Long-Term Sequential Decision Making Under Risk

Irmaan (Mohammad) Mirzanejad[1] **Nadjet Bourdache**[1] **Abdel-illah Mouaddib**[1]

[1]Université de Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, France

Abstract

We study finite-horizon MDP planning under *root-based* (resolute) risk objectives that apply a rank-dependent functional to the distribution of total returns. Such objectives are non-linear in the return distribution and generally break Bellman optimality, so direct optimization by scenario-tree enumeration is intractable. We propose **ERQDP**, an enumeration-free and sampling-free method that solves a rank–quantile surrogate via exact DP (Dynamic Programming), evaluates candidate policies exactly by DP over return Probability Mass Functions (PMFs) on a discretized return grid (with an explicit rounding bound), and refines the surrogate in an anytime loop that reports an explicit upper–lower gap (certificate) for the target objective up to discretization budgets. Across tested benchmarks, ERQDP returns certified solutions or explicit residual gaps, enables fast risk-parameter sweeps with substantial runtime gains, and supports both risk-averse and risk-seeking behaviors.

1 INTRODUCTION

Decision making under risk refers to settings in which the possible outcomes of an action (and their probabilities) are known or modeled. Even in the one-shot case, the notion of an "optimal" decision is subjective: two agents facing the same lottery can prefer different options depending on their attitude toward risk. For instance, a guaranteed gain of 10 may be preferred to a lottery paying 50 with probability 0.2 by a risk-averse agent, while a risk-seeking agent may prefer the lottery, and a risk-neutral agent would be indifferent (both have an expectation of 10). This subjectivity motivates models that go beyond expectation by evaluating the *entire* outcome distribution, including tail-oriented criteria such as Value-at-Risk/Conditional Value-at-Risk (VaR/CVaR) and more general rank-based (distortion) aggregators that reweight quantiles across a spectrum of outcomes [Yaari, 1987, Schmeidler, 1989, Wang, 1996, Bäuerle and Jaśkiewicz, 2024].

Sequential decision making under risk is strictly more challenging. Actions are chosen repeatedly, each choice shapes future uncertainty, and the number of possible consequence paths (trajectories) as well as the number of possible strategies grows combinatorially with the horizon. Markov Decision Processes (MDPs) provide a standard formalism for such problems [Puterman, 1994]. Under the expected-return criterion, dynamic programming (DP) applies because expectation is linear and satisfies the Bellman optimality principle. However, in domains such as finance, healthcare planning, and safety-critical operations, the distributional shape of the long-term return (including low-probability adverse outcomes) often matters at least as much as the mean, motivating risk-sensitive MDP solutions [Kolobov et al., 2012, Ahmadi et al., 2021, Bäuerle and Jaśkiewicz, 2024]. Most distributional criteria are *non-linear* in the return distribution and therefore do not admit a straightforward Bellman decomposition.

Example 1. Consider the following two-stage decision problem: an agent is at time 0 at state s_0 and can either invest in protection (sure reward -5) or skip it (reward 0). If invest, it arrives at s_1 and gets a certain reward of -5, otherwise it arrives at s_2 and gets a certain reward of 0. In s_1 (respectively s_2), the agent can only select cautious (respectively aggressive) mode and gets an uncertain reward: 20 with probability 0.99 and -110 otherwise (respectively 40 with probability 0.9 and -100 otherwise). This MDP is graphically given in Figure 1 and the lotteries associated to both strategies are given in Table 1. We can see that the short-term strategy π_{short} has a 10% catastrophic outcome, whereas the long term one π_{long} reduces catastrophe to 1% which is somehow safer and could (depending on the preferences) compensate the lower rewards (compared to π_{short}).

This example illustrates why short-term screening and

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4543–4560.

Plan	Outcomes (total return)	Prob.	$\mathbb{E}[\cdot]$	$\mathrm{CVaR}_{0.1}[\cdot]$
π_{long} (invest $\to$ cautious)	$20,\ -110$	$0.99,\ 0.01$	18.7	7
π_{short} (skip $\to$ aggressive)	$40,\ -100$	$0.9,\ 0.1$	26	-100

Table 1: The lotteries of Example 1 and their evaluations.

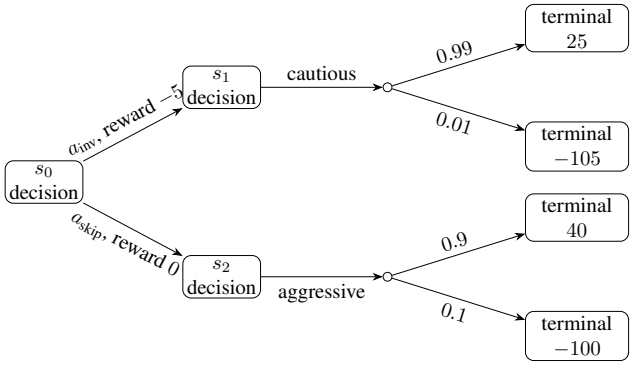

Figure 1: The MDP of Example 1.

trajectory-level risk evaluation can lead to different policies. Furthermore, the determination of optimal strategies in both cases are quite different. While short-term screening only requires local optimal choices (with dynamic programming for example), at the root level, risk-sensitive planning selects a single policy at the initial state and evaluates it by applying a risk functional to the induced distribution of *total* (long-term) returns. This couples exponentially many trajectories and generally breaks Bellman optimality. A naive optimization approach enumerates policies and scenario trees and then optimize over the enumerated solutions, but this quickly becomes intractable as the horizon grows. Existing approaches often restore tractability by changing the formulation (e.g., state augmentation or robust/dual reformulations), restricting the risk model to DP-friendly classes, or relying on approximations and sampling [Howard and Matheson, 1972, Chow et al., 2015, Lin Hau et al., 2023a, Bäuerle and Jaśkiewicz, 2024, Ávila Pires et al., 2025], but such reformulations or restrictions do not always apply.

Moreover, while CVaR is widely used as a principled tail criterion (e.g., [Rigter et al., 2022]), it is inherently pessimistic (it only summarizes the worst outcomes) and cannot express the full spectrum of behaviors that applications may require, including more nuanced tradeoffs across the entire distribution. This motivates flexible rank-based aggregators such as the Yaari model [Yaari, 1987] also known as WOWA (Weighted Ordered Weighted Average) in the multicriteria decision making context [Yager, 1988, Torra, 1997].

To study long-term risk in sequential decision-making problems, we focus in particular on non-linear models that can describe a wide range of preferences w.r.t risk, thereby requiring specific algorithmic solutions. In this purpose, we introduce **ERQDP** (Exact Rank–Quantile Dynamic Programming), an enumeration-free framework for finite-horizon MDPs with resolute (root-level) distributional risk. It combines three ingredients:

- **Surrogate-to-exact risk planning:** solve a resolute risk objective via a distribution surrogate with resolution K, using DP to obtain a surrogate-optimal policy.
- **Exact distributional evaluation:** compute the trajectory-return distribution of any policy by DP and evaluate CVaR/WOWA exactly on that distribution.
- **Anytime certified refinement:** increase K and accept only root-improving updates, reporting an explicit optimality-gap/certificate signal.

2 RELATED WORK

A central distinction in risk-sensitive sequential decision making is between *time-consistent* (nested) criteria and *resolute* (root-based) criteria that apply a risk functional once to the total return distribution [Bäuerle and Jaśkiewicz, 2024]. The latter is often preferred when the goal is to encode a trajectory-level preference, but it is generally non-linear in the induced return distribution and therefore violates Bellman optimality. Root-based **VaR** (Value at Risk) selects a return threshold at a prescribed probability level and is not additive, so standard DP is inapplicable in general [Xia and Pan, 2025]. Root-based **CVaR** (Conditional Value at Risk, a.k.a. Expected Shortfall) averages the worst tail mass and is coherent [Artzner et al., 1999, Rockafellar and Uryasev, 2000], yet it remains time inconsistent under conditioning and typically regains tractability only through reformulation, e.g., via state augmentation or related constructions [Bäuerle and Ott, 2011, Bäuerle and Rieder, 2014, Ding and Feinberg, 2022] and robust/dual viewpoints [Chow et al., 2015]. While widely used, CVaR is intrinsically tail-focused and can be overly pessimistic when preferences across the full distribution matter.

A classical route is *expected utility*: transform outcomes through a utility function and maximize expectation. In sequential problems, applying utility *locally* to one-step rewards preserves DP but changes semantics, whereas applying it *globally* to the induced lottery over total returns matches resolute (root-level) preferences but typically breaks Bellman separability. A canonical DP-friendly instance is **exponential/entropic utility**, which admits transformed Bellman recursions [Howard and Matheson, 1972, Jaquette, 1976, Bäuerle and Jaśkiewicz, 2024]. More broadly, **OCE** unifies many convex criteria (including entropic risk and CVaR) and supports both recursive (time-consistent) and "risk outside recursion" formulations [Ben-Tal and Teboulle, 1986, Bäuerle and Jaśkiewicz, 2024]; this has led to DP-style algorithms for entropic criteria such as ERM and EVaR [Lin Hau et al., 2023b, Su et al., 2025, Marthe et al., 2025]. In model-free RL, coherent/convex risk criteria are often optimized via sampling-based meth-

ods [Tamar et al., 2017, Yu and Ying, 2023, Coache and Jaimungal, 2024], with complementary analyses for entropic objectives [Zhang et al., 2024, Mortensen and Talebi, 2025] and dynamic time-consistent risk measures [Yu and Shen, 2022]; these are indispensable when the model is unknown, but they typically either adopt time-consistent semantics or provide sample-based approximations rather than certified root-level optimality for static (resolute) objectives. Finally, although expected-utility models are relatively effective in terms of preference modelling, they do not, however, allow to express some (more or less) sophisticated and complex behaviours like, for example, the behaviour highlighted by the well-known Allais paradox [Allais, 1953] (see [Gonzales and Perny, 2020] for more limitations of this model).

When preferences must be controlled *across* the return distribution (not only in an extreme tail), **distortion** and **spectral** criteria reweight quantiles through a probability-weighting (distortion) function [Yaari, 1987, Schmeidler, 1989, Wang, 1996]. In decision theory, related rank-dependent utility models distort probabilities inside an expected-utility construction; in this paper we work with distortion *risk measures* applied directly to return distributions, which share similar ingredients but are not the same object. The **WOWA** operator provides a flexible rank-based aggregation mechanism [Yager, 1988, Torra, 1997] that can express behaviors that neither expected utility (too restrictive to target specific parts of the distribution) nor CVaR (tail-only) can capture. However, in sequential problems the dependence on the *ordering* of trajectory returns makes optimization difficult: small policy changes can reshuffle ranks, breaking Bellman separability and complicating search over policies, motivating heavy MILP/search approaches for even simpler problems: for example decision trees [Jeantet and Spanjaard, 2011, Jeantet et al., 2012], or multiobjective MDPs where the reward for each objective is obtained by expectation Ogryczak et al. [2013].

A complementary route is **distributional dynamic programming**, which propagates return distributions (or sufficient statistics) and then evaluates functionals of these distributions [Ávila Pires et al., 2025]. A key caution is that enforcing a DP decomposition for static/root objectives by discretizing "risk levels" can be inexact and yield policies that are suboptimal for the intended static objective [Lin Hau et al., 2023a]; related dependence on the initial state also appears for non-linear scalarizations in multiobjective MDPs [Ogryczak et al., 2013]. ERQDP avoids this pitfall by using DP only on an explicitly defined rank–quantile surrogate model, while evaluating the root objective *exactly* for any explicit policy via Probability Mass Function (PMF) DP.

Positioning of this paper. The literature exposes a recurring tradeoff between *tractability* and *objective fidelity*. DP-friendly criteria preserve Bellman structure but either restrict the preference model or change semantics to time-consistent nested risk [Howard and Matheson, 1972, Ben-Tal and Teboulle, 1986, Bäuerle and Jaśkiewicz, 2024]. Root-based quantiles/CVaR match trajectory-level evaluation but often regain tractability through augmentation or robust/dual reformulations [Artzner et al., 1999, Rockafellar and Uryasev, 2000, Chow et al., 2015, Bäuerle and Ott, 2011, Ding and Feinberg, 2022, Xia and Pan, 2025]. Distortion and rank-based aggregators offer spectrum-wide control but their non-linearity breaks Bellman optimality and can lead to hard optimization problems or heavy search/MILP schemes [Jeantet and Spanjaard, 2011, Jeantet et al., 2012, Ogryczak et al., 2013].

To the best of our knowledge, ERQDP is the first DP-faithful planning framework for finite-horizon MDPs with root-level risk consideration that satisfies, at the same time: (i) the consideration of a wide range of risk preferences (from tail-dominated to spectrum-wide); (ii) the scalability by avoiding trajectory enumeration and sampling approximations; and (iii) a theoretical guarantee of optimality.

3 METHODOLOGY

3.1 PRELIMINARIES: MDPS, TRAJECTORIES, AND RANK-DEPENDENT MEASURES

MDP, return, and resolute objective. We consider a finite-horizon discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, H)$ and use standard time-augmentation so a deterministic non-stationary policy is a mapping $\pi : \mathcal{S} \times \{0, \dots, H-1\} \to \mathcal{A}$. Given s_0, π induces a trajectory $\tau = (s_0, a_0, \dots, s_H)$ with $a_t = \pi(s_t, t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$, and discounted return

$$g(\tau) \triangleq \sum_{t=0}^{H-1} \gamma^t \, r(s_t, a_t, s_{t+1}). \tag{1}$$

Resolute planning commits to one policy at the root and maximizes a risk functional of the resulting return distribution:

$$\pi^\star \in \arg\max_{\pi \in \Pi_{\mathrm{DM}}} \rho(G^\pi(s_0)) \,. \tag{2}$$

Distortion functions and rank-dependent risk measures. A broad class of law-invariant risk functionals is given by distortion (rank-dependent) measures [Schmeidler, 1989, Yaari, 1987, Wang, 1996]. A distortion is a nondecreasing map $\phi : [0,1] \to [0,1]$ with $\phi(0) = 0$ and $\phi(1) = 1$, which reweights *ranks* (cumulative probability levels) and thus encodes attitudes toward favorable vs. unfavorable outcomes [Schmeidler, 1989, Yaari, 1987]. In our finite-horizon setting, returns are discrete: for X with outcomes $x_1 \geq \cdots \geq x_N$ (best to worst), probabilities m_i, and cumulative masses $C_i = \sum_{j=1}^{i} m_j$ ($C_0 = 0$), the distortion value is

$$\rho_\phi(X) = \sum_{i=1}^{N} x_i \big(\phi(C_i) - \phi(C_{i-1})\big). \tag{3}$$

Plan	$\mathrm{CVaR}_{0.10}$	$\mathrm{CVaR}_{0.05}$	$\mathrm{WOWA}_{\beta=1}$	$\mathrm{WOWA}_{\beta=2}$
π_{long}	7	-6	18.7	17.4
π_{short}	-100	-100	26.0	13.4

Table 2: Risk evaluations for Example 1 using total returns. WOWA uses $\phi_\beta(p) = p^\beta$ with the best-to-worst convention.

We instantiate (3) with two criteria: CVaR (tail-only) and WOWA (spectrum-wide).

Conditional Value at Risk (CVaR). For $\alpha \in (0, 1]$, $\mathrm{CVaR}_\alpha(X)$ is the average of the worst α mass of X:

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{\alpha} \int_{1-\alpha}^{1} Q_X(p)\, dp, \tag{4}$$

where under the best-to-worst convention

$$Q_X(p) \triangleq F_X^{-1}(1-p), \qquad p \in [0, 1], \tag{5}$$

with $F_X^{-1}(p) \triangleq \inf\{x \in \mathbb{R} : \ F_X(x) \geq p\}$ [Dhaene et al., 2012]. CVaR admits classical coherent/spectral representations [Artzner et al., 1999, Rockafellar and Uryasev, 2000, Acerbi and Tasche, 2002, Acerbi, 2002] and is itself a distortion risk measure with distortion

$$\phi_\alpha(p) = \begin{cases} 0, & p \leq 1-\alpha, \\ \frac{p-(1-\alpha)}{\alpha}, & p > 1-\alpha, \end{cases} \tag{6}$$

so (3) applies directly.

Weighted Ordered Weighted Average (WOWA). WOWA is a flexible rank-based aggregation operator [Yager, 1988, Torra, 1997] that can be written in the distortion form (3) for a suitable ϕ. When ϕ is strictly increasing, ρ_ϕ is monotone w.r.t. first-order stochastic dominance; CVaR violates this because ϕ_α is flat on $[0, 1-\alpha]$ [Yaari, 1987]. Different shapes recover different behaviors: $\phi(u) = u$ yields expectation, convex ϕ yields risk aversion, and concave ϕ yields risk seeking [Chew et al., 1987, Yaari, 1987]. Accordingly, *standard* (lower-tail) CVaR is inherently risk-averse, while risk seeking can be modeled via concave distortions (e.g., WOWA with $\beta < 1$) or via optimistic CVaR variants such as OCVaR that optimize the upper tail [Ávila Pires et al., 2025].

Revisiting Example 1 Table 2 illustrates how tail-only and spectrum-wide criteria can induce different preferences. For WOWA we use the power distortion $\phi_\beta(p) = p^\beta$ and report two settings, $\beta \in \{1, 2\}$ (with $\beta = 1$ recovering expectation), alongside CVaR_α for $\alpha \in \{0.10, 0.05\}$.

Finally, note that naively solving (2) by enumerating policies is infeasible due to the combinatorial policy space, and even evaluating a single policy by expanding its scenario tree is exponential in the horizon H. This motivates algorithms that optimize (2) efficiently. [Strotz, 1955, Ruszczyński, 2010, Bäuerle and Ott, 2011, Chow et al., 2015, Bäuerle and Jaśkiewicz, 2024].

3.2 ERQDP: CERTIFIED RANK–QUANTILE DYNAMIC PROGRAMMING

We now introduce **ERQDP**, which optimizes (2) through three coupled steps:

(A) *Surrogate-global planning:* replace the full ranked return distribution by a K-slice summary on the rank (cumulative probability) axis, and solve this induced surrogate problem exactly by backward induction (**RQDP**).

(B) *Exact root evaluation (no trajectory enumeration):* for any explicit policy, compute the return PMF by dynamic programming on a discretized return grid, and evaluate the true root objective (up to a controlled rounding error).

(C) *Anytime refinement with root-safe updates:* increase K to refine the surrogate resolution, but accept a new policy only if its *exactly evaluated* root objective improves, ensuring monotone improvement under resolute semantics.

(A) Rank–Quantile surrogate and Dynamic Programming (RQDP). We define a surrogate approximation of (3) to make the optimization feasible. The main idea is that, instead of enumerating and sorting the full return distribution of a policy, we approximate it by grouping outcomes according to their *rank* i.e., their position along the cumulative probability axis into K slices.

Rank axis, grid, and slices Recall that rank-dependent criteria weight outcomes according to their position in the sorted list. Under the best-to-worst convention (5), the variable $p \in [0, 1]$ plays the role of a *rank (cumulative probability) level*: $p = 0$ indexes the best end of the distribution (top quantiles) and increasing p moves toward worse quantiles, with $p = 1$ indexing the worst end. It is not a probability of a single outcome; it is the cumulative-mass parameter used to traverse the sorted return distribution.

For a fixed resolution K, we discretize this rank axis into a grid

$$0 = p_0 < p_1 < \cdots < p_K = 1,$$

typically the uniform grid $p_k = k/K$, in which case $q_k = 1/K$ for all k. Each interval $(p_{k-1}, p_k]$ is a *slice* with probability mass (quota) $q_k \triangleq p_k - p_{k-1}$, so $\sum_{k=1}^{K} q_k = 1$. The integer K controls the resolution: larger K yields finer rank slices.

Piecewise-linear distortion on the grid. Let ϕ_K be the piecewise-linear interpolant of ϕ on this grid and define the slice weights

$$\Delta\phi_k \triangleq \phi_K(p_k) - \phi_K(p_{k-1}), \qquad k = 1, \ldots, K.$$

Then the rank-dependent objective of (3) reduces to the weighted sum

$$\rho_{\phi_K}(X) = \sum_{k=1}^{K} \Delta\phi_k \, \bar{Q}_X(k) = \langle \Delta\phi^{(K)}, \bar{Q}_X \rangle, \quad (7)$$

With $\Delta\phi^{(K)} \triangleq (\Delta\phi_1, \ldots, \Delta\phi_K) \in \mathbb{R}^K$, we summarize a return lottery X by a K-vector $\bar{Q}_X \in \mathbb{R}^K$ that averages the best-to-worst quantile function Q_X over K rank slices. Let $0 = p_0 < p_1 < \cdots < p_K = 1$ be a rank grid and $q_k \triangleq p_k - p_{k-1}$ the mass of slice k. For each $k \in \{1, \ldots, K\}$ we define the *slice-average quantile*

$$\bar{Q}_X(k) \triangleq \frac{1}{q_k} \int_{p_{k-1}}^{p_k} Q_X(p)\, dp. \quad (8)$$

In words, $\bar{Q}_X(k)$ is the average return level among outcomes whose *rank* (cumulative probability position in the best-to-worst ordering) lies in the interval $(p_{k-1}, p_k]$.

Since X is discrete in our setting, Q_X is a step function. Writing the distinct outcomes of X as $x_1 \geq \cdots \geq x_N$ with probabilities m_i and cumulative masses $C_i = \sum_{j=1}^{i} m_j$ (with $C_0 = 0$), the integral reduces to the finite overlap sum

$$\bar{Q}_X(k) = \frac{1}{q_k} \sum_{i=1}^{N} x_i \, \lambda\big((C_{i-1}, C_i] \cap (p_{k-1}, p_k]\big), \quad k = 1, \ldots, K. \quad (9)$$

where $\lambda(\cdot)$ denotes interval length on $[0, 1]$ (i.e., how much probability mass of outcome x_i falls inside slice k). This is exactly what the quota-filling step in our RQDP backup computes. Therefore, optimizing ρ_{ϕ_K} can be done over the K-dimensional summary $\bar{Q}_X$ rather than the full (return) distribution. An explicit $K{=}2$ quota-filling computation is given in Appendix A.

Surrogate planning problem In (2) we therefore replace ρ by the surrogate formulation (7):

$$\pi_K^\star \in \arg\max_{\pi \in \Pi_{\mathrm{DM}}} \rho_{\phi_K}(G^\pi(s_0)). \quad (10)$$

Given the slice-average vector $\bar{Q}_X \in \mathbb{R}^K$, the surrogate evaluation $\rho_{\phi_K}(X) = \langle \Delta\phi^{(K)}, \bar{Q}_X \rangle$ is a *linear functional of this K-vector*. Dynamic programming applies because we define an induced K-slice *surrogate return model* whose "state" is exactly this vector summary and whose backup computes the correct parent summary from successor summaries.

In Algorithm 1, we associate to each state s (and remaining horizon h) a K-vector $V_h(s) \in \mathbb{R}^K$, interpreted as the slice-average summary (8) of the h-step discounted return distribution from s under the surrogate semantics.

Backup For each state s, and given successor summaries $V_{h-1}(s') \in \mathbb{R}^K$, RQDP computes for each action a a candidate parent vector $\bar{Q}_h(s, a)$ by:

1. **atomize** successor slices into a list of value–mass pairs. *Atomization* means converting the successor K-slice summaries into a finite list of value–mass pairs ("atoms") whose mixture defines the one-step-ahead surrogate return distribution. Let $t \triangleq H - h$ denote the stage index from the root (so the one-step reward is discounted by γ^t). Then $\{(x_{s',k}, m_{s',k})\}$ that defines a discrete distribution over returns, where

$$x_{s',k} = \gamma^t \, r(s, a, s') + V_{h-1}(s')(k), \quad (11)$$

$$m_{s',k} = \Pr(s' \mid s, a)\, q_k. \quad (12)$$

   Each atom $(x_{s',k}, m_{s',k})$ represents the entire rank slice k of the successor distribution collapsed to its slice-average value. Importantly, parent slices are *not* computed "diagonally" in k: after all atoms are pooled across all successors and all k, we re-sort them globally and quota-fill the parent slices. Thus, $\bar{Q}_h(s, a)(k)$ can depend on many child indices k' through this re-ranking step.

2. **sort and quota-fill.** Sort all atoms by value $x_{s',k}$ from best to worst. Initialize remaining slice capacities $\mathrm{cap}_k \leftarrow q_k$ for $k = 1, \ldots, K$ and initialize slice accumulators $\mathrm{num}_k \leftarrow 0$. Scan the sorted atoms in order; for the current atom with remaining mass m and value x, repeatedly: allocate $\delta \leftarrow \min\{m, \mathrm{cap}_k\}$ to the current slice k, update $\mathrm{num}_k \leftarrow \mathrm{num}_k + \delta\, x$, $\mathrm{cap}_k \leftarrow \mathrm{cap}_k - \delta$, and $m \leftarrow m - \delta$; when $\mathrm{cap}_k = 0$, increment k. At the end, define $\bar{Q}_h(s, a)(k) \triangleq \mathrm{num}_k / q_k$.

The presented algorithm is optimal for the surrogate model: We justify that Algorithm 1 solves the induced K-slice surrogate problem (10). The only technical point is that the Backup step (atomize, sort, quota-fill) computes the slice-average vector $\bar{Q}$ of the discrete distribution represented by the atoms; this is formalized in Lemma 1. The surrogate optimality then follows by standard finite-horizon backward induction.

Lemma 1 (Quota filling computes slice-average quantiles)**.** *Fix a rank grid $0 = p_0 < p_1 < \cdots < p_K = 1$, with slice masses $q_k = p_k - p_{k-1}$. For any discrete return distribution with sorted outcomes $x_1 \geq \cdots \geq x_N$, probabilities m_i, and cumulative masses $C_i = \sum_{j=1}^{i} m_j$, the sorting and quota-filling procedure returns exactly the slice-average vector $\bar{Q}_X$ defined in* (8)–(9)*.*

Proof. The best-to-worst quantile function is constant and equal to x_i on each interval $(C_{i-1}, C_i]$. Quota filling scans these same intervals in sorted order and allocates to slice k exactly the overlap $(C_{i-1}, C_i] \cap (p_{k-1}, p_k]$. Thus each slice accumulator is precisely the numerator of (9). Full details are given in Appendix F. □

Algorithm 1 RQDP: Rank–Quantile DP (fixed K)

Require: distortion ϕ (evaluated on grid points), horizon H, rank resolution K

Define rank grid $0 = p_0 < \cdots < p_K = 1$; quotas $q_k \leftarrow p_k - p_{k-1}$
Define slice weights $\Delta\phi_k \leftarrow \phi(p_k) - \phi(p_{k-1})$ for $k = 1, \ldots, K$
for all $s \in \mathcal{S}$ **do**
    $V_0(s) \leftarrow \mathbf{0} \in \mathbb{R}^K$
end for
for $h = 1$ to H **do** ▷ h = remaining steps
    $t \leftarrow H - h$ ▷ stage index from the root (for discounting)
    for all $s \in \mathcal{S}$ **do**
        if terminal(s) **then**
            $V_h(s) \leftarrow \mathbf{0}$; **continue**
        end if
        for all $a \in \mathcal{A}(s)$ **do**
            $\bar{Q}_h(s,a) \leftarrow \textsc{Backup}(s, a, t, V_{h-1}(\cdot), q)$
        end for
        $\pi_h(s) \in \arg\max_{a \in \mathcal{A}(s)} \langle \Delta\phi, \bar{Q}_h(s,a) \rangle$
        $V_h(s) \leftarrow \bar{Q}_h(s, \pi_h(s))$
    end for
end for
return $\pi = \{\pi_h\}_{h=1}^{H}$ (and optionally $V_H(\cdot)$)

Proposition 1 (Surrogate optimality of RQDP)**.** *For fixed K and the induced piecewise-linear distortion ϕ_K, Algorithm 1 returns a deterministic Markov policy that maximizes the surrogate root objective in* (10)*.*

Proof. For each state, action, and remaining horizon, Lemma 1 shows that the backup computes the exact K-slice summary of the induced surrogate return lottery. Since the surrogate objective is linear in this summary, $\langle \Delta\phi^{(K)}, \bar{Q} \rangle$, selecting the action with maximal score is Bellman-optimal for the induced surrogate model. Finite-horizon backward induction then proves surrogate optimality. Full details are in Appendix F. □

(B) Exact enumeration-free policy evaluation via PMF dynamic programming We nee to evaluate a candidate policy π (obtained with Algorithm 1) under the resolute measure (3). This deterministic method computes the induced return distribution of $G^\pi(s_0)$ *without* trajectory enumeration, Monte Carlo rollouts, or stochastic sampling, and then evaluates it using ρ.

We evaluate a fixed policy π by dynamic programming over return PMFs on a discretized grid of step $1/c$ (returns rounded to $\{j/c : j \in \mathbb{Z}\}$), restricted to the bounded interval $[G_{\min}, G_{\max}]$. Using global bounds $G_{\min} \le G^\pi(s_0) \le G_{\max}$ (e.g., from reward bounds), we restrict to indices $j \in \{\lceil cG_{\min} \rceil, \ldots, \lfloor cG_{\max} \rfloor\}$. Here $c \in \mathbb{N}$ controls resolution: larger c gives a finer grid resolution(step size $1/c$).

For a state s and a remaining horizon h, let $\mathrm{PMF}_h^\pi(s)$ denote the distribution of the remaining h-step discounted return from s when following the policy π.

Here δ_0 denotes the Dirac mass at return 0 (probability 1 at grid index 0). With base $\mathrm{PMF}_0^\pi(s) = \delta_0$, we propagate PMFs backward by mixing shifted successor distributions:

$$\mathrm{PMF}_h^\pi(s) = \sum_{s'} P(s' \mid s, \pi_h(s)) \, \mathrm{Shift}\Big(\mathrm{PMF}_{h-1}^\pi(s'), \; \mathrm{round}\big(c\,\gamma^t\, r(s, \pi_h(s), s')\big)\Big) \tag{13}$$

where $t = H - h$ is the stage index from the root. Here $\mathrm{Shift}(\cdot, k)$ shifts a PMF by k integer grid points, and $\mathrm{round}(\cdot)$ rounds to the nearest integer.

Proposition 2 (Enumeration-free PMF evaluation and rounding guarantees)**.** *For a fixed policy π and discretization parameter c, the recursion* (13) *computes the distribution of the rounded return $\tilde{G}^\pi(s_0)$ on the grid $\{j/c\}$. If all discounted rewards $\gamma^t r(s_t, a_t, s_{t+1})$ lie on the grid (exact alignment), then $\tilde{G}^\pi(s_0) = G^\pi(s_0)$.*

Otherwise, for every trajectory,

$$\big|G^\pi(s_0) - \tilde{G}^\pi(s_0)\big| \;\le\; \frac{H}{2c} \quad \textit{a.s.}$$

and for any distortion ϕ,

$$\big|\rho_\phi(G^\pi(s_0)) - \rho_\phi(\tilde{G}^\pi(s_0))\big| \;\le\; \frac{H}{2c}.$$

Proof. For a fixed policy, the return distribution satisfies a standard law-of-total probability recursion over the next state. Equation (13) implements this recursion by shifting each successor PMF by the rounded one-step discounted reward and mixing by transition probabilities. Each rounded reward differs from the true discounted reward by at most $1/(2c)$; summing over H stages gives the trajectory-wise bound $H/(2c)$. Since distortion values are monotone weighted integrals of quantiles, the same uniform error bounds the risk value. Full details are in Appendix F. □

(C) Anytime refinement and root-safe improvements. ERQDP is an anytime loop over increasing rank resolutions K (e.g., doubling). For each K update:

after running (A) and (B): (1) perform root-safe improvement sweeps that consider one-step deviations (take an action a outside the considered policy and then follow the incumbent policy) Meaning, at a selected state–stage pair (s, h), we try each $a \in \mathcal{A}(s)$ as a one-step deviation: take a at (s, h), then follow the incumbent policy for the remaining

step and accept a changed policy only if it strictly improves the *root* objective, avoiding any reliance on time-consistency [Strotz, 1955, Ruszczyński, 2010]; (2) prune actions using surrogate envelopes to reduce computation in large action spaces.

ERQDP (Algorithm 2) For any explicit policy π, let

$$J(\pi) \triangleq \rho_\phi\Big(\tilde{G}^\pi(s_0)\Big)$$

denote the root objective evaluated on the return grid (Proposition 2); in particular $J_K \triangleq J(\pi_K)$. Let ϕ_K be the K-slice surrogate distortion and define the surrogate-optimal value

$$\widehat{J}_K^\star \triangleq \max_{\pi\in\Pi_{\mathrm{DM}}} \rho_{\phi_K}(G^\pi(s_0)),$$

which is returned by RQDP (Proposition 1). We maintain a lower bound $LB \triangleq \max\{J(\pi)$ over evaluated policies $\pi\}$ and an upper bound

$$UB \triangleq \widehat{J}_K^\star + \mathcal{E}(K,c),$$
$$\mathcal{E}(K,c) \triangleq (G_{\max}-G_{\min})\,\delta_K + \frac{H}{2c}.$$

where $\delta_K = \|\phi-\phi_K\|_\infty$. Then UB upper-bounds the optimal root value for ρ_ϕ up to the discretization/rounding budgets, and the stopping test $UB - LB \le \varepsilon$ yields the reported certificate gap. The entire ERQDP algorithm is summarized in Algorithm 2.

In this algorithm, there are two independent sources of approximation: (i) rank discretization through ϕ_K (and the K-slice surrogate), and (ii) return-grid rounding in PMF evaluation (Proposition 2). The following result shows that the distance between the approximated evaluation and the exact one is bounded. A standard stability bound for distortion integrals controls sensitivity to the distortion function. Let $\delta_K = \|\phi-\phi_K\|_\infty$.

On the uniform grid $p_k = k/K$, δ_K decreases as $O(1/K)$ for Lipschitz ϕ (and more generally $\delta_K \le L\max_k(p_k - p_{k-1})$).

Theorem 1 (Distortion stability under $\|\cdot\|_\infty$)**.** *If $X \in [m,M]$ almost surely and ϕ,ψ are distortions, then*

$$\big|\rho_\phi(X)-\rho_\psi(X)\big| \le (M-m)\,\|\phi-\psi\|_\infty. \tag{14}$$

Proof. Write the distortion value as the Stieltjes integral $\rho_\phi(X) = \int_0^1 Q_X(p)\,d\phi(p)$. Then the difference between two distortions is

$$\rho_\phi(X)-\rho_\psi(X) = \int_0^1 Q_X(p)\,d(\phi-\psi)(p).$$

Since Q_X is monotone and ranges over an interval of length at most $M-m$, this integral is bounded by $(M-m)\|\phi-\psi\|_\infty$. Full details are in Appendix F. □

Algorithm 2 ERQDP

Require: distortion ϕ, horizon H, start state s_0, initial rank resolution K_0, max resolution $K_{\max}$, tolerance ε, return-grid scale c

$K \leftarrow K_0$
Initialize an incumbent policy π_{best}
$LB \leftarrow$ exact root value of π_{best} via PMF-DP (grid scale c)
repeat
  (A) Build ϕ_K (the K-slice surrogate of ϕ) and run RQDP to get a candidate policy π_K
  $\widehat{J}_K^\star \leftarrow$ surrogate root score of π_K (ret. by RQDP)
  (B) $J_K \leftarrow$ exact root value of π_K (original MDP) via PMF-DP (grid scale c)
  if $J_K > LB$ **then** $\pi_{\text{best}} \leftarrow \pi_K$; $LB \leftarrow J_K$
  end if
  (C) Try local policy changes around π_{best}; accept only if the exact root value improves; update LB
  Stopping / refine
  $UB \leftarrow \widehat{J}_K^\star + \mathcal{E}(K,c)$
  if $UB - LB \le \varepsilon$ **or** $K = K_{\max}$ **then**
    break
  end if
  $K \leftarrow \min\{2K,\ K_{\max}\}$
until false
return $\pi_{\text{best}}, LB, UB$ ▷ or certificate $UB-LB$

In practice, we compute $\delta_K = \|\phi-\phi_K\|_\infty$ either by a provable per-interval maximization (for common distortions such as the power family and the CVaR kink), or by deterministic grid evaluation on $[0,1]$ with a conservative margin. Combined with global return bounds $[G_{\min}, G_{\max}]$ (e.g., if one-step rewards satisfy $r \in [r_{\min}, r_{\max}]$ then $G_{\min} = \sum_{t=0}^{H-1}\gamma^t r_{\min}$ and $G_{\max} = \sum_{t=0}^{H-1}\gamma^t r_{\max}$) and the PMF rounding budget of Proposition 2, this yields a gap diagnostic. We use the term *ε-certificate* only in a **surrogate-scoped** sense: it becomes formal when δ_K is provably upper-bounded and when the objective is interpreted on the induced K-slice surrogate model; for the original objective it remains an *a posteriori* diagnostic absent additional bounds on surrogate-to-true compression.

Distance to optimality and certificate interpretation. Let $\pi_K^\star$ be the surrogate-optimal policy returned by RQDP for ϕ_K, and let $\pi^\star$ be an optimal policy for the original distortion ϕ. By Theorem 1 and Proposition 2, for any evaluated policy π,

$$\big|\rho_\phi(\pi)-\rho_{\phi_K}(\pi)\big| \le \mathcal{E}(K,c).$$

Since $\pi_K^\star$ maximizes the surrogate objective exactly, the optimality loss of the incumbent is controlled by

$$\rho_\phi(\pi^\star)-\rho_\phi(\pi_{\text{best}}) \le \Big(\rho_{\phi_K}(\pi_K^\star)-\rho_\phi(\pi_{\text{best}})\Big)+\mathcal{E}(K,c).$$

This is the reason ERQDP reports the deterministic gap $\mathrm{UB}-\mathrm{LB}$. If the gap is below the target tolerance, the incum-

bent is certified under the selected discretization budgets; otherwise, ERQDP still returns the policy together with an explicit residual optimality-gap bound. See Appendix F.

4 EXPERIMENTS

Common protocol. All experiments use finite-horizon tabular MDPs and evaluate a policy by the distribution of total return from a specified initial state. ERQDP planning, policy evaluation, and certificate computation are deterministic: they use rank–quantile dynamic programming and PMF-DP, not Monte Carlo rollouts or trajectory sampling. Throughout, "Opt. gap" denotes the final deterministic upper–lower optimality-gap bound returned by ERQDP under the stated discretization budgets. Values below the run tolerance correspond to certified rows; larger values remain explicit residual bounds. Runtime is end-to-end planning time. We use three benchmark families as follows.

4.1 BETTING GAME (BG)

BG is a ten-stage wealth-allocation problem from Rigter et al. [2022], with horizon $H = 10$ and discount factor $\gamma = 1$. A state (t, m) consists on one time step $t \in \{0, \dots, 10\}$ and a current wealth $m \in \{0, \dots, 100\}$; the initial state is $(0, 5)$. At each nonterminal stage the agent chooses a bet $b \in \{0, \dots, 5\}$, clipped by available wealth. The transition increases wealth by b with probability 0.7, by $10b$ with probability 0.05, and decreases it by b with probability 0.25, with wealth clipped to $[0, 100]$. The terminal cost is the shortfall $C := 100 - m_T$; our implementation uses the equivalent final reward $m_T - 100$, so minimizing cost is equivalent to maximizing return. We report mean cost and tail cost $\mathrm{CVaR}_\alpha(C)$. ERQDP tracks an upper bound on the optimal risk value and reports the final policy gap. If this gap is below the strict run tolerance, the row is certified; otherwise the same number is an explicit residual bound. Rigter et al. report CVaR-based planning, so WOWA objective values are not directly comparable to their optimized criterion. We therefore compare policies under common cost metrics, regardless of the objective used to obtain them. Table 3 summarizes representative WOWA settings spanning risk-seeking ($\beta < 1$), risk-neutral ($\beta = 1$), and risk-averse ($\beta > 1$). Additional BG results are in Appendix C.

4.2 CERTIFIED PLANNING ON GRIDWORLD (GW) RETURN-RISK METRICS

GW is a stochastic navigation benchmark on a 5×5 grid with row-major cell indices $0, \dots, 24$, start cell 0, horizon $H = 20$, and discount $\gamma = 0.9$. The agent chooses one of four compass actions. With probability $1 - p_{\mathrm{slip}}$ the intended action is executed, and with probability p_{slip} the executed action is redrawn uniformly from the four compass moves.

Table 3: BG (cost; lower is better). Policies tuned for different objectives are evaluated under common cost metrics. WOWA objective values are *not* comparable to CVaR objectives, so comparison is via mean / CVaR / VaR.

Source	Measure	Tuned for	Mean	$\mathrm{CVaR}_{0.2}$	$\mathrm{VaR}_{0.2}$	Opt. gap
Rigter et al.	EV	–	58.58	97.41	–	–
Rigter et al.	CVaR-WC	$\alpha = 0.2$	82.86	91.84	–	–
Rigter et al.	CVaR-EV	$\alpha = 0.2$	75.33	91.79	87	–
ERQDP	CVaR	$\alpha = 0.2$	95	95	95	1.4e-14
ERQDP	WOWA	$\beta = 0.5$	61.73	99.71	94	9.4e-03
ERQDP	WOWA	$\beta = 1$	61.65	97.91	87	**0**
ERQDP	WOWA	$\beta = 2$	62.03	96.59	88	**0**

Table 4: GW CVaR sweep. Positive deltas mean ERQDP achieves *higher* tail-return and/or mean return. Speedup is Ávila Pires et al. [2025] reported training time divided by ERQDP solve time, using medians across instances.

α	Med. opt. gap ↓	Solve(s) ↓	Speedup(×) ↑	$\Delta\mathrm{CVaR}_\alpha$ ↑	$\Delta\mathbb{E}[G]$ ↑
0.01	8.46e-02	724.0	6.44	-0.58	-10.08
0.05	1.98e-02	1219.7	1.04	-4.42	-10.96
0.10	0.0e+00	212.0	78.49	+7.16	+6.11
0.25	4.80e-03	7.49	259.47	+5.02	+2.82
0.50	3.41e-03	1.40	928.18	+9.42	+9.47
1.00	4.34e-03	1.70	772.77	+9.73	+9.33

Entering a goal cell terminates the episode with reward $+50$; entering a penalty cell gives the negative reward listed thereafter. We report four matched instances, denoted GW1 to GW4: GW1 (respectively GW2, GW3 and GW4) use goal cell 21 (respectively 21, 21 and 24) with penalties on cells $\{10, 12\}$ (respectively $\{10, 12\}$, $\{12\}$, and $\{16\}$) corresponding to reward -15 (respectively -15, -5 and -30), and slip probability of 0.1 (respectively 0.3, 0.1, and 0.1). Appendix B.1 gives the exact layouts.

We compare ERQDP with the reported Ávila Pires et al. [2025] results on shared return-risk metrics. ERQDP is a model-based planner/certifier, so the reported speedup compares baseline training time with ERQDP solve time rather than identical runtime modes. Table 4 reports the median ERQDP optimality gap, solve time, speedup, and median differences in CVaR_α and mean return, with positive differences favoring ERQDP. Table 4 also shows the main pattern. For $\alpha \geq 0.10$, ERQDP has zero or small median optimality gaps, large training-time/solve-time speedups, and positive median tail-return gains while preserving or improving mean return. At extreme tails ($\alpha \leq 0.05$), the problem becomes more demanding; the reported gap quantifies the remaining bound at the tested resolution. Beyond lower-tail CVaR, the same ERQDP pipeline supports WOWA distortions for risk-seeking, risk-neutral, and risk-averse behavior. Appendix B.2 compares this spectrum-wide modeling option with OCVaR, an optimistic upper-tail CVaR variant.

4.3 INVENTORY CONTROL (IC)

Inventory Control (IC) is reported in detail in Appendix D, because the main comparison with Rigter et al. [2022] is summarized by a single cost table there. IC is a ten-period replenishment problem with capacity $N = 20$, discount $\gamma = 1$, state (t, n, d_{prev}), initial inventory 10, and initial previous demand 10. At each period the agent chooses an order quantity $a \in \{0, \dots, 20 - n\}$, equivalently implemented with 21 action indices and capacity clipping. Demand follows a clipped random walk with increments in $\{-5, \dots, 5\}$, and the policy trades purchasing and holding costs against revenue and rare high-cost outcomes. On this benchmark, ERQDP-CVaR at $\alpha = 0.02$ attains $\text{CVaR}_{0.02}$ cost 373.26 (about 3.4% lower than the best Rigter et al. CVaR baseline), while additionally outputting an explicit a posteriori optimality-gap bound.

4.4 PARAMETER SENSITIVITY AND SCALABILITY

Our algorithm contains two discretization parameters that play different roles. The rank resolution K controls the number of slices used by RQDP on the cumulative-probability axis. Increasing K tightens the approximation of the distortion function and usually decreases the certificate gap, but it increases the cost of the RQDP backup because each successor contributes K atoms. The return-grid scale c controls deterministic policy evaluation: returns are rounded to multiples of $1/c$, and Proposition 2 gives the worst-case rounding contribution $H/(2c)$. Thus, K primarily affects the surrogate optimality envelope, whereas c affects PMF evaluation accuracy and memory. When reporting normalized gaps, we use the return $\text{span} = G_{\max} - G_{\min}$, where $G_{\min}$ and $G_{\max}$ are the global cumulative-return bounds.

For fixed K and c, ERQDP remains polynomial in the tabular model size and horizon and avoids exponential trajectory or policy enumeration. If B is the maximum number of successors per state–action pair, each RQDP backup sorts BK atoms and costs $O(\text{BK}\log(\text{BK}))$ time. A full finite-horizon RQDP sweep costs $O(H|\mathcal{S}||\mathcal{A}|\text{BK}\log(\text{BK}))$ time, with $O(|\mathcal{S}|K)$ memory when only two horizon layers are kept. PMF evaluation for a fixed policy costs $O(H|\mathcal{S}|\text{BL})$ time and $O(|\mathcal{S}|L)$ memory, where L is the number of reachable return-grid values. This is more expensive than scalar expected-return DP, whose backup stores one value per state, but it replaces exponential scenario-tree enumeration with dynamic programming over rank slices and return-grid distributions.

The tabular sizes in our experiments range from 25-state GridWorld (GW) instances with horizon $H = 20$ and four actions, to Betting Game (BG) with horizon $H = 10$, 1011 time-augmented states, and six bet actions, and Inventory Control (IC) with horizon $H = 10$, 4411 time-augmented states, and up to 21 order actions. Table 5 summarizes representative sensitivity checks for these benchmark families.

Table 5: Representative discretization sensitivity. Gaps are percentages of the return span. Δ_K compares the smallest and largest tested $K_{\max}$; Δ_c is the gap variation at fixed rank budget. The GW row reports medians over GW1–GW4.

Bench.	Risk setting	Δ_K	Δ_c
BG	CVaR $\alpha = .20$	$1.25 \to 0.078$	0.234
BG	WOWA $\beta = 2$	$0.015 \to 0.006$	0.006
GW	CVaR $\alpha = .10$	$3.74 \to 0.11$	0.15
IC	CVaR $\alpha = .02$	$15.8 \to 1.78$	2.06

The sensitivity results should be read as a compute-accuracy trade-off. The rank resolution K is the main dial: increasing $K_{\max}$ consistently tightens the reported residual gap, with some runs crossing the strict tolerance and others stopping with an explicit remaining bound. In contrast, varying c over $\{25, 50, 100, 200\}$ changes the certificate gap only marginally in these runs, indicating that return-grid rounding is not the limiting factor once c is moderately large.

Our practical rule is therefore simple: set c so that $H/(2c)$ is below the desired numerical tolerance, then increase K by doubling until the certificate target is met or the remaining residual gap is acceptable. When the tested budget is insufficient, ERQDP reports the residual gap, allowing the remaining possible suboptimality due to approximation to be bounded rather than left unspecified.

5 CONCLUSION

We introduced ERQDP for finite-horizon tabular MDP planning under resolute root-level risk objectives. It preserves trajectory-level risk semantics while avoiding trajectory enumeration and sampling: it solves a rank–quantile surrogate by dynamic programming, evaluates explicit policies by deterministic PMF-DP up to a stated rounding budget, and reports a deterministic upper–lower optimality-gap bound. When this gap is below tolerance, ERQDP certifies the returned policy; otherwise it returns a residual bound under the chosen discretization budgets. Experiments on Betting Game, GridWorld, and Inventory Control illustrate the compute–accuracy tradeoff, support CVaR and WOWA in one pipeline, and show small gaps with fast risk-parameter sweeps on matched GridWorld comparisons. The sensitivity study confirms that rank resolution is the main certificate-tightening parameter, whereas PMF scale has little effect once return-grid rounding is negligible. Future work includes adaptive rank grids, sharper distortion-error bounds, and tighter surrogate-to-original certificates.

REFERENCES

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

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Mohamadreza Ahmadi, Anushri Dixit, Joel W. Burdick, and Aaron D. Ames. Risk-averse stochastic shortest path planning. In *Proceedings of the 60th IEEE Conference on Decision and Control (CDC)*, pages 5199–5204. IEEE, 2021. doi: 10.1109/CDC45484.2021.9683527.

Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21(4):503–546, 1953.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Bernardo Ávila Pires, Mark Rowland, Diana Borsa, Zhaohan Daniel Guo, Khimya Khetarpal, André Barreto, David Abel, Rémi Munos, and Will Dabney. Optimizing return distributions with distributional dynamic programming. *Journal of Machine Learning Research*, 26(185): 1–90, 2025.

Nicole Bäuerle and Anna Jaśkiewicz. Markov decision processes with risk-sensitive criteria: an overview. *Mathematical Methods of Operations Research*, 99:141–178, 2024.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Aharon Ben-Tal and Marc Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32(11):1445–1466, 1986. doi: 10.1287/mnsc.32.11.1445.

Soo Hong Chew, Edi Karni, and Zvi Safra. Risk aversion in the theory of expected utility with rank dependent probabilities. *Journal of Economic Theory*, 42(2):370–381, 1987. doi: 10.1016/0022-0531(87)90093-7.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: A CVaR optimization approach. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, 2015.

Anthony Coache and Sebastian Jaimungal. Reinforcement learning with dynamic convex risk measures. *Mathematical Finance*, 34(2):557–587, 2024. doi: 10.1111/mafi.12388. Published online 2023-04-17.

Jan Dhaene, Alexander Kukush, Dennis Linders, and Qihe Tang. Remarks on quantiles and distortion risk measures. *European Actuarial Journal*, 2:319–328, 2012. doi: 10.1007/s13385-012-0058-0.

Rui Ding and Eugene A. Feinberg. CVaR optimization for MDPs: Existence and computation of optimal policies. *SIGMETRICS Performance Evaluation Review*, 50(2): 39–41, aug 2022.

Christophe Gonzales and Patrice Perny. Decision under uncertainty. In *A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning*, pages 549–586. Springer, 2020.

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

Stratton C. Jaquette. A utility criterion for Markov decision processes. *Management Science*, 23(1):43–49, 1976.

Gildas Jeantet and Olivier Spanjaard. Computing rank dependent utility in graphical models for sequential decision problems. *Artificial Intelligence*, 175(7–8):1366–1389, 2011. doi: 10.1016/j.artint.2010.11.019.

Gildas Jeantet, Patrice Perny, and Olivier Spanjaard. Sequential decision making with rank dependent utility: A minimax regret approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1931–1937, 2012. doi: 10.1609/aaai.v26i1.8399.

Andrey Kolobov, Mausam, and Daniel S. Weld. A theory of goal-oriented MDPs with dead ends. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 438–447, 2012.

Jia Lin Hau, Erick Delage, Mohammad Ghavamzadeh, and Marek Petrik. On dynamic programming decompositions of static risk measures in markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 51734–51757, 2023a. Also available as arXiv:2304.12477.

Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted MDPs. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pages 47–76, 2023b.

Alexandre Marthe, Samuel Bounan, Aurélien Garivier, and Claire Vernade. Efficient risk-sensitive planning via entropic risk measures. arXiv preprint, 2025. Also circulated as HAL preprint hal-04967311v2.

Oliver Mortensen and Mohammad Sadegh Talebi. Recursive entropic risk optimization in discounted MDPs: Sample complexity bounds with a generative model. arXiv preprint, 2025.

Włodzimierz Ogryczak, Patrice Perny, and Paul Weng. A compromise programming approach to multiobjective Markov decision processes. *International Journal of Information Technology & Decision Making*, 12(05):1021–1053, 2013.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Marc Rigter, Paul Duckworth, Bruno Lacerda, and Nick Hawes. Planning for risk-aversion and expected value in MDPs. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, volume 32, pages 307–315, 2022. doi: 10.1609/icaps.v32i1.19814.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-Risk. *Journal of Risk*, 2(3):21–41, 2000.

Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010.

David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.

Robert H. Strotz. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, 23(3): 165–180, 1955.

Xihong Su, Marek Petrik, and Julien Grand-Clément. Risk-averse total-reward MDPs with ERM and EVaR. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20646–20654, 2025. doi: 10.1609/aaai.v39i19.34275.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7): 3323–3338, 2017. doi: 10.1109/TAC.2016.2644871.

Vicenç Torra. The weighted OWA operator. *International Journal of Intelligent Systems*, 12(2):153–166, 1997.

Shaun S. Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin*, 26(1):71–92, 1996.

Li Xia and Jinyan Pan. Markov decision processes with value-at-Risk criterion. arXiv preprint, 2025.

Menahem E. Yaari. The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, 55(1): 95–115, 1987.

Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1): 183–190, 1988.

Xian Yu and Siqian Shen. Risk-averse reinforcement learning via dynamic time-consistent risk measures. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, pages 2307–2312. IEEE, 2022. doi: 10.1109/CDC51059.2022.9992450.

Xian Yu and Lei Ying. On the global convergence of risk-averse policy gradient methods with expected conditional risk measures. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40425–40451. PMLR, 2023.

Dake Zhang, Boxiang Lyu, Shuang Qiu, Mladen Kolar, and Tong Zhang. Pessimism meets risk: Risk-sensitive offline reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 59459–59489. PMLR, 2024.

Long-Term Sequential Decision Making Under Risk (Supplementary Material)

Irmaan (Mohammad) Mirzanejad[1] **Nadjet Bourdache**[1] **Abdel-illah Mouaddib**[1]

[1]Université de Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, France

A TOY SLICE EXAMPLE FOR QUOTA FILLING

Consider a lottery with outcomes $x_1 \geq x_2 \geq x_3 \geq x_4 \geq x_5$ and probabilities $m = (0.1, 0.2, 0.2, 0.3, 0.2)$, so $C = (0.1, 0.3, 0.5, 0.8, 1)$. Let $K = 2$ with grid $p_0 = 0$, $p_1 = 0.5$, $p_2 = 1$. Then slice 1 averages the top half-mass: $\bar{Q}_X(1) = \frac{1}{0.5}\big(0.1x_1 + 0.2x_2 + 0.2x_3\big)$, and slice 2 averages the bottom half-mass: $\bar{Q}_X(2) = \frac{1}{0.5}\big(0.3x_4 + 0.2x_5\big)$.

B GRIDWORLD COMPARISON DETAILS

This section documents the precise aggregation procedure used to produce Table 4.

B.1 GRIDWORLD LAYOUTS AND SETTINGS

The four GridWorld instances used in the matched comparisons are denoted GW1–GW4. They use row-major cell numbering, start at cell 0, and follow the slip rule stated in Section 4.2. Table 6 lists the exact settings, and Figure 2 shows the corresponding reward layouts.

(1) Certificate-backed planning and an explicit "how close" signal. For all moderate risk levels and the risk-neutral setting ($\alpha \geq 0.10$, including $\alpha = 1$), every matched GW instance has an ERQDP gap below the prespecified tolerance. This matters because ERQDP reports either an ε-certificate or a residual optimality-gap bound, whereas the Ávila Pires et al. [2025] results report empirical performance without a solver-side certificate.

(2) Deployment-style speed for risk sweeps. For $\alpha \geq 0.10$, the reported training-time/solve-time speedups range from $78\times$ to $928\times$, with ERQDP solve times dropping to single-digit seconds for $\alpha \geq 0.25$. This is useful for risk sweeps because ERQDP directly re-solves each setting without a training loop.

(3) Tail-return behavior. For $\alpha \geq 0.10$, ERQDP has positive median $\Delta\mathrm{CVaR}_\alpha$ for every $\alpha \in 0.10, 0.25, 0.50, 1.00$, while also improving or largely preserving mean return. Thus, in the moderate-to-risk-neutral regime, the certificate-backed solutions are not obtained by simply sacrificing average performance; it often yields *Pareto-improvements* (higher tail-return *and* higher mean) at moderate-to-risk-neutral settings, where tail estimates are statistically stable.

Extreme tails ($\alpha \leq 0.05$): tradeoff between guarantees and computation. At the most extreme tails ($\alpha = 0.01, 0.05$), ERQDP is less frequently certified (under prespecified $varepsilon$)and the baseline can show better estimated tail-return. This regime is intrinsically hard: the objective concentrates on a vanishing fraction of trajectories, amplifying discretization demands (and, for any empirical method, requiring large sample sizes to stabilize tail estimates). Importantly, ERQDP still provides a principled *compute–guarantee dial*: increasing internal resolution tightens the envelope, and the remaining gap quantifies exactly how far we are from a certificate when we stop at a compute budget.

Beyond CVaR: distortion-risk sweep with WOWA ($\beta \geq 1$). While the Ávila Pires et al. [2025] GridWorld results are reported for CVaR, ERQDP supports a broader family of distortions under the same certified DP framework. We include

Table 6: GridWorld settings used in the main comparison. All instances are 5×5 grids with start cell 0, horizon $H = 20$, and discount $\gamma = 0.9$.

Instance	Slip p_{slip}	Goal cell/reward	Penalty cells/reward	Compared setting
GW1	0.1	$21/+50$	$10, 12/-15$	CVaR sweep
GW2	0.3	$21/+50$	$10, 12/-15$	CVaR sweep
GW3	0.1	$21/+50$	$12/-5$	CVaR sweep
GW4	0.1	$24/+50$	$16/-30$	CVaR sweep

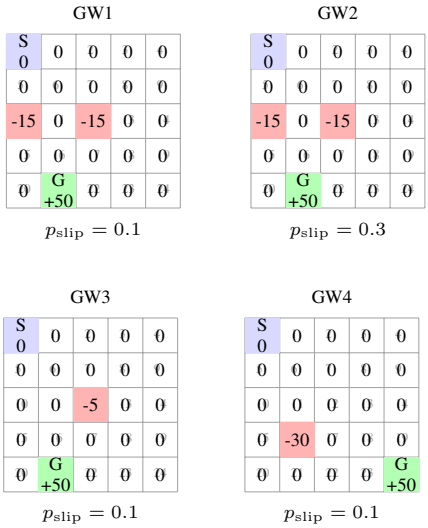

Figure 2: GridWorld reward layouts for GW1–GW4. Ordinary cells have reward 0; S is the start cell and G the goal cell. Small gray labels show row-major cell indices.

WOWA with $\beta \geq 1$ (risk-neutral at $\beta = 1$; more tail-sensitive as β increases), enabling risk-aversion sweeps without changing the algorithmic pipeline or losing the certificate mechanism.

Risk-seeking comparison. Ávila Pires et al. [2025] proposes OCVaR as a risk-seeking variant. On the four matched GridWorlds (GW1–GW4), ERQDP-WOWA (e.g., β=0.9) matches OCVaR's optimistic upper-tail score at τ=0.1 (median $\Delta\text{OCVaR}_{0.10} \approx 0$) while improving mean return (median $+3.03$) and the lower tail (median $+21.07$ in $\text{CVaR}_{0.10}$), with fewer penalty visits, and runs 28–41$\times$ faster (median 33.85$\times$).In these matched instances, OCVaR preserves the optimistic-tail score but can reduce lower-tail return, whereas ERQDP-WOWA with $\beta < 1$ attains comparable optimistic-tail values while maintaining stronger lower-tail metrics and solver-side optimality gaps/certificates.

Matched comparison. Table 4 aggregates only entries with the same GridWorld instance and risk parameter in both methods.

Metrics and sign convention. All comparisons are in return units, so higher is better. We use lower-tail CVaR_α and report median differences $\Delta\text{CVaR}_\alpha := \text{CVaR}_\alpha^{\text{ERQDP}} - \text{CVaR}_\alpha^{\text{DM}}$ and $\Delta\mathbb{E}[G] := \mathbb{E}[G]^{\text{ERQDP}} - \mathbb{E}[G]^{\text{DM}}$.

Speedup. ERQDP time is end-to-end planning wall-clock time. Since the Ávila Pires et al. [2025] baseline is training-based, we report speedup as

$$\text{Speedup} = \frac{\text{reported baseline training time}}{\text{ERQDP solve time}},$$

and take the median over matched instances for each α. Thus the timing comparison is a training-time/solve-time diagnostic, not an identical-runtime-mode comparison.

Optimality-gap interpretation. ERQDP reports a final upper–lower optimality-gap bound for the returned policy. Table 4 summarizes this quantity by its median over matched instances for each α. At extreme tails ($\alpha \leq 0.05$), tighter bounds

Table 7: Risk-seeking summary over the four matched GridWorld environments (GW1–GW4). Values are medians over GW1–GW4. Higher return is better; lower penalty-visits is better. Speedup compares Ávila Pires et al. [2025] reported training time with ERQDP solve time.

Method	Param	Median $\mathbb{E}[G] \uparrow$	Median $\mathrm{OCVaR}_{0.10} \uparrow$	Median $\mathrm{CVaR}_{0.10} \uparrow$	Median Penalty $\downarrow$	Time(s) $\downarrow$
OCVaR	$\tau = 0.1$	23.13	32.81	-4.27	0.238	1751 (train)
OCVaR	$\tau = 0.9$	23.90	32.59	1.86	0.224	1751 (train)
ERQDP-WOWA	$\beta = 0.9$	26.82	32.80	16.80	0.030	51.8 (solve)
ERQDP-WOWA	$\beta = 0.3$	26.82	32.80	16.80	0.030	805 (solve)

Table 8: Per-environment risk-seeking comparison: Ávila Pires et al. [2025] OCVaR ($\tau = 0.1$, seed-mean) vs. ERQDP-WOWA ($\beta = 0.9$). Speedup is reported training time divided by ERQDP solve time; Gap is the ERQDP optimality-gap bound.

Env	$\mathbb{E}[G]_{\mathrm{DM}}$	$\mathrm{CVaR}_{0.10,\mathrm{DM}}$	$\mathrm{OCVaR}_{0.10,\mathrm{DM}}$	$\mathbb{E}[G]_{\mathrm{ER}}$	$\mathrm{CVaR}_{0.10,\mathrm{ER}}$	$\mathrm{OCVaR}_{0.10,\mathrm{ER}}$	Speedup($\times$)	Gap
GW1	26.51	-9.25	32.81	30.11	17.21	32.80	32.62	4.85e-3
GW2	17.90	-24.17	32.81	23.53	4.59	32.80	27.95	4.41e-3
GW3	30.37	23.41	32.81	31.48	25.23	32.80	40.92	4.89e-3
GW4	19.75	0.71	23.91	22.20	16.39	23.91	35.07	4.18e-3

require higher rank resolution; otherwise ERQDP reports the remaining residual bound, whereas the baseline reports empirical performance without a solver-side optimality certificate.

Moderate-to-risk-neutral regime. For $\alpha \geq 0.10$, Table 4 shows small median optimality gaps, large training-time/solve-time speedups, and positive median $\Delta\mathrm{CVaR}_\alpha$, with typically positive $\Delta\mathbb{E}[G]$.

B.2 RISK-SEEKING OBJECTIVES: OCVAR VS. ERQDP-WOWA

Setup. We compare OCVaR and ERQDP-WOWA on the four matched GridWorld instances GW1–GW4. Table 7 summarizes medians, and Table 8 gives the per-instance rows.

What is "risk-seeking" here (empirical, from the reported metrics). Standard CVaR targets the *lower* tail and is inherently risk-averse. Ávila Pires et al. [2025]'s OCVaR is designed to act as an *optimistic* (upper-tail) functional. This is directly visible in the reported metrics: on GW1/2/8, the reported $\mathrm{OCVaR}_{0.10}$ saturates at the environment's maximum return, while the lower-tail $\mathrm{CVaR}_{0.10}$ can become strongly negative under risk-seeking tuning. In contrast, ERQDP achieves risk-seeking behavior through WOWA with $\beta < 1$ (optimistic distortion), while still producing an ε-certificate (or a final gap).

Interpretation (what the results actually show). On these four shared environments, OCVaR can indeed behave "risk-seeking" in the sense of maximizing an upper-tail score ($\mathrm{OCVaR}_{0.10}$), but this comes with a measurable downside: the lower-tail return ($\mathrm{CVaR}_{0.10}$) degrades sharply (median becomes negative at $\tau = 0.1$) and penalty exposure increases. ERQDP-WOWA ($\beta < 1$) matches the same upper-tail saturation where it is achievable, while simultaneously maintaining a much stronger lower tail (median $\mathrm{CVaR}_{0.10} \approx 16.8$) and fewer penalty visits.

Certified optimality and speed. A key advantage of ERQDP is that it reports either an explicit ε-certificate for the returned policy or, when the strict prespecified tolerance is not reached, the final residual gap. For both $\beta = 0.9$ and $\beta = 0.3$, ERQDP certifies all four shared environments. The speedups compare Ávila Pires et al. [2025]'s reported training runtime with ERQDP solve time on the same tasks; they should therefore be read as training-time/solve-time diagnostics, not as a same-mode planner-runtime comparison. (see Table 8) It can be seen while more extreme optimism ($\beta = 0.3$) increases runtime due to tighter certification (needing larger internal resolution), yet still preserves the certificate mechanism.

Table 9: BG: ERQDP-WOWA sweep (cost; lower is better). We report both the optimized WOWA objective value (not comparable to CVaR) and common cost-risk metrics (VaR/CVaR) for cross-method risk-handling comparison.

β	Mean	$\mathrm{CVaR}_{0.2}$	$\mathrm{VaR}_{0.2}$	$\mathrm{CVaR}_{0.02}$	$\mathrm{VaR}_{0.02}$	WOWA obj.	Gap	K	Time(s)
0.1	61.86	99.97	97	100	100	11.41	7.3e-02	17000	1621.1
0.5	61.73	99.71	94	100	100	42.13	9.4e-03	17000	1586.6
1	61.65	97.91	87	100	100	61.92	0.0e+00	256	82.4
2	62.03	96.59	88	100	100	78.66	0.0e+00	1024	141.7
3	95	95	95	95	95	95	4.4e-05	4096	428.6
5	95	95	95	95	95	95	1.5e-04	4096	364.7

Table 10: BG extended results (ERQDP). Mean is reported with 95% CI in brackets. VaR/CVaR are computed on the cost distribution (i.e., the upper tail of cost, equivalent to the lower tail of return).

Risk	Param	Mean [CI]	$\mathrm{VaR}_{0.02}$	$\mathrm{CVaR}_{0.02}$	$\mathrm{VaR}_{0.2}$	$\mathrm{CVaR}_{0.2}$	Gap	K	Time(s)
CVaR	$\alpha = 0.02$	95.00 [95.00,95.00]	95.00	95.00	95.00	95.00	2.8e-14	17000	1596.36
CVaR	$\alpha = 0.2$	95.00 [95.00,95.00]	95.00	95.00	95.00	95.00	1.4e-14	17000	1447.72
CVaR	$\alpha = 1$	61.65 [61.24,62.07]	100.00	100.00	87.00	97.91	0.0e+00	256	60.47
WOWA	$\beta = 0.1$	61.86 [61.43,62.30]	100.00	100.00	97.00	99.98	7.3e-02	17000	1621.07
WOWA	$\beta = 0.5$	61.73 [61.30,62.16]	100.00	100.00	94.00	99.71	9.4e-03	17000	1586.61
WOWA	$\beta = 1$	61.65 [61.24,62.07]	100.00	100.00	87.00	97.91	0.0e+00	256	82.41
WOWA	$\beta = 2$	62.03 [61.62,62.44]	100.00	100.00	88.00	96.59	0.0e+00	1024	141.69
WOWA	$\beta = 3$	95.00 [95.00,95.00]	95.00	95.00	95.00	95.00	4.4e-05	4096	428.59
WOWA	$\beta = 5$	95.00 [95.00,95.00]	95.00	95.00	95.00	95.00	1.5e-04	4096	364.66

C EXTENDED BG RESULTS AND ERQDP CERTIFICATES/GAPS

Table 9 reports the WOWA sweep, and Table 10 gives the extended ERQDP certificate/gaps and rollout summaries.

The extended table includes 95% confidence intervals over 20,000 rollouts, tail metrics, the final certification/optimality gap, and the rank-grid size K.

Takeaway. Lower α (CVaR) or higher β (WOWA) pushes toward policies that reduce catastrophic outcomes, which may increase average cost. Conversely, $\beta < 1$ is risk-seeking: it emphasizes best-case outcomes, yielding optimistic distorted values even when average cost stays similar.

D INVENTORY CONTROL (IC): DETAILED COMPARISON TO RIGTER ET AL. (ICAPS'22)

Setup and metric. Following Rigter et al., IC is evaluated over 10 stages with N=20 and the episode cost defined as $C = 400 - \sum_{t=0}^{H-1} \mathrm{profit}_t$ (thus $C < 400$ indicates net profit). We report $\mathrm{CVaR}_{0.02}$ (mean cost of the worst 2% runs) and mean cost.

Results. Table 11 shows that ERQDP improves the $\mathrm{CVaR}_{0.02}$ objective substantially, corresponding to roughly doubling the implied worst-tail profit (since tail-profit $= 400 - \mathrm{CVaR}_{0.02}$).

A posteriori optimality-gap bound (certificate). Beyond the empirical scores, ERQDP logs explicit bounds for the optimized risk objective: an upper bound UB_{opt}, a certified lower bound $LB_{\mathrm{opt}}^{\mathrm{cert}}$, and the achieved policy value $V^{\pi}(s_0)$, inducing a computable policy-gap bound

$$\mathrm{policy\_gap} = UB_{\mathrm{opt}} - V^{\pi}(s_0).$$

For IC at α=0.02 (with $K_{\max}$=512), ERQDP reports $UB_{\mathrm{opt}} = -287.747$, $LB_{\mathrm{opt}}^{\mathrm{cert}} = -458.747$, $V^{\pi}(s_0) = -373.264$, yielding *policy gap = 85.52* ; i.e., an explicit bound on suboptimality for the returned policy. Rigter et al. do not provide an

Table 11: Inventory Control (IC). Lower cost is better. Tail-profit is $400 - \text{CVaR}_{0.02}$.

Method	$\text{CVaR}_{0.02}$ cost ↓	Mean cost ↓	Tail-profit ↑	Gap bound?
Rigter EV	416.42	235.62	−16.42	–
Rigter CVaR-WC (α=0.02)	386.49	286.18	13.51	–
Rigter CVaR-EV (α=0.02)	386.92	250.38	13.08	–
ERQDP-CVaR (α=0.02)	**373.26**	250.29	**26.74**	✓

analogous bound in their evaluation protocol.

E DISCRETIZATION SENSITIVITY

This section reports the sensitivity checks summarized in Table 5. The experiments use the same ERQDP pipeline as the main results and vary only discretization budgets. Table 12 varies the rank budget $K_{\max}$, while Table 13 varies the PMF grid scale c. BG denotes Betting Game, GW denotes GridWorld, and IC denotes Inventory Control; a star denotes certification at the target tolerance. The qualitative conclusion is that K controls the certificate gap, while c is stable once the return grid is fine enough.

Table 12: Representative sensitivity to the rank resolution K. Larger K tightens the rank approximation and typically decreases the certificate gap, at the cost of additional runtime. Gap/span is reported as a percentage of the return span; $\star$ marks rows whose gap is below the target tolerance (prespecified ε).

Environment	Risk	$K_{\max}$	final K	gap/span	time (s)
Betting Game	CVaR $\alpha = 0.2$	64	64	1.250%	9.3
Betting Game	CVaR $\alpha = 0.2$	128	128	0.938%	18.2
Betting Game	CVaR $\alpha = 0.2$	256	256	0.312%	28.0
Betting Game	CVaR $\alpha = 0.2$	512	512	0.234%	48.5
Betting Game	CVaR $\alpha = 0.2$	1024	1024	0.078%*	131.0
Betting Game	WOWA $\beta = 2$	64	64	0.015%*	19.3
Betting Game	WOWA $\beta = 2$	128	128	0.006%*	26.6
Betting Game	WOWA $\beta = 2$	256	128	0.006%*	35.3
Betting Game	WOWA $\beta = 2$	512	128	0.006%*	32.0
Betting Game	WOWA $\beta = 2$	1024	128	0.006%*	28.7
GridWorld (median over GW1-4)	CVaR $\alpha = 0.1$	64	64	3.736%	2.8
GridWorld (median over GW1-4)	CVaR $\alpha = 0.1$	128	128	1.223%	4.7
GridWorld (median over GW1-4)	CVaR $\alpha = 0.1$	256	256	0.741%	7.0
GridWorld (median over GW1-4)	CVaR $\alpha = 0.1$	512	512	0.152%	12.0
GridWorld (median over GW1-4)	CVaR $\alpha = 0.1$	1024	512–1024	0.112%	18.9
Inventory Control	CVaR $\alpha = 0.02$	64	64	15.758%	839.6
Inventory Control	CVaR $\alpha = 0.02$	128	128	9.628%	1240.0
Inventory Control	CVaR $\alpha = 0.02$	256	256	2.063%	1960.0
Inventory Control	CVaR $\alpha = 0.02$	512	512	1.782%	3260.0

Table 13: Representative sensitivity to the PMF return-grid scale c at fixed rank budget. In these runs, changing c over $\{25, 50, 100, 200\}$ has little effect on the certificate gap, indicating that the rank approximation, rather than return rounding, is the dominant limiting factor.

Environment.	Risk	c	final K	gap/span	time (s)
Betting Game	CVaR $\alpha = 0.2$	25	512	0.234%	48.9
Betting Game	CVaR $\alpha = 0.2$	50	512	0.234%	46.2
Betting Game	CVaR $\alpha = 0.2$	100	512	0.234%	48.7
Betting Game	CVaR $\alpha = 0.2$	200	512	0.234%	51.4
Betting Game	WOWA $\beta = 2$	25	128	0.006%*	27.0
Betting Game	WOWA $\beta = 2$	50	128	0.006%*	25.7
Betting Game	WOWA $\beta = 2$	100	128	0.006%*	31.4
Betting Game	WOWA $\beta = 2$	200	128	0.006%*	29.7
GridWorld (median over 4)	CVaR $\alpha = 0.1$	25	512	0.152%	10.7
GridWorld (median over 4)	CVaR $\alpha = 0.1$	50	512	0.152%	9.0
GridWorld (median over 4)	CVaR $\alpha = 0.1$	100	512	0.152%	10.2
GridWorld (median over 4)	CVaR $\alpha = 0.1$	200	512	0.152%	11.0
Inventory Control	CVaR $\alpha = 0.02$	25	256	2.063%	1830.0
Inventory Control	CVaR $\alpha = 0.02$	50	256	2.063%	1790.0
Inventory Control	CVaR $\alpha = 0.02$	100	256	2.063%	1940.0
Inventory Control	CVaR $\alpha = 0.02$	200	256	2.063%	2030.0

F PROOFS

F.1 PROOF OF LEMMA 1

Proof. Let the sorted outcomes be $x_1 \geq \cdots \geq x_N$ with masses m_i and cumulative masses $C_i = \sum_{j=1}^{i} m_j$. The best-to-worst quantile function is constant and equal to x_i on the interval $(C_{i-1}, C_i]$. For a grid slice $(p_{k-1}, p_k]$, the slice-average quantile is therefore

$$\bar{Q}_X(k) = \frac{1}{q_k} \sum_{i=1}^{N} x_i \, \lambda\big((C_{i-1}, C_i] \cap (p_{k-1}, p_k]\big).$$

The quota-filling procedure scans the same outcome intervals in decreasing value order and allocates exactly the intersection length above to slice k. Its accumulator for slice k is consequently the numerator of this display, and division by q_k gives exactly $\bar{Q}_X(k)$. This holds for every slice, so the returned vector is the slice-average quantile vector. □

F.2 PROOF OF PROPOSITION 1

Proof. We prove the claim by induction on the remaining horizon h. For $h = 0$, all policies induce the zero return and the algorithm stores $V_0(s) = \mathbf{0}$, which is optimal. Suppose the claim holds for $h - 1$. For a fixed state s and action a, the successor summaries $V_{h-1}(s')$ are, by the induction hypothesis, the optimal surrogate summaries for the continuation problems. The backup forms the discrete mixture over discounted one-step rewards plus successor slice values. By Lemma 1, the sort-and-quota-fill operation computes the exact K-slice summary of this induced surrogate lottery. Since the surrogate objective is linear in the slice summary, namely $\langle \Delta\phi^{(K)}, \bar{Q} \rangle$, selecting an action that maximizes this score gives an optimal h-step surrogate action at s. Applying this argument for all states and all horizons proves that Algorithm 1 returns a deterministic Markov policy maximizing the induced K-slice surrogate objective. □

F.3 PROOF OF PROPOSITION 2

Proof. For a fixed policy, the return distribution satisfies the usual law of total probability over the first transition. The recursion in Eq. (13) applies exactly this decomposition: for each successor s', it shifts the already-computed PMF of the remaining return by the rounded one-step discounted reward, weights it by $P(s' \mid s, \pi_h(s))$, and sums over successors. Induction on h proves that the recursion computes the PMF of the rounded return $\tilde{G}^{\pi}(s_0)$.

If every discounted one-step reward is aligned with the grid, the shift is exact at every step, so $\tilde{G}^{\pi}(s_0) = G^{\pi}(s_0)$. Otherwise each rounded one-step reward differs from the true discounted reward by at most $1/(2c)$. Along any trajectory, summing over H stages gives $|G^{\pi}(s_0) - \tilde{G}^{\pi}(s_0)| \leq H/(2c)$ almost surely. If two random variables differ almost surely by at most η, then their quantile functions differ uniformly by at most η. A distortion risk functional is a monotone weighted integral of quantiles with total weight one, hence its value changes by at most η. Taking $\eta = H/(2c)$ gives the stated bound. □

F.4 PROOF OF THEOREM 1

Proof. Let Q_X denote the best-to-worst quantile function of X. The distortion value can be written as a Stieltjes integral $\rho_{\phi}(X) = \int_0^1 Q_X(p)\, d\phi(p)$. Therefore

$$\rho_{\phi}(X) - \rho_{\psi}(X) = \int_0^1 Q_X(p)\, d(\phi - \psi)(p).$$

Because $X \in [m, M]$ almost surely, Q_X is monotone with total variation at most $M - m$. Since $\phi(0) = \psi(0)$ and $\phi(1) = \psi(1)$ for distortions, integration by parts gives

$$\begin{aligned} \left| \int_0^1 Q_X\, d(\phi - \psi) \right| &= \left| \int_0^1 (\phi - \psi)\, dQ_X \right| \\ &\leq \|\phi - \psi\|_{\infty} (M - m), \end{aligned}$$

which proves the claim. □