Optimistic Reinforcement Learning with Quantile Objectives

Mohammad Alipour-Vaezi
Virginia Tech, USA

Huaiyang Zhong
Virginia Tech, USA

Kwok-Leung Tsui
University of Texas at Arlington, USA

Sajad Khodadadian
Virginia Tech, USA

Abstract

Reinforcement Learning (RL) has achieved tremendous success in recent years. However, the classical foundations of RL do not account for the risk sensitivity of the objective function, which is critical in various fields, including healthcare, finance, etc. A popular approach to incorporate risk sensitivity is to optimize a specific quantile of the cumulative reward distribution. In this paper, we develop UCB–QRL, an optimistic learning algorithm for the τ-quantile objective in finite-horizon Markov decision processes (MDPs). UCB–QRL is an iterative algorithm in which, at each iteration, we first estimate the underlying transition probability and then optimize the quantile value function over a confidence ball around this estimate. Here, we show that UCB–QRL yields high-probability regret bounds $\mathcal{O}\left((2/\kappa)^H H\sqrt{SATH\log(2SATH/\delta)}\right)$ in the episodic setting with S states, A actions, T episodes, and H horizons. Here, $\kappa > 0$ is a problem-dependent constant that captures the sensitivity of the underlying MDP's quantile value.

1 INTRODUCTION

Reinforcement learning (RL) provides a general framework for sequential decision making by learning policies through interaction with an unknown environment (Sutton, Barto, et al. 1998). Over the past decade, RL—often coupled with powerful function approximators such as deep neural networks, linear models, splines, and even quantum circuits—has revolutionized our ability to solve complex, high-dimensional decision-making problems. This synergy has enabled RL agents to achieve superhuman performance in games, competitive results in robotic control and locomotion, and large-scale deployment in recommendation systems and operations research (Elfwing et al. 2017; Busoniu et al. 2017; Shakya et al. 2023; F. Zhang et al. 2020). Despite these advances, classical RL methods face an important limitation: they typically optimize expected return and are therefore risk–neutral (Moos et al. 2022). This highlights the need for formulations that account for variability and reliability of returns while maintaining data efficiency and principled exploration.

Upper Confidence Bound (UCB) RL operationalizes optimism in the face of uncertainty. At the beginning of each episode, the learner forms confidence regions around the empirical transition model, plans in the most favorable (optimistic) Markov Decision Process (MDP) inside those regions, and executes the resulting greedy policy. This plan–act–learn loop achieves near–minimax regret for finite episodic MDPs under the expectation objective (Azar et al. 2017; Auer et al. 2008).

Classical UCB methods are intrinsically risk–neutral: they optimize expected return (Liu et al. 2020). In safety–critical control, service-level guarantees, and finance, tail performance (e.g., high-percentile delivery time or loss) is paramount (Q. Yang et al. 2023). Quantile or Value-at-Risk (VaR)-based objectives capture such requirements directly. The Quantile Markov Decision Process (QMDP) furnishes a backward dynamic program for τ–quantile values, where the quantile operator is defined as $Q_\tau(X) := \inf\{x \in \mathbb{R} : \mathbb{P}(X \le x) \ge \tau\}$ for the (left-continuous) τ–quantile, via an operator on next-step quantile value maps (Li

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

et al. 2022).

A central obstacle in applying the quantile objective in RL is analytical: quantiles are nonlinear and can change abruptly with small distributional perturbations. In contrast to expectation operators, the quantile backup lacks smoothness and convexity, so classical optimism analyses that linearize value differences fail. We overcome this by (i) expressing the one–step QMDP backup as the quantile of a *continuation–mixture* random variable, and (ii) proving a sharp, local Lipschitz property of this mapping under a benign *quantile margin* at level τ. Informally, if the CDF of the continuation–mixture has a jump of size at least κ, then the τ–quantile is $(2/\kappa)$–Lipschitz with respect to the 1-Wasserstein distance. Coupled with a TV$\to W_1$ bound for mixtures and a coupling argument that uses a single auxiliary uniform variable to "align" next-state randomness, this yields a clean propagation inequality that is directly analogous to expectation-based analyses with explicit κ–dependence.

Contributions.

1. **Algorithmic framework.** We introduce UCB–QRL, an optimism-based learning algorithm for quantile objectives in finite-horizon MDPs. Each episode estimates the transition model, builds an ℓ_1 confidence set, and plans in the most favorable model via a quantile-aware planner.

2. **High-probability regret guarantees.** We study the convergence of UCB–QRL algorithm and establish a high-probability regret bound of order $\tilde{\mathcal{O}}\big((\frac{2}{\kappa})^H H\sqrt{SATH}\big)$ in the episodic setting where κ is a problem-dependent constant that captures the sensitivity of the underlying MDP's quantile value.

3. **Analytical toolkit** We develop new machinery, including (i) the continuation–mixture representation, and (ii) a coupling argument which allows us to handle the nonlinearity of the quantile value function.

2 RELATED WORK

2.1 Risk-Sensitive MDPs

Risk-sensitive objectives in sequential decision making have been studied under several paradigms. Early work on percentile/quantile criteria analyzed existence, structure, and computation in controlled Markov processes with known dynamics, including shortest-path and service-level formulations (Filar et al. 1995; Delage and Mannor 2010). The Quantile MDP (QMDP) framework formalizes dynamic programming for fixed quantile levels and establishes a backward recursion and planning algorithms under known kernels (Li et al. 2022). Closely related, distributional RL propagates the full return distribution and has yielded practical quantile parameterizations such as QR-DQN and IQN (Bellemare et al. 2017; Dabney, Rowland, et al. 2018; Dabney, Ostrovski, et al. 2018; Rowland et al. 2018; D. Yang et al. 2019). While distributional RL methods typically optimize expectation, their estimators provide tools for learning quantile slices.

Beyond quantiles, classical risk-sensitive control optimizes exponential-utility (entropic) criteria leading to modified Bellman equations and dynamic consistency (Howard and Matheson 1972). Mean–variance MDPs study return–variance trade-offs but face time-inconsistency without special structure (Sobel 1982; Mannor and Tsitsiklis 2011; Guo et al. 2012). Coherent risk measures—especially Conditional Value-at-Risk (CVaR) (Rockafellar, Uryasev, et al. 2000; "Conditional value-at-risk for general loss distributions" 2002)—enable convex surrogates and have been widely explored in RL via value-based, policy-gradient, and actor–critic methods as either objectives or constraints (Chow and Ghavamzadeh 2014; Tamar et al. 2015; Prashanth 2014). Constrained MDPs (CMDPs) and safe RL incorporate chance- or CVaR-type constraints using Lagrangian, primal–dual, or Lyapunov approaches (Altman 2021; Chow, Ghavamzadeh, et al. 2015; Q. Zhang et al. 2024; M et al. 2022; Ahmadi et al. 2020). These lines of research are largely complementary to our setting, which maximizes a fixed quantile objective rather than enforcing it as a constraint, and thus requires handling the non-smooth, set-valued nature of the quantile backup itself. Methodologically, quantile regression (Koenker and Bassett Jr 1978) underlies many practical estimators used by distributional/quantile RL, but most of this literature does not address online regret with unknown transitions (Dabney, Rowland, et al. 2018; Dabney, Ostrovski, et al. 2018; D. Yang et al. 2019).

2.2 Optimism and Upper Confidence Bounds (UCB)

Optimism in the face of uncertainty provides near-minimax regret guarantees for expectation-maximizing RL by planning in confidence sets built around empirical transition models. In average-reward communicating MDPs, the UCRL2 algorithm achieves $\tilde{\mathcal{O}}(DS\sqrt{AT})$-type guarantees via ℓ_1 confidence sets and Extended Value Iteration (EVI) (Jaksch et al. 2010). In finite-horizon problems, UCBVI algorithm attains $\tilde{\mathcal{O}}(H\sqrt{SAT})$ with Hoeffding bonuses

and $\tilde{\mathcal{O}}(\sqrt{HSAT})$ with Bernstein bonuses (Azar et al. 2017). Robust and distributionally robust MDPs planning against uncertainty sets at decision time, yielding max–min or ambiguity-aware backups that are algorithmically akin to optimistic EVI subroutines (Yu and H. Xu 2015; Goyal and Grand-Clément 2022; Q. Xu et al. 2016; Deo 2025).

Adapting optimism to nonlinear, tail-focused criteria poses additional challenges: quantile objectives are non-smooth and can change abruptly under small distributional perturbations, so linear value-difference decompositions used for expectation do not directly apply. In one-step settings, bandit studies have designed risk-aware indices for VaR/CVaR and general risk measures (Sani et al. 2012; Galichet et al. 2013; Cassel et al. 2023), clarifying how confidence design must reflect tail sensitivity. Extending these ideas to MDPs requires new contraction/sensitivity tools for the backup operator.

3 PRELIMINARIES

We consider a finite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P^\star, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H denotes the horizon, $P_h^\star(\cdot \mid s,a)$ represents the true transition kernel, and $r(s,a) \in [0,1]$ is the reward function for all states s and actions a. We assume a finite-dimensional state-action space, and we denote $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$.

For a policy π, transition kernel P, horizon step $h \in \{0,\dots,H\}$, and any quantile level $q \in (0,1)$, the q–quantile value at state s is defined as

$$V_{q,h}^{\pi,P}(s) \coloneqq \inf\Bigg\{ x \in \mathbb{R} : \mathbb{P}\Bigg(\sum_{k=h}^{H-1} r_k(S_k, A_k) \le x \;\Big|\; S_h = s,\ A_k = \pi_k(S_k), \\ S_{k+1} \sim P_k(\cdot \mid S_k, A_k), \\ k = h,\dots,H-1 \Bigg) \ge q \Bigg\}.$$

Throughout this paper, our goal is to maximize the quantile objective defined as the τ-quantile of the return distribution for a fixed target level $\tau \in (0,1)$. In particular, we consider the optimization

$$\max_{\pi}\ V_{\tau,0}^{\pi,P^\star}(\bar{s}), \tag{1}$$

where $\bar{s} = S_0^t$ is a fixed initial state. We denote by $\pi^\star$ the maximizer in Equation (1) and consider $V_{\tau,0}^\star \equiv V_{\tau,0}^{\pi^\star,P^\star}$.

4 FINITE HORIZON UCB–QRL

This section provides a detailed introduction to UCB–QRL algorithm. We first impose an assumption on the underlying MDP in order to control the sensitivity of the quantile backup.

Definition 1 (Continuation–mixture)**.** *Fix a step h, a state–action (s,a), a policy sequence π, and a transition kernel P. Let $p \coloneqq P_h(\cdot \mid s,a) \in \Delta^S$, and define $p_i \coloneqq P_h(s_i \mid s,a)$. We define the continuation–mixture random variable $Z_{s,a,h}(p; V_{\cdot,h+1}^{\pi,P})$ such that for any $x \in \mathbb{R}$*

$$\mathbb{P}\Big(Z_{s,a,h}\big(p; V_{\cdot,h+1}^{\pi,P}\big) \le x \Big) = \sum_{i=1}^{S} p_i\, \phi_i(x),$$

where

$$\phi_i(x) \coloneqq \sup\Big\{ q \in [0,1] :\ V_{q,h+1}^{\pi,P}(s_i) \le x \Big\}.$$

Intuitively, the continuation–mixture bundles next-state quantile value maps into a single scalar random variable whose q–quantile equals the one-step backup. This reduction lets us control quantile sensitivity via transportation distances on distributions rather than directly on set-valued quantile correspondences. It noteworthi,

Quantile backups are non-smooth and may change discontinuously with small perturbations of the transition law, which prevents the standard expectation-based linearization used in UCB analyses. To obtain a tractable stability estimate, we assume a jump ("margin") at the operative quantile of the continuation–mixture (Definition 1). We adopt a margin because we use the left-continuous quantile; and as a result, no two-sided density lower bound is required.

Assumption 1 (Uniform quantile margin)**.** *For each (h,s,a) and any deterministic policy π consider the continuation-mixture $Z_{s,a,h}\big(P_h^\star(\cdot \mid s,a); V_{\cdot,h+1}^{\pi,P^\star}\big)$. There exists $\kappa \in (0,1]$ such that for every $q \in [0,1]$,*

$$\mathbb{P}\Big(Z_{s,a,h}(P_h^\star(\cdot|s,a); V_{\cdot,h+1}^{\pi,P^\star}) \le c_q^\star \Big) - \\ \lim_{\epsilon \downarrow 0} \mathbb{P}\Big(Z_{s,a,h}(P_h^\star(\cdot|s,a); V_{\cdot,h+1}^{\pi,P^\star}) \le c_q^\star - \epsilon \Big) \ \ge\ \kappa.$$

where $c_q^\star \coloneqq Q_q\big(Z_{s,a,h}(P_h^\star(\cdot \mid s,a); V_{\cdot,h+1}^{\pi,P^\star})\big)$.

Since the state and action sets are finite and $H < \infty$, the collection of deterministic (time–dependent) policies is finite. With deterministic rewards and finite H, the continuation–mixture $Z_{s,a,h}\big(P_h^\star(\cdot \mid s,a); V_{\cdot,h+1}^{\pi,P^\star}\big)$ has finite support for every (s,a,h,π), so its CDF has a positive jump at each support point. Taking the

minimum jump over this finite family yields a uniform margin $\kappa \in (0,1]$, so Assumption 1 is satisfied.

We begin by fixing a designated start state $\bar{s} \in \mathcal{S}$. Each episode $t \in \{0,1,\dots,T-1\}$ starts at $S_0^t = \bar{s}$, and within each episode, steps are indexed by $h \in \{0,\dots,H-1\}$.

Morover, let $N_h^t(s,a)$ and $N_h^t(s,a,s')$ denote visit and transition counts up to (but excluding) episode t.

Next, consider a fixed confidence level $\delta \in (0,1)$. Let T denote the number of episodes, and H the number of horizons.

We introduce a universal constant

$$c \;\geq\; \frac{\max\left\{2, \sqrt{2\log\left(\frac{SATH(2^S-2)}{\delta}\right)}\right\}}{\sqrt{\log\frac{2SATH}{\delta}}},$$

and define the confidence radius as

$$f_\delta(n) = c\sqrt{\frac{\log\frac{2SATH}{\delta}}{\max\{1,n\}}}. \tag{2}$$

Using this radius, we form an empirical confidence set

$$\mathcal{C}_\delta^t := \Big\{P:\ \|P_h(\cdot|s,a) - \widehat{P}_h^t(\cdot|s,a)\|_1 \leq f_\delta(N_h^t(s,a)), \quad \forall s,a,h\Big\},$$

This set contains all transition kernels that are statistically plausible given the data observed up to episode t. On the global confidence event $\mathcal{E}_\delta$, which for all (s,a,h,t) is defined as

$$\mathcal{E}_\delta:\ \|P_h^\star(\cdot\mid s,a) - \widehat{P}_h^t(\cdot\mid s,a)\|_1 \leq f_\delta\big(N_h^t(s,a)\big),$$

we have $P^\star \in \mathcal{C}_\delta^t$ simultaneously for all t.

In parallel, we define the set of models that respect the quantile margin assumption (Assumption 1):

$$\mathcal{C}_\kappa := \Big\{P : \mathbb{P}\Big(Z_{s,a,h}(P_h(\cdot|s,a); V_{\cdot,h+1}^{\pi,P}) \leq c_{q,s,a,h}^{\pi,P}\Big) - \lim_{\epsilon\downarrow 0}\mathbb{P}\Big(Z_{s,a,h}(P_h(\cdot|s,a); V_{\cdot,h+1}^{\pi,P}) \leq c_{q,s,a,h}^{\pi,P} - \epsilon\Big) \;\geq\; \kappa, \quad \forall s,a,h, \text{deterministic } \pi\Big\},$$

where $c_{q,s,a,h}^{\pi,P} := Q_q\big(Z_{s,a,h}(P_h(\cdot\mid s,a); V_{\cdot,h+1}^{\pi,P})\big)$.

By construction, the true kernel $P^\star$ lies in $\mathcal{C}_\kappa$, since it satisfies the margin condition by Assumption 1. With high probability, it also belongs to $\mathcal{C}_\delta^t$ for all t. Thus, $P^\star \in \mathcal{C}_\delta^t \cap \mathcal{C}_\kappa$ with high probability.

These two sets together form the foundation of our learning algorithm, UCB–QRL, which adopts the optimism-in-the-face-of-uncertainty principle.

At the start of episode t, we (i) form an ℓ_1 confidence region $\mathcal{C}_\delta^{t+1}$ around the empirical kernel using the radius in Equation (2); (ii) intersect it with the margin-respecting models $\mathcal{C}_\kappa$; and (iii) plan optimistically over the intersection $\bar{\mathcal{C}}_{\delta,\kappa}^{t+1} := \mathcal{C}_\delta^{t+1} \cap \mathcal{C}_\kappa$ to obtain a policy–model pair (π^{t+1}, P^{t+1}) that maximizes the τ–quantile value at the start state. This "estimate → certify → plan" structure mirrors UCB in expectation-based RL, but the planner is quantile-aware.

Algorithm 1 UCB–QRL

1: **Input:** quantile level $\tau \in (0,1)$, confidence level $\delta \in (0,1)$
2: **Initialize:** counts $N_h^0(s,a,s') \leftarrow 0$ for all h and (s,a,s'); choose any policy π^0
3: **for** $t = 0,1,\dots,T-1$ **do**
4: *Start:* $S_0^t \leftarrow \bar{s}$
5: *Roll out under* π^t*:* generate $\big(S_h^t, A_h^t, S_{h+1}^t\big)_{h=0}^{H-1}$
6: *Update per–step counts:* for each h,

$$N_h^{t+1}(s,a,s') = \sum_{i=0}^{t} \mathbb{1}\big\{\, S_h^i = s, A_h^i = a, S_{h+1}^i = s'\big\}$$

$$N_h^{t+1}(s,a) = \sum_{i=0}^{t} \mathbb{1}\big\{\, S_h^i = s, A_h^i = a\big\}.$$

7: *Update empirical model:*

$$\widehat{P}_h^{t+1}(s'|s,a) \leftarrow \frac{N_h^{t+1}(s,a,s')}{\max\{1, N_h^{t+1}(s,a)\}}$$

8: *Build confidence sets:*

$$\bar{\mathcal{C}}_{\delta,\kappa}^{t+1} \leftarrow \mathcal{C}_\delta^{t+1} \cap \mathcal{C}_\kappa$$

9: *Optimistic re-planning*

$$\pi^{t+1} \;\in\; \arg\max_\pi \max_{P \in \bar{\mathcal{C}}_{\delta,\kappa}^{t+1}} V_{\tau,0}^{\pi,P},$$

and set P^{t+1} be any maximizer in $\bar{\mathcal{C}}_{\delta,\kappa}^{t+1}$.

10: **end for**

In Algorithm 1, lines 1–3 set the quantile target and initialize counts and policy. Lines 4–7 roll out one episode under the current policy and update per-step counts and the empirical kernel. Lines 8–10 form the confidence region $\bar{\mathcal{C}}_{\delta,\kappa}^{t+1}$ and re-plan: the inner maximization over $P \in \bar{\mathcal{C}}_{\delta,\kappa}^{t+1}$ implements optimism for the quantile (not the expectation), while the outer maxi-

mization over π produces the next policy.

Having specified the algorithm, we now turn to its performance analysis.

Our objective is to measure how much reward is lost by following UCB–QRL compared to the optimal τ–quantile policy in the true environment. This gap is captured by the notion of *quantile regret*:

$$\mathrm{Reg}_\tau(T) \;=\; \sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^\star,P^\star}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big). \quad (3)$$

where $\pi^\star$ is the optimal τ–quantile policy under the true kernel $P^\star$.

Note that τ is fixed in the regret definition, while other quantile levels $q \in (0,1)$ are only used internally to describe the full quantile value maps inside the dynamic program.

Because quantiles can be discontinuous in the underlying distribution, the regret analysis would be more delicate than in the expectation case. To control this, we invoke the margin condition (Assumption 1), which ensures a mild local regularity of the CDF. Under this assumption, we obtain the following high-probability regret bound.

Theorem 1 (High-probability Quantile-Regret)**.** *Let Assumption 1 hold with parameter $\kappa > 0$. Then for* UCB–QRL *with confidence radii shown in Equation* (2)*, with probability at least $1-2\delta$,*

$$\mathrm{Reg}_\tau(T) \;\le 2cH\left(\frac{2}{\kappa}\right)^H \sqrt{SATH\;\log\frac{2SATH}{\delta}} \;+\; \left(\frac{2}{\kappa}\right)^H H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}T\log\tfrac{2}{\delta}}\quad. \quad (4)$$

Theorem 1 provides a *high-probability* upper bound on the quantile regret $\mathrm{Reg}_\tau(T)$ defined in (3). Ignoring logarithmic factors and constants, the bound scales as

$$\mathrm{Reg}_\tau(T) \;=\; \tilde{\mathcal{O}}\left(\left(\tfrac{2}{\kappa}\right)^H H\sqrt{SATH}\right),$$

which is sublinear in the number of episodes T (specifically, $\sqrt{T}$-growth). So UCB–QRL learns a policy whose τ-quantile return approaches that of the optimal τ-quantile policy under $P^\star$.

It is instructive to compare this rate to optimism-based, risk-neutral algorithms such as UCBVI and UCB–RL2, which optimize the expected return and achieve $\tilde{\mathcal{O}}(\sqrt{HSAT})$ regret in finite-horizon tabular MDPs (up to refined Bernstein improvements). Our result preserves the same desirable sublinear $\sqrt{T}$ dependence and the standard $\sqrt{SA}$ scaling, but differs in a key way: it exhibits an explicit *quantile-sensitivity* factor $\left(\frac{2}{\kappa}\right)^H$. This factor has no analogue in expectation-based analyses and reflects a genuine difficulty of quantile objectives.

The appearance of κ is intrinsic to quantile control. Unlike expectation, the quantile operator $Q_\tau(\cdot)$ is nonlinear and can be highly sensitive to small distributional perturbations: if the CDF is nearly flat around level τ, tiny transition-model errors can induce large shifts in the backed-up quantile value. Assumption 1 enforces a jump (margin) of size at least κ at the τ-quantile of each continuation–mixture. Under this condition, Q_τ becomes locally Lipschitz with constant on the order of $2/\kappa$, so each Bellman-style backup can amplify estimation errors by at most a factor $2/\kappa$. Propagating this stability bound across H stages yields the compounded factor $\left(\frac{2}{\kappa}\right)^H$. When κ is moderate (the τ-quantile is well supported), the bound approaches the familiar risk-neutral scaling; when κ is small (a fragile tail), the result correctly reflects that learning a tail-optimal policy is substantially harder.

To the best of our knowledge, this is the first finite-time, high-probability regret guarantee for a quantile objective reinforcement learning setup.

Proof sketch. The proof follows the standard optimism in the face of uncertainty template, adapted to quantile backups via a local Lipschitz property under a margin.

(1) Optimism reduces regret to model error. We define the regret as

$$\mathrm{Reg}_\tau(T) = \sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^\star,P^\star}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big)$$

By adding and subtracting $V_{\tau,0}^{\pi^t,P^t}(\bar{s})$, by high-probability optimism at episode t with probability at least $1-\delta$,

$$\mathrm{Reg}_\tau(T) \;\le\; \sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^t,P^t}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big).$$

Thus we only need to control, per episode t, $V_{\tau,0}^{\pi^t,P^t}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})$.

(2) One–step decomposition (model vs. propagation). By adding and subtracting $V_{\tau,0}^{\pi^t,P^t}(\bar{s})$, by high-probability optimism at episode t with probability at least $1-\delta$,

$$V_{q,h}^{\pi,P}(s) \;=\; r_h(s,a) + Q_q\Big(Z_{s,a,h}\big(P_h(\cdot\mid s,a);\, V_{\cdot,h+1}^{\pi,P}\big)\Big),$$

where $Z_{s,a,h}(\cdot;\cdot)$ is the continuation–mixture variable (Definition. 1). Since the immediate reward cancels, we get the standard "model vs. propagation" split:

$$\Delta_h^t(q) \;:= \left|V_{q,h}^{\pi^t,P^t}(S_h^t) - V_{q,h}^{\pi^t,P^\star}(S_h^t)\right| \;\le\; \underbrace{\left|Q_q\big(Z(P_h^t;V_{\cdot,h+1}^{\pi^t,P^t})\big) - Q_q\big(Z(P_h^\star;V_{\cdot,h+1}^{\pi^t,P^t})\big)\right|}_{\text{model term}} + \underbrace{\left|Q_q\big(Z(P_h^\star;V_{\cdot,h+1}^{\pi^t,P^t})\big) - Q_q\big(Z(P_h^\star;V_{\cdot,h+1}^{\pi^t,P^\star})\big)\right|}_{\text{propagation term}}.$$

(3) Model term: local Lipschitz under a margin. Assumption 1 postulates a jump of size κ at the quantile of the continuation–mixture under the true kernel. Under this benign margin, the quantile operator is locally Lipschitz in W_1 distance with constant $2/\kappa$ (by quantile sensitivity under a jump margin). We then control W_1 by TV on a bounded interval and TV of mixtures by ℓ_1 distance of mixing weights. This yields the pointwise Lipschitz estimate:

$$\text{model term} \;\le\; \frac{H}{\kappa}\left\|P_h^t(\cdot\mid S_h^t,A_h^t) - P_h^\star(\cdot\mid S_h^t,A_h^t)\right\|_1.$$

On the global confidence event built from Weissman's inequality and the confidence radius (presented in Equation (2)), a triangle inequality gives

$$\|P_h^t - P_h^\star\|_1 \le 2f_\delta(N_h^t(S_h^t,A_h^t))$$

simultaneously for all (t,h).

(4) Propagation term: auxiliary–uniform coupling. Let $\mathcal{F}_h^t$ be the σ-field generated by all observations up to step h of episode t. Throughout the proof we work on the intersection of the optimism event (Algorithm 1 selects an optimistic model) and the confidence event $\mathcal{E}_\delta$, which together hold with probability at least $1-2\delta$.

Condition on $\mathcal{F}_h^t$ and couple the two continuation mixtures via the same pair (S_{h+1}^t, U), where $S_{h+1}^t \sim P_h^\star(\cdot\mid S_h^t,A_h^t)$ and $U\sim\mathrm{Unif}[0,1]$. We show,

$$\begin{aligned} Z\big(P_h^\star;V_{\cdot,h+1}^{\pi^t,P^t}\big) &\overset{d}{=} V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t),\\ Z\big(P_h^\star;V_{\cdot,h+1}^{\pi^t,P^\star}\big) &\overset{d}{=} V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)\\ &\text{conditionally on}\quad \mathcal{F}_h^t. \end{aligned} \tag{5}$$

Applying quantile sensitivity under a jump margin at level q and then taking a supremum over $q'\in[0,1]$ gives

$$\begin{aligned} &\text{propagation term} \;\le\\ &\frac{2}{\kappa}\,\mathbb{E}\Big[\,\big|V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)\big|\;\Big|\;\mathcal{F}_h^t\Big] \le\\ &\frac{2}{\kappa}\,\mathbb{E}\Big[\sup_{q'\in[0,1]}\big|V_{q',h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{q',h+1}^{\pi^t,P^\star}(S_{h+1}^t)\big|\;\Big|\;\mathcal{F}_h^t\Big] \end{aligned}$$

(5) One–step recursion and scaled supermartingale. Define

$$W_h^t(s) := \sup_{q'\in[0,1]}\big|V_{q',h}^{\pi^t,P^t}(s) - V_{q',h}^{\pi^t,P^\star}(s)\big|.$$

Steps (3)–(4) yield

$$W_h^t(S_h^t) \;\le\; \frac{H}{\kappa}\,\|P_h^t - P_h^\star\|_1 + \frac{2}{\kappa}\,\mathbb{E}\big[W_{h+1}^t(S_{h+1}^t)\mid\mathcal{F}_h^t\big].$$

Define the scaled potential

$$Y_h^t := (2/\kappa)^h W_h^t(S_h^t).$$

Then

$$\begin{aligned} &Y_h^t \;\le\; \frac{H}{\kappa}\left(\frac{2}{\kappa}\right)^h\|P_h^t - P_h^\star\|_1 \;+\; \mathbb{E}\big[Y_{h+1}^t\mid\mathcal{F}_h^t\big],\\ &\text{with}\quad |Y_{h+1}^t - \mathbb{E}[Y_{h+1}^t\mid\mathcal{F}_h^t]| \;\le\; \big(\tfrac{2}{\kappa}\big)^{h+1}H. \end{aligned}$$

Unrolling the recursion over $h=0,\dots,H-1$ (using $W_H^t\equiv 0$) gives, for each episode t,

$$\begin{aligned} &W_0^t(\bar s) \;\le\; \frac{H}{\kappa}\sum_{h=0}^{H-1}\big(\tfrac{2}{\kappa}\big)^h\|P_h^t - P_h^\star\|_1 \;-\; \sum_{h=0}^{H-1}\xi_{h+1}^t,\\ &\xi_{h+1}^t := Y_{h+1}^t - \mathbb{E}[Y_{h+1}^t\mid\mathcal{F}_h^t]. \end{aligned}$$

Summing over episodes $t=0,\dots,T-1$, substituting the confidence radii bound $\|P_h^t - P_h^\star\|_1 \le 2f_\delta(N_h^t(S_h^t,A_h^t))$, and using $V_{\tau,0}^{\pi^t,P^t}(\bar s) - V_{\tau,0}^{\pi^t,P^\star}(\bar s) \le W_0^t(\bar s)$ yields

$$\sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^t,P^t}(\bar s) - V_{\tau,0}^{\pi^t,P^\star}(\bar s)\Big) \;\le\; \underbrace{\frac{2H}{\kappa}\sum_{t,h}\big(\tfrac{2}{\kappa}\big)^h f_\delta\big(N_h^t(S_h^t,A_h^t)\big)}_{\text{radius/visit term}} \;-\; \underbrace{\sum_{t,h}\xi_{h+1}^t}_{\text{martingale term}}.$$

(6) Concentration and counting. By the standard visit-count argument,

$$\sum_{t,h} f_\delta\big(N_h^t(S_h^t,A_h^t)\big) \;\le\; c\sqrt{SATH\,\log\tfrac{2SATH}{\delta}},$$

and therefore

$$\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\big(\tfrac{2}{\kappa}\big)^h f_\delta\big(N_h^t(S_h^t,A_h^t)\big) \;\le\; \big(\tfrac{2}{\kappa}\big)^{H-1}\,2c\sqrt{SATH\,\log\tfrac{2SATH}{\delta}}.$$

A one-sided Azuma–Hoeffding bound with time-varying ranges and the bounded increment $|\xi_{h+1}^t| \le \big(\frac{2}{\kappa}\big)^{h+1}H$ give, with probability $\ge 1-\delta$,

$$-\sum_{t,h}\xi_{h+1}^t \;\le\; \big(\tfrac{2}{\kappa}\big)^H\,H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}\,T\,\log\tfrac{2}{\delta}}.$$

(7) Putting it together. Combining Steps (1)–(6), using $\|P_h^t - P_h^\star\|_1 \le 2f_\delta(\cdot)$ inside the unrolled recursion, and a union bound over the optimism/confidence events yields, with probability at least $1-2\delta$,

$$\mathrm{Reg}_\tau(T) \le 2cH\left(\frac{2}{\kappa}\right)^H \sqrt{SATH\ \log\frac{2SATH}{\delta}} + \left(\frac{2}{\kappa}\right)^H H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}T\log\tfrac{2}{\delta}}\quad .$$

which is exactly as Equation (4). This concludes the proof sketch. Full proofs of all auxiliary lemmas and constant tracking are comprehensively provided in the supplementary material. □

5 DISCUSSION AND CONCLUSION

In this section, first, we analyze the regret bound in Theorem 1 and situate it within the existing literature, emphasizing where our guarantees align with or improve upon prior results in optimistic and risk-sensitive reinforcement learning. Second, we examine the computational profile of Algorithm 1, and practical implementation choices. Third, we discuss our underlying assumption and discuss the practical implications of our results.

5.1 Toolkit novelty

In risk-neutral finite-horizon RL (e.g., UCBVI), the analysis relies critically on the linearity of expectation: one can decompose value differences through the Bellman recursion and obtain a direct telescoping argument in which transition estimation errors enter additively and propagate smoothly across stages. In contrast, for the τ-quantile objective, the Bellman-type backup involves the nonlinear and potentially discontinuous operator $Q_\tau(\cdot)$, so small perturbations of the transition kernel can cause abrupt shifts in the backed-up value; consequently, the standard expectation-based linearization and telescoping steps do not apply. Our paper therefore introduces tools that are specific to quantile control: (i) the continuation–mixture random variable $Z_{s,a,h}(\cdot;\cdot)$ (Definition 1), which rewrites the one-step QMDP backup as a scalar quantile of a mixture distribution; (ii) a margin-based local sensitivity bound under the jump condition (Assumption 1) that quantifies how transition perturbations affect the return quantile, yielding an explicit $(2/\kappa)$-Lipschitz dependence; and (iii) a tailored coupling argument (5) based on a single auxiliary uniform variable.

5.2 Computational Aspects

In the setting that we have the knowledge of the transition probability $P^\star$, Li et al. (2022) develops a backward dynamic program that computes $V^\star_{\tau,0}(s)$ as follows:

$$V^\star_{\tau,h}(s) = \max_{a\in\mathcal{A}} \mathsf{OPT}\big(s,\tau,a;V^\star_{\cdot,h+1},P^\star_h(\cdot\mid s,a)\big),\quad h=0,\dots,H-1,\quad V^\star_{\tau,H}(s)\equiv 0. \tag{6}$$

Here, for any distribution P, we define

$$\mathsf{OPT}(s,\tau,a;V_{\cdot,h+1},P) = \max_{q\in[0,1]^S} \min_{i:\ q_i\neq 1} V_{q_i,h+1}(s_i) \quad \text{s.t.} \quad \sum_{i=1}^S P(s_i)q_i \le \tau. \tag{7}$$

where $s_i \in \mathcal{S}$ denotes the i'th state. Note that since $\tau\in(0,1)$, we can always find $q\in[0,1]^S$ such that this constraint is satisfied. For the intuition on the backward dynamic programming presented above, please refer to Appendix C.

While the above recursion provides an exact planner, its literal implementation is computationally demanding. Exact quantile planning in Equation (7) requires maintaining an explicit piecewise representation of the $V^\star_{\cdot,h}(s)$ at each (s,h), and the number of breakpoints can grow exponentially with horizon. In our learning setting, the same quantile dynamic program must be solved repeatedly inside the optimistic re-planning step (Line 9) in addition to the maximization over model $P\in\bar{\mathcal{C}}^{t+1}_{\delta,\kappa}$ at every episode. Consequently, Algorithm 1 is statistically principled but does not yield a computationally tractable procedure under an exact quantile planner.

For non-quantile value functions where we define

$$V_0^{\pi,P}(s) = \mathbb{E}\left[\sum_{k=0}^{H-1} r_k(S_k,A_k)\,\middle|\, S_0=s\right],$$

Jaksch et al. (2010) introduced Extended Value Iteration (EVI) in the average-reward setting: at each (s,a), the optimistic kernel is chosen inside the ℓ_1 confidence slice to maximize the next-step value (effectively transporting probability mass toward higher-value states under the ℓ_1 budget), and the resulting optimistic MDP is solved by value iteration to obtain the greedy policy. In finite-horizon problems, the analogous optimistic planning pass is the UCBVI backward recursion of Azar et al. (2017), which can be viewed as an EVI-style update rolled over stages $h=H-1,\dots,0$.

In the UCB–QRL algorithm, Line 9 computes π^{t+1} and selects P^{t+1} as the joint maximizer of $V^{\pi,P}_{\tau,0}$. One

can develop a combination of the backward dynamic program developed by (Li et al. 2022) and the EVI algorithm developed in (Jaksch et al. 2010) to replace line 9 of Algorithm 1. This is a future direction of this research study.

5.3 Comparison with the Prior Work

Theorem 1 shows that UCB–QRL achieves high-probability quantile regret that scales as

$$\mathrm{Reg}_\tau(T) \;=\; \tilde{\mathcal{O}}\Big(\big(\tfrac{2}{\kappa}\big)^H H\sqrt{SATH}\Big),$$

up to logarithmic factors and an additive martingale term specified in Equation (4). In contrast to risk-neutral UCB results (e.g., $\tilde{\mathcal{O}}(H\sqrt{SAT})$ for UCBVI with Hoeffding bonuses and $\tilde{\mathcal{O}}(\sqrt{HSAT})$ with Bernstein bonuses (Azar et al. 2017)), and, in average-reward communicating MDPs, $\tilde{\mathcal{O}}(DS\sqrt{AT})$ for UCRL2 where D is the MDP diameter (Jaksch et al. 2010).

Our analysis exhibits an explicit sensitivity to the *quantile margin* κ. This is unavoidable for quantiles: when the CDF at level τ is flat (small κ), tiny model errors can shift Q_τ substantially, and exploration must compensate accordingly.

The factor $(2/\kappa)$ arises from iterating a local sensitivity bound across H stages.This reflects worst-case compounding under the margin assumption. Whether this dependence can be improved is open.

In many problems, margins are heterogeneous across stages and states; refined, stagewise analyses that track realized occupancy and local margins can shrink the compounding (e.g., from exponential in H to polynomial or linear in effective horizon), at the expense of heavier notation. Developing such adaptive-margin bounds is a promising direction.

We acknowledge that the factor $\big(\frac{2}{\kappa}\big)^H$ in Theorem 1 can be exponentially large in the horizon. While tightening constant factors is not the focus of this work, we conjecture that some exponential horizon dependence is unavoidable in the worst case for fixed-quantile control. A heuristic justification comes from the close relationship between quantiles and the classical entropic-risk (exponential-utility) objective

$$U_\beta(X) \;\coloneqq\; \frac{1}{\beta}\log\mathbb{E}\big[e^{\beta X}\big], \qquad \beta\neq 0.$$

For any $\beta > 0$ and any $t \in \mathbb{R}$, Markov's inequality yields

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}\big(e^{\beta X} \geq e^{\beta t}\big)\\ &\leq \frac{\mathbb{E}[e^{\beta X}]}{e^{\beta t}} = \exp\big(\beta U_\beta(X) - \beta t\big).\end{aligned}$$

Choosing t so that the right-hand side equals $1-\tau$ gives

$$\begin{aligned}t \;&= U_\beta(X) + \frac{1}{\beta}\log\frac{1}{1-\tau}\\ &\Longrightarrow \mathbb{P}(X \leq t) \geq \tau\\ &\Longrightarrow Q_\tau(X) \leq U_\beta(X) + \frac{1}{\beta}\log\frac{1}{1-\tau}.\end{aligned}$$

Similarly, applying the same argument to $-X$ gives, for any $\beta > 0$,

$$\begin{aligned}\mathbb{P}(X \leq t) &= \mathbb{P}\big(e^{-\beta X} \geq e^{-\beta t}\big)\\ &\leq \frac{\mathbb{E}[e^{-\beta X}]}{e^{-\beta t}} = \exp\big(\beta U_\beta(-X) + \beta t\big),\end{aligned}$$

and choosing $t = -U_\beta(-X) - \frac{1}{\beta}\log\frac{1}{\tau}$ ensures $\mathbb{P}(X \leq t) \leq \tau$, hence $Q_\tau(X) \geq t$. Therefore, for every $\beta > 0$,

$$\begin{aligned}-U_\beta(-X) \;-\; \frac{1}{\beta}\log\frac{1}{\tau} \;&\leq Q_\tau(X) \quad \text{and}\\ U_\beta(X) \;+\; \frac{1}{\beta}\log\frac{1}{1-\tau} \;&\geq Q_\tau(X). \qquad (8)\end{aligned}$$

Equation (8) shows that fixed-quantile control is tightly coupled to entropic-risk control (up to β-dependent additive terms). Since existing regret analyses for entropic-risk RL exhibit exponential dependence on the horizon/effective horizon (see, e.g., (Fei et al. 2020; Liang and Luo 2024; Ding et al. 2023)), it is plausible that worst-case quantile regret bounds must also incur exponential horizon dependence unless additional structure is imposed (e.g., stronger regularity or stagewise margins).

Hau et al. (2024) propose a dynamic-programming decomposition for VaR in MDPs and a model-free VaR-Q-learning algorithm that does not assume known transitions and avoids saddle-point solvers. They prove convergence of their algorithm to a unique fixed point induced by a κ-soft quantile loss; their analysis is presented for finite-horizon, time-indexed control. Our work is model-based and provides a nonasymptotic, high-probability regret bound for episodic, tabular QMDPs under a quantile margin. Hau et al. provide convergence guarantees (no regret rates) for a model-free Q-learning scheme tailored to VaR. Methodologically, they define a quantile-aware Q operator and a soft-quantile loss to ensure uniqueness of the fixed point, whereas we construct ℓ_1 confidence sets and control the quantile backup via the continuation–mixture sensitivity bound (yielding the explicit $(2/\kappa)^H$ dependence). The DP decomposition and risk-indexed operator from (Hau et al. 2024) suggest model-free extensions of UCB–QRL. Conversely, our sensitivity tools and confidence design provide a path toward

finite-sample, high-probability guarantees for VaR-style Q-learning under margin conditions. Establishing nonasymptotic regret bounds for model-free VaR control remains open.

5.4 Assumption Discussion

Assumption 1 requires that each continuation–mixture $Z_{s,a,h}$ (Definition 1) has a jump of size at least κ at the τ-quantile under $P^\star$. In our finite, deterministic-reward setting, these mixtures are discrete, hence a (problem-dependent) $\kappa > 0$ always exists, though it may be small. Analytically, κ controls the local Lipschitz constant of the quantile backup with respect to the Wasserstein distance (W_1), which is the key to the propagation inequality. Practically, larger margins arise when next-state value distributions allocate non-negligible mass exactly at the operative quantile; small margins indicate intrinsically fragile tails and make tail-optimal learning harder.

Our analysis makes a simplifying assumption. We work in the finite tabular setting with episodic horizons, leaving extensions to function approximation (linear, kernelized, or neural) as an open question that will require new concentration tools for distributional value errors. Rewards are assumed deterministic given (s, a, s'); when rewards are stochastic, the continuation–mixture can absorb the noise without altering the quantile-optimality structure (Li et al. 2022). Finally, the algorithm itself does not require κ, but the bound does. Developing data-driven methods to estimate local margins and adapt bonuses accordingly could yield sharper, data-dependent guarantees.

Quantile-optimal learning directly targets tail risk and is natural for safety-critical domains such as service-level guarantees, latency minimization, and adverse-event prevention. The explicit κ-dependence aligns theory with practice: when the operative quantile is well supported, learning is efficient; when the tail is thin, the bound correctly reflects the increased difficulty. This observation motivates several future directions, including relaxing the quantile target, adopting smoother risk measures such as CVaR, or incorporating problem-specific structure through priors. Beyond this, Bernstein-type confidence bonuses may remove the extra $\sqrt{H}$ factor, and the continuation–mixture framework may extend to other coherent risk objectives; a formal analysis is left to future work. Combining distributional critics with optimism-based exploration in large-scale problems, as well as extending our techniques to infinite-horizon discounted or average-reward settings, remains an important challenge for future work.

We conjecture that Assumption 1 is necessary to achieve a sublinear regret for Quantile MDPs. In particular, we believe that if κ is not known, and the optimization in line 9 is performed over $\mathcal{C}_\delta^t$ only, there exists an instance of MDP and τ for which the Algorithm 1 will cause a linear regret.

In practice, where κ is not known, one can start from an initial $\kappa_0 \in (0, 1)$ point. As the algorithm runs, if empirical regret appears linear, gradually reduce κ until the algorithm starts converging.

Finally, two natural extensions are left open.

1. *Discounted, infinite-horizon quantile control:* develop an optimistic algorithm and analysis for the discounted objective with $\gamma \in (0, 1)$, including a discounted continuation–mixture operator, an appropriate contraction/sensitivity inequality for the quantile backup, and stationary ℓ_1-confidence sets for model uncertainty.

2. *Lower bounds for regret:* establish minimax and instance-dependent lower bounds for quantile-regret under margin assumptions to determine which dependencies on κ, S, A, and horizon (or effective horizon) are improvable versus unavoidable.

References

Sutton, R. S., A. G. Barto, et al. (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge.

Elfwing, S., E. Uchibe, and K. Doya (2017). "Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning". In: *Neural networks : the official journal of the International Neural Network Society* 107, pp. 3–11. DOI: 10.1016/j.neunet.2017.12.012.

Busoniu, L., R. Babuska, B. De Schutter, and D. Ernst (2017). *Reinforcement learning and dynamic programming using function approximators*. CRC press.

Shakya, A., G. Pillai, and S. Chakrabarty (2023). "Reinforcement learning algorithms: A brief survey". In: *Expert Syst. Appl.* 231, p. 120495. DOI: 10.1016/j.eswa.2023.120495.

Zhang, F., J. Li, and Z. Li (2020). "A TD3-based multi-agent deep reinforcement learning method in mixed cooperation-competition environment". In: *Neurocomputing* 411, pp. 206–215. DOI: 10.1016/j.neucom.2020.05.097.

Moos, J., K. Hansel, H. Abdulsamad, S. Stark, D. Clever, and J. Peters (2022). "Robust Reinforcement Learning: A Review of Foundations and Recent Advances". In: *Mach. Learn. Knowl. Extr.* 4, pp. 276–315. DOI: 10.3390/make4010013.

Azar, M. G., I. Osband, and R. Munos (2017). "Minimax regret bounds for reinforcement learning". In: *International conference on machine learning*. PMLR, pp. 263–272.

Auer, P., T. Jaksch, and R. Ortner (2008). "Near-optimal regret bounds for reinforcement learning". In: *Advances in neural information processing systems* 21.

Liu, X., M. Derakhshani, S. Lambotharan, and M. Van der Schaar (2020). "Risk-aware multi-armed bandits with refined upper confidence bounds". In: *IEEE Signal Processing Letters* 28, pp. 269–273.

Yang, Q., T. D. Simão, S. H. Tindemans, and M. T. Spaan (2023). "Safety-constrained reinforcement learning with a distributional safety critic". In: *Machine Learning* 112.3, pp. 859–887.

Li, X., H. Zhong, and M. L. Brandeau (2022). "Quantile Markov decision processes". In: *Operations research* 70.3, pp. 1428–1447.

Filar, J. A., D. Krass, and K. W. Ross (1995). "Percentile performance criteria for limiting average Markov decision processes". In: *IEEE Transactions on Automatic Control* 40.1, pp. 2–10.

Delage, E. and S. Mannor (2010). "Percentile optimization for Markov decision processes with parameter uncertainty". In: *Operations research* 58.1, pp. 203–213.

Bellemare, M. G., W. Dabney, and R. Munos (2017). "A distributional perspective on reinforcement learning". In: *International conference on machine learning*. PMLR, pp. 449–458.

Dabney, W., M. Rowland, M. Bellemare, and R. Munos (2018). "Distributional reinforcement learning with quantile regression". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Dabney, W., G. Ostrovski, D. Silver, and R. Munos (2018). "Implicit quantile networks for distributional reinforcement learning". In: *International conference on machine learning*. PMLR, pp. 1096–1105.

Rowland, M., M. G. Bellemare, W. Dabney, R. Munos, and Y. Teh (2018). "An Analysis of Categorical Distributional Reinforcement Learning". In: pp. 29–37.

Yang, D., L. Zhao, Z. Lin, T. Qin, J. Bian, and T.-Y. Liu (2019). "Fully parameterized quantile function for distributional reinforcement learning". In: *Advances in neural information processing systems* 32.

Howard, R. A. and J. E. Matheson (1972). "Risk-sensitive Markov decision processes". In: *Management science* 18.7, pp. 356–369.

Sobel, M. J. (1982). "The variance of discounted Markov decision processes". In: *Journal of Applied Probability* 19.4, pp. 794–802.

Mannor, S. and J. Tsitsiklis (2011). "Mean-variance optimization in Markov decision processes". In: *arXiv preprint arXiv:1104.5601*.

Guo, X., L. Ye, and G. Yin (2012). "A mean–variance optimization problem for discounted Markov decision processes". In: *European Journal of Operational Research* 220.2, pp. 423–429.

Rockafellar, R. T., S. Uryasev, et al. (2000). "Optimization of conditional value-at-risk". In: *Journal of risk* 2, pp. 21–42.

"Conditional value-at-risk for general loss distributions" (2002). In: *Journal of Banking & Finance* 26.7, pp. 1443–1471.

Chow, Y. and M. Ghavamzadeh (2014). "Algorithms for CVaR optimization in MDPs". In: *Advances in neural information processing systems* 27.

Tamar, A., Y. Glassner, and S. Mannor (2015). "Optimizing the CVaR via sampling". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1.

Prashanth, L. (2014). "Policy gradients for CVaR-constrained MDPs". In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 155–169.

Altman, E. (2021). *Constrained Markov decision processes*. Routledge.

Chow, Y., M. Ghavamzadeh, L. Janson, and M. Pavone (2015). "Risk-Constrained Reinforcement Learning with Percentile Risk Criteria". In: *ArXiv* abs/1512.01629.

Zhang, Q. et al. (2024). "CVaR-Constrained Policy Optimization for Safe Reinforcement Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 36, pp. 830–841. DOI: 10.1109/TNNLS.2023.3331304.

M, U. K., S. Bhat, V. Kavitha, and N. Hemachandra (2022). "Approximate solutions to constrained risk-sensitive Markov decision processes". In: *Eur. J. Oper. Res.* 310, pp. 249–267. DOI: 10.1016/j.ejor.2023.02.039.

Ahmadi, M., U. Rosolia, M. Ingham, R. Murray, and A. Ames (2020). "Constrained Risk-Averse Markov Decision Processes". In: *ArXiv* abs/2012.02423. DOI: 10.1609/aaai.v35i13.17393.

Koenker, R. and G. Bassett Jr (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Jaksch, T., R. Ortner, and P. Auer (2010). "Near-optimal Regret Bounds for Reinforcement Learning". In: *Journal of Machine Learning Research* 11, pp. 1563–1600.

Yu, P. and H. Xu (2015). "Distributionally Robust Counterpart in Markov Decision Processes". In: *IEEE Transactions on Automatic Control* 61, pp. 2538–2543. DOI: 10.1109/TAC.2015.2495174.

Goyal, V. and J. Grand-Clément (2022). "Robust Markov Decision Processes: Beyond Rectangularity". In: *Math. Oper. Res.* 48, pp. 203–226. DOI: 10.1287/moor.2022.1259.

Xu, Q., X. Liu, C. Jiang, and K. Yu (2016). "Quantile autoregression neural network model with applications to evaluating value at risk". In: *Appl. Soft Comput.* 49, pp. 1–12. DOI: 10.1016/j.asoc.2016.08.003.

Deo, A. (2025). "On Design of Representative Distributionally Robust Formulations for Evaluation of Tail Risk Measures". In: *arXiv preprint arXiv:2506.16230*.

Sani, A., A. Lazaric, and R. Munos (2012). "Risk-aversion in multi-armed bandits". In: *Advances in neural information processing systems* 25.

Galichet, N., M. Sebag, and O. Teytaud (2013). "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits". In: *Asian conference on machine learning*. PMLR, pp. 245–260.

Cassel, A., S. Mannor, and A. Zeevi (2023). "A general framework for bandit problems beyond cumulative objectives". In: *Mathematics of Operations Research* 48.4, pp. 2196–2232.

Fei, Y., Z. Yang, Y. Chen, Z. Wang, and Q. Xie (2020). "Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret". In: *Advances in Neural Information Processing Systems* 33, pp. 22384–22395.

Liang, H. and Z.-Q. Luo (2024). "Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds". In: *Journal of Machine Learning Research* 25.221, pp. 1–56.

Ding, Y., M. Jin, and J. Lavaei (2023). "Non-stationary risk-sensitive reinforcement learning: Near-optimal dynamic regret, adaptive detection, and separation design". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 6, pp. 7405–7413.

Hau, J. L., E. Delage, E. Derman, M. Ghavamzadeh, and M. Petrik (2024). "Q-learning for quantile MDPs: A decomposition, performance, and convergence analysis". In: *arXiv preprint arXiv:2410.24128*.

Villani, C. (2009). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer.

Wang, K., N. Kallus, and W. Sun (2023). "Near-minimax-optimal risk-sensitive reinforcement learning with cvar". In: *International Conference on Machine Learning*. PMLR, pp. 35864–35907.

Gibbs, A. L. and F. E. Su (2002). "On choosing and bounding probability metrics". In: *International statistical review* 70.3, pp. 419–435.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A PROOF OF THEOREM 1

. We have

$$\begin{aligned}\mathrm{Reg}_\tau(T) &= \sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^\star,P^\star}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big)\\ &= \sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^\star,P^\star}(\bar{s}) - V_{\tau,0}^{\pi^t,P^t}(\bar{s}) + V_{\tau,0}^{\pi^t,P^t}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big). \qquad \text{(Adding and subtracting } V_{\tau,0}^{\pi^t,P^t}(\bar{s}))\end{aligned}$$

By Lemma 1 at $s=\bar{s}$, with probability at least $1-\delta$, for all $t\geq 0$ we have

$$V_{\tau,0}^{\pi^\star,P^\star}(\bar{s}) \leq V_{\tau,0}^{\pi^t,P^t}(\bar{s}),$$

which implies that with probability at least $1-\delta$,

$$\mathrm{Reg}_\tau(T) \;\leq\; \sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^t,P^t}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big).$$

For any policy π, any state s, horizon h, any quantile level q and transition probability P, by Lemma 3, the QMDP Bellman recursion gives

$$V_{q,h}^{\pi,P}(s) \;=\; r_h(s,a_h) \;+\; Q_q\Big(Z_{s,a_h,h}\big(P_h(\cdot\mid s,a_h);\; V_{\cdot,h+1}^{\pi,P}\big)\Big), \tag{9}$$

where $a_h=\pi_h(s)$.

Applying Equation (9) with P^t (optimistic kernel) and $P^\star$ (true kernel), by Lemma 3 the difference is

$$\begin{aligned}\Delta_h^t(q) \;&:=\; \Big|V_{q,h}^{\pi^t,P^t}(S_h^t) - V_{q,h}^{\pi^t,P^\star}(S_h^t)\Big|\\ &=\; \Big|r_h(S_h^t,A_h^t) + Q_q\Big(Z_{S_h^t,A_h^t,h}(P_h^t(\cdot\mid S_h^t,A_h^t); V_{\cdot,h+1}^{\pi^t,P^t})\Big) - r_h(S_h^t,A_h^t) - Q_q\Big(Z_{S_h^t,A_h^t,h}(P_h^\star(\cdot\mid S_h^t,A_h^t); V_{\cdot,h+1}^{\pi^t,P^\star})\Big)\Big|.\end{aligned}$$

Here, we have $A_h^t=\pi_h^t(S_h^t)$. It's noteworthy that according to Lemma 5, π^t is a deterministic policy and hence denoting A_h^t as above is appropriate.

Since the immediate reward $r_h(S_h^t,a_h)$ does not depend on P, it cancels out exactly. Hence

$$\Delta_h^t(q) \;=\; \Big|Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^t(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^t}\big)\Big)\Big| \;-\; Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^\star(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^\star}\big)\Big). \tag{10}$$

Adding and subtracting $Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^\star(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^t}\big)\Big)$, we can write

$$\begin{aligned}\Delta_h^t(q) = \Bigg|&\underbrace{\Big[Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^t(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^t}\big)\Big) - Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^\star(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^t}\big)\Big)\Big]}_{\text{model term}}\\ &+\underbrace{\Big[Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^\star(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^t}\big)\Big) - Q_q\Big(Z_{S_h^t,A_h^t,h}\big(P_h^\star(\cdot\mid S_h^t,A_h^t);\; V_{\cdot,h+1}^{\pi^t,P^\star}\big)\Big)\Big]}_{\text{propagation term}}\Bigg|.\end{aligned} \tag{11}$$

The *model term* isolates the effect of using P_h^t instead of $P_h^\star$ while freezing the continuation map; by Lemma 6 and Assumption 1,

$$\text{model term} \;\le\; \frac{H}{\kappa}\,\big\|P_h^t(\cdot|S_h^t,A_h^t) - P_h^\star(\cdot|S_h^t,A_h^t)\big\|_1.$$

Let $p^\star := P_h^\star(\cdot \mid S_h^t, A_h^t)$ and draw $S_{h+1}^t \sim p^\star$. Write

$$\text{propagation term} \;=\; \Big|Q_q\Big(Z_{S_h^t,A_h^t,h}\big(p^\star;\, V_{\cdot,h+1}^{\pi^t,P^t}\big)\Big) \;-\; Q_q\Big(Z_{S_h^t,A_h^t,h}\big(p^\star;\, V_{\cdot,h+1}^{\pi^t,P^\star}\big)\Big)\Big|, \tag{12}$$

Define the continuation–mixture variables

$$Z^t \;:=\; Z_{S_h^t,A_h^t,h}\big(p^\star;\, V_{\cdot,h+1}^{\pi^t,P^t}\big), \qquad Z^\star \;:=\; Z_{S_h^t,A_h^t,h}\big(p^\star;\, V_{\cdot,h+1}^{\pi^t,P^\star}\big).$$

By Definition 1, Z^t and $Z^\star$ has CDF $\Phi_{p^\star}^t(y) = \sum_i p_i^\star\,\phi_i^t(y)$ and $\Phi_{p^\star}^\star(y) = \sum_i p_i^\star\,\phi_i^\star(y)$ respectively with $\phi_i^{\cdot}(y) = \sup\{q \in [0,1] : V_{q,h+1}^{\pi^t,P^{\cdot}}(s_i) \le y\}$. Let $U \sim \mathrm{Unif}[0,1]$ be independent of S_{h+1}^t. Let $\mathcal{F}_h^t$ be the σ-algebra up to step h of episode t. By Lemma 10, $Z^t \overset{d}{=} V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t)$ and $Z^\star \overset{d}{=} V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)$ *conditionally on* $\mathcal{F}_h^t$. Combining with Equation (12) and the quantile form above, we obtain

$$\text{propagation term} \;=\; \Big|Q_q\Big(V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t)\Big) \;-\; Q_q\Big(V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)\Big)\Big| \qquad \text{(in law, conditionally on } \mathcal{F}_h^t\text{)}. \tag{13}$$

By Lemma 7, applied at level q with margin parameter $\kappa > 0$ from Assumption 1,

$$\big|Q_q(Z^t) - Q_q(Z^\star)\big| \;\le\; \frac{2}{\kappa}\,W_1\Big(\mathcal{L}(Z^t \mid \mathcal{F}_h^t),\; \mathcal{L}(Z^\star \mid \mathcal{F}_h^t)\Big). \tag{14}$$

By the primal (Kantorovich) formulation of W_1 Villani 2009,

$$W_1(\mu,\nu) \;=\; \inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|\,d\pi(x,y), \tag{15}$$

so for *any* coupling π of the two laws we have $W_1(\mu,\nu) \le \mathbb{E}_\pi[\,|X-Y|\,]$. We take the explicit coupling that uses the same randomness (S_{h+1}^t, U) to generate $X = V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t)$ and $Y = V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)$ conditionally on $\mathcal{F}_h^t$, which yields

$$W_1(\cdot,\cdot) \;\le\; \mathbb{E}\Big[\,\big|V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)\big| \;\Big|\; \mathcal{F}_h^t\Big]. \tag{16}$$

Combining Equations (14) and (16) with Equatin (13) yields the conditional bound

$$\begin{aligned}\text{propagation term} \;&\le\; \frac{2}{\kappa}\,\mathbb{E}\Big[\,\big|V_{U,h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{U,h+1}^{\pi^t,P^\star}(S_{h+1}^t)\big| \;\Big|\; \mathcal{F}_h^t\Big] \\ &\le\; \frac{2}{\kappa}\,\mathbb{E}\Big[\sup_{q'\in[0,1]}\big|V_{q',h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{q',h+1}^{\pi^t,P^\star}(S_{h+1}^t)\big|\Big|\; \mathcal{F}_h^t\Big]\end{aligned} \tag{17}$$

where the last inequality is due to $U \sim \mathrm{Unif}[0,1]$.

Combining the propagation term with the model-term bound gives, and conditioning on $\mathcal{F}_h^t$,

$$\Delta_h^t(q) \;\le\; \frac{H}{\kappa}\,\big\|P_h^t(\cdot \mid S_h^t,A_h^t) - P_h^\star(\cdot \mid S_h^t,A_h^t)\big\|_1 \;+\; \frac{2}{\kappa}\,\mathbb{E}\Big[\sup_{q'\in[0,1]}\big|V_{q',h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{q',h+1}^{\pi^t,P^\star}(S_{h+1}^t)\big| \;\Big|\; \mathcal{F}_h^t\Big]. \tag{18}$$

Define

$$W_h^t(S_h^t) := \sup_{q'\in[0,1]}\Big|V_{q',h+1}^{\pi^t,P^t}(S_{h+1}^t) - V_{q',h+1}^{\pi^t,P^\star}(S_{h+1}^t)\Big|$$

Taking the supremum over q on the left hand side of Equation (18), we get

$$W_h^t(S_h^t) \;\le\; \frac{H}{\kappa}\,\big\|P_h^t(\cdot \mid S_h^t,A_h^t) - P_h^\star(\cdot \mid S_h^t,A_h^t)\big\|_1 \;+\; \frac{2}{\kappa}\,\mathbb{E}\Big[\,W_{h+1}^t(S_{h+1}^t) \;\Big|\; \mathcal{F}_h^t\Big]. \tag{19}$$

Introduce the scaled potential

$$Y_h^t \;\coloneqq\; \left(\frac{2}{\kappa}\right)^h W_h^t(S_h^t), \qquad \xi_{h+1}^t \;\coloneqq\; Y_{h+1}^t - \mathbb{E}\big[Y_{h+1}^t \mid \mathcal{F}_h^t\big].$$

Then Equation (19) is equivalent to the *one-step supermartingale* inequality

$$Y_h^t \;\le\; \frac{H}{\kappa}\left(\frac{2}{\kappa}\right)^h \big\|P_h^t(\cdot \mid S_h^t, A_h^t) - P_h^\star(\cdot \mid S_h^t, A_h^t)\big\|_1 \;+\; \mathbb{E}\big[Y_{h+1}^t \mid \mathcal{F}_h^t\big], \qquad |\xi_{h+1}^t| \;\le\; \left(\frac{2}{\kappa}\right)^{h+1} H. \tag{20}$$

Unrolling Equation (20) over $h = 0, \dots, H-1$ and using $W_H^t \equiv 0$ (hence $Y_H^t \equiv 0$) yields

$$Y_0^t \;\le\; \frac{H}{\kappa}\sum_{h=0}^{H-1}\left(\frac{2}{\kappa}\right)^h \big\|P_h^t(\cdot \mid S_h^t, A_h^t) - P_h^\star(\cdot \mid S_h^t, A_h^t)\big\|_1 \;-\; \sum_{h=0}^{H-1}\xi_{h+1}^t. \tag{21}$$

Therefore,

$$W_0^t(\bar{s}) \;\le\; \frac{H}{\kappa}\sum_{h=0}^{H-1}\left(\frac{2}{\kappa}\right)^h \big\|P_h^t(\cdot \mid S_h^t, A_h^t) - P_h^\star(\cdot \mid S_h^t, A_h^t)\big\|_1 \;-\; \sum_{h=0}^{H-1}\xi_{h+1}^t.$$

By a union bound and Lemma 2, with probability at least $1-\delta$ the confidence events

$$\big\|P_h^\star(\cdot \mid s,a) - \widehat{P}_h^t(\cdot \mid s,a)\big\|_1 \le f_\delta\big(N_h^t(s,a)\big), \qquad \big\|P_h^t(\cdot \mid s,a) - \widehat{P}_h^t(\cdot \mid s,a)\big\|_1 \le f_\delta\big(N_h^t(s,a)\big) \tag{22}$$

hold simultaneously for all (t,h,s,a); hence, by the triangle inequality,

$$\big\|P_h^t(\cdot \mid S_h^t, A_h^t) - P_h^\star(\cdot \mid S_h^t, A_h^t)\big\|_1 \;\le\; 2\, f_\delta\big(N_h^t(S_h^t, A_h^t)\big). \tag{23}$$

Therefore with probability at least $1-\delta$ one can rewrite the bound of $W_0^t(\bar{s})$ as

$$W_0^t(\bar{s}) \;\le\; \frac{H}{\kappa}\sum_{h=0}^{H-1}\left(\frac{2}{\kappa}\right)^h 2\, f_\delta\big(N_h^t(S_h^t, A_h^t)\big) \;-\; \sum_{h=0}^{H-1}\xi_{h+1}^t.$$

Summing over $t = 0, \dots, T-1$

$$\sum_{t=0}^{T-1} W_0^t(\bar{s}) \;\le\; \frac{H}{\kappa}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\left(\frac{2}{\kappa}\right)^h 2\, f_\delta\big(N_h^t(S_h^t, A_h^t)\big) \;-\; \sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\xi_{h+1}^t.$$

Moreover, $\{\xi_h^t\}$ is a martingale-difference sequence with $|\xi_h^t| \;\le\; \big(\frac{2}{\kappa}\big)^h H$, so we apply the *one-sided* Azuma–Hoeffding inequality for martingale differences with non-identical bounds: for all $\lambda > 0$, we have

$$\mathbb{P}\left(-\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\xi_{h+1}^t \ge \lambda\right) \;\le\; \exp\Big(-\frac{\lambda^2}{2\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\big(\big(\frac{2}{\kappa}\big)^h H\big)^2}\Big).$$

Furthermore, we have

$$\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\left(\left(\frac{2}{\kappa}\right)^h H\right)^2 = T\,H^2\sum_{h=0}^{H-1}\left(\tfrac{2}{\kappa}\right)^{2h} = T\,H^2\frac{(2/\kappa)^{2H}-1}{(4/\kappa^2)-1} = T\,H^2\Big(\tfrac{2}{\kappa}\Big)^{2H}\frac{1-(\kappa/2)^{2H}}{(4/\kappa^2)-1}.$$

Applying the one-sided inequality with the display above and setting the right-hand side to δ yields, with probability at least $1-\delta$,

$$-\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\xi_h^t \;\le\; \left(\frac{2}{\kappa}\right)^H H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}T\log\tfrac{1}{\delta}}. \tag{24}$$

Moreover, using Elliptical Potential Lemma Wang et al. 2023, Lemma G.12, we have

$$\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^t(S_h^t,A_h^t)}} \le \sqrt{SAHT\log T}.$$

Hence,

$$\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\left(\frac{2}{\kappa}\right)^h f_\delta\big(N_h^t(S_h^t,A_h^t)\big) \;\le\; \left(\frac{2}{\kappa}\right)^{H-1} 2c\sqrt{SATH\;\log\frac{2SATH}{\delta}}. \tag{25}$$

Combining Equations (24) and (25) we have with probability at least $1-\delta$,

$$\sum_{t=0}^{T-1} W_0^t(\bar{s}) \;\le\; \frac{H}{\kappa}\left(\frac{2}{\kappa}\right)^{H-1} 2c\sqrt{SATH\;\log\frac{2SATH}{\delta}} + \left(\frac{2}{\kappa}\right)^{H} H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}T\log\tfrac{2}{\delta}}.$$

Noting $\sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^t,P^t}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big) \le \sum_{t=0}^{T-1} W_0^t(\bar{s})$, with probability at least $1-\delta$,

$$\sum_{t=0}^{T-1}\Big(V_{\tau,0}^{\pi^t,P^t}(\bar{s}) - V_{\tau,0}^{\pi^t,P^\star}(\bar{s})\Big) \;\le\; \frac{H}{\kappa}\left(\frac{2}{\kappa}\right)^{H-1} 2c\sqrt{SATH\;\log\frac{2SATH}{\delta}} + \left(\frac{2}{\kappa}\right)^{H} H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}T\log\tfrac{2}{\delta}}.$$

Recalling that with probability at least $1-\delta$, $\mathrm{Reg}_\tau(T) \le \sum_{t=0}^{T-1}\Delta_0^t$, using union bound we conclude that with probability at least $1-2\delta$,

$$\mathrm{Reg}_\tau(T) \;\le\; 2cH\left(\frac{2}{\kappa}\right)^{H}\sqrt{SATH\;\log\frac{2SATH}{\delta}} \;+\; \left(\frac{2}{\kappa}\right)^{H} H\sqrt{\frac{2\big(1-(\kappa/2)^{2H}\big)}{\frac{4}{\kappa^2}-1}T\log\tfrac{2}{\delta}}. \tag{26}$$

This completes the proof. □

B TECHNICAL LEMMAS

Lemma 1 (High-probability optimism at episode t)**.** *Consider Algorithm 1. For every episode t and state s, with probability at least $1-\delta$, we have*

$$V_{\tau,0}^{\pi^\star,P^\star}(s) \;\le\; V_{\tau,0}^{\pi^t,P^t}(s).$$

Proof of Lemma 1. Define the "good event"

$$\mathcal{E}_\delta \;\coloneqq\; \bigcap_{t=0}^{T-1}\bigcap_{h=0}^{H-1}\bigcap_{s,a}\mathcal{G}_{t,h}(s,a),$$

where (t,h,s,a), define

$$\mathcal{G}_{t,h}(s,a) \;\coloneqq\; \Big\{\big\|P_h^\star(\cdot\mid s,a) - \widehat{P}_h^t(\cdot\mid s,a)\big\|_1 \;\le\; f_\delta\big(N_h^t(s,a)\big)\Big\}.$$

By Lemma 2, for any $\varepsilon>0$,

$$\mathbb{P}\Big(\big\|P^\star(\cdot\mid s,a) - \widehat{P}_h^t(\cdot\mid s,a)\big\|_1 \ge \varepsilon \;\Big|\; N_h^t(s,a)=n\Big) \;\le\; (2^S-2)\,\exp\Big(-\frac{n\varepsilon^2}{2}\Big).$$

Choosing $\varepsilon = f_\delta(n) = c\sqrt{\frac{\log\frac{2SATH}{\delta}}{\max\{1,n\}}}$ (and $n\ge 1$),

$$\mathbb{P}\big(\mathcal{G}_{t,h}(s,a)^{\mathsf{c}} \mid N_h^t(s,a)=n\big) \;\le\; (2^S-2)\left(\tfrac{2SATH}{\delta}\right)^{-c^2/2}.$$

For $n = 0$, the event $\mathcal{G}_{t,h}(s,a)$ holds due to $c \geq \frac{2}{\sqrt{\log \frac{2SATH}{\delta}}}$. Hence, unconditionally,

$$\mathbb{P}\big(\mathcal{G}_{t,h}(s,a)^{\mathsf{c}}\big) \;\leq\; (2^S - 2)\left(\tfrac{2SATH}{\delta}\right)^{-c^2/2}.$$

Applying a union bound over all $t \in \{0,\dots,T-1\}$, $h \in \{0,\dots,H-1\}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ gives

$$\mathbb{P}\big((\mathcal{E}_\delta)^{\mathsf{c}}\big) \;\leq\; THSA\,(2^S - 2)\left(\tfrac{2SATH}{\delta}\right)^{-c^2/2}.$$

Due to $c \geq \sqrt{2\frac{\log\left(\frac{SATH(2^S-2)}{\delta}\right)}{\log\left(\frac{2SATH}{\delta}\right)}}$, we have

$$\mathbb{P}(\mathcal{E}_\delta) \;\geq\; 1 - \delta.$$

Furthermore, on $\mathcal{E}_\delta$ we have, for every (t,h,s,a),

$$\big\|P_h^\star(\cdot \mid s,a) - \widehat{P}_h^t(\cdot \mid s,a)\big\|_1 \;\leq\; f_\delta\big(N_h^t(s,a)\big).$$

Comparing with the confidence set definition in (2), this is exactly the membership condition $P^\star \in \bar{\mathcal{C}}_{\delta,\kappa}^t$ for all $0 \leq t \leq T-1$. Therefore, since the optimistic planner selects π^t and $P^t \in \bar{\mathcal{C}}_{\delta,\kappa}^t$ to maximize the quantile value,

$$V_{\tau,0}^{\pi^\star,P^\star}(s) \;\leq\; \max_{\pi}\max_{P \in \bar{\mathcal{C}}_{\delta,\kappa}^t} V_{\tau,0}^{\pi,P}(s) \;=\; V_{\tau,0}^{\pi^t,P^t}(s).$$

This proves the lemma. □

Lemma 2 (Weissman's ℓ_1 concentration)**.** *Let $X_1,\dots,X_n$ be i.i.d. on $[S] \coloneqq \{1,\dots,S\}$ with $\mathbb{P}(X_1 = i) = p_i$. Define the empirical distribution $\widehat{p}_i \coloneqq \frac{1}{n}\sum_{t=1}^n \mathbb{1}\{X_t = i\}$. Then*

$$\mathbb{P}\big(\|\widehat{p} - p\|_1 \geq \varepsilon\big) \;\leq\; (2^S - 2)\,\exp\!\left(-\frac{n\varepsilon^2}{2}\right).$$

Proof of Lemma 2. For any $x \in \mathbb{R}^S$,

$$\|x\|_1 \;=\; \max_{v \in \{-1,+1\}^S} v^\top x.$$

If, in addition, $\sum_i x_i = 0$, then the maximizers cannot be $v = \mathbf{1}$ or $v = -\mathbf{1}$ (since $v^\top x = 0$ for those two), where $\mathbf{1}$ is an all-one vector. Hence

$$\|x\|_1 \;=\; \max_{v \in \mathcal{V}} v^\top x, \qquad \mathcal{V} \;\coloneqq\; \{-1,+1\}^S \setminus \{\mathbf{1}, -\mathbf{1}\}, \tag{27}$$

and consequently

$$\big\{x : \|x\|_1 \geq \varepsilon\big\} \;\subseteq\; \bigcup_{v \in \mathcal{V}} \big\{x : v^\top x \geq \varepsilon\big\}.$$

Fix $v \in \mathcal{V}$ and define $Y_t^{(v)} \coloneqq v_{X_t} \in \{-1,+1\}$. Then

$$v^\top \widehat{p} \;=\; \sum_{i=1}^S v_i \widehat{p}_i \;=\; \frac{1}{n}\sum_{t=1}^n v_{X_t} \;=\; \frac{1}{n}\sum_{t=1}^n Y_t^{(v)}, \qquad \mathbb{E}\,Y_t^{(v)} \;=\; \sum_{i=1}^S p_i v_i \;=\; v^\top p.$$

Hence

$$v^\top(\widehat{p} - p) \;=\; \frac{1}{n}\sum_{t=1}^n \big(Y_t^{(v)} - \mathbb{E}Y_t^{(v)}\big),$$

a mean of i.i.d. centered random variables taking values in $[-1,1]$.

By Hoeffding's inequality,

$$\mathbb{P}\big(v^\top(\widehat{p} - p) \geq \varepsilon\big) \;\leq\; \exp\!\Big(-\frac{2n\varepsilon^2}{(1-(-1))^2}\Big) \;=\; \exp\!\left(-\frac{n\varepsilon^2}{2}\right). \tag{28}$$

Combining Equations (27), the union bound, and Equation (28),

$$\mathbb{P}\big(\|\widehat{p}-p\|_1 \geq \varepsilon\big) \;\leq\; \sum_{v\in\mathcal{V}} \mathbb{P}\big(v^\top(\widehat{p}-p) \geq \varepsilon\big) \;\leq\; |\mathcal{V}|\,\exp\!\left(-\frac{n\varepsilon^2}{2}\right).$$

Since $|\mathcal{V}| = 2^S - 2$, the stated bound follows. □

For a probability measure μ on $\mathbb{R}$, $F_\mu(x) \coloneqq \mu((-\infty, x])$ and the left-continuous quantile function is $F_\mu^{-1}(u) \coloneqq \inf\{x \in \mathbb{R} : F_\mu(x) \geq u\}, u \in (0,1]$. We write $Q_\tau(\mu) \equiv F_\mu^{-1}(\tau)$ for the τ-quantile. If F_μ jumps at $x^\star$, then $F_\mu^{-1}(u) = x^\star$ for all $u \in (F_\mu(x^\star-), F_\mu(x^\star)]$.

Lemma 3 (Bellman evaluation identity for the τ–quantile)**.** *Fix a policy π, a kernel P, a stage $h \in \{0,\dots,H-1\}$, and a state s. Let $a \coloneqq \pi_h(s)$ and $p \coloneqq P_h(\cdot \mid s,a) \in \Delta^S$. Then*

$$V_{\tau,h}^{\pi,P}(s) \;=\; r_h(s,a) \;+\; Q_\tau\Big(Z_{s,a,h}\big(p;\, V_{\cdot,h+1}^{\pi,P}\big)\Big).$$

Proof of Lemma 3. Fix h, s, and let $a \coloneqq \pi_h(s)$ and $p \coloneqq P_h(\cdot \mid s,a) \in \Delta^S$. By definition of the (left–continuous) τ–quantile,

$$V_{\tau,h}^{\pi,P}(s) = \inf\left\{x :\; \mathbb{P}\Big(\sum_{k=h}^{H-1} r_k(S_k,\pi_k(S_k)) \leq x \;\Big|\; S_h = s\Big) \;\geq \tau\right\}.$$

Write $r_h \coloneqq r_h(s,a)$. Conditioning on the next state S_{h+1} and using the Markov property, for every $x \in \mathbb{R}$,

$$\mathbb{P}\Big(\sum_{k=h}^{H-1} r_k(S_k,\pi_k(S_k)) \leq x \;\Big|\; S_h = s\Big) = \sum_{i=1}^{S} p_i\, \mathbb{P}\Big(\sum_{k=h+1}^{H-1} r_k(S_k,\pi_k(S_k)) \leq x - r_h \;\Big|\; S_{h+1} = s_i\Big),$$

where $p_i = P_h(s_i|s,a)$.

Let F_i be the CDF of the $(h{+}1)$–to–$(H{-}1)$ return starting from s_i under (π, P):

$$F_i(t) \;\coloneqq\; \mathbb{P}\Big(\sum_{k=h+1}^{H-1} r_k(S_k,\pi_k(S_k)) \leq t \;\Big|\; S_{h+1} = s_i\Big).$$

By definition of the QMDP quantile map, for each $q \in (0,1)$

$$V_{q,h+1}^{\pi,P}(s_i) \;=\; F_i^{-1}(q) \quad \text{(left–continuous quantile)}.$$

By Lemma 4

$$F_i(t) = \sup\{q \in [0,1] :\; F_i^{-1}(q) \leq t\}.$$

Applying it with $F_i^{-1}(q) = V_{q,h+1}^{\pi,P}(s_i)$ yields

$$\phi_i(t) \;\coloneqq\; \sup\{q \in [0,1] :\; V_{q,h+1}^{\pi,P}(s_i) \leq t\} \;=\; \sup\{q \in [0,1] : F_i^{-1}(q) \leq t\} \;=\; F_i(t),$$

i.e.,

$$\mathbb{P}\Big(\sum_{k=h+1}^{H-1} r_k(S_k,\pi_k(S_k)) \leq t \;\Big|\; S_{h+1} = s_i\Big) = \phi_i(t) \quad \text{for all } t \in \mathbb{R}.$$

Recall $p_i = P_h(s_i \mid s,a)$, hence

$$\Phi_p(t) \;\coloneqq\; \sum_{i=1}^{S} p_i\,\phi_i(t) = \sum_{i=1}^{S} p_i\,F_i(t).$$

Therefore the CDF of the h–step return at s is $x \mapsto \Phi_p(x - r_h)$, and

$$V_{\tau,h}^{\pi,P}(s) = \inf\{x :\; \Phi_p(x - r_h) \geq \tau\} = r_h + \inf\{y :\; \Phi_p(y) \geq \tau\}.$$

By Definition 1, $\Phi_p(\cdot)$ is the CDF of $Z_{s,a,h}\big(p; V^{\pi,P}_{\cdot,h+1}\big)$, so

$$\inf\{y:\ \Phi_p(y) \geq \tau\} = Q_\tau\Big(Z_{s,a,h}\big(p; V^{\pi,P}_{\cdot,h+1}\big)\Big).$$

Therefore, $V^{\pi,P}_{\tau,h}(s) = r_h(s,a) + Q_\tau\Big(Z_{s,a,h}\big(p; V^{\pi,P}_{\cdot,h+1}\big)\Big)$. □

Lemma 4 (Right identity for the left–continuous quantile)**.** *Let F be a CDF on $\mathbb{R}$ and $F^{-1}(q) = \inf\{x : F(x) \geq q\}$ its left–continuous quantile. For every $t \in \mathbb{R}$,*

$$\{q \in [0,1]:\ F^{-1}(q) \leq t\} = [0, F(t)] \quad \textit{and hence} \quad \sup\{q \in [0,1]:\ F^{-1}(q) \leq t\} = F(t).$$

Proof of Lemma 4. Define $B_t \coloneqq \{q \in [0,1]:\ F^{-1}(q) \leq t\}$. We show the pointwise equivalence

$$q \in B_t \quad \Longleftrightarrow \quad q \leq F(t),$$

which immediately yields $B_t = [0, F(t)]$ and the stated supremum.

($\Rightarrow$) Assume $q \in B_t$, i.e., $F^{-1}(q) \leq t$. Set $y \coloneqq F^{-1}(q)$. By definition, $y = \inf\{x : F(x) \geq q\}$. Because F is right–continuous and nondecreasing, the set $\{x : F(x) \geq q\}$ is right–closed; hence $F(y) \geq q$.[1] Since $y \leq t$ and F is nondecreasing, $F(t) \geq F(y) \geq q$, i.e., $q \leq F(t)$.

($\Leftarrow$) Assume $q \leq F(t)$. Then $t \in \{x : F(x) \geq q\}$, so this set is nonempty and $F^{-1}(q) = \inf\{x : F(x) \geq q\} \leq t$. Thus $q \in B_t$.

Combining the two directions gives $B_t = [0, F(t)]$. Taking the supremum over B_t yields $\sup B_t = F(t)$. □

Lemma 5 (Deterministic optimality for QMDP)**.** *Fix a kernel P. In the recursion* (6)*, for every (s,h) there exists a* deterministic *maximizer $a^\star \in \arg\max_a Q_\tau\Big(Z_{s,a,h}\big(P_h(\cdot \mid s,a);\ V^\star_{\cdot,h+1}\big)\Big)$. Consequently, there exists an optimal* deterministic Markov *policy.*

Proof of Lemma 5. Fix a stage h, a state s, a kernel P, and a continuation map $V_{\cdot,h+1}$. For each action $a \in \mathcal{A}$ let

$$F_a(t) \ \coloneqq\ \mathbb{P}\Big(Z_{s,a,h}\big(P_h(\cdot \mid s,a);\ V_{\cdot,h+1}\big) \leq t\Big), \qquad m_a \ \coloneqq\ Q_\tau\Big(Z_{s,a,h}\big(P_h(\cdot \mid s,a);\ V_{\cdot,h+1}\big)\Big).$$

For any mixed action $\mu \in \Delta^{\mathcal{A}}$, define the mixture CDF $F_\mu(t) \coloneqq \sum_{a\in\mathcal{A}} \mu(a)\, F_a(t)$ (draw $A \sim \mu$, then sample the continuation under A). Let $M \coloneqq \max_a m_a$ and fix any a. By definition $m_a = \inf\{x :\ F_a(x) \geq \tau\}$, so $M \geq m_a$ implies $F_a(M) \geq \tau$ (monotonicity of F_a suffices; left–continuity is not needed here). Hence

$$F_\mu(M) \ =\ \sum_a \mu(a) F_a(M) \ \geq\ \sum_a \mu(a)\, \tau \ =\ \tau,$$

so by the definition of the (left–continuous) τ–quantile, $Q_\tau(F_\mu) \leq M = \max_a m_a$.

Proceed by backward induction on h. For $h = H$ the claim is trivial. Assume an optimal continuation map $V^\star_{\cdot,h+1}$ has been realized at step $h{+}1$ (this is the inductive hypothesis provided by the Bellman program (6)). Consider any state s at step h. By Lemma 3,

$$V^{\pi,P}_{\tau,h}(s) = r_h(s,a) \ +\ Q_\tau\Big(Z_{s,a,h}\big(P_h(\cdot \mid s,a);\ V^{\pi,P}_{\cdot,h+1}\big)\Big).$$

Evaluating the optimality backup with continuation fixed at $V^\star_{\cdot,h+1}$ amounts to comparing the collection $\{Q_\tau(F_a)\}_{a\in\mathcal{A}}$ defined in Lemma 5. By that lemma, randomization over actions cannot exceed $\max_a Q_\tau(F_a)$, hence a single action $a^\star$ attains the maximum. Define $\pi^\star_h(s)$ to pick such an $a^\star$ (break ties deterministically). Doing this for every state produces a deterministic Markov decision rule at step h; composing with the inductively optimal rules from steps $h{+}1, \ldots, H{-}1$ yields a deterministic Markov policy that attains the optimal values $V^\star_{\tau,h}$. This completes the induction. □

[1] Equivalently: for every $\varepsilon > 0$, $y{+}\varepsilon$ belongs to the set, so $F(y{+}\varepsilon) \geq q$; right–continuity gives $F(y) = \lim_{\varepsilon\downarrow 0} F(y{+}\varepsilon) \geq q$.

Lemma 6 (Local Lipschitz of $Z_{s,a,h}(p; V^{\pi,P}_{q,h+1})$ in p)**.** *Fix a step h and state–action (s,a), and let $V^{\pi,P}_{q,h+1} : \mathcal{S} \times [0,1] \to [0,H]$ be the next-step quantile value map. Let $Z_{s,a,h}(p; V^{\pi,P}_{q,h+1})$ be the continuation-mixture variable of Definition 1. Then for all $p \in \Delta^S$,*

$$\left|Q_\tau\Big(Z_{s,a,h}\big(p; V^{\pi,P}_{\cdot,h+1}\big)\Big) - Q_\tau\Big(Z_{s,a,h}\big(P^\star; V^{\pi,P}_{\cdot,h+1}\big)\Big)\right| \;\le\; \frac{H}{\kappa}\,\|p - P^\star\|_1.$$

provided Assumption 1 holds for $Z_{s,a,h}(P^\star; g)$ with parameter $\kappa > 0$.

Proof of Lemma 6. Let $\mu_{P^\star}$ and μ_p be the laws of $Z(P^\star)$ and $Z(p)$, respectively. Moreover, for probability measures μ, ν on $\mathbb{R}$, Total Variation (TV) and 1-Wasserstein (W_1) can be defined as presented in Equations ((29)) and ((30)) respectively Gibbs and Su 2002.

$$\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)| = \tfrac{1}{2} \sup_{\|f\|_\infty \le 1} \Big|\int f\, d(\mu - \nu)\Big|. \tag{29}$$

$$W_1(\mu,\nu) = \sup_{\mathrm{Lip}(f)\le 1} \int f\, d(\mu - \nu) = \int_0^1 \big|F_\mu^{-1}(u) - F_\nu^{-1}(u)\big|\, du. \tag{30}$$

Assumption 1 provides only a *one-sided* lower bound on the mass at the τ-quantile (total jump $\ge \kappa$), so Lemma 7 yields

$$\big|Q_\tau(\mu_p) - Q_\tau(\mu_{P^\star})\big| \;\le\; \frac{2}{\kappa}\, W_1(\mu_{P^\star}, \mu_p).$$

Also, by Lemma 8 and Lemma 9,

$$W_1(\mu_{P^\star}, \mu_p) \le \tfrac{H}{2}\, \|p - P^\star\|_1.$$

Therefore,

$$\big|Q_\tau(\mu_p) - Q_\tau(\mu_{P^\star})\big| \;\le\; \frac{H}{\kappa}\, \|p - P^\star\|_1.$$

□

Lemma 7 (Quantile sensitivity under a one-sided jump margin)**.** *Let μ be a probability measure on $[0,H]$ with CDF F and τ-quantile $x^\star := Q_\tau(\mu)$ such that $F(x^\star) - F(x^{\star-}) = \kappa$. Then for any probability measure ν on $[0,H]$,*

$$\big|Q_\tau(\nu) - Q_\tau(\mu)\big| \;\le\; \frac{2}{\kappa}\, W_1(\mu,\nu).$$

Proof of Lemma 7. Let F and G be the CDFs of μ and ν, and let $Q_\mu(u) \coloneqq \inf\{x : F(x) \ge u\}$ and $Q_\nu(u) \coloneqq \inf\{x : G(x) \ge u\}$ be their left–continuous quantile functions. By Equation (30)

$$W_1(\mu,\nu) = \int_0^1 \big|Q_\mu(u) - Q_\nu(u)\big|\, du.$$

Let $x^\star \coloneqq Q_\mu(\tau)$. By assumption, F has a jump of size at least κ at $x^\star$:

$$L \coloneqq F(x^{\star-}), \qquad U \coloneqq F(x^\star), \qquad U - L \ge \kappa, \qquad \tau \in (L, U].$$

Hence

$$Q_\mu(u) \equiv x^\star \quad \text{for all } u \in (L, U].$$

Let $v \coloneqq Q_\nu(\tau)$ and $d \coloneqq v - x^\star$. Monotonicity of Q_ν yields

$$u \le \tau \;\Rightarrow\; Q_\nu(u) \le v, \qquad u > \tau \;\Rightarrow\; Q_\nu(u) \ge v.$$

Case $d \ge 0$. For any $u \in (\tau, U]$, since $Q_\nu(u) \ge v \ge x^\star$,

$$\big|x^\star - Q_\nu(u)\big| = Q_\nu(u) - x^\star \;\ge\; v - x^\star \;=\; |d|,$$

so

$$W_1(\mu,\nu) \;\geq\; \int_{\tau}^{U} \left|x^\star - Q_\nu(u)\right| du \;\geq\; (U-\tau)\,|d|.$$

Case $d<0$. For any $u\in(L,\tau)$, since $Q_\nu(u)\leq v\leq x^\star$,

$$\left|x^\star - Q_\nu(u)\right| = x^\star - Q_\nu(u) \;\geq\; x^\star - v \;=\; |d|,$$

and therefore

$$W_1(\mu,\nu) \;\geq\; \int_{L}^{\tau} \left|x^\star - Q_\nu(u)\right| du \;\geq\; (\tau-L)\,|d|.$$

Combining the two cases gives

$$W_1(\mu,\nu) \;\geq\; \Big(\mathbf{1}_{\{d\geq 0\}}(U-\tau) + \mathbf{1}_{\{d<0\}}(\tau-L)\Big)\,|d|.$$

Since $(\tau-L)+(U-\tau)=U-L\geq\kappa$, at least one of the two side lengths is $\geq\kappa/2$. Thus

$$W_1(\mu,\nu) \;\geq\; \frac{\kappa}{2}\,|d| \quad\Rightarrow\quad |d| \;\leq\; \frac{2}{\kappa}\,W_1(\mu,\nu).$$

Recalling $d=Q_\nu(\tau)-Q_\mu(\tau)$ finishes the proof. □

Lemma 8 (Mixture total variation vs. mixing weights)**.** *Let* $\{\mu_i\}_{i=1}^S$ *be probability measures on* $[0,H]$. *For* $p,p'\in\Delta^S$ *define the mixtures* $\mu_p=\sum_{i=1}^S p_i\,\mu_i$ *and* $\mu_{p'}=\sum_{i=1}^S p'_i\,\mu_i$. *Then*

$$\mathrm{TV}(\mu_p,\mu_{p'}) \;\leq\; \tfrac{1}{2}\,\|p-p'\|_1.$$

Proof of Lemma 8. Recall Equation (29): for probability measures μ,ν,

$$\mathrm{TV}(\mu,\nu) \;=\; \tfrac{1}{2}\sup_{\|f\|_\infty\leq 1}\Big|\int f\,d(\mu-\nu)\Big|.$$

Write the signed measure difference of the two mixtures as

$$\mu_p-\mu_{p'} \;=\; \sum_{i=1}^{S}(p_i-p'_i)\,\mu_i.$$

For any measurable f with $\|f\|_\infty\leq 1$,

$$\Big|\int f\,d(\mu_p-\mu_{p'})\Big| \;=\; \Big|\sum_{i=1}^{S}(p_i-p'_i)\int f\,d\mu_i\Big| \;\leq\; \sum_{i=1}^{S}|p_i-p'_i|\,\Big|\int f\,d\mu_i\Big|.$$

Since each μ_i is a probability measure and $\|f\|_\infty\leq 1$,

$$\Big|\int f\,d\mu_i\Big| \;\leq\; \int |f|\,d\mu_i \;\leq\; 1.$$

Therefore

$$\Big|\int f\,d(\mu_p-\mu_{p'})\Big| \;\leq\; \sum_{i=1}^{S}|p_i-p'_i|.$$

Taking the supremum over all $\|f\|_\infty\leq 1$ and multiplying by $\frac{1}{2}$ yields $\mathrm{TV}(\mu_p,\mu_{p'})\leq\frac{1}{2}\|p-p'\|_1$, as claimed. □

Lemma 9 (TV controls W_1 on $[0,H]$)**.** *If* μ,ν *are probability measures supported on* $[0,H]$, *then*

$$W_1(\mu,\nu) \;\leq\; H\,\mathrm{TV}(\mu,\nu).$$

Proof of Lemma 9. Let $\sigma := \mu - \nu$ and write its Jordan decomposition $\sigma = \sigma^- + (\sigma - \sigma^-) = \sigma^+ - \sigma^-$ with $\sigma^+([0,H]) = \sigma^-([0,H]) = \mathrm{TV}(\mu,\nu) =: m$. By the Kantorovich–Rubinstein duality (presented in Equation ((30))),

$$W_1(\mu,\nu) = \sup_{\mathrm{Lip}(f)\le 1} \int f \, d\sigma.$$

For any bounded f,

$$\int f \, d\sigma = \int f \, d\sigma^+ - \int f \, d\sigma^- \le (\sup f)\, m - (\inf f)\, m = (\sup f - \inf f)\, m.$$

If $\mathrm{Lip}(f) \le 1$ on $[0,H]$, then $\sup f - \inf f \le H$; hence $W_1(\mu,\nu) \le H\, m = H\, \mathrm{TV}(\mu,\nu)$. □

Lemma 10 (Auxiliary–uniform representation of the continuation mixture)**.** *Fix episode t, step h, and the realized pair (S_h^t, a_h) with $a_h = \pi_h^t(S_h^t)$. Let $p^\star := P_h^\star(\cdot \mid S_h^t, a_h)$ and condition on $\mathcal{F}_h^t$ so that $S_{h+1}^t \sim p^\star$ is the only randomness going forward at step h. For any kernel P and policy π^t, define the continuation–mixture variable*

$$Z^{(P)} \;:=\; Z_{S_h^t, a_h, h}\big(p^\star;\, V_{\cdot,h+1}^{\pi^t,P}\big).$$

Let $U \sim \mathrm{Unif}[0,1]$ be independent of $\mathcal{F}_h^t$ and of S_{h+1}^t. Then, conditionally on $\mathcal{F}_h^t$,

$$Z^{(P)} \;\overset{d}{=}\; V_{U,h+1}^{\pi^t,P}(S_{h+1}^t).$$

Proof of Lemma 10. Fix P and abbreviate $V_{q,i} := V_{q,h+1}^{\pi^t,P}(s_i)$. For each next state s_i, let F_i denote the CDF of the $(h{+}1)$–to–$(H{-}1)$ return under (π^t, P) starting from s_i. By definition of the QMDP quantile map, $V_{q,i} = F_i^{-1}(q)$ where the inverse is left–continuous.

By Definition 1, the CDF of $Z^{(P)}$ (given $\mathcal{F}_h^t$) is

$$\Phi_{p^\star}(t) \;=\; \sum_i p_i^\star\, \phi_i(t), \qquad \phi_i(t) := \sup\{q \in [0,1] :\, V_{q,i} \le t\}.$$

By Lemma 4, for every i we have $\phi_i(t) = \sup\{q : F_i^{-1}(q) \le t\} = F_i(t)$. Hence

$$\Phi_{p^\star}(t) \;=\; \sum_i p_i^\star\, F_i(t). \tag{31}$$

Now consider $Y^{(P)} := V_{U,h+1}^{\pi^t,P}(S_{h+1}^t) = F_{S_{h+1}^t}^{-1}(U)$. Condition on the event $\{S_{h+1}^t = s_i\}$. Since U is independent and uniform, Lemma 4 gives

$$\mathbb{P}\big(Y^{(P)} \le t \;\big|\; \mathcal{F}_h^t,\, S_{h+1}^t = s_i\big) = \mathbb{P}\big(F_i^{-1}(U) \le t\big) = \mathbb{P}\big(U \le F_i(t)\big) = F_i(t).$$

Taking the conditional expectation over $S_{h+1}^t \sim p^\star$ yields

$$\mathbb{P}\big(Y^{(P)} \le t \;\big|\; \mathcal{F}_h^t\big) = \sum_i p_i^\star\, F_i(t) = \Phi_{p^\star}(t) = \mathbb{P}\big(Z^{(P)} \le t \;\big|\; \mathcal{F}_h^t\big),$$

where we used Equation (31) in the second equality and Definition 1 in the last. Thus $Y^{(P)}$ and $Z^{(P)}$ have the same conditional CDF given $\mathcal{F}_h^t$, i.e., the same conditional law. Therefore, conditionally on $\mathcal{F}_h^t$, $Z^{(P)} \overset{d}{=} V_{U,h+1}^{\pi^t,P}(S_{h+1}^t)$, as claimed. □

C Illustration of the Quantile Backup

To make the recursion in (6)–(7) more concrete, we provide a simple two-state, two-action illustration, adapted in spirit from the backward-dynamic-program schematic proposed by Li et al. (2022). Figure 1 summarizes how the continuation information from the two next states is combined through the optimization over q to produce a one-step quantile backup.

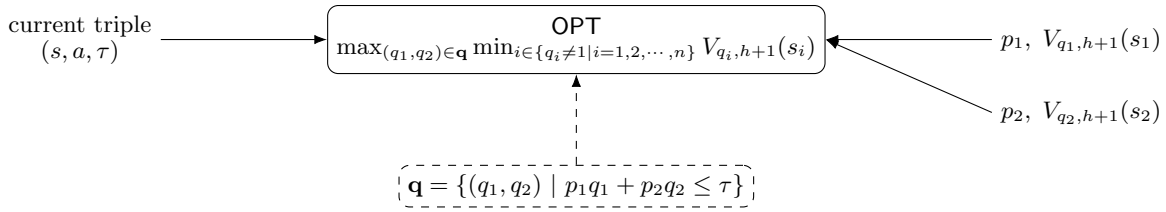

Figure 1: Schematic of a one-step quantile backup.

To visualize this optimization, consider the two-state case $\mathcal{S} = \{s_1, s_2\}$, and write

$$p_1 \coloneqq P(s_1 \mid s, a), \qquad p_2 \coloneqq P(s_2 \mid s, a), \qquad p_1 + p_2 = 1.$$

Then the OPT problem reduces to

$$\mathsf{OPT}\big(s, \tau, a; V_{\cdot,h+1}, P\big) = \max_{q_1,q_2\in[0,1]} \min_{i:\, q_i\neq 1} V_{q_i,h+1}(s_i) \quad \text{s.t.} \quad p_1q_1 + p_2q_2 \leq \tau. \tag{32}$$

The key point is that τ is a total quantile budget, and the variables q_1 and q_2 specify how that budget is allocated across the two next-state branches. The constraint $p_1q_1 + p_2q_2 \leq \tau$ determines which allocations are feasible, and the objective then selects the feasible allocation that maximizes the bottleneck continuation value $\min_{i:\, q_i\neq 1} V_{q_i,h+1}(s_i)$.

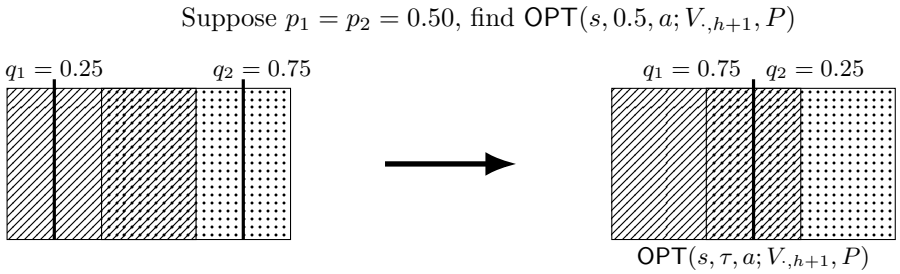

Figure 2: Two feasible allocations on the boundary $0.5q_1 + 0.5q_2 = 0.5$, that is, $q_1 + q_2 = 1$. The OPT subproblem compares such feasible allocations through the objective $\min_{i:\, q_i\neq 1} V_{q_i,h+1}(s_i)$.

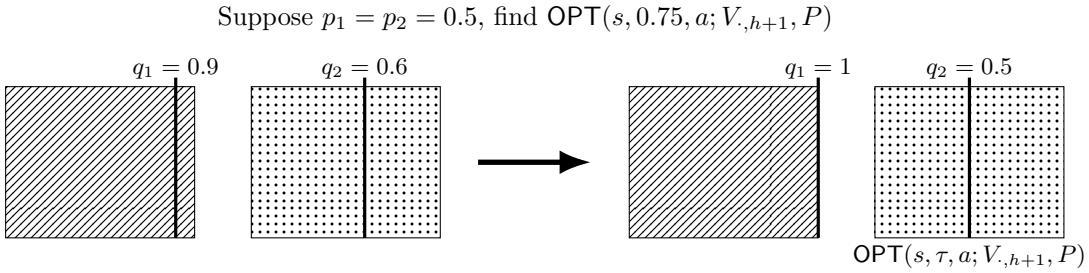

Figure 3: Two feasible allocations when $0.5q_1 + 0.5q_2 \leq 0.75$, that is, $q_1 + q_2 \leq 1.5$. Because the feasible set is larger than in Figure 2, the OPT subproblem can explore more aggressive allocations.

Figure 2 corresponds to the case $\tau = 0.5$. Since $p_1 = p_2 = 0.5$, the feasibility constraint becomes

$$0.5q_1 + 0.5q_2 \leq 0.5 \qquad \Longleftrightarrow \qquad q_1 + q_2 \leq 1.$$

Thus, boundary allocations such as $(q_1, q_2) = (0.25, 0.75)$ and $(q_1, q_2) = (0.75, 0.25)$ are feasible. The value of $\mathsf{OPT}(s, 0.5, a; V_{\cdot,h+1}, P)$ is obtained by evaluating

$$\min_{i:\, q_i\neq 1} V_{q_i,h+1}(s_i)$$

over all such feasible choices and selecting the maximizing one.

Figure 3 corresponds to the case $\tau = 0.75$. The constraint becomes

$$0.5q_1 + 0.5q_2 \leq 0.75 \qquad \Longleftrightarrow \qquad q_1 + q_2 \leq 1.5.$$

Hence the feasible region is larger. For example, both $(q_1, q_2) = (0.5, 0.5)$ and $(q_1, q_2) = (1, 0.5)$ are feasible. The second case is especially useful for interpreting the objective, because when $q_1 = 1$, the first branch is excluded from the bottleneck minimum by the definition $\min_{i:\, q_i \neq 1} V_{q_i, h+1}(s_i)$, so only the second branch remains active in the minimum.

These figures therefore illustrate the role of OPT in (6)–(7): the one-step quantile backup is obtained by solving a constrained allocation problem over the branchwise quantile levels q_i, with the transition probabilities p_i acting as weights in the budget constraint. This is the quantile counterpart of the one-step averaging step in the classical expectation-based Bellman recursion.