 4 of Bertsekas and Tsitsiklis (1996), which we summarize in this section. The Q-learning algorithm, consider the following iteration for some random sequence $\tilde{z}_i \colon \Omega \to \mathbb{R}^{\mathcal{N}}$ where $\mathcal{N} = \{1, \ldots, n\}$ defined as

$$\begin{aligned} \tilde{z}_{i+1}(b) &= (1 - \tilde{\theta}_i(b)) \cdot \tilde{z}_i(b) + \tilde{\theta}_i(b) \cdot ((H\tilde{z}_i)(b) + \tilde{\phi}_i(b)), \qquad i = 0, 1, \ldots, \\ &= \tilde{z}_i(b) + \tilde{\theta}_i(b) \cdot ((H\tilde{z}_i)(b) + \tilde{\phi}_i(b) - \tilde{z}_i(b)), \qquad i = 0, 1, \ldots, \end{aligned} \tag{25}$$

for all $b \in \mathcal{N}$, where $H \colon \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$ is some possibly non-linear operator, $\tilde{\theta}_i \colon \Omega \to \mathbb{R}_+$ is a step size, and $\tilde{\phi}_i \colon \Omega \to \mathbb{R}^{\mathcal{N}}$ is some random noise sequence. The *random* history $\mathcal{F}_i$ at iteration $i = 1, \ldots$ is denoted by

$$\mathcal{F}_i = \left(\tilde{z}_0, \ldots, \tilde{z}_i, \tilde{\phi}_0, \ldots, \tilde{\phi}_{i-1}, \tilde{\theta}_0, \ldots, \tilde{\theta}_i\right).$$

To study the convergence of the algorithm, we may also need to define a weighted maximum norm for $x \in \mathbb{R}^{\mathcal{N}}$ and weights $w \in \mathbb{R}^{\mathcal{N}}_{++}$, i.e. $w \in \mathbb{R}^{\mathcal{N}}$ with $w(b) > 0$ for all $b \in \mathcal{N}$, as

$$\|x\|_w \;:=\; \max_{b \in \mathcal{N}} \frac{|x(b)|}{w(b)}.$$

The weighted maximum norm is useful when analyzing the convergence of non-discounted MDPs. Its importance is in the fact that a non-negative matrix with a sub-unit spectral radius is a contraction in the weighted norm, but may not be a contraction in the plain maximum norm.

Definition C.8 (Star-contraction)**.** An operator $H \colon \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$ is a *weighted maximum norm star-contraction* if there exist $z^\star \in \mathbb{R}^{\mathcal{N}}$, $w \in \mathbb{R}^{\mathcal{N}}_{++}$, $\chi \in [0, 1)$ such that

$$\|Hz - z^\star\|_w \;\le\; \chi \cdot \|z - z^\star\|_w.$$

Note that the original name for star-contraction is pseudo-contraction. We use the term star-contraction because of its close resemblance to star-convexity.

The following assumption on the noise of the stochastic process will be needed to ensure the convergence of our algorithm.

Assumption C.9 (Assm. 4.3 Bertsekas and Tsitsiklis (1996))**.** The noise terms in (25) satisfy for each $i = 1, 2, \ldots$ that:

(1) Random errors are conditionally unbiased, a.s.:

$$\mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i] = 0, \qquad \forall b \in 1{:}n,$$

(2) There exists a norm $\|\cdot\|$ on $\mathbb{R}^{\mathcal{N}}$ and $c, g \in \mathbb{R}$ such that

$$\mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i] \le c + g \cdot \|\tilde{z}_i\|^2, \qquad \forall b \in 1{:}n, \quad \text{a.s.}$$

Proposition C.10. *[Proposition 4.4 Bertsekas and Tsitsiklis (1996)] Let $\tilde{z}_i, i = 1, \ldots$ be the sequence generated by the iteration in (25). Assume that*

1. *The step-sizes $\tilde{\theta}_i, \forall i = 1, \ldots$ satisfy almost surely that $\tilde{\theta}_i \ge 0$ and*

$$\sum_{t=0}^{\infty} \tilde{\theta}_i(b) = \infty, \quad \sum_{i=0}^{\infty} \tilde{\theta}_i^2(b) < \infty, \qquad \forall b \in \mathcal{N}.$$

2. *The noise terms $\tilde{\phi}_i, i = 1, \ldots$ satisfy Assumption C.9.*

3. *The operator H in (25) is a weighted maximum norm star-contraction as in Definition C.8.*

Then, $\tilde{z}_i$ converges to $z^\star$, a fixed point of H, with probability 1:

$$\mathbb{P}\left[\lim_{i \to \infty} \tilde{z}_i = z^\star\right] = 1.$$

C.7.3 Operator Definitions

We will need the following operators for any $\xi > 0$ and $\tilde{j}' \sim U([J-1])$. Let $b = (t, s, j, a)$ for $t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}$:

$$\begin{aligned}
(Gq)(b) &:= \begin{cases} q(b) - \underline{R} \cdot t & \text{if } j = 0 \vee t = 0, \\ \frac{d}{dx}\, \mathbb{E}^{a,s}\left[\ell^{\kappa}_{j/J}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q(t-1, \tilde{s}_1, \tilde{j}', a') - x\right)\right]\Big|_{x = q(t,s,j,a)} & \text{otherwise,} \end{cases} \\
&= \begin{cases} q(b) - \underline{R} \cdot t & \text{if } j = 0 \vee t = 0, \\ -\mathbb{E}^{a,s}\left[\partial \ell^{\kappa}_{\frac{j}{J}}\left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q(t-1, \tilde{s}_1, \tilde{j}', a') - q(t,s,j,a)\right)\right] & \text{otherwise,} \end{cases} \\
(G_{s'}q)(b) &:= \begin{cases} q(b) - \underline{R} \cdot t & \text{if } j = 0 \vee t = 0, \\ -\mathbb{E}\left[\partial \ell^{\kappa}_{\frac{j}{J}}\left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q(t-1, s', \tilde{j}', a') - q(t,s,j,a)\right)\right] & \text{otherwise,} \end{cases} \\
Hq &:= q - \xi \cdot Gq, \\
H_{s'}q &:= q - \xi \cdot G_{s'}q.
\end{aligned} \tag{26}$$

Consider a random sequence of inputs $((\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i, \tilde{s}'_i), \tilde{\beta}_i, \underline{\tilde{q}}_i)_{i=0}^{\infty}$ in Algorithm 2. We can define a real-valued random variable $\tilde{\phi}$ for $t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}$ as

$$
\begin{aligned}
\tilde{\phi}_i(t,s,j,a) &:= \begin{cases} (H_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(t,s,j,a) - (H\underline{\tilde{q}}^{\mathrm{d}}_i)(t,s,j,a) & \text{if } (\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i) = (t,s,j,a), \\ 0 & \text{otherwise.} \end{cases} \\
\tilde{\theta}_i(t,s,j,a) &:= \begin{cases} \frac{\tilde{\beta}_i}{\xi} & \text{if } (\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i) = (t,s,j,a), \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
\tag{27}
$$

Lemma C.11. *The random sequence of iterations followed by Algorithm 2 satisfies*

$$
\begin{aligned}
\underline{\tilde{q}}^{\mathrm{d}}_0(t,s,j,a) &= t\underline{R}, \ \ a.s. \ \ , \\
\underline{\tilde{q}}^{\mathrm{d}}_{i+1}(t,s,j,a) &= \underline{\tilde{q}}^{\mathrm{d}}_i(t,s,j,a) + \tilde{\theta}_i(t,s,j,a) \cdot (H\underline{\tilde{q}}^{\mathrm{d}}_i + \tilde{\phi}_i - \underline{\tilde{q}}^{\mathrm{d}}_i)(t,s,j,a), \quad \forall i \in \mathbb{N}, \ \ a.s.,
\end{aligned}
$$

where the terms are defined in (26) and (27).

Proof. We prove the claim by induction on i. The base case holds immediately from the definition. To prove the inductive case, suppose that $i \in \mathbb{N}$ and we prove the result in the following cases.

Case 1a: Suppose that $\tilde{b} = (\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i) = (t,s,j,a) = b$ and $t_i > 0, j_i > 0$, then by algebraic manipulation:

$$
\begin{aligned}
\underline{\tilde{q}}^{\mathrm{d}}_{i+1}(b) &= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) + \tilde{\phi}_i(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) + (H_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - (H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((H_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((\underline{\tilde{q}}^{\mathrm{d}}_i - \xi G_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \frac{\beta_i}{J} \sum_{j' \in [J-1]} \partial \ell^{\kappa}_{\frac{j}{J}} \Big(r(s,a) + \gamma \max_{a' \in \mathcal{A}} \underline{\tilde{q}}^{\mathrm{d}}_i(t-1, \tilde{s}'_i, j', a') - \underline{\tilde{q}}^{\mathrm{d}}_i(b) \Big).
\end{aligned}
$$

Case 1b: Suppose that $\tilde{b} = (\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i) = (t,s,j,a) = b$ and $t_i = 0 \vee j_i = 0$, then by algebraic manipulation:

$$
\begin{aligned}
\underline{\tilde{q}}^{\mathrm{d}}_{i+1}(b) &= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) + \tilde{\phi}_i(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) + (H_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - (H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((H_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)((\underline{\tilde{q}}^{\mathrm{d}}_i - \xi G_{\tilde{s}'_i}\underline{\tilde{q}}^{\mathrm{d}}_i)(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b)(\underline{\tilde{q}}^{\mathrm{d}}_i(b) - \xi\underline{\tilde{q}}^{\mathrm{d}}_i(b) + \xi\underline{R}t - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) - \theta_i(b)\xi(\underline{\tilde{q}}^{\mathrm{d}}_i(b) - \underline{R}t) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) - \beta_i(\underline{\tilde{q}}^{\mathrm{d}}_i(b) - \underline{R}t).
\end{aligned}
$$

Case 2: Suppose that $\tilde{b} = (\tilde{t}_i, \tilde{s}_i, \tilde{j}_i, \tilde{a}_i) \neq (t,s,j,a) = b$, then by algebraic manipulation, the algorithm does not change the q-function:

$$
\begin{aligned}
\underline{\tilde{q}}^{\mathrm{d}}_{i+1}(b) &= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + \theta_i(b) \cdot ((H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) + \tilde{\phi}_i(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b) + 0 \cdot ((H\underline{\tilde{q}}^{\mathrm{d}}_i)(b) + \tilde{\phi}_i(b) - \underline{\tilde{q}}^{\mathrm{d}}_i(b)) \\
&= \underline{\tilde{q}}^{\mathrm{d}}_i(b).
\end{aligned}
$$

□

C.7.4 Operator H is a contraction

We use the following weights $w \in \mathbb{R}^{[T] \times \mathcal{S} \times [J-1] \times \mathcal{A}}$ with a weighted max norm to prove the contraction properties in this section: where the weights are defined as

$$
w(t,s,j,a) := 2^t, \quad \forall t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}. \tag{28}
$$

Lemma C.12. *The operator* $\underline{B}^{\mathrm{d}}_{\kappa}$ *is a weighted max norm contraction for* w *defined in* (28)*:*

$$\|\underline{B}^{\mathrm{d}}_{\kappa}x - \underline{B}^{\mathrm{d}}_{\kappa}y\|_w \;\le\; \frac{1}{2}\|x-y\|_w, \quad \forall x,y \in \mathbb{R}^{\mathcal{S}}.$$

Proof. The operator $\underline{B}^{\mathrm{d}}_{\kappa}$ is equivalently defined using a shortfall risk measure $\hat{\mathfrak{q}}^{\kappa}_{\alpha}$ from Lemma C.7 for each $b=(t,s,j,a)$ as

$$(\underline{B}^{\mathrm{d}}_{\kappa}q)(b) := \begin{cases} \underline{R}\cdot t & \text{if } j=0 \vee t=0,\\ (\hat{\mathfrak{q}}^{\kappa}_{\frac{j}{J}})^{a,s}\left[r(s,a)+\gamma \max_{a'\in\mathcal{A}} q(t-1,\tilde{s}_1,\tilde{j}',a')\right] & \text{otherwise,}\end{cases}$$

We analyze the following two cases.

Case 1: For each $t\in 1{:}T, s\in\mathcal{S}, j\in 1{:}(J-1), a\in\mathcal{A}$:

$$\begin{aligned}
0 &\le \frac{1}{w(t,s,j,a)}\left|(\underline{B}^{\mathrm{d}}_{\kappa}x)(t,s,j,a)-(\underline{B}^{\mathrm{d}}_{\kappa}y)(t,s,j,a)\right|\\
&\overset{(a)}{=} \frac{1}{w(t,s,j,a)}\left|(\hat{\mathfrak{q}}^{\kappa}_{j/J})^{a,s}\left[\max_{a'\in\mathcal{A}}\gamma x(t-1,\tilde{s}_1,\tilde{j}',a')\right]-(\hat{\mathfrak{q}}^{\kappa}_{j/J})^{a,s}\left[\max_{a'\in\mathcal{A}}\gamma y(t-1,\tilde{s}_1,\tilde{j}',a')\right]\right|\\
&\overset{(b)}{\le} \frac{1}{w(t,s,j,a)}\max_{a'\in\mathcal{A}^{\mathcal{S}\times[J-1]}}\left|(\hat{\mathfrak{q}}^{\kappa}_{j/J})^{a,s}\left[\gamma x(t-1,\tilde{s}_1,\tilde{j}',a'(\tilde{s}_1,\tilde{j}'))\right]-(\hat{\mathfrak{q}}^{\kappa}_{j/J})^{a,s}\left[\gamma y(t-1,\tilde{s}_1,\tilde{j}',a'(\tilde{s}_1,\tilde{j}'))\right]\right|\\
&\overset{(c)}{\le} \frac{1}{w(t,s,j,a)}\cdot\max_{s'\in\mathcal{S},j'\in[J-1],a'\in\mathcal{A}}|\gamma x(t-1,s',j',a')-\gamma y(t-1,s',j',a')|\\
&= \frac{1}{w(t,s,j,a)}\cdot\gamma\cdot\max_{s'\in\mathcal{S},j'\in[J-1],a'\in\mathcal{A}}|x(t-1,s',j',a')-y(t-1,s',j',a')|\\
&\overset{(d)}{\le} \max_{s'\in\mathcal{S},j'\in[J-1],a'\in\mathcal{A}}\frac{|x(t-1,s',j',a')-y(t-1,s',j',a')|}{2w(t-1,s',j',a')}.
\end{aligned}$$

In step (a), we use the translation invariance of the risk measure to cancel out $r(s,a)$. Step (b) follows by upper bounding the difference using the monotonicity of $\hat{\mathfrak{q}}^{\kappa}_{\frac{j}{J}}$, step (c) follows from Lemmas C.2 and C.7, step (d) follows from the definition of w, its dependence on t only, and $\gamma\le 1$.

Case 2: For $t=0\vee j=0$:

$$\frac{1}{w(t,s,j,a)}\left|(\underline{B}^{\mathrm{d}}_{\kappa}x)(t,s,j,a)-(\underline{B}^{\mathrm{d}}_{\kappa}y)(t,s,j,a)\right| = \frac{1}{2^t}|\underline{R}t-\underline{R}t| = 0.$$

Then, using the equalities above and $\mathcal{K}=[T]\times\mathcal{S}\times[J-1]\times\mathcal{A}$ we get that

$$\begin{aligned}
\|\underline{B}^{\mathrm{d}}_{\kappa}x-\underline{B}^{\mathrm{d}}_{\kappa}y\|_w &= \max_{b\in\mathcal{K}}\frac{1}{w(b)}\left|(\underline{B}^{\mathrm{d}}_{\kappa}x)(b)-(\underline{B}^{\mathrm{d}}_{\kappa}y)(b)\right|\\
&\le \max_{t\in 1:T,s'\in\mathcal{S},j'\in 1:J-1,a'\in\mathcal{A}}\frac{|x(t-1,s',j',a')-y(t-1,s',j',a')|}{2\cdot w(t-1,s',j',a')}\\
&\le \max_{t\in[T],s'\in\mathcal{S},j'\in[J-1],a'\in\mathcal{A}}\frac{|x(t,s',j',a')-y(t,s',j',a')|}{2\cdot w(t,s',j',a')}\\
&= \frac{1}{2}\|x-y\|_w.
\end{aligned}$$

□

The following lemma establishes that the gradient update is a convex combination of the starting value and an optimal solution for an appropriate step size.

Lemma C.13. *Suppose that* $f\colon\mathbb{R}\to\mathbb{R}$ *is a differentiable* μ*-strongly convex function with an* L*-Lispchitz continuous gradient. Consider* $x_i\in\mathbb{R}$ *and a gradient update for any step size* $\xi\in(0,{}^{1}\!/_{L}]$*:*

$$x_{i+1} := x_i - \xi\cdot f'(x_i).$$

Then $\exists l \in [{}^1\!/_L, {}^1\!/_\mu]$ *such that* ${}^\xi\!/_l \in (0,1]$ *and*

$$x_{i+1} = (1 - {}^\xi\!/_l) \cdot x_i + {}^\xi\!/_l \cdot x^\star,$$

where $x^\star = \operatorname{argmin}_{x \in \mathbb{R}} f(x)$ *(unique from strong convexity).*

Proof. Assume that $f'(x_i) \neq 0$ hence $x_i \neq x^\star$; otherwise the result holds trivially. Then construct $l \neq 0$ as

$$l \;:=\; \frac{x_i - x^\star}{f'(x_i)}.$$

Substituting the definition of l into the gradient update, we get that

$$x_{i+1} \;:=\; x_i - \xi \cdot f'(x_i) = (1 - {}^\xi\!/_l)\, x_i + {}^\xi\!/_l \cdot x^\star,$$

as desired.

It remains to show that $l \in [{}^1\!/_L, {}^1\!/_\mu]$ and ${}^\xi\!/_l \in (0,1]$. Using that $f'(x^\star) = 0$ and strong convexity (Nesterov, 2018, Thm. 2.1.10) and Lipschitz continuity of the derivative:

$$\frac{1}{L}|f'(x_i)| = \frac{1}{L}|f'(x_i) - f'(x^\star)| \;\leq\; |x_i - x^\star| \;\leq\; \frac{1}{\mu}|f'(x_i) - f'(x^\star)| = \frac{1}{\mu}|f'(x_i)|. \tag{29}$$

Next, we analyze two cases.

Case 1: Suppose that $x^\star > x_i$. Then $f'(x_i) > f'(x^\star) = 0$ because f' is increasing for a strongly convex f, and therefore, (29) becomes

$$\begin{array}{rcl} -\frac{1}{L} f'(x_i) \leq & x^\star - x_i & \leq -\frac{1}{\mu} f'(x_i), \\ -\frac{1}{L} \geq & \frac{x^\star - x_i}{f'(x_i)} & \geq -\frac{1}{\mu}, \\ -\frac{1}{L} \geq & -l & \geq -\frac{1}{\mu}, \\ \frac{1}{L} \leq & l & \leq \frac{1}{\mu}. \end{array}$$

In addition, $\xi \in (0, {}^1\!/_L] \implies {}^\xi\!/_l \in (0,1]$.

Case 2: Suppose that $x^\star < x_i$. Then $f'(x_i) > f'(x^\star) = 0$ because f' is increasing for a strongly convex f, and therefore, (29) becomes

$$\begin{array}{rcl} \frac{1}{L} f'(x_i) \leq & x_i - x^\star & \leq \frac{1}{\mu} f'(x_i), \\ \frac{1}{L} \leq & \frac{x_i - x^\star}{f'(x_i)} & \leq \frac{1}{\mu}, \\ \frac{1}{L} \leq & l & \leq \frac{1}{\mu}. \end{array}$$

In addition, $\xi \in (0, {}^1\!/_L] \implies {}^\xi\!/_l \in (0,1]$. □

Theorem C.14. *A fixed point* $x^\star = \underline{B}^{\mathrm{d}}_\kappa x^\star$ *exists and satisfies* $x^\star = Hx^\star$. *Let* $\bar{\mu} := J^{-1}\kappa$, $\bar{L} := \kappa^{-1}$, *and* $\xi \in (0, \min(1, {}^1\!/_{\bar{L}}))$ *in the definition of* H. *Then,* H *is a weighted max norm star contraction:*

$$\|Hx - Hx^\star\|_w \;\leq\; \left(1 - \frac{\bar{\mu}\xi}{2}\right) \cdot \|x - x^\star\|_w,$$

for w *defined in* (28).

Proof. The fixed point $x^\star$ exists from Lemma C.12 and the Banach fixed point theorem. The operator H takes a gradient step towards $\underline{B}^{\mathrm{d}}_\kappa$.

Case 1: Fix some $b = (t, s, j, a)$ with $t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}$ and suppose that $t > 0$ and $j > 0$. Fix some q and define

$$f(y) = \mathbb{E}^{s,a}\left[\ell^\kappa_{j/J}\left(r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} q(t-1, \tilde{s}_1, \tilde{j}', a') - y\right)\right], \tag{30}$$

The function f is strongly convex with Lipschitz gradient with parameters $\bar{\mu}$ and $\bar{L}$ based on Lemma 4.6 and since

$$\bar{\mu} \leq \min\left\{\frac{j}{J}, 1 - \frac{j}{J}\right\}\kappa \leq \max\left\{\frac{j}{J}, 1 - \frac{j}{J}\right\}\kappa^{-1} \leq \bar{L}, \forall j \in 1{:}J-1.$$

Let $y^\star \in \arg\min_{y\in\mathbb{R}} f(y)$. Then, $\exists l \in [\frac{1}{L}, \frac{1}{\bar{\mu}}]$ such that

$$\begin{aligned}(Hq)(b) &= (q - \xi G q)(b) = q(b) - \xi \cdot f'(q(b)) = (1 - \xi/l) \cdot q(b) + \xi/l \cdot y^\star \\ &= (1 - \xi/l) \cdot q(b) + \xi/l \cdot (\underline{B}^{\mathrm{d}}_{\kappa} q)(b),\end{aligned} \tag{31}$$

from algebraic manipulation and application of Lemma C.13 to the function f in (30) which satisfies the requisite strong convexity and Lipschitz continuity properties.

The fixed point of $x^\star$ of $\underline{B}^{\mathrm{d}}_{\kappa}$ is a fixed point of H from (31)

$$(Hx^\star)(b) = \left(1 - \frac{\xi}{l}\right) x^\star(b) + \frac{\xi}{l}(\underline{B}^{\mathrm{d}}_{\kappa} x^\star)(b) = \left(1 - \frac{\xi}{l}\right) x^\star(b) + \frac{\xi}{l}(x^\star)(b) = x^\star(b).$$

Finally, we get using (31) that

$$\begin{aligned}|(Hx)(b) - (Hx^\star)(b)| &= |(1 - \xi/l)x(b) + \xi/l(\underline{B}^{\mathrm{d}}_{\kappa} x)(b) - ((1 - \xi/l)x^\star(b) + \xi/l(\underline{B}^{\mathrm{d}}_{\kappa} x^\star)(b))| \\ &= |(1 - \xi/l)(x - x^\star)(b) + \xi/l(\underline{B}^{\mathrm{d}}_{\kappa} x - \underline{B}^{\mathrm{d}}_{\kappa} x^\star)(b)| \\ &\le (1 - \xi/l)|(x - x^\star)(b)| + \xi/l|(\underline{B}^{\mathrm{d}}_{\kappa} x - \underline{B}^{\mathrm{d}}_{\kappa} x^\star)(b)| \\ &\le (1 - \xi/l)|(x - x^\star)(b)| + \frac{1}{2}\xi/l|(x - x^\star)(b)| \\ &= (1 - \xi/2l)|(x - x^\star)(b)|.\end{aligned}$$

Here, we used the triangle inequality for absolute values and the fact that $x^\star$ is a fixed point of $\underline{B}^{\mathrm{d}}_{\kappa}$ and Lemma C.12. Hence,

$$\frac{1}{w(b)}|(Hx - Hx^\star)(b)| \le \left(1 - \frac{\bar{\mu}\xi}{2}\right) \frac{1}{w(b)}|(x - x^\star)(b)|. \tag{32}$$

Case 2: Fix some $b = (t, s, j, a)$ with $t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}$ and suppose that $t = 0 \vee j = 0$. Then, from $x^\star(b) = \underline{R} \cdot t$, we have that if $\xi < 1$:

$$\begin{aligned}|(Hx)(b) - (Hx^\star)(b)| &= |x(b) - \xi(x(b) - \underline{R} \cdot t) - \underline{R} \cdot t| \\ &= (1 - \xi)|x(b) - \underline{R} \cdot t| = (1 - \xi)|x(b) - x^\star(b)|.\end{aligned}$$

Hence,

$$\frac{1}{w(b)}|(Hx - Hx^\star)(b)| \le (1 - \xi)\, \frac{1}{w(b)}|(x - x^\star)(b)| \le (1 - \bar{\mu}\xi/2)\, \frac{1}{w(b)}|(x - x^\star)(b)| \tag{33}$$

since $\bar{\mu}\xi/2 \le \kappa J^{-1}\xi/2 \le \kappa\xi \le \xi$ because $\kappa \le 1$.

Conclusion: Putting (32) and (33) together with the definition of the weighted norm, we get the desired star contraction rate. □

C.7.5 Noise Properties

The history $\mathcal{F}_i$ at an iteration $i \in \mathbb{N}$ is defined as

$$\mathcal{F}_i := \left(\tilde{\underline{q}}^{\mathrm{d}}_0, \ldots, \tilde{\underline{q}}^{\mathrm{d}}_i, \tilde{\phi}_0, \ldots, \tilde{\phi}_{i-1}, \tilde{\theta}_0, \ldots, \tilde{\theta}_i\right). \tag{34}$$

Recall from Assumption 4.7 that

$$\mathcal{G}_i := (\tilde{\beta}_l, (\tilde{t}_l, \tilde{s}_l, \tilde{j}_l, \tilde{a}_l, \tilde{s}'_l))_{l=0}^{i},$$

and

$$\mathbb{P}\left[\tilde{s}'_i = s' \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i\right] = p(\tilde{s}_i, \tilde{a}_i, s'),\ \forall s' \in \mathcal{S},$$

almost surely, where $\mathcal{G}_{i-1} := (\tilde{\beta}_l, (\tilde{t}_l, \tilde{s}_l, \tilde{j}_l, \tilde{a}_l, \tilde{s}'_l))_{l=0}^{i-1}$.

Lemma C.15. *Let Ω be an appropriate sample space. Then for each $\omega_1, \omega_2 \in \Omega$ and $i = 1, \dots$:*

$$(\mathcal{G}_{i-1}(\omega_1) = \mathcal{G}_{i-1}(\omega_2)) \wedge (\tilde{b}_i(\omega_1) = \tilde{b}_i(\omega_2)) \wedge (\tilde{\beta}_i(\omega_1) = \tilde{\beta}_i(\omega_2))$$
$$\Longrightarrow$$
$$\mathcal{F}_i(\omega_1) = \mathcal{F}_i(\omega_2) = \bar{\mathcal{F}}_i(\mathcal{G}_{i-1}(\omega_1), \tilde{b}_i(\omega_1), \tilde{\beta}_i(\omega_1)), \text{ a.s.}$$

for some $\bar{\mathcal{F}}_i$ operator that maps a tuple $((\beta_l, (t_l, s_l, j_l, a_l, s_l'))_{l=0}^{i-1}, (t_i, s_i, j_i, a_i), \beta_i)$ to some $\left(\underline{q}_0^{\mathrm{d}}, \dots, \underline{q}_i^{\mathrm{d}}, \phi_0, \dots, \phi_{i-1}, \theta_0, \dots, \theta_i\right)$.

Proof. We proceed by induction on i. To prove the base step for $i = 0$:

$$\mathcal{F}_0(\omega_1) = (\tilde{\underline{q}}_0^{\mathrm{d}}(\omega_1), \tilde{\theta}_0(\omega_1)) = (t\underline{R}, \tilde{\theta}_0(\omega_1)) = (t\underline{R}, \tilde{\theta}_0(\omega_2)) = \mathcal{F}_0(\omega_2).$$

Here, $\tilde{\theta}_0(\omega_1) = \tilde{\theta}_0(\omega_2)$ because $\tilde{\beta}_0(\omega_1) = \tilde{\beta}_0(\omega_2)$, and $\tilde{b}_0(\omega_1) = \tilde{b}_0(\omega_2)$.

To prove the inductive step, assume that the property holds for i and prove it for $i+1$. That is, suppose that $l = i + 1$

$$(\mathcal{G}_{l-1}(\omega_1) = \mathcal{G}_{l-1}(\omega_2)) \wedge (\tilde{b}_l(\omega_1) = \tilde{b}_l(\omega_2)) \wedge (\tilde{\beta}_l(\omega_1) = \tilde{\beta}_l(\omega_2)).$$

Then from the inductive assumption:

$$\mathcal{F}_l = \mathcal{F}_l, \quad \forall\, l = 1, \dots, i,$$

and for $b = \tilde{b}_{i+1}(\omega_1) = \tilde{b}_{i+1}(\omega_2)$:

$$\begin{aligned}
(\tilde{\phi}_i(b))(\omega_1) &= (H_{\tilde{s}_i'(\omega_1)} \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_1))(b) - (H \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_1))(b) \\
&= (H_{\tilde{s}_i'(\omega_2)} \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_2))(b) - (H \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_2))(b) \\
&= (\tilde{\phi}_i(b))(\omega_2). \\
(\tilde{\theta}_{i+1}(b))(\omega_1) &= \frac{\tilde{\beta}_{i+1}(\omega_1)}{\xi} = \frac{\tilde{\beta}_{i+1}(\omega_2)}{\xi} = (\tilde{\theta}_{i+1}(b))(\omega_2),
\end{aligned}$$

and for $b \neq \tilde{b}_{i+1}(\omega_1) = \tilde{b}_{i+1}(\omega_2)$:

$$(\tilde{\phi}_i(b))(\omega_1) = 0 = (\tilde{\phi}_i(b))(\omega_2).$$
$$(\tilde{\theta}_{i+1}(b))(\omega_1) = 0 = (\tilde{\theta}_{i+1}(b))(\omega_2).$$

In addition,

$$\begin{aligned}
(\tilde{\underline{q}}_{i+1}^{\mathrm{d}}(b))(\omega_1) &= (\tilde{\underline{q}}_i^{\mathrm{d}}(b))(\omega_1) + (\tilde{\theta}_i(b))(\omega_1) \cdot (H \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_1) + \tilde{\phi}_i(\omega_1) - \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_1))(t, s, j, a) \\
&= (\tilde{\underline{q}}_i^{\mathrm{d}}(b))(\omega_2) + (\tilde{\theta}_i(b))(\omega_2) \cdot (H \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_2) + \tilde{\phi}_i(\omega_2) - \tilde{\underline{q}}_i^{\mathrm{d}}(\omega_2))(b), \\
&= (\tilde{\underline{q}}_{i+1}^{\mathrm{d}}(b))(\omega_2).
\end{aligned}$$

Putting the inequalities above together we get the desired equality:

$$\begin{aligned}
\mathcal{F}_{i+1}(\omega_1) &= \left(\tilde{\underline{q}}_0^{\mathrm{d}}(\omega_1), \dots, \tilde{\underline{q}}_{i+1}^{\mathrm{d}}(\omega_1), \tilde{\phi}_0(\omega_1), \dots, \tilde{\phi}_i(\omega_1), \tilde{\theta}_0(\omega_1), \dots, \tilde{\theta}_{i+1}(\omega_1)\right) \\
&= \left(\tilde{\underline{q}}_0^{\mathrm{d}}(\omega_2), \dots, \tilde{\underline{q}}_{i+1}^{\mathrm{d}}(\omega_2), \tilde{\phi}_0(\omega_2), \dots, \tilde{\phi}_i(\omega_2), \tilde{\theta}_0(\omega_1), \dots, \tilde{\theta}_{i+1}(\omega_1)\right) \\
&= \mathcal{F}_{i+1}(\omega_2).
\end{aligned}$$

□

Lemma C.16. *Under Assumption 4.7:*

$$\mathbb{P}\left[\tilde{s}_i' = s' \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i, \mathcal{F}_i\right] = p(\tilde{s}_i, \tilde{a}_i, s'), \ \text{a.s.},$$

for each $s' \in \mathcal{S}$ and $i \in \mathbb{N}$.

Proof. Using Lemma C.15, we have that

$$\mathbb{P}\left[\mathcal{F}_i = \bar{\mathcal{F}}(\mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i) \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i\right] = 1.$$

Hence, from Assumption 4.7 and the law of total probability, for each $s' \in \mathcal{S}, i \in \mathbb{N}$:

$$\begin{aligned}
p(\tilde{s}_i, \tilde{a}_i, s') &= \mathbb{P}\left[\tilde{s}'_i = s' \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i\right] \\
&= \mathbb{P}\left[\tilde{s}'_i = s' \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i, \mathcal{F}_i = \bar{\mathcal{F}}(\mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i)\right] \mathbb{P}\left[\mathcal{F}_i = \bar{\mathcal{F}}(\mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i) \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i)\right] \\
&= \mathbb{P}\left[\tilde{s}'_i = s' \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i, \mathcal{F}_i = \bar{\mathcal{F}}(\mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i)\right] \\
&= \mathbb{P}\left[\tilde{s}'_i = s' \mid \mathcal{G}_{i-1}, \tilde{b}_i, \tilde{\beta}_i, \mathcal{F}_i\right] \text{ a.s.}
\end{aligned}$$

□

Lemma C.17. *The noise $\tilde{\phi}_i$ in (27) satisfies almost surely*

$$\mathbb{E}[\tilde{\phi}_i(t,s,j,a) \mid \mathcal{F}_i] = 0, \quad \forall t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}, i \in \mathbb{N},$$

where $\mathcal{F}_i$ is the history defined in (34).

Proof. Let $b := (t,s,j,a)$ and $i \in \mathbb{N}$ be arbitrary. We decompose the expectation using the law of total expectation to get thta

$$\mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i] = \mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i, \tilde{b}_i \neq b] \cdot \mathbb{P}[\tilde{b}_i \neq b \mid \mathcal{F}_i] + \mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \cdot \mathbb{P}[\tilde{b}_i = b \mid \mathcal{F}_i] \text{ a.s.,} \tag{35}$$

where $\tilde{b}_i := (\tilde{s}_i, \tilde{a}_i, \tilde{t}_i, \tilde{j}_i)$.

The first r.h.s. term in (35) is, from the definition of $\tilde{\phi}_i(b)$,

$$\mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i, \tilde{b}_i \neq b] = \mathbb{E}[0 \mid \mathcal{F}_i, \tilde{b}_i \neq b] = 0, \quad \text{a.s..} \tag{36}$$

We now analyze two cases to evaluate the second r.h.s. term in (35).

Case 1: $j > 0 \wedge t > 0$. Then almost surely:

$$\begin{aligned}
\mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i, \tilde{b}_i = b] &= \mathbb{E}[(H_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b) - (H \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= \xi \cdot \mathbb{E}[-(G_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= \xi \cdot \mathbb{E}[-(G_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= \xi \cdot \mathbb{E}[\mathbb{E}[-(G_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b, \tilde{\beta}_i, \mathcal{G}_{i-1}] + (G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&\overset{\text{(a)}}{=} \xi \cdot \mathbb{E}[\mathbb{E}^{a,s}[-(G_{\tilde{s}_1} \tilde{\underline{q}}^{\mathrm{d}}_i)(b)] + (G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= \xi \cdot \mathbb{E}[-(G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= 0.
\end{aligned} \tag{37}$$

To clarify, when $\tilde{s}_1$ is used in an expectation with a superscript, such as $\mathbb{E}^{a,s}$, then it does not represent a sample $\tilde{s}_i$ with $i = 1$, but instead it represents the transition from $\tilde{s}_0 = s$ to $\tilde{s}_1$ distributed as $p(s, a, \cdot)$.

Step (a) above follows from Lemma C.16 given that the randomness of $(G_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b)$ only comes from $\tilde{s}'_i$ when conditioning on $\mathcal{F}_i, \tilde{b}_i = b, \tilde{\beta}_i$, and $\mathcal{G}_{i-1}$.

Case 2: $j = 0 \vee t = 0$. Directly from the definition of the operators in (26):

$$\begin{aligned}
\mathbb{E}[\tilde{\phi}_i(b) \mid \mathcal{F}_i, \tilde{b}_i = b] &= \mathbb{E}[(H_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b) - (H \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= \xi \cdot \mathbb{E}[-(G_{\tilde{s}'_i} \tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G \tilde{\underline{q}}^{\mathrm{d}}_i)(b) \mid \mathcal{F}_i, \tilde{b}_i = b] \\
&= 0.
\end{aligned} \tag{38}$$

Substituting (36), and the appropriate case, (37) or (38), into (35) proves the desired equality. □

Lemma C.18. *The noise* $\tilde{\phi}_i$ *in* (27) *satisfies*

$$\mathbb{E}[(\tilde{\phi}_i(t,s,j,a))^2 \mid \mathcal{F}_i] \le c + g\|\tilde{\underline{q}}^{\mathrm{d}}_i\|_\infty^2, \quad \forall t \in [T], s \in \mathcal{S}, j \in [J-1], a \in \mathcal{A}, i \in \mathbb{N},$$

almost surely, for some $c, g \in \mathbb{R}_+$ *where* $\mathcal{F}_i$ *is the history defined in* (34).

Proof. Let $b := (t,s,j,a)$ and $i \in \mathbb{N}$ be arbitrary. We decompose the expectation using the law of total expectation to get almost surely

$$\mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i] = \mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i, \tilde{b}_i \neq b] \cdot \mathbb{P}[\tilde{b}_i \neq b \mid \mathcal{F}_i] + \mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i, \tilde{b}_i = b] \cdot \mathbb{P}[\tilde{b}_i = b \mid \mathcal{F}_i], \tag{39}$$

where $\tilde{b}_i := (\tilde{t}_i, \tilde{s}_i, \tilde{a}_i, \tilde{j}_i, \tilde{a}_i)$.

The first r.h.s. term in (39) is, from the definition of $\tilde{\phi}_i(b)$,

$$\mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i, \tilde{b}_i \neq b] = \mathbb{E}[0 \mid \mathcal{F}_i, \tilde{b}_i \neq b] = 0, \quad \text{a.s.}. \tag{40}$$

We now analyze two cases to evaluate the second r.h.s. term in (39).

Case 1: Assume that $j > 0, t > 0$. Then from the definitions of the operators in (26):

$$\begin{aligned}
\mathbb{E}[(\tilde{\phi}_i(b))^2 \mid \mathcal{F}_i, \tilde{b}_i = b] &= \mathbb{E}\left[\left((H_{\tilde{s}'_i}\tilde{\underline{q}}^{\mathrm{d}}_i)(b) - (H\tilde{\underline{q}}^{\mathrm{d}}_i)(b)\right)^2 \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&= \xi^2 \mathbb{E}\left[\left(-(G_{\tilde{s}'_i}\tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G\tilde{\underline{q}}^{\mathrm{d}}_i)(b)\right)^2 \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&= \xi^2 \mathbb{E}\left[\mathbb{E}\left[\left(-(G_{\tilde{s}'_i}\tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G\tilde{\underline{q}}^{\mathrm{d}}_i)(b)\right)^2 \mid \mathcal{F}_i, \tilde{b}_i = b, \tilde{\beta}_i, \mathcal{G}_{i-1}\right] \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(a)}}{=} \xi^2 \mathbb{E}\left[\mathbb{E}^{a,s}\left[\left((G_{\tilde{s}_1}\tilde{\underline{q}}^{\mathrm{d}}_i)(b) - (G\tilde{\underline{q}}^{\mathrm{d}}_i)(b)\right)^2\right] \mid \mathcal{F}_i, \tilde{b}_i = b\right].
\end{aligned}$$

To clarify, when $\tilde{s}_1$ is used in an expectation with a superscript, such as $\mathbb{E}^{a,s}$, then it does not represent a sample $\tilde{s}_i$ with $i = 1$, but instead it represents the transition from $\tilde{s}_0 = s$ to $\tilde{s}_1$ distributed as $p(s,a,\cdot)$.

Step (a) above follows from Lemma C.16 given that the randomness of $-(G_{\tilde{s}'_i}\tilde{\underline{q}}^{\mathrm{d}}_i)(b) + (G\tilde{\underline{q}}^{\mathrm{d}}_i)(b)$ only comes from $\tilde{s}'_i$ when conditioning on $\mathcal{F}_i, \tilde{b}_i = b, \tilde{\beta}_i$, and $\mathcal{G}_{i-1}$. Then, continuing the derivation:

$$\begin{aligned}
\mathbb{E}[(\tilde{\phi}_i(b))^2 \mid \mathcal{F}_i, \tilde{b}_i = b] &= \xi^2 \mathbb{E}\left[\mathbb{E}^{a,s}\left[\left((G_{\tilde{s}_1}\tilde{\underline{q}}^{\mathrm{d}}_i)(b) - (G\tilde{\underline{q}}^{\mathrm{d}}_i)(b)\right)^2\right] \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(a)}}{=} \xi^2 \mathbb{E}\left[\mathbb{E}^{a,s}\left[\left(\mathbb{E}\left[\partial\ell^\kappa_{\frac{j}{J}}\left(\tilde{\delta}_i(\tilde{s}_1,\tilde{j}')\right) |\tilde{s}_1\right] - \mathbb{E}^{a,s}\left[\partial\ell^\kappa_{\frac{j}{J}}\left(\tilde{\delta}_i(\tilde{s}_1,\tilde{j}')\right)\right]\right)^2\right] \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(b)}}{=} \xi^2 \mathbb{E}\left[\left(\mathbb{E}^{a,s}\left[\left(\mathbb{E}\left[\partial\ell^\kappa_{\frac{j}{J}}\left(\tilde{\delta}_i(\tilde{s}_1,\tilde{j}')\right) \mid \tilde{s}_1\right]\right)^2\right] - \left(\mathbb{E}^{a,s}\left[\partial\ell^\kappa_{\frac{j}{J}}\left(\tilde{\delta}_i(\tilde{s}_1,\tilde{j}')\right)\right]\right)^2\right) \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\le \xi^2 \mathbb{E}\left[\mathbb{E}^{a,s}\left[\left(\mathbb{E}\left[\partial\ell^\kappa_{\frac{j}{J}}\left(\tilde{\delta}_i(\tilde{s}_1,\tilde{j}')\right) \mid \tilde{s}_1\right]\right)^2\right] \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(c)}}{\le} \xi^2 \mathbb{E}\left[\max_{j' \in [J-1], s' \in \mathcal{S}} \partial\ell^\kappa_{\frac{j}{J}}(\tilde{\delta}_i(s',j'))^2 \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(d)}}{\le} \xi^2 \mathbb{E}\left[\max_{j' \in [J-1], s' \in \mathcal{S}} \left(|\partial\ell^\kappa_{\frac{j}{J}}(\tilde{\delta}_i(s',j')) - \partial\ell^\kappa_{\frac{j}{J}}(0)|\right)^2 \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(e)}}{\le} \xi^2 \mathbb{E}\left[\max_{j' \in [J-1], s' \in \mathcal{S}} \left(\max\left\{\frac{j}{J}, 1 - \frac{j}{J}\right\}\kappa^{-1}|\tilde{\delta}_i(s',j')|\right)^2 \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\le \xi^2\kappa^{-2} \cdot \mathbb{E}\left[\max_{j' \in [J-1], s' \in \mathcal{S}} \tilde{\delta}_i(s',j')^2 \mid \mathcal{F}_i, \tilde{b}_i = b\right] \\
&\overset{\text{(f)}}{\le} \xi^2\kappa^{-2} \cdot (2\|r\|_\infty^2 + 8 \cdot \|\tilde{\underline{q}}^{\mathrm{d}}_i\|_\infty^2).
\end{aligned}$$

Step (a) above follows by substituting $(G_{\tilde{s}_1}\underline{\tilde{q}}_i^{\mathrm{d}})(b) - (G\underline{\tilde{q}}_i^{\mathrm{d}})(b)$ and replacing

$$\tilde{\delta}_i(s', j') := r(s,a) + \gamma \max_{a' \in \mathcal{A}} \underline{\tilde{q}}_i^{\mathrm{d}}(t-1, s', j', a') - \underline{\tilde{q}}_i^{\mathrm{d}}(t, s, j, a). \tag{41}$$

The equality in step (b) holds because for a random variable $\tilde{x} := \mathbb{E}[\partial \ell^{\kappa}_{\frac{j}{J}} \left(\delta_i(\tilde{s}_1, \tilde{j}')\right) | \tilde{s}_1]$, the variance satisfies $\mathbb{E}[(\tilde{x} - \mathbb{E}[\tilde{x}])^2] = \mathbb{E}[\tilde{x}^2] - (\mathbb{E}[\tilde{x}])^2$. Step (c) upper bounds the expectation by a supremum, and step (d) uses $\partial \ell^{\kappa}_{\frac{j}{J}}(0) = 0$ from the definition in (14). Step (e) uses Lemma C.5 to bound the derivative difference as a function of the step size. Finally, step (f) derives the final upper bound since

$$\|r\|_\infty = \max_{s \in \mathcal{S}, a \in \mathcal{A}} |r(s,a)|, \qquad \|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty = \max_{b \in \mathcal{K}} |\underline{\tilde{q}}_i^{\mathrm{d}}(b)|,$$

where $\mathcal{K} = [T] \times \mathcal{S} \times [J-1] \times \mathcal{A}$ and

$$\begin{aligned} \max_{j' \in [J-1], s' \in \mathcal{S}} \tilde{\delta}_i(s', j')^2 &\le (\|r\|_\infty + 2\|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty)^2 \le (\|r\|_\infty + 2\|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty)^2 + (\|r\|_\infty - 2\|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty)^2 \\ &= 2\|r\|_\infty^2 + 8\|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty^2, \end{aligned}$$

and because $\underline{\tilde{q}}_i^{\mathrm{d}}$ is measurable on $\mathcal{F}_i$.

Case 2: Assume that $t = 0 \vee j = 0$. Directly from the definition of the operators in (26):

$$\begin{aligned} \mathbb{E}[(\tilde{\phi}_i(b))^2 \mid \mathcal{F}_i, \tilde{b}_i = b] &= \mathbb{E}[((H_{\tilde{s}_i'}\underline{\tilde{q}}_i^{\mathrm{d}})(b) - (H\underline{\tilde{q}}_i^{\mathrm{d}})(b))^2 \mid \mathcal{F}_i, \tilde{b}_i = b] \\ &= \xi \cdot \mathbb{E}[(-(G_{\tilde{s}_i'}\underline{\tilde{q}}_i^{\mathrm{d}})(b) + (G\underline{\tilde{q}}_i^{\mathrm{d}})(b))^2 \mid \mathcal{F}_i, \tilde{b}_i = b] \\ &= 0. \end{aligned}$$

Finally, we can confirm that

$$\begin{aligned} \mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i] &= \mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i, \tilde{b}_i \neq b] \cdot \mathbb{P}[\tilde{b}_i \neq b \mid \mathcal{F}_i] + \mathbb{E}[\tilde{\phi}_i(b)^2 \mid \mathcal{F}_i, \tilde{b}_i = b] \cdot \mathbb{P}[\tilde{b}_i = b \mid \mathcal{F}_i] \\ &\le 0 \cdot \mathbb{P}[\tilde{b}_i \neq b \mid \mathcal{F}_i] + \xi^2 \kappa^{-2} \cdot (2\|r\|_\infty^2 + 8 \cdot \|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty^2) \cdot \mathbb{P}[\tilde{b}_i = b \mid \mathcal{F}_i] \\ &\le \xi^2 \kappa^{-2} \cdot (2\|r\|_\infty^2 + 8 \cdot \|\underline{\tilde{q}}_i^{\mathrm{d}}\|_\infty^2). \end{aligned}$$

□

C.7.6 Main Proof

Proof of Theorem 4.8. We verify that the sequence of our q-learning iterates satisfies the properties in Proposition C.10.

- The step size condition in Theorem 4.8 guarantees that we satisfy property (a) in Proposition C.10
- We satisfy property (b) in Proposition C.10 because
  - Lemma C.17 shows that we satisfy property (1) in Assumption C.9
  - Lemma C.18 shows that we satisfy property (2) in Assumption C.9
- Theorem C.14 shows that we satisfy property (c) in Proposition C.10

□

D EMPIRICAL RESULTS

The code used to generate all plots can be found in https://github.com/MonkieDein/DRA-Q-LA.

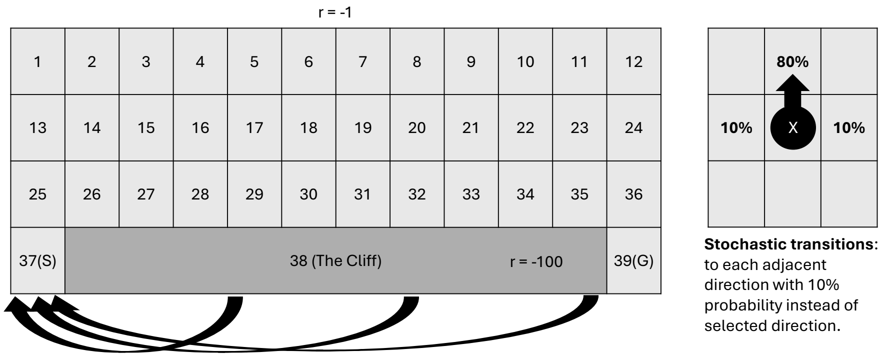

Figure 5: The cliffwalk domain

D.1 Domain Details

For each domain, we provide CSV files and julia JLD files in the supplementary material with the exact specifications of the domains we use. Domain detail for six out of seven of our domains include (Machine Replacement (MR), Gambler's Ruin (GR), Inventory1 (INV1), Inventory2 (INV2), Riverswim (RS) and Population Management (POP)) can be found in (Hau et al., 2023b, Appx. E). The Cliffwalk (CW) domain is similar to the one described in (Sutton and Barto, 2018, Ex. 6.6), with a minor modification. In this version, the agent transitions to each adjacent direction with a 10%-probability instead of always following the selected direction (see Fig. 5). The initial state s_0 specification can be found in Table 2. We initialize all environments with a discount factor of $\gamma = 0.9$ and a horizon $T = 100$.

	MR	GR	INV1	INV2	RS	POP	CW
s_0	1	5	10	20	9	44	37

Table 2: Initial state for each domain

D.2 Algorithmic Details

Algorithm 1 Line 6 is implemented with $\epsilon = 10^{-14}$ to account for the non-associative property of floating point arithmetic with $\tau = \frac{\underline{q}^{\mathrm{d}}_t(s,j,a^\star)-r}{\gamma}$ as:

$$j \leftarrow \operatorname{argmin}\left\{ j' \in [J-1] \mid \max_{a'\in\mathcal{A}} \underline{q}^{\mathrm{d}}_{t-1}(s',j',a') \geq \tau - \epsilon|\tau| \right\} .$$

Algorithm 2 Without loss of generality, the VaR-Q-value function is trained with a standardized scaled reward function $\hat{r}(s,a) \leftarrow \frac{r(s,a)-\underline{R}}{\bar{R}-\underline{R}}$. We also remove the time indices to reduce computational overhead and initialize the q-value function with $\hat{q}^{\mathrm{d}} \leftarrow (1-\gamma)^{-1}$. The VaR-Q-value function is then unscaled via $\tilde{q}^{\mathrm{d}} \leftarrow \hat{q}^{\mathrm{d}} \cdot (\bar{R} - \underline{R}) + \frac{\underline{R}}{1-\gamma}$ before being compared with the DP variant $\underline{q}^{\mathrm{d}}$. The learning rate is defined as $\beta_i \leftarrow 100 \cdot (0.1^{i \cdot 0.0003})$ for i-th occurrence of sample (s,a) across all domains. For a fair comparison between domains, we sample a transition for every (s,a)-pair at each iteration.

Nested VaR (nVaR) also known as dynamic VaR, nVaR is solved via the following DP for each $s \in \mathcal{S}$ and $t \in [T-1]$ as

$$v_{t+1}(s) = \max_{a\in\mathcal{A}} \mathrm{VaR}_{\alpha_0}\left[r(s,a) + \gamma \cdot v_t(\tilde{s}')\right],$$

where $\tilde{s}' \sim p(s,a,\cdot)$. Then, we evaluate a greedy policy $\pi_t \colon \mathcal{S} \to \mathcal{A}, k \in [T-1]$ constructed to satisfy

$$\pi_k(s) \in \operatorname*{argmax}_{a\in\mathcal{A}} \mathrm{VaR}_{\alpha_0}\left[r(s,a) + \gamma \cdot v_{T-k}(\tilde{s}')\right].$$

Distributional VaR (VaR-IQN) It uses the Markov action-selection strategy proposed by Dabney et al. (2018a); Keramati et al. (2020) with $J = 4096$ uniform quantile discretization (Dabney et al., 2018b; Rowland et al., 2024):

$$q_{t+1}(s,\alpha_j,a) = \mathrm{VaR}_{\alpha_j}\left[r(s,a) + \gamma \max_{a'\in\mathcal{A}} \mathrm{VaR}_{\alpha_0}\left[q_t(\tilde{s}',\tilde{u},a')\right]\right] \qquad \forall \alpha_j = \frac{2j+1}{J} \text{ where } j \in [J-1]\,,$$

where $\tilde{u}$ refer to the discretized uniform distribution satisfy $\mathbb{P}[\tilde{u} = \frac{2j+1}{J}] = \frac{1}{J}\ \forall j \in [J-1]$ follows greedy policy

$$\pi_k(s) \in \operatorname*{argmax}_{a\in\mathcal{A}} \mathrm{VaR}_{\alpha_0}\left[q_{T-k}(s,\tilde{u},a)\right].$$

In contrast to our algorithm, an optimal action a is selected w.r.t the initial Markov risk level of interest α_0 instead of the quantile-dependent risk level α_j (Lim and Malik, 2022).

Conditional Value at Risk (CVaR) Follows the dynamic program described in (Bäuerle and Ott, 2011), that solves a bi-level optimization

$$\begin{aligned}\max_{\pi\in\Pi_{\mathrm{HD}}^t} \mathrm{CVaR}_\alpha^{\pi,s}\left[\sum_{k=0}^{t-1}\gamma^k \cdot r(\tilde{s}_k,\tilde{a}_k)\right] &:= \max_{\pi\in\Pi_{\mathrm{HD}}^t}\sup_{z\in\mathbb{R}}(z - \alpha^{-1}\mathbb{E}[z - \sum_{k=0}^{t-1}\gamma^k \cdot r(\tilde{s}_k,\tilde{a}_k)]_+) \\ &= \sup_{z\in\mathbb{R}}(z - \alpha^{-1}\min_{\pi\in\Pi_{\mathrm{HD}}^t}\mathbb{E}[z - \sum_{k=0}^{t-1}\gamma^k \cdot r(\tilde{s}_k,\tilde{a}_k)]_+)\,,\end{aligned}$$

given $z \in \mathbb{R}$, a recursive function with memoization is used to solve the inner optimization for $t \in 1:T$ as

$$v_t^\star(s,z) := \min_{\pi\in\Pi_{\mathrm{HD}}^t}\mathbb{E}\left[z - \sum_{k=0}^{t-1}\gamma^k \cdot r(\tilde{s}_k,\tilde{a}_k)\right]_+ = \gamma \cdot \min_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} v_{t-1}^\star(s', \frac{z - r(s,a)}{\gamma}) \cdot p(s,a,s')\,,$$

with the base case $v_0^\star(s,z) = \mathbb{E}\left[z - 0\right]_+ = \max\{0,z\}$.

Since it is not computationally feasible to solve for all $z \in \mathbb{R}$, with the same intuition presented in Section 4.1, we consider an approximation z by rounding it up to a precision determined by the span of the return range to under-approximate the CVaR return. Specifically, we round z to $d = 5 - \lceil\log_{10}(\bar{R} - \underline{R}) - \log_{10}(1-\gamma)\rceil$ decimal places, ensuring an accuracy of five significant digits relative to the return range.

Furthermore, in the discounted setting, it suffices to compute the value function for a closed set $\mathcal{Z}_t \subset [L,U]$ for an MDP with horizon t where $L = \lceil\frac{\underline{\mathrm{R}}(1-\gamma^t)}{1-\gamma}\rceil_d$ and $U = \lceil\frac{\bar{R}(1-\gamma^t)}{1-\gamma}\rceil_d$, because the value function and behavior of z outside this range, i.e. $z \in (-\infty, L) \cup (U,\infty)$ are well defined. Specifically,

- For $z \le L$, the value function for all states is given by $\min_\pi \mathbb{E}\left[z - \tilde{x}^\pi\right]_+ = 0$.
- For $z \ge U$, the value function for all states satisfies $\min_\pi \mathbb{E}\left[z - \tilde{x}^\pi\right]_+ = (z - U) + \min_\pi \mathbb{E}\left[U - \tilde{x}^\pi\right]_+ = (z-U) + v_t^\star(s,U)$.

This approach can also be extended to the non-discounted finite-horizon case, where the bounds simplify to $L = t\underline{\mathrm{R}}$ and $U = t\bar{R}$.

Quantile Based CVaR (Chow) We implemented the algorithm described in (Chow et al., 2015; Hau et al., 2023a) with $J = 4096$ uniform discretization as

$$(B_{\max}q)(s,\alpha_j,a) := r(s,a) + \gamma \cdot \min_{\zeta\in\mathcal{Z}_{sa}(\alpha_j)}\sum_{s'\in\mathcal{S}}\zeta_{s'} \cdot \max_{a'\in\mathcal{A}} q(s', \frac{\alpha_j\zeta_{s'}}{p(s,a,s')}, a') \qquad \forall \alpha_j = \frac{j}{J} \text{ where } j \in [J]\,,$$

$$\mathcal{Z}_{sa}(\alpha) \;:=\; \{\xi \in \Delta_{\mathcal{S}} \mid \alpha\zeta_{s'} \le p(s,a,s')\}\,.$$

which follow a greedy history-dependent policy described in (Chow et al., 2015). We use the same discretization level as our algorithm. It is important to note that this algorithm could over-approximate the true static CVaR value function due to the duality gap, so it may perform badly (Hau et al., 2023a).

EVaR Algorithm described in (Hau et al., 2023b) and known to perform well when evaluated with static CVaR and EVaR. We implemented the algorithm with the time-dependent policy described there, and with fixed ERM discretization $\beta_j = 100 \cdot (0.99^j)\ \forall j \in [3000]$, instead of a domain-dependent discretization.

D.3 More empirical results

Fig. 6 extends Fig. 2 and demonstrates that the performance of $\underline{\pi}$ across all the domains to understand how the selection of $\underline{q}^{\mathrm{d}}$ in Algorithm 1 contribute to the quality of the solution. More specifically, $\underline{\pi}$ lies within $[\underline{q}^{\mathrm{d}}, \bar{q}^{\mathrm{d}}]$, whereas $\bar{\pi}$ may performs worse than $\underline{q}^{\mathrm{d}}$. Furthermore, as the discretization level increases, the bounding gap $\bar{q}^{\mathrm{d}} - \underline{q}^{\mathrm{d}}$ shrinks, suggesting that $\underline{\pi}$ converges to $\pi^\star$.

Fig. 7 extends Fig. 1 and compares our algorithm with other related algorithms (detailed in Appendix D.2) for all the domains on quantile levels $\alpha_0 \in \{0.05, 0.15, \ldots, 0.85, 0.95\}$. As we can see, our algorithm consistently outperforms all other algorithms across all tested domains and quantile levels.

Fig. 8 shows that for all the domains, both the value function and the performance of the policy for the κ-soft quantile Q-learning (Algorithm 2) with $\kappa \in \{10^{-4}, 10^{-8}, 10^{-12}, 0\}$ and uniform discretization of $J = 256$ converges to the DP variant Eq. (12) after $20,000 \mathit{iterations}$. Not only the value functions for the Q-learning converges closely to the DP's value function, the performance of the policy computed from the Q-learning value function also matches the policy from the DP variant.

Fig. 9 extends Fig. 3 and demonstrates for all the domains that the value function achieve from κ-soft quantile Q-learning (Algorithm 2) for $\kappa \in \{10^{-4}, 10^{-8}, 10^{-12}, 0\}$ and uniform discretization of $J = 256$ converges to the DP variant Eq. (12). The Wasserstein-1 distance of quantile function defined as

$$W_1(\underline{q}^{\mathrm{d}}, \tilde{\underline{q}}^{\mathrm{d}}) := \frac{1}{J} \sum_{j \in [J-1]} |\max_{a \in \mathcal{A}} \{\underline{q}^{\mathrm{d}}(s_0, j, a)\} - \max_{a \in \mathcal{A}} \{\tilde{\underline{q}}^{\mathrm{d}}(s_0, j, a)\}|.$$

We use it to evaluate the differences between two value functions. From Fig. 9 we can see that for all domains, the Wasserstein-1 distances of the value functions are far apart at the beginning and quickly converge to zero as the number of samples for each (s, a) pair increase.

E COMPARISON OF VAR-Q-LEARNING AND IQN

We now compare the Implicit Quantile Network (IQN) Q-learning algorithm proposed in Dabney et al. (2018a) to a variant of our Q-learning algorithm that stochastically approximates the expected value operation over the sampling of next risk $j' \sim U([J-1])$ using K' sampled risk level. Specifically, we focus on a version of IQN Q-learning that considers a finite horizon problem (using an additional t state dimension), models the state-value function using a piecewise constant function of the risk level, i.e. $q(t, s, \alpha, a) := \hat{q}^{\mathrm{d}}(t, s, \lfloor \alpha J \rfloor, a)$ with $\hat{q}^{\mathrm{d}} \in \mathbb{R}^{[T] \times \mathcal{S} \times [J-1] \times \mathcal{A}}$, and models the risk aversion using a distorted risk measure parameterized by some non-decreasing $\beta_{IQN} \colon [0,1] \to [0,1]$ and implied $\Gamma(j) := \mathbb{E}[\beta_{IQN}(\tilde{u}) \mid j \leq \tilde{u} \leq (j+1)/J]$ for $\tilde{u} \sim U([0,1])$. Algorithms 3 and 4 present in a comparable format how a quantile MDP and IQN approach compute their respective loss when updating their respective approximate state-value functions.

The biggest distinction between the two algorithms lies in the computation of the action or actions associated to state s'. On one hand, IQN seeks for each sampled (s, t, a, r, s') tuple a single action that captures a form of risk aversion portrayed by $\operatorname{argmax}_{a' \in \mathcal{A}} \mathbb{E}[\hat{q}^{\mathrm{d}}(t-1, s', \beta_{IQN}(\tilde{u}), a')] = \operatorname{argmax}_{a' \in \mathcal{A}} (1/J) \sum_{j'=0}^{J-1} \Gamma(j) \hat{q}^{\mathrm{d}}(t-1, s', j', a')$, where $\tilde{u} \sim U([0,1])$. On the other hand, the variant of our Quantile Q-learning algorithm seeks an optimal action for each sampled next state quantile level $j'_{k'}$. The latter reflects our finding that the quantile MDP can be solved by solving a nested VaR DP where the risk level is independently sampled from a uniform distribution at each time step. In comparison, it is not clear what criterion of optimality is satisfied by the policy evaluated by IQN; see Lim and Malik (2022) for a discussion regarding the case where $\beta_{IQN}(\cdot)$ reflect a CVaR measure.

As part of the finer differences between the two algorithms, one can observe that our quantile Q-learning employs our κ-soft quantile loss, whereas IQN uses a Huber quantile loss, denoted by $\ell^h_\alpha(\cdot)$. We also use the quantile loss function associated to the discretized level $\lfloor \tau_k J \rfloor / J$ instead of using τ_k directly in order to guarantee a conservative approximation (instead of an estimation) of the value-at-risk of level τ_k. We finally handle samples with $\tau_k < 1/J$ differently than the rest given that our Algorithm 2 prescribes steering the value of $\underline{q}^{\mathrm{d}}(t, s, 0, a)$

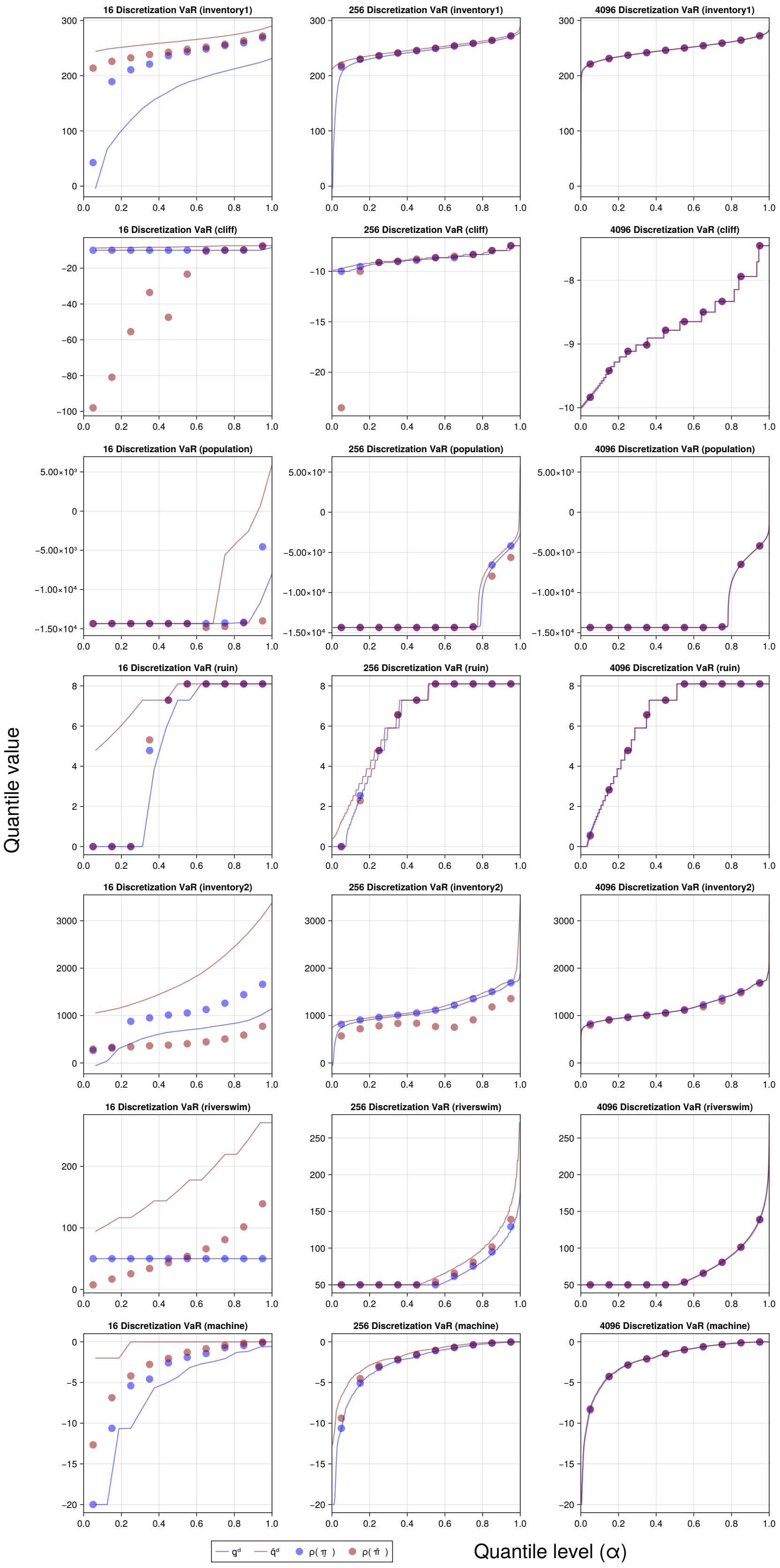

Figure 6: Approximation bound q^{d} and its respective π policy performance

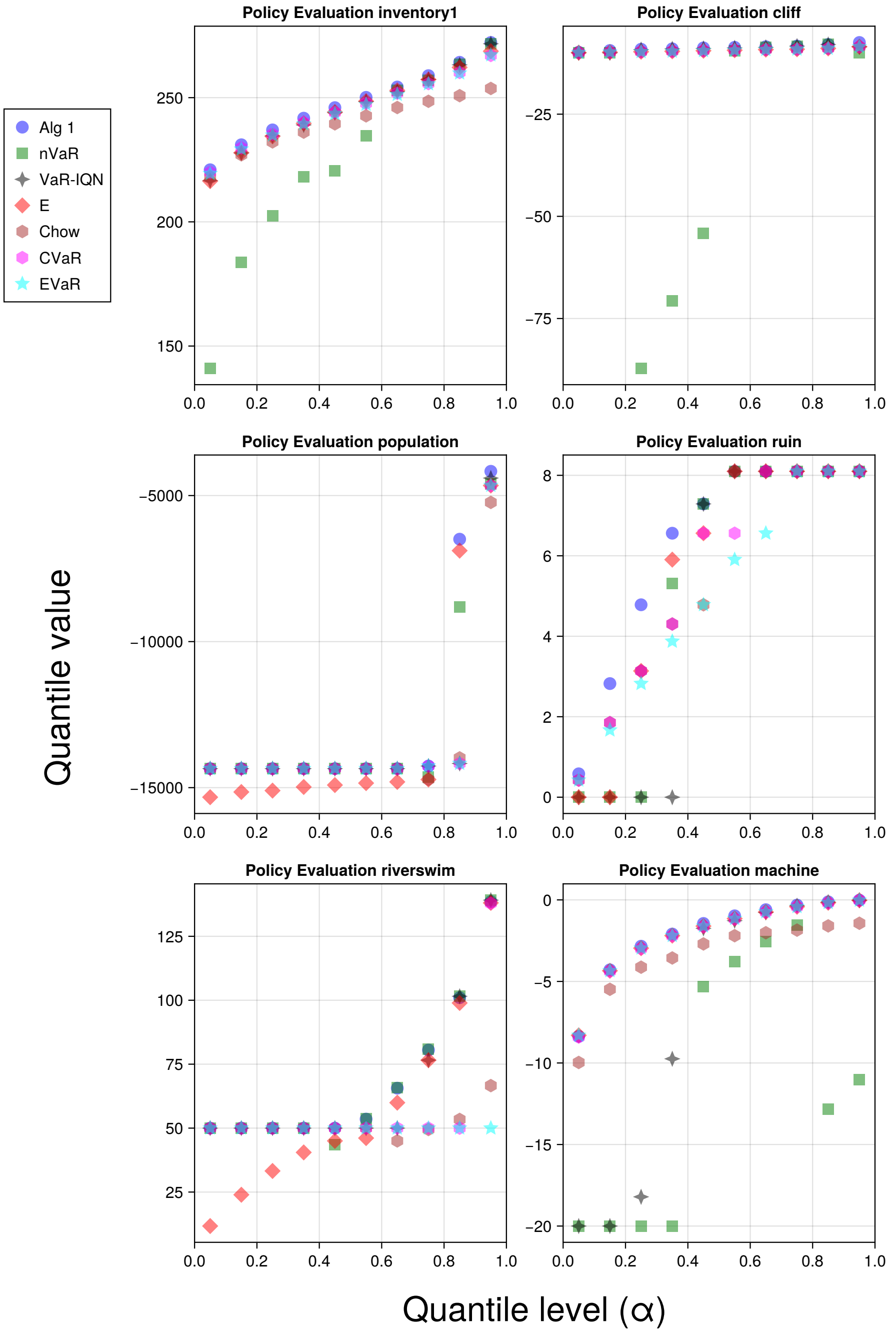

Figure 7: Policy performance evaluation $\rho(\pi)$

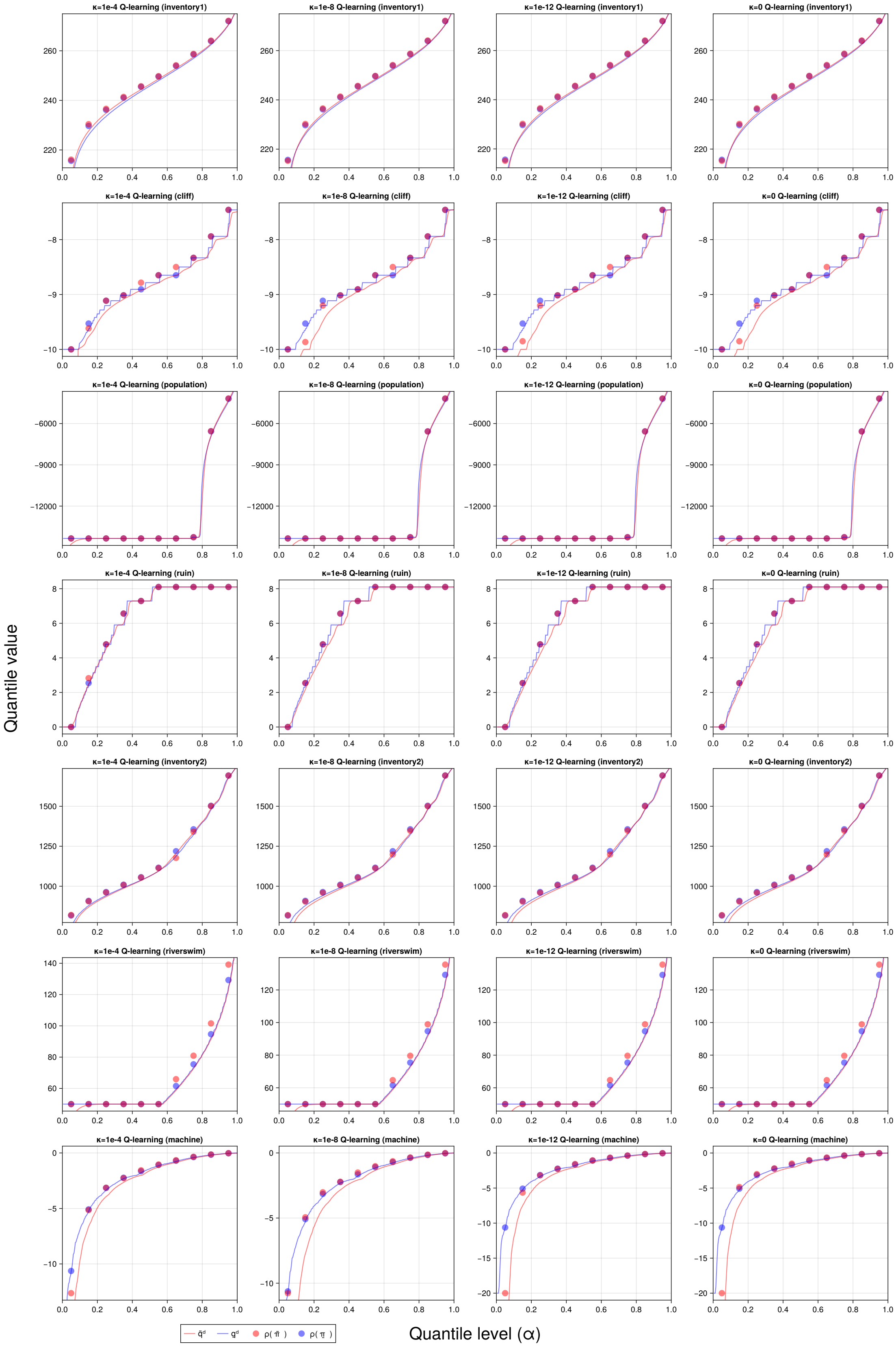

Figure 8: Q learning $\tilde{q}$ vs DP $\underline{q}$ value function and policy performance after 20,000 iterations ($J = 256$)

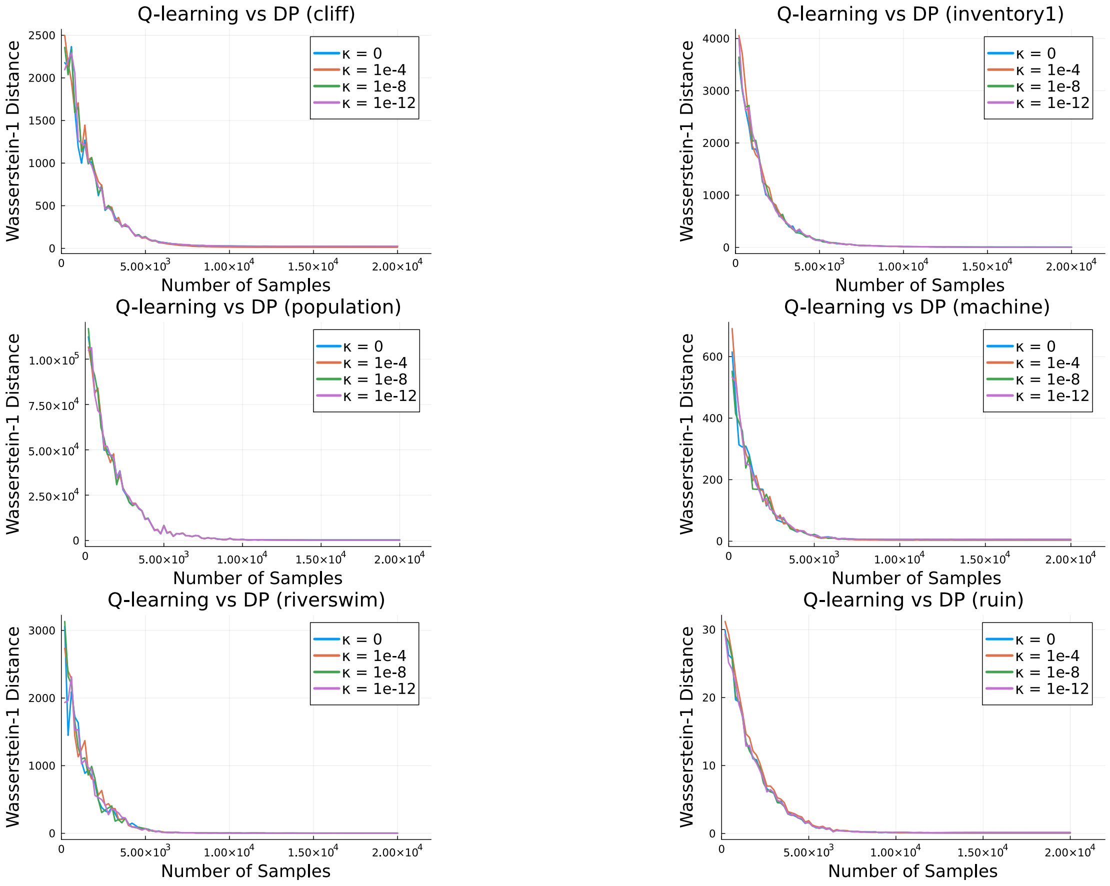

Figure 9: Q-learning vs DP value function Wasserstein distance.

towards the value of $\underline{R}t$ in order to produce a natural lower bound on the value-at-risk for risk levels in that lower range.

Algorithm 3: Risk Sampled-based Quantile Q-learning Loss (adapted from Algorithm 2)

```
Require: $K$, $K'$, $\kappa$, and functions $\underline{q}^{\mathrm{d}}$
Input: $t$, $s$, $a$, $s'$
# Sample current quantile thresholds
$\tau_k \sim U([0,1]), \quad 1 \le k \le K$
# Sample next quantile thresholds
$\tau'_{k'} \sim U([0,1]), \quad 1 \le k' \le K'$
# Compute greedy next quantile-based actions
$\boldsymbol{a^\star_{k'} \leftarrow \mathrm{argmax}_{a' \in \mathcal{A}} \underline{q}^{\mathrm{d}}(t-1, s', \lfloor \tau'_{k'} J \rfloor, a'), \quad 1 \le k' \le K'}$
# Compute distributional temporal differences
$\delta_{kk'} \leftarrow r(s,a) + \gamma \underline{q}^{\mathrm{d}}(t-1, s', \lfloor \tau'_{k'} J \rfloor, \boldsymbol{a^\star_{k'}}) - \underline{q}^{\mathrm{d}}(t, s, \lfloor \tau_k J \rfloor, a), \quad 1 \le k \le K,\ 1 \le k' \le K'$
# Compute $\kappa$-soft quantile loss
Output: $\sum_{k=1}^{K} (1/K) \sum_{k'=1}^{K'} \ell^{\boldsymbol{\kappa}}_{\lfloor \boldsymbol{\tau_k J} \rfloor / \boldsymbol{J}}(\delta_{kk'}) \cdot \mathbb{I}_{\boldsymbol{\tau_k \in [1/J, 1]}} + \boldsymbol{(\underline{q}^{\mathrm{d}}(t, s, 0, a) - \underline{R}t)^2} \cdot \mathbb{I}_{\boldsymbol{\tau_k \in [0, 1/J)}}$
```

Algorithm 4: Implicit Quantile Network Loss (adapted from Dabney et al. (2018a))

```
Require: $K$, $K'$, $h$, and functions $\Gamma$, $\hat{q}^{\mathrm{d}}$
Input: $t$, $s$, $a$, $s'$
# Sample current quantile thresholds
$\tau_k \sim U([0,1]), \quad 1 \le k \le K$
# Sample next quantile thresholds
$\tau'_{k'} \sim U([0,1]), \quad 1 \le k' \le K'$
# Compute greedy next uniform action
$\boldsymbol{a^\star \leftarrow \mathrm{argmax}_{a' \in \mathcal{A}} (1/J) \sum_{j'=0}^{J-1} \Gamma(j') \hat{q}^{\mathrm{d}}(t-1, s', j', a')}$
# Compute distributional temporal differences
$\delta_{kk'} \leftarrow r(s,a) + \gamma \hat{q}^{\mathrm{d}}(t-1, s', \lfloor \tau'_{k'} J \rfloor, \boldsymbol{a^\star}) - \hat{q}^{\mathrm{d}}(t, s, \lfloor \tau_k J \rfloor, a), \quad 1 \le k \le K,\ 1 \le k' \le K'$
# Compute Huber quantile loss
Output: $\sum_{k=1}^{K} (1/K') \sum_{k'=1}^{K'} \ell^{\boldsymbol{h}}_{\boldsymbol{\tau_k}}(\delta_{kk'})$
```