On Dynamic Programming Decompositions of Static Risk Measures in Markov Decision Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Optimizing static risk-averse objectives in Markov decision processes is difficult because they do not admit standard dynamic programming equations common in Reinforcement Learning (RL) algorithms. Dynamic programming decompositions that augment the state space with discrete risk levels have recently gained popularity in the RL community. Prior work has shown that these decompositions are optimal when the risk level is discretized sufficiently. However, we show that these popular decompositions for Conditional-Value-at-Risk (CVaR) and Entropic-Value-at-Risk (EVaR) are inherently suboptimal regardless of the discretization level. In particular, we show that a saddle point property assumed to hold in prior literature may be violated. However, a decomposition does hold for Value-at-Risk and our proof demonstrates how this risk measure differs from CVaR and EVaR. Our findings are significant because risk-averse algorithms are used in high-stake environments, making their correctness much more critical.

1 Introduction

Risk-averse reinforcement learning (RL) seeks to provide a risk-averse policy for high stake real-world decision problems. These high-stake domains include autonomous driving (Jin et al., 2019; Sharma et al., 2020), robot collision avoidance (Ahmadi et al., 2021; Hakobyan and Yang, 2021), liver transplant timing (Köse, 2016), HIV treatment (Keramati et al., 2020; Zhong, 2020), unmanned aerial vehicle (UAV) (Choudhry et al., 2021), and investment liquidation (Min et al., 2022), to name a few. Because these domains call for reliable solutions, risk-averse algorithms must be based on solid theoretical foundations. This is one reason why monetary risk measures, such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), have become pervasive in risk-averse RL (Prashanth and Fu, 2022). Indeed, risk measures such as CVaR are known to be coherent (Artzner et al., 1999) with respect to a set of fundamental axioms that define how risk should be quantified and have been adopted as gold standards in banking regulations (Basel Committee on Banking Supervision, 2019).

Introducing risk-averse objectives in Markov decision processes (MDPs), the primary model used in RL, is challenging. Dynamic programming, the linchpin of most RL algorithms, cannot be used directly to optimize a risk measure like VaR or CVaR in MDPs. One line of work tackles this challenge by exploiting the primal representation of risk measures and augmenting the state space of their dynamic programs (DPs) with an additional parameter that typically represents the total cumulative reward up to the current point (Bäuerle and Ott, 2011; Boda et al., 2004; Chow and Ghavamzadeh, 2014; Filar et al., 1995; Hau et al., 2023; Lin et al., 2003; Wu and Lin, 1999; Xu and Mannor, 2011). Even when the original MDP is finite, this DP requires computing the value function for a continuous state space, and thus, has been considered inefficient in practice (Chapman et al., 2022; Chow et al., 2015; Li et al., 2022).

Another line of recent work leverages the dual representation to produce a *risk-level decomposition* of risk measures (Pflug and Pichler, 2016). Using this decomposition, numerous authors have derived

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

DPs for common risk measures and integrated them within various RL algorithms (Chapman et al., 2019, 2022; Chow et al., 2015; Ding and Feinberg, 2022; Ni and Lai, 2022; Rigter et al., 2021; Stanko and Macek, 2019). Although this risk-level decomposition requires augmenting the state space with a continuous parameter, this parameter is naturally bounded between 0 and 1. It has been generally accepted, with several *tentative* proofs supporting this claim (Chow et al., 2015; Li et al., 2022), that these DPs recover the optimal policy if we can discretize the augmented state space sufficiently finely. Moreover, it is believed that one can use the optimal value function from this DP to recover the policies that are optimal for the full range of risk levels.

In this paper, we make a surprising discovery that numerous claims of optimality of risk-level decompositions published in the past several years are incorrect. Even when one discretizes the augmented state space arbitrarily finely, most risk-level DPs are not guaranteed to recover the optimal value function and policy. There are several reasons why existing arguments fail. As the most common reason, several papers assume that a certain saddle point property holds, either explicitly (Chow et al., 2015) or implicitly (Ding and Feinberg, 2022; Li et al., 2022). We show that this property does not generally hold, invalidating the optimality of DPs, as hinted at in Chapman et al. (2019, 2022). This finding directly refutes the claimed or hypothesized optimality of algorithms proposed in many recent research papers and pre-prints, such as Chapman et al. (2019); Chow et al. (2015); Ding and Feinberg (2022); Li et al. (2022); Rigter et al. (2021); Stanko and Macek (2019). Our results also affect applications of these algorithms, such as automated vehicle motion planning (Jin et al., 2019). We also identify gaps in related decompositions (Li et al., 2022; Ni and Lai, 2022) and propose how to fix them.

We make the following contributions in this paper. *First*, we show in Section 3 that the popular DP for optimizing CVaR in MDPs may not recover the optimal value function and policy regardless of how finely one discretizes the risk level in the augmented states. This method was first proposed in Chow et al. (2015) but adopted widely afterwards (Chapman et al., 2019; Ding and Feinberg, 2022; Li et al., 2022; Rigter et al., 2021; Stanko and Macek, 2019). The simple counterexample in this section contradicts the optimality claims in Chow et al. (2015); Ding and Feinberg (2022); Li et al. (2022). We hypothesize that prior work missed this issue, because the CVaR DP works for policy evaluation and only fails when one uses it to optimize policies based on the "risk-to-go value function". Therefore, our results do not contradict the original decomposition in Pflug and Pichler (2016) that only applies to policy evaluation. We give a new independent and simple proof that the CVaR decomposition indeed works when evaluating a fixed policy.

Second, we show in Section 4 that the DP for optimizing the Entropic-Value-at-Risk in MDPs, proposed by Ni and Lai (2022), does not compute the correct value function even when the policy is fixed. Although EVaR has not been as popular as CVaR, it has been gaining attention in recent years (Hau et al., 2023). We give an example that contradicts the correctness claims of the risk-level decomposition for EVaR in Ni and Lai (2022). The gap that we identify with this objective applies to both policy evaluation and policy optimization. Furthermore, we prove a new, correct EVaR decomposition for policy evaluation. Unfortunately, the EVaR decomposition fails and is sub-optimal when is applied to policy optimization, similar to CVaR.

Third, we propose an *optimal dynamic program* for policy optimization of VaR in Section 5. Our DP is based on a risk-level decomposition that closely resembles the quantile MDP decomposition in Li et al. (2022) but corrects for several technical inaccuracies. The derivation shows why VaR stands apart from coherent risk measures like CVaR and EVaR. VaR is unique in that the decomposition can be constructed directly from the primal formulation of the risk measure, which avoids the complications that arise in the robust formulations used in CVaR and EVaR decompositions.

It is important to note that the correctness of DPs that augment the state space with the accumulated rewards is unaffected by our results (Bäuerle and Ott, 2011; Chow and Ghavamzadeh, 2014; Chow et al., 2018; Hau et al., 2023). These DPs use the *primal* risk measure representation and do not suffer from the same saddle point issue as the augmentation methods that use the *dual* representation of the risk measures, such as the one in Chow et al. (2015).

2 Preliminaries

This section summarizes relevant properties of monetary risk measures and outlines how they are typically used in the context of solving MDPs.

Monetary Risk Measures We restrict our attention to probability spaces with a finite outcome space Ω such that $|\Omega| = m$ for some $m \in \mathbb{N}$. We use $\mathbb{X} = \mathbb{R}^m$ to denote the space of real-valued random variables. To improve the clarity of probabilistic claims, we always adorn random variables with a tilde, such as $\tilde{x} \in \mathbb{X}$. In finite spaces, we can represent any random variable $\tilde{x} \in \mathbb{X}$ as a vector $\boldsymbol{x} \in \mathbb{R}^m$. We also use $\boldsymbol{q} \in \Delta_m$ to represent a probability distribution over Ω where Δ_m represents the m-dimensional probability simplex. Using this notation, we can write that $\mathbb{E}[\tilde{x}] = \boldsymbol{q}^\top \boldsymbol{x}$.

A *monetary risk measure* $\psi \colon \mathbb{X} \to \mathbb{R}$ assigns a real value to each real-valued random variable in a way that it is monotone and cash-invariant (Follmer and Schied, 2016; Shapiro et al., 2014). A risk measure can be seen as a generalization of the expectation operator $\mathbb{E}[\cdot]$ that also takes into account the uncertainty in the random variable. In this work, we define all risk measures for random variables $\tilde{x}$ that represent *rewards*. Thus, the risk-averse decision-maker aims to choose actions that maximize the value of the risk measure, i.e., a higher value of risk measure represents a lower exposure to risk.

We consider three monetary risk measures common in RL. Perhaps the most well-known measure is *Value-at-Risk* (VaR), which is defined for a risk-level $\alpha \in [0, 1]$ and a random variable $\tilde{x} \in \mathbb{X}$ in modern literature as (e.g., Follmer and Schied 2016; Shapiro et al. 2014)

$$\mathrm{VaR}_\alpha[\tilde{x}] \;=\; \sup\;\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} < z] \le \alpha\} \;=\; \inf\;\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} \le z] > \alpha\}\,. \tag{1}$$

Note that $\mathrm{VaR}_1[\tilde{x}] = \infty$. The equality between the two definition holds, for example, by Follmer and Schied (2016, remark A.20). Another popular risk measure is the *Conditional-value-at-Risk* (CVaR), which is defined for a risk level $\alpha \in [0, 1]$ and a random variable $\tilde{x} \in \mathbb{X}$ distributed as $\tilde{x} \sim \boldsymbol{q}$ as (e.g., Follmer and Schied 2016, definition 11.8 and Shapiro et al. 2014, eq. 6.23)

$$\mathrm{CVaR}_\alpha[\tilde{x}] \;=\; \sup_{z \in \mathbb{R}}\;\left(z - \alpha^{-1}\mathbb{E}\left[z - \tilde{x}\right]_+\right) \;=\; \inf_{\boldsymbol{\xi} \in \Delta_m}\;\left\{\boldsymbol{\xi}^\top \boldsymbol{x} \mid \alpha \cdot \boldsymbol{\xi} \le \boldsymbol{q}\right\}, \tag{2}$$

with $\mathrm{CVaR}_0[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$ and $\mathrm{CVaR}_1[\tilde{x}] = \mathbb{E}[\tilde{x}]$. Finally, the *entropic value at risk* (EVaR), with $\mathrm{EVaR}_0[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$ and $\mathrm{EVaR}_1[\tilde{x}] = \mathbb{E}[\tilde{x}]$, is defined for $\alpha \in (0, 1]$ as (Ahmadi-Javid, 2012)

$$\mathrm{EVaR}_\alpha[\tilde{x}] = \sup_{\beta > 0} \frac{1}{\beta}\left(-\log \alpha^{-1}\mathbb{E}\big[\exp\left(-\beta\tilde{x}\right)\big]\right) = \inf_{\boldsymbol{\xi} \in \Delta_m : \boldsymbol{\xi} \ll \boldsymbol{q}}\left\{\boldsymbol{\xi}^T \boldsymbol{x} \mid \mathrm{KL}(\boldsymbol{\xi} \| \boldsymbol{q}) \le -\log \alpha\right\}, \tag{3}$$

where KL is the standard KL-divergence defined for each $\boldsymbol{x}, \boldsymbol{y} \in \Delta_m$ as $\mathrm{KL}(\boldsymbol{x} \| \boldsymbol{y}) = \sum_{\omega \in \Omega} x_\omega \log\left(x_\omega / y_\omega\right)$. This definition is valid only when $\boldsymbol{x}$ is absolutely continuous with respect to $\boldsymbol{y}$, which is denoted as $\boldsymbol{x} \ll \boldsymbol{y}$ and corresponds to $y_\omega = 0 \Rightarrow x_\omega = 0$ for each $\omega \in \Omega$.

Risk Averse MDPs A Markov decision process (MDP) is a sequential decision model that underlies most of RL (Puterman, 2005). We consider finite MDPs with states $\mathcal{S} = \{s_1, \ldots, s_S\}$ and actions $\mathcal{A} = \{a_1, \ldots, a_A\}$. After taking an action in a state, the agent transitions to a next state according to a transition probability function $p \colon \mathcal{S} \times \mathcal{A} \to \Delta_S$ such that $p(s, a, s')$ represents the transition probability from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ after taking $a \in \mathcal{A}$. We use $\boldsymbol{p}_{s,a} = p(s, a, \cdot) \in \Delta_S$ to denote the vector of transition probabilities. The initial state $\tilde{s}_0$ is distributed according to $\hat{\boldsymbol{p}} \in \Delta_S$. To avoid divisions by 0 that are not central to our claims, we assume that $\hat{p}_s > 0$ for each $s \in \mathcal{S}$. Finally, the reward function is $r \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, where $r(s, a, s')$ represents the deterministic reward associated with the transition to s' from s after taking an action a.

The most-general solution to an MDP is a *history-dependent randomized* policy π which maps a sequence of observed states and actions $s^0, a^0, s^1, a^1, \ldots, s^t$ to a distribution over the next action a^t. It is well-known that with risk-neutral objectives, there always exists an optimal stationary (depends only on the last state) deterministic policy (Puterman, 2005). When the objective is risk-averse, like VaR, or CVaR, there may not exist an optimal stationary or deterministic policy. Hence, we use the symbol Π to denote the set of history-dependent randomized policies in the remainder of the paper.

This paper focuses on the *finite-horizon objective* in which the agent aims to compute policies that optimize the sum of rewards over a known horizon T. We further restrict our attention to the objective with horizon $T = 1$. It turns out that having a single time-step is sufficient to derive our counterexamples to existing dynamic programs. Moreover, deriving the decompositions with $T = 1$ makes it possible to avoid technicalities caused by history-dependent policies, which could distract us from the main ideas presented in this work. Our results an be extended to general horizons $T > 1$ and the discounted infinite-horizon objectives using standard techniques (Chow et al., 2015).

With horizon $T = 1$, the set of randomized history-dependent policies is $\Pi = \{\boldsymbol{\pi} \colon \mathcal{S} \to \Delta_A\}$. The symbol $\pi(s, a)$ denotes the probability of an action a in a state s, and $\boldsymbol{\pi}(s) = \pi(s, \cdot) \in \Delta_A$ denotes the A-dimensional vector of action probabilities in a state s. Given a risk measure ψ with a risk level $\alpha \in [0,1]$, the finite-horizon risk-averse value of a policy $\pi \in \Pi$ is computed as

$$v_0^\pi(\alpha) := \psi_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}\left[r(\tilde{s},\tilde{a},\tilde{s}')\right], \tag{4}$$

where the superscript in $\psi_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}$ specifies the distribution of the random action. Throughout the paper, we generally use $\tilde{s}$ to denote the random state at time $t=0$ and $\tilde{s}'$ to denote the random state at time $t=1$. In risk-neutral objectives, when $\psi=\mathbb{E}$, one can use the tower property of the expectation operator and define a value function v_t for each time step t (Puterman, 2005), but this property does not hold in most static risk measures (Hau et al., 2023). The term *policy evaluation* in the remainder of the paper refers to computing the value in (4).

The goal in an MDP is to compute an *optimal* value function and a policy that attains it. In risk-averse MDPs, this goal is formalized as the following risk-averse optimization

$$v_0^\star(\alpha) := \max_{\pi\in\Pi} v_0^\pi(\alpha) = \max_{\pi\in\Pi} \psi_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}\left[r(\tilde{s},\tilde{a},\tilde{s}')\right], \tag{5}$$

with the *optimal policy* $\pi^\star$ being any policy that attains the maximum in (5). As with policy evaluation, when $\psi=\mathbb{E}$, the optimal value function $v_t^\star$ can be defined for each time-step t (Puterman, 2005), but this is impossible in general for common risk measures, like VaR and CVaR. The term *policy optimization* in the remainder of the paper refers to computing the value and the maximizer in (5).

In the remainder of the paper, we study dynamic programming algorithms proposed to solve the policy evaluation problem in (4) and policy optimization problem in (5). In general, these algorithms build on risk-level decomposition (Pflug and Pichler, 2016) of risk measures to define a value function $v_t^\pi(s,\alpha)$ for each time step $t\in[T]$, state $s\in\mathcal{S}$, and risk-level $\alpha\in[0,1]$ (Chow et al., 2015). The value function represents the risk-adjusted sum of rewards that can be obtained if starting in a state $s\in\mathcal{S}$ at time t and a risk level α. For example, one would define the value function as $v_1^\pi(s,\alpha) := \psi_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(s)}[r(s,\tilde{a},\tilde{s}')]$ and compute v_0^π using a Bellman operator T_α^π as $v_0^\pi(\alpha) = (T^\pi v_1^\pi)(\alpha)$. In risk-neutral objectives, the Bellman operator is defined as $(T^\pi(v_1^\pi))(\alpha) = \mathbb{E}[v_1^\pi(\tilde{s},\alpha)]$, with $\alpha\in\{1\}$, but in risk-averse formulations the operator definition is more complex. The remainder of the paper discusses the decompositions and the operator for CVaR, EVaR, and VaR risk measures respectively.

3 CVaR: Decomposition Fails in Policy Optimization

In this section, we show that a common CVaR decomposition proposed in Chow et al. (2015) and used to optimize risk-averse policies is inherently sub-optimal regardless of how closely one discretizes the state space. The following proposition represents one of the key results used to decompose the risk measure in multi-stage decision-making.

Proposition 3.1 (lemma 22 in Pflug and Pichler 2016)**.** *Suppose that $\pi\in\Pi$ and $\tilde{s}\sim\hat{\boldsymbol{p}}$, $\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})$, $\tilde{s}'\sim\boldsymbol{p}_{s,a}$. Then,*

$$\mathrm{CVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\right] \quad = \quad \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \zeta_s\, \mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right], \tag{6}$$

where the state s on the right-hand side is not random and

$$\mathcal{Z}_{\mathrm{C}} \;=\; \{\boldsymbol{\zeta}\in\Delta_S \mid \alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}\}\,. \tag{7}$$

The notation in Proposition 3.1 differs superficially from lemma 22 in Pflug and Pichler (2016). Specifically, our CVaR is defined for rewards rather than costs, the meaning of our α corresponds to $1-\alpha$ in Pflug and Pichler (2016), and we use $\xi_s = z_s\hat{p}_s$ as the optimization variable. We include a simple proof of Proposition 3.1 for completeness in Appendix A.1.

The decomposition in Proposition 3.1 is important because it shows that the CVaR evaluation can be formulated as a dynamic program. The theorem shows that CVaR at time $t=0$ decomposes into a convex combination of CVaR values at time $t=1$. Recursively repeating this process, one can formulate a dynamic program for any finite time horizon T. Because the risk-level at time $t=1$ differs from the level at $t=0$ and depends on the optimal $\boldsymbol{\zeta}$, one must compute CVaR values for all (or many) risk-levels $\alpha\in[0,1]$ at time $t=1$. As a result, the dynamic program includes an additional state variable that represents the current risk level.

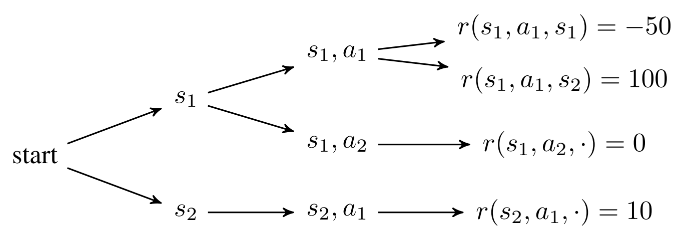

Figure 1: Rewards of MDP M_{C} used in the proof of Theorem 3.2. The dot indicates that the rewards are independent of the next state.

Chow et al. (2015) proposed to adapt the decomposition in Proposition 3.1 to policy optimization as

$$\begin{aligned}\max_{\pi\in\Pi}\ \mathrm{CVaR}_{\alpha}^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] \ &= \ \max_{\pi\in\Pi}\min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\left(\mathrm{CVaR}_{\alpha\xi_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{\pi}(s)}[r(s,\tilde{a},\tilde{s}')\mid\tilde{s}=s]\right) \\ &\overset{??}{=} \ \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\left(\max_{\boldsymbol{d}\in\Delta_{\mathcal{A}}}\mathrm{CVaR}_{\alpha\xi_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{d}}[r(s,\tilde{a},\tilde{s}')\mid\tilde{s}=s]\right).\end{aligned} \tag{8}$$

They used the decomposition in (8) to formulate a dynamic program with the current risk-level as an additional state variable. We prove in the following theorem that the second equality in (8) marked with question marks is false in general.

Theorem 3.2. *There exists an MDP and a risk level $\alpha\in[0,1]$ such that*

$$\max_{\pi\in\Pi}\ \mathrm{CVaR}_{\alpha}^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] < \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\left(\max_{\boldsymbol{d}\in\Delta_{\mathcal{A}}}\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{d}}[r(s,\tilde{a},\tilde{s}')\mid\tilde{s}=s]\right). \tag{9}$$

Before proving Theorem 3.2, we discuss its implications. First, Theorem 3.2 contradicts theorems 5 and 7 in Chow et al. (2015) and shows that their algorithm is inherently sub-optimal regardless of the resolution of the discretization. Theorem 3.2 also contradicts the optimality of the accelerated dynamic program proposed in Stanko and Macek (2019). The result of (Chow et al., 2015) was exploited as is in Chapman et al. (2019), Ding and Feinberg (2022), and Jin et al. (2019) to propose DP reductions, and extended, without proof, in (Rigter et al., 2021) to the context of a Bayesian MDP.

Finally, it is important to emphasize that Theorem 3.2 only applies to the policy optimization setting and does not contradict Proposition 3.1, which holds for the evaluation of policies that assign the same action distribution to each history of states and actions (i.e., policies that are independent of the hypothesized values of $\boldsymbol{\zeta}$).

Proof. Let $\alpha = 0.5$ and consider the MDP M_{C} in Figure 1. In state s_1, both actions a_1 and a_2 are available, and in state s_2, only action a_1 is available. The MDP's rewards are

$$\begin{aligned} &r(s_1,a_1,s_1) = -50, && r(s_1,a_1,s_2) = 100,\\ &r(s_1,a_2,s_1) = r(s_1,a_2,s_2) = 0, && r(s_2,a_1,s_1) = r(s_2,a_1,s_2) = 10\,.\end{aligned}$$

The transition probabilities in M_{C} are

$$p(s_1,a_1,s_1) = 0.4, \qquad\qquad p(s_1,a_1,s_2) = 0.6\,,$$

and the initial distribution is uniform: $\hat{p}_{s_1} = \hat{p}_{s_2} = 0.5$.

To simplify the notation, we define $\theta_\pi\colon \mathcal{Z}_{\mathrm{C}}\to\mathbb{R}$ for each $\pi\in\Pi$ and $\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}$ as

$$\theta_\pi(\boldsymbol{\zeta}) \ = \ \sum_{s\in\mathcal{S}}\zeta_s\,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{\pi}(s)}[r(s,\tilde{a},\tilde{s}')\mid\tilde{s}=s]\,.$$

Because CVaR is convex in the distribution (Delage et al., 2019) and any distribution for $r(\tilde{s},\tilde{a},\tilde{s}')$ obtained from a policy $\pi\in\Pi$ is a mixture of the distributions of $r(\tilde{s},a_1,\tilde{s}')$ and $r(\tilde{s},a_2,\tilde{s}')$, it is sufficient to consider only *deterministic* policies (there exists an optimal deterministic policy). Thus,

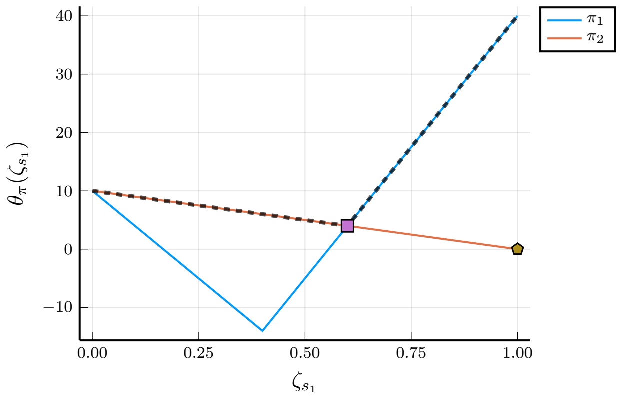

Figure 2: The functions $\theta_{\pi_1}(\cdot)$ and $\theta_{\pi_2}(\cdot)$ used in the CVaR counterexample in the proof of Theorem 3.2. The dashed line shows the function $\zeta_{s_1} \mapsto \max_{\pi\in\{\pi_1,\pi_2\}} \theta_\pi([\zeta_{s_1}, 1-\zeta_{s_1}])$.

we can reformulate the *left-hand side* of (9) in terms of $\theta_\pi(\boldsymbol{\zeta})$ as

$$\begin{aligned}\max_{\boldsymbol{\pi}\in\Pi} \mathrm{CVaR}_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] &= \max_{\pi\in\{\pi_1,\pi_2\}} \mathrm{CVaR}_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] \\ &= \max_{\pi\in\{\pi_1,\pi_2\}} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \zeta_s \cdot \mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{\pi}(s)}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \\ &= \max_{\pi\in\{\pi_1,\pi_2\}} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \theta_\pi(\boldsymbol{\zeta}),\end{aligned}$$

with $\pi_1(s,a_1) = 1-\pi_1(s,a_2) = 1$ and $\pi_2(s,a_2) = 1-\pi_2(s,a_1) = 1$, for all $s\in\mathcal{S}$. The functions $\theta_{\pi_1}(\cdot)$ and $\theta_{\pi_2}(\cdot)$ are depicted in Figure 2. Similarly, the *right-hand side* of (9) can be expressed using the convexity of CVaR in the distribution by algebraic manipulation as

$$\min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \zeta_s \max_{\boldsymbol{d}\in\Delta_A} \mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{d}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] = \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \max_{\pi\in\{\pi_1,\pi_2\}} \theta_\pi(\boldsymbol{\zeta}).$$

Using the notation introduced above and the sufficiency of optimizing over deterministic policies only, the inequality in (9) becomes

$$\max_{\pi\in\{\pi_1,\pi_2\}} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \theta_\pi(\boldsymbol{\zeta}) \;<\; \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \max_{\pi\in\{\pi_1,\pi_2\}} \theta_\pi(\boldsymbol{\zeta}). \tag{10}$$

Figure 2 demonstrates the inequality in (10) numerically, with the rectangle representing the left-hand side maximum and the pentagon representing the right-hand side minimum. The dashed line represents the function $\boldsymbol{\zeta} \mapsto \max_{\pi\in\{\pi_1,\pi_2\}} \theta_\pi(\boldsymbol{\zeta})$.

To show the strict inequality in (10) formally, we evaluate the functions $\theta_{\pi_1}(\cdot)$ and $\theta_{\pi_2}(\cdot)$ for MDP M_{C}. The function $\theta_{\pi_2}(\cdot)$ is linear because the CVaR applies to a constant, and CVaR is translation invariant. The function $\theta_{\pi_1}(\cdot)$ is piecewise-linear and convex, and its slope can be computed using the subgradient that for each $s\in\mathcal{S}$ and $\hat{\boldsymbol{\zeta}}\in\mathcal{Z}_{\mathrm{C}}$ satisfies (Chow et al., 2015)

$$\partial_{\zeta_s}\, \hat{\zeta}_s\, \mathrm{CVaR}_{\alpha\hat{p}_s^{-1}\hat{\zeta}_s}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right] \;\ni\; \mathrm{VaR}_{\alpha\hat{p}_s^{-1}\hat{\zeta}_s}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right].$$

Simple algebraic manipulation then shows that

$$\theta_{\pi_1}(\boldsymbol{\zeta}) \;=\; \max\{10-60\,\zeta_{s_1},\; 90\,\zeta_{s_1}-50\}, \qquad\qquad \theta_{\pi_2}(\boldsymbol{\zeta}) \;=\; 10-10\,\zeta_{s_1},$$

and $\mathcal{Z}_{\mathrm{C}} = \Delta_{\mathcal{S}}$, which implies that $\zeta_{s_1}\in(0,1)$. Therefore, by algebraic manipulation, we get the desired strict inequality

$$0 \;=\; \max_{\pi\in\{\pi_1,\pi_2\}} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \theta_\pi(\boldsymbol{\zeta}) \;<\; \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \max_{\pi\in\{\pi_1,\pi_2\}} \theta_\pi(\boldsymbol{\zeta}) \;=\; 4,$$

where 0 and 4 are represented by the pentagon and rectangle in Figure 2, respectively. □

In summary, the decomposition in Proposition 3.1 cannot be exploited in policy optimization because the inequality in the derivation above may not be tight:

$$
\begin{aligned}
\max_{\pi\in\Pi} \mathrm{CVaR}_{\alpha}^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}') \mid \tilde{s}=s] &= \max_{\pi\in\Pi}\min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \zeta_s \,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \\
&\le \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\max_{\pi\in\Pi} \sum_{s\in\mathcal{S}} \zeta_s \,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \\
&= \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \zeta_s \max_{\boldsymbol{d}\in\Delta_A} \mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}^{\tilde{a}\sim\boldsymbol{d}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s]\,,
\end{aligned}
$$

where the last equality follows from the interchangeability property of optimization and expected value (Shapiro et al., 2014, Theorem 7.92).

It is finally worth noting that we omit to comment on the validity of the CVaR decomposition in Li et al. (2022) given that it considers a different measure than the CVaR defined in Equation (1). Namely their measure takes the form: $\widetilde{\mathrm{CVaR}}_\alpha[\tilde{x}] := \inf_{z\in\mathbb{R}} \left(z+(1-\alpha)^{-1}\mathbb{E}\left[\tilde{x}-z\right]_+\right) = -\mathrm{CVaR}_{1-\alpha}\left[-\tilde{x}\right]$, which is not a coherent risk measure.

4 EVaR: Decomposition Fails for Policy Evaluation

In this section. we show that a decomposition for EVaR proposed in Ni and Lai (2022) is inexact even when considering the policy evaluation setting. Ni and Lai (2022) recently proposed a decomposition of EVaR for a fixed $\pi\in\Pi$ with $\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})$ and a risk level $\alpha\in(0,1]$ as

$$
\mathrm{EVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\right] \overset{??}{=} \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s \,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right]\,, \tag{11}
$$

where $\tilde{s}\sim\hat{\boldsymbol{p}}$, $\tilde{s}'\sim p(\tilde{s},\tilde{a},\cdot)$, and

$$
\mathcal{Z}_{\mathrm{E}} = \left\{\boldsymbol{\xi}\in\Delta_{\mathcal{S}} \mid \sum_{s\in\mathcal{S}} \xi_s \log(\xi_s/\hat{p}_s) \le -\log\alpha\,,\quad \overbrace{\alpha\cdot\boldsymbol{\xi}\le\hat{\boldsymbol{p}}}^{\text{implicit in Ni and Lai (2022)}} \right\}. \tag{12}
$$

Note that we use variables $\xi_s = z_s\hat{p}_s$ in comparison with z_s in Ni and Lai (2022).

The constraint $\alpha\cdot\boldsymbol{\xi}\le\hat{\boldsymbol{p}}$ in (12) was not stated explicitly in Ni and Lai (2022) but is necessary because $\mathrm{EVaR}_{\alpha'}\left[\cdot\right]$ is defined only for $\alpha'\in[0,1]$. When $\alpha'=\alpha\xi_s\hat{p}_s^{-1}$ in (11) it must also satisfy for each $s\in\mathcal{S}$ that

$$
\alpha'\le 1 \quad\Leftrightarrow\quad \alpha\xi_s\hat{p}_s^{-1}\le 1 \quad\Leftrightarrow\quad \alpha\cdot\xi_s\le\hat{p}_s\,.
$$

This additional constraint on $\boldsymbol{\xi}$ implies that $\mathcal{Z}_{\mathrm{E}}\subseteq\mathcal{Z}_{\mathrm{C}}$, for the $\mathcal{Z}_{\mathrm{C}}$ defined in (7).

We claim in the following theorem (see Appendix A.2 for a proof) that the equality in (11) does not hold even in the policy evaluation setting.

Theorem 4.1. *There exists an MDP with a single action and $\alpha\in(0,1]$ such that*

$$
\mathrm{EVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right] < \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s \,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,a_1,\tilde{s}') \mid \tilde{s}=s\right]\,, \tag{13}
$$

the set $\mathcal{Z}_{\mathrm{E}}$ defined by (12)*.*

Theorem 4.1 demonstrates a stronger failure mode than the one in Theorem 3.2 (for CVaR policy optimization), since it applies to both policy evaluation and policy optimization settings.

We propose a correct decomposition of EVaR in the following theorem and employ it to establish that the decomposition in (11) overestimates the actual value of EVaR (see Appendix A.3 for a proof).

Theorem 4.2. *Given any finite MDP with horizon $T=1$ and $\alpha\in(0,1]$, we have that*

$$
\mathrm{EVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\right] = \inf_{\boldsymbol{\zeta}\in(0,1]^S,\,\boldsymbol{\xi}\in\mathcal{Z}'_{\mathrm{E}}(\boldsymbol{\zeta})} \sum_{s\in\mathcal{S}} \xi_s \,\mathrm{EVaR}_{\zeta_s}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right]\,,
$$

where

$$
\mathcal{Z}'_{\mathrm{E}}(\boldsymbol{\zeta}) = \left\{\boldsymbol{\xi}\in\Delta_S \mid \boldsymbol{\xi}\ll\hat{\boldsymbol{p}},\ \sum_{s\in\mathcal{S}} \xi_s\big(\log(\xi_s/\hat{p}_s)-\log(\zeta_s)\big)\le-\log\alpha\right\}.
$$

Moreover, EVaR can be upper-bounded as

$$
\mathrm{EVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\right] \le \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s \,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right]. \tag{14}
$$

5 VaR: Decomposition Holds for Policy Evaluation and Optimization

In this section, we discuss a dynamic program decomposition for VaR whose decomposition resembles those for CVaR and EVaR described in Sections 3 and 4. We provide a new proof of the VaR decomposition to elucidate the differences that make it optimal in contrast to CVaR and EVaR decompositions. Our VaR decomposition closely resembles the quantile MDP approach in Li et al. (2022) with a few technical modifications that can significantly impact the computed value.

To contrast the typical definition of VaR with the quantile definition in Li et al. (2022), it is helpful to summarize how VaR is related to the quantile of a random variable. Let $\mathfrak{q} \in \mathbb{R}$ define as the α-*quantile* of $\tilde{x} \in \mathbb{X}$ when

$$\mathbb{P}\left[\tilde{x} \leq \mathfrak{q}\right] \geq \alpha \quad \text{and} \quad \mathbb{P}\left[\tilde{x} < \mathfrak{q}\right] \leq \alpha\,. \tag{15}$$

In general, the set of quantiles is an interval $[\mathfrak{q}^-_{\tilde{x}}(\alpha), \mathfrak{q}^+_{\tilde{x}}(\alpha)]$ with the bounds computed as (Follmer and Schied, 2016, appendix A.3)

$$\begin{aligned}
\mathfrak{q}^-_{\tilde{x}}(\alpha) &= \sup\left\{z \mid \mathbb{P}\left[\tilde{x} < z\right] < \alpha\right\} = \inf\left\{z \mid \mathbb{P}\left[\tilde{x} \leq z\right] \geq \alpha\right\}\\
\mathfrak{q}^+_{\tilde{x}}(\alpha) &= \inf\left\{z \mid \mathbb{P}\left[\tilde{x} \leq z\right] > \alpha\right\} = \sup\left\{z \mid \mathbb{P}\left[\tilde{x} < z\right] \leq \alpha\right\}.
\end{aligned}$$

Note that when the distribution of $\tilde{x}$ is absolutely continuous (atomless), then $\mathfrak{q}^+_{\tilde{x}}(\alpha) = \mathfrak{q}^-_{\tilde{x}}(\alpha)$ and the quantile is unique. The following example illustrates a simple setting in which the quantile is not unique.

Example 1 (Bernoulli random variable). Consider a Bernoulli random variable $\tilde{e}$ such that $\tilde{e} = 1$ and $\tilde{e} = 0$ with equal (50%) probabilities. Then, any value $q \in [0, 1]$ is a valid 0.5-quantile because

$$\begin{aligned}
\mathfrak{q}^-_{\tilde{e}}(0.5) &= \inf_{z\in\mathbb{R}} \left\{z \mid \mathbb{P}\left[\tilde{e} \leq z\right] \geq 0.5\right\} = \inf_{z\in\mathbb{R}} \left\{z \mid z \geq 0\right\} = 0\\
\mathfrak{q}^+_{\tilde{e}}(0.5) &= \sup_{z\in\mathbb{R}} \left\{z \mid \mathbb{P}\left[\tilde{e} \geq z\right] \geq 0.5\right\} = \sup_{z\in\mathbb{R}} \left\{z \mid z \leq 1\right\} = 1.
\end{aligned}$$

The objective in Li et al. (2022) is to maximize the quantile operator $Q_\alpha \colon \mathbb{X} \to \mathbb{R}$ defined for a reward random variable $\tilde{x} \in \mathbb{X}$ and a risk level $\alpha \in [0, 1]$ as

$$Q_\alpha(\tilde{x}) \;=\; \inf_{z\in\mathbb{R}} \left\{z \mid \mathbb{P}\left[\tilde{x} \leq z\right] \geq \alpha\right\}\,. \tag{16}$$

The quantile operator Q_α and VaR differ in which quantile of the random variable they consider:

$$Q_\alpha(\tilde{x}) \;=\; \mathfrak{q}^-_{\tilde{x}}(\alpha), \qquad \text{but} \qquad \mathrm{VaR}_\alpha\left[\tilde{x}\right] \;=\; \mathfrak{q}^+_{\tilde{x}}(\alpha)\,. \tag{17}$$

As a result, the quantile MDP objective in (16) coincides with the VaR value only when the quantile is unique, which is not always the case, as shown in Example 1.

Theorem 5.1. *Let $\tilde{y} \colon \Omega \to [N]$ be a random variable distributed as $\hat{\boldsymbol{p}} = (\hat{p}_i)_{i=1}^N$ with $\hat{p}_i > 0$. Then for any random variable $\tilde{x} \in \mathbb{X}$, we have*

$$\mathrm{VaR}_\alpha\left[\tilde{x}\right] \quad = \quad \sup_{\boldsymbol{\zeta}\in\Delta_N} \left\{\min_i \; \mathrm{VaR}_{\alpha\zeta_i\hat{p}_i^{-1}}\left[\tilde{x} \mid \tilde{y} = i\right] \mid \alpha\cdot\boldsymbol{\zeta} \leq \hat{\boldsymbol{p}}\right\}\,, \tag{18}$$

where we interpret the minimum to evaluate to ∞ if all terms are infinite, which only occurs if $\alpha = 1$.

Proof. We first decompose VaR using the definition in (1) as

$$\begin{aligned}
\mathrm{VaR}_\alpha\left[\tilde{x}\right] = \sup_{z\in\mathbb{R}} \left\{z \mid \mathbb{P}\left[\tilde{x} < z\right] \leq \alpha\right\} &\overset{\text{(a)}}{=} \sup_{z\in\mathbb{R}} \left\{z \mid \sum_{i=1}^N \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right]\hat{p}_i \leq \alpha\right\}\\
&\overset{\text{(b)}}{=} \sup_{z\in\mathbb{R},\; \boldsymbol{\zeta}\in[0,1]^N} \left\{z \mid \sum_{i=1}^N \zeta_i\hat{p}_i \leq \alpha,\; \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] \leq \zeta_i,\; \forall i \in [N]\right\}\\
&\overset{\text{(c)}}{=} \sup_{z\in\mathbb{R},\; \boldsymbol{\zeta}\in[0,1]^N} \left\{z \mid z \leq \mathrm{VaR}_{\zeta_i}\left[\tilde{x} \mid \tilde{y} = i\right],\; \forall i \in [N],\; \sum_{i=1}^N \zeta_i\hat{p}_i \leq \alpha\right\}\\
&\overset{\text{(d)}}{=} \sup_{\boldsymbol{\zeta}\in[0,1]^N} \left\{\sup_{z\in\mathbb{R}} \left\{z \mid z \leq \mathrm{VaR}_{\zeta_i}\left[\tilde{x} \mid \tilde{y} = i\right], \forall i \in [N]\right\} \mid \sum_{i=1}^N \zeta_i\hat{p}_i \leq \alpha\right\}\\
&\overset{\text{(e)}}{=} \sup_{\boldsymbol{\zeta}\in[0,1]^N} \left\{\min_{i\in[N]} \; \mathrm{VaR}_{\zeta_i}\left[\tilde{x} \mid \tilde{y} = i\right] \mid \sum_{i=1}^N \zeta_i\hat{p}_i \leq \alpha\right\}.
\end{aligned}$$

We decompose the probability $\mathbb{P}[\tilde{x} < z]$ in terms of the conditional probabilities $\mathbb{P}[\tilde{x} < z \mid \tilde{y} = i]$ in step (a) and then lower-bound them by an auxiliary variable ζ_i in step (b). In step (c), we exploit the following equivalence:

$$\mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] \le \zeta_i \quad \Leftrightarrow \quad z \le \mathrm{VaR}_{\zeta_i}[\tilde{x} \mid \tilde{y} = i]$$

The direction $\Leftarrow$ in the above equivalence follows immediately from the fact that $\mathrm{VaR}_{\zeta_i}[\tilde{x} \mid \tilde{y} = i]$ is a ζ_i-quantile and satisfies (15), i.e.,

$$z \le \mathrm{VaR}_{\zeta_i}[\tilde{x} \mid \tilde{y} = i] = q^+_{\tilde{x}}(\zeta_i \mid \tilde{y} = i) \Rightarrow \mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] \le \mathbb{P}\left[\tilde{x} < q^+_{\tilde{x}}(\zeta_i \mid \tilde{y} = i) \mid \tilde{y} = i\right] \le \zeta_i.$$

The direction $\Rightarrow$ follows from the definition of VaR (see Eq. 1), which implies that VaR upper-bounds any z that satisfies the left-hand condition:

$$\mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] \le \zeta_i \;\Rightarrow\; \mathrm{VaR}_{\zeta_i}[\tilde{x} \mid \tilde{y} = i] = \sup\,\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] \le \zeta_i\} \ge z.$$

In step (e), we solve for z. Finally, the form in (18) follows by replacing each ζ_i by $\alpha\zeta_i\hat{p}_i^{-1}$. □

Focusing on the finite MDP with horizon $T = 1$, we can show that the decomposition proposed in Theorem 5.1 is amenable to policy optimization. The main difference between the VaR decomposition and CVaR is that the former VaR was expressed as a supremum instead of an infimum over quantile levels $\boldsymbol{\zeta}$. For VaR, changing the order of maximum (π) and supremum ($\boldsymbol{\zeta}$) does not suffer from a potential gap, but changing the order of maximum (π) and infimum/minimum ($\boldsymbol{\zeta}$) in CVaR does suffer from such a gap as shown in Theorem 3.2.

The following theorem (proved in Appendix A.4) summarizes the decomposition for VaR.

Theorem 5.2. *Given any finite MDP with horizon $T = 1$ and $\alpha \in [0, 1]$, we have*

$$\max_{\pi \in \Pi} \mathrm{VaR}^{\tilde{a} \sim \boldsymbol{\pi}(\tilde{s})}_{\alpha}[r(\tilde{s}, \tilde{a}, \tilde{s}')] = \sup_{\boldsymbol{\zeta} \in \Delta_S} \left\{ \min_{s \in \mathcal{S}} \max_{\boldsymbol{d} \in \Delta_A} \mathrm{VaR}^{\tilde{a} \sim \boldsymbol{d}}_{\alpha\zeta_s\hat{p}_s^{-1}} \left[r(s, \tilde{a}, \tilde{s}') \mid \tilde{s} = s\right] \mid \alpha \cdot \boldsymbol{\zeta} \le \hat{\boldsymbol{p}} \right\}.$$

For completeness, we also present the valid decomposition for the lower quantile MDP (see Appendix A.5 for a proof).

Proposition 5.3. *Given any finite MDP with horizon $T = 1$ and some $\alpha \in [0, 1]$, we have that:*

$$\max_{\pi \in \Pi} Q^{\tilde{a} \sim \boldsymbol{\pi}(\tilde{s})}_{\alpha}(r(\tilde{s}, \tilde{a}, \tilde{s}')) = \sup_{\boldsymbol{\zeta} \in [0,1]^S} \left\{ \min_{s \in \mathcal{S}: \zeta_s < 1} \max_{\boldsymbol{d} \in \Delta_A} Q^{\tilde{a} \sim \boldsymbol{d}}_{\zeta_s}(r(s, \tilde{a}, \tilde{s}') \mid \tilde{s} = s) \mid \sum_{s=1}^{S} \zeta_s \hat{p}_s < \alpha \right\}.$$

We note that the difference with the result presented in (Li et al., 2022) resides in the constraint imposed on $\boldsymbol{\zeta}$ that replaces the weak inequality with a strict one. In fact, this strict versus weak inequality is the main distinguishing factor between the decompositions for the lower and upper quantile.

6 Conclusion

This paper shows that a popular decomposition approach to solving MDPs with CVaR and EVaR objectives is suboptimal despite the claims to the contrary. This suboptimality arises from a saddle-point gap when *optimizing policy*. We also prove that a similar decomposition approach is optimal for policy optimization and evaluation when solving MDPs with the VaR objective. The decomposition is optimal because VaR does not involve the same saddle point problem as CVaR and EVaR.

Our findings are significant because practitioners who make risk-averse decisions in high-stakes scenarios need to have confidence in the correctness of the algorithms they use. Our work raises awareness that popular static CVaR and EVaR MDP algorithms are suboptimal, and their analyses are inaccurate. We hope the results we present in our paper will increase the scrutiny of dynamic programming methods for risk-averse MDPs and motivate research into alternative approaches, such as the parametric dynamic programs.

References

Mohamadreza Ahmadi, Xiaobin Xiong, and Aaron D Ames. Risk-averse control via CVaR barrier functions: Application to bipedal robot locomotion. *IEEE Control Systems Letters*, 6:878–883, 2021.

A. Ahmadi-Javid. Entropic Value-at-Risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

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

Basel Committee on Banking Supervision. Minimum capital requirements for market risk. In *Basel III: international regulatory framework for banks*, 2019.

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

Kang Boda, Jerzy A Filar, Yuanlie Lin, and Lieneke Spanjers. Stochastic target hitting time and the problem of early retirement. *IEEE Transactions on Automatic Control*, 49(3):409–419, 2004.

Margaret P Chapman, Jonathan Lacotte, Aviv Tamar, Donggun Lee, Kevin M Smith, Victoria Cheng, Jaime F Fisac, Susmit Jha, Marco Pavone, and Claire J Tomlin. A risk-sensitive finite-time reachability approach for safety of stochastic dynamic systems. In *2019 American Control Conference (ACC)*, pages 2958–2963. IEEE, 2019.

Margaret P Chapman, Riccardo Bonalli, Kevin M Smith, Insoon Yang, Marco Pavone, and Claire J Tomlin. Risk-sensitive safety analysis using conditional value-at-risk. *IEEE Transactions on Automatic Control*, 67(12):6521–6536, 2022.

Arnav Choudhry, Brady Moon, Jay Patrikar, Constantine Samaras, and Sebastian Scherer. CVaR-based flight energy risk assessment for multirotor uavs using a deep energy model. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 262–268. IEEE, 2021.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Neural Information Processing Systems (NIPS)*, 2014.

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

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18: 1–51, 2018.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2$^{\text{nd}}$ edition, 2006.

Erick Delage, Daniel Kuhn, and Wolfram Wiesemann. "Dice"-sion-making under uncertainty: When can a random decision reduce risk? *Management Science*, 65(7):3282–3301, 2019.

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

Jerzy A. Filar, Dmitry Krass, and Keith W. Ross. Percentile Performance Criteria For Limiting Average Markov Decision Processes. *IEEE Transactions on Automatic Control*, 40(1):2–10, 1995.

Hans Follmer and Alexander Schied. *Stochastic Finance: Introduction in Discrete Time*. De Gruyter Graduate, fourth edition, 2016.

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

Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted MDPs. In *Artificial Intelligence and Statistics (AISTATS)*, 2023.

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

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a CVaR policy. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4436–4443, 2020.

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

Xiaocheng Li, Huaiyang Zhong, and Margaret L. Brandeau. Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447, 2022.

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

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

Xinyi Ni and Lifeng Lai. Risk-sensitive reinforcement learning via Entropic-VaR optimization. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 953–959. IEEE, 2022.

Georg Ch Pflug and Alois Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. *Mathematics of Operations Research*, 41(2):682–699, 2016.

L. A Prashanth and Michael Fu. Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends in Machine Learning*, 15(5):537–693, 2022.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse Bayes-adaptive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1142–1154, 2021.

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

Alexander Shapiro. Interchangeability principle and dynamic equations in risk averse stochastic programming. *Operations Research Letters*, 45(4):377–381, 2017.

Vishnu D Sharma, Maymoonah Toubeh, Lifeng Zhou, and Pratap Tokekar. Risk-aware planning and assignment for ground vehicles using uncertain perception from aerial vehicles. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11763–11769. IEEE, 2020.

Silvestr Stanko and Karel Macek. Risk-averse distributional reinforcement learning: A CVaR optimization approach. In *International Joint Conference on Computational Intelligence*, pages 412–423, 2019.

Congbin Wu and Yuanlie Lin. Minimizing risk models in Markov decision processes with policies depending on target values. *Journal of mathematical analysis and applications*, 231(1):47–67, 1999.

Huan Xu and Shie Mannor. Probabilistic goal Markov decision processes. In *International Joint Conference on Artificial Intelligence*, 2011.

Huaiyang Zhong. *Decision Making for Disease Treatment: Operations Research and Data Analytic Modeling*. Stanford University, 2020.

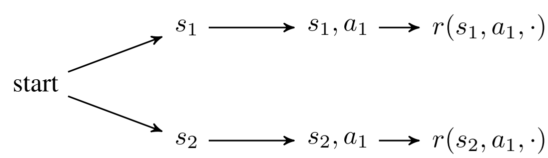

Figure 3: Rewards of the MDP M_{E} used in the proof of Theorem 4.1. The dot indicates that the rewards are independent of the next state.

A Proofs

A.1 Proof of Proposition 3.1

Suppose that $\alpha > 0$; the decomposition for $\alpha = 0$ holds readily because $\mathrm{CVaR}_0[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$.

To streamline the notation, Define a random variable $\tilde{x} = r(\tilde{s}, \tilde{a}, \tilde{s}')$ over a random space $\Omega = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ with a probability distribution $\boldsymbol{q} \in \Delta_m$ such that $q_{s,a,s'} = \hat{p}_s \cdot \pi(s,a) \cdot p(s,a,s')$. The value $\boldsymbol{x}$ is the vector representation of the random variable $\tilde{x}$ and $\boldsymbol{\xi}_s = \boldsymbol{\xi}_{s,\cdot,\cdot} \in \mathbb{R}^{S\cdot A}$ for $\boldsymbol{\xi} \in \mathbb{R}^m$ is a vector that corresponds to the subset of the elements of Ω in which the first element is some $s \in \mathcal{S}$. The vectors $\boldsymbol{x}_s = \boldsymbol{x}_{s,\cdot,\cdot} \in \mathbb{R}^{S\cdot A}$ and $\boldsymbol{q}_s = \boldsymbol{q}_{s,\cdot,\cdot} \in \mathbb{R}^{S\cdot A}$ are defined analogously to $\boldsymbol{\xi}_s$.

Starting with the CVaR definition in (2) and introducing an auxiliary variable $\boldsymbol{\zeta}$ we get that

$$
\begin{aligned}
\mathrm{CVaR}_\alpha[\tilde{x}] = \min_{\boldsymbol{\xi}\in\Delta_m}\left\{\boldsymbol{x}^\top\boldsymbol{\xi} \mid \alpha\boldsymbol{\xi} \le \boldsymbol{q}\right\} &= \min_{\boldsymbol{\xi}\in\Delta_m,\boldsymbol{\zeta}\in\mathbb{R}^S}\left\{\boldsymbol{x}^\top\boldsymbol{\xi} \mid \alpha\boldsymbol{\xi} \le \boldsymbol{q},\, \zeta_s = \mathbf{1}^\top\boldsymbol{\xi}_s, \forall s \in \mathcal{S}\right\} \\
&= \min_{\boldsymbol{\xi}\in\Delta_m,\boldsymbol{\zeta}\in\Delta_S}\left\{\boldsymbol{x}^\top\boldsymbol{\xi} \mid \alpha\boldsymbol{\xi} \le \boldsymbol{q},\, \zeta_s = \mathbf{1}^\top\boldsymbol{\xi}_s,\, \alpha\zeta_s \le \hat{p}_s,\, \forall s \in \mathcal{S}\right\} \\
&= \min_{\boldsymbol{\xi}\in\mathbb{R}^\Omega_+,\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\left\{\boldsymbol{x}^\top\boldsymbol{\xi} \mid \alpha\boldsymbol{\xi} \le \boldsymbol{q},\, \zeta_s = \mathbf{1}^\top\boldsymbol{\xi}_s,\, \forall s \in \mathcal{S}\right\} .
\end{aligned}
$$

In the derivation above, we replaced the infimum by a minimum because Ω is finite, introduced a new variable $\boldsymbol{\zeta}$, derived implied constraints on $\boldsymbol{\zeta}$, and then dropped superfluous constraints on $\boldsymbol{\xi}$. Continuing with the derivation above and noticing that the constraints on each $\boldsymbol{\xi}_s$ are independent given $\boldsymbol{\zeta}$, we get that

$$
\begin{aligned}
\mathrm{CVaR}_\alpha[\tilde{x}] &= \min_{\boldsymbol{\xi}\in\mathbb{R}^\Omega_+,\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\left\{\sum_{s\in\mathcal{S}}\boldsymbol{x}_s^\top\boldsymbol{\xi}_s \mid \alpha\boldsymbol{\xi} \le \boldsymbol{q},\, \zeta_s = \mathbf{1}^\top\boldsymbol{\xi}_s,\, \forall s \in \mathcal{S}\right\} \\
&\overset{\text{(a)}}{=} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\inf_{\boldsymbol{\xi}_s\in\mathbb{R}^{\Omega_s}_+}\left\{\boldsymbol{x}_s^\top\boldsymbol{\xi}_s \mid \alpha\boldsymbol{\xi}_s \le \boldsymbol{q}_s,\, \zeta_s = \mathbf{1}^\top\boldsymbol{\xi}_s\right\} \\
&\overset{\text{(b)}}{=} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\cdot\min_{\boldsymbol{\chi}\in\Delta_{S\cdot A}}\left\{\boldsymbol{x}_s^\top\boldsymbol{\chi} \mid \alpha\hat{p}_s^{-1}\zeta_s\chi_{a,s'} \le \hat{p}_s^{-1}q_{s,a,s'}, \forall a \in \mathcal{A}, s' \in \mathcal{S}\right\} \\
&\overset{\text{(c)}}{=} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\cdot\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[\tilde{x} \mid \tilde{s} = s] .
\end{aligned}
$$

The step (a) follows from the interchangeability principle (Shapiro et al., 2014, theorem 7.92), and the step (b) follows by substituting $\xi_{s,a,s'} = \zeta_s\chi_{a,s'}$ taking care when $\zeta_s = 0$ and multiplying both sides of the inequality by $\hat{p}_s^{-1} > 0$. Finally, in step (c), the random variable $\tilde{x} = r(\tilde{s}, \tilde{a}, \tilde{s}')$ conditional on $\tilde{s} = s$ is distributed according to $q_{s,a,s'}\hat{p}_s^{-1}$ and the equality follows from the definition of CVaR in (2). □

A.2 Proof of Theorem 4.1

Consider an MDP M_{E} depicted in Figure 3 with $\mathcal{S} = \{s_1, s_2\}$ and $\mathcal{A} = \{a_1\}$ and a reward function $r(s_1, a_1, \cdot) = 1$ and $r(s_2, a_1, \cdot) = 0$. We abbreviate the rewards to $r(s_1)$ and $r(s_2)$ because they only depend on the originating state. The initial distribution is $\hat{p}_{s_1} = \hat{p}_{s_2} = 0.5$. We finally let $\alpha = 0.75$.

Because $\mathcal{Z}_{\mathrm{E}} \subseteq \mathcal{Z}_{\mathrm{C}}$, the right-hand side of (13) can be lower-bounded by CVaR as

$$
\begin{aligned}
\min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}}\sum_{s\in\mathcal{S}}\xi_s\,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}[r(s,a_1,\tilde{s}')] &= \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}}\sum_{s\in\mathcal{S}}\xi_s r(s) \\
&\ge \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\xi_s r(s) = \mathrm{CVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] .
\end{aligned} \tag{19}
$$

The first equality holds from the positive homogeneity and cash invariance properties of EVaR, and the last equality follows from the dual representation of CVaR (Follmer and Schied, 2016).

Because $\mathrm{EVaR}_\alpha[\tilde{x}] \le \mathrm{CVaR}_\alpha[\tilde{x}]$ for each $\alpha \in [0,1]$ and $\tilde{x} \in \mathbb{X}$ (see (Ahmadi-Javid, 2012, Proposition 3.2)), we can further lower-bound (19) as

$$\mathrm{EVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] \;\le\; \mathrm{CVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] \;\le\; \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}[r(s,a_1,\tilde{s}')] \,. \tag{20}$$

Therefore, (13) holds with an inequality.

To prove by contradiction that the inequality in (13) is strict, suppose that

$$\mathrm{EVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] \;=\; \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}[r(s,a_1,\tilde{s}')] \,. \tag{21}$$

Equalities (21) and (20) imply that $\mathrm{EVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] = \mathrm{CVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')]$ which is false in general (Ahmadi-Javid, 2012).

We now show that EVaR does not equal CVaR even for the categorical distribution of $\tilde{s}$. The CVaR of the return in M_{E} reduces from (2) to

$$\mathrm{CVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] = \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \xi_s r(s) = \max\left\{0,\, \frac{\hat{p}_{s_1}+\alpha-1}{\alpha}\right\} . \tag{22}$$

Since $1-\alpha = 0.25 < 0.5 = \hat{p}_{s_1}$, then the optimal $\boldsymbol{\xi}^\star$ in (22) is

$$\boldsymbol{\xi}^\star = \begin{pmatrix} \frac{\hat{p}_{s_1}+\alpha-1}{\alpha} \\ \frac{1-\hat{p}_{s_1}}{\alpha} \end{pmatrix} .$$

Since $\mathrm{KL}(\boldsymbol{\xi}^\star\|\hat{\boldsymbol{p}}) < -\log\alpha$. we have that $\boldsymbol{\xi}^\star$ is in the relative interior of the EVaR feasible region in (3), and, therefore, there exists an $\epsilon > 0$ such that

$$\mathrm{EVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] = \mathrm{CVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] - \epsilon < \mathrm{CVaR}_\alpha[r(\tilde{s},a_1,\tilde{s}')] \,,$$

which proves the desired inequality. $\square$

A.3 Proof of Theorem 4.2

We start by proposing a new decomposition for EVaR.

Proposition A.1. *Given a random variable $\tilde{x}\in\mathbb{X}$ and a discrete variable $\tilde{y}\colon \Omega \to \mathcal{N} = \{1,\dots,N\}$, with probabilities denoted as $\{\hat{p}_i\}_{i=1}^N$, for any $\alpha\in(0,1]$ we have that*

$$\mathrm{EVaR}_\alpha[\tilde{x}] \quad = \quad \inf_{\boldsymbol{\zeta}\in(0,1]^N} \min_{\boldsymbol{\xi}\in\mathcal{Z}'_{\mathrm{E}}(\boldsymbol{\zeta})} \sum_i \xi_i\,\mathrm{EVaR}_{\zeta_i}[\tilde{x}\mid\tilde{y}=i] \,,$$

where

$$\mathcal{Z}'_{\mathrm{E}}(\boldsymbol{\zeta}) = \left\{\boldsymbol{\xi}\in\Delta_N \mid \boldsymbol{\xi}\ll\hat{\boldsymbol{p}}, \sum_{i=1}^N \xi_i(\log(\xi_i/\hat{p}_i) - \log(\zeta_i)) \le -\log\alpha\right\} .$$

Proof. Let $\boldsymbol{q}$ denote the joint probability distribution of $\tilde{x}$ and $\tilde{y}$. The proof exploits the chain rule of relative entropy (e.g., Cover and Thomas (2006, theorem 2.5.3)), which states that for any probability distributions $\boldsymbol{\eta},\boldsymbol{q}\in\Delta_\Omega$ with $\boldsymbol{\eta}\ll\boldsymbol{q}$

$$\mathrm{KL}(\boldsymbol{\eta}\|\boldsymbol{q}) = \mathrm{KL}(\boldsymbol{\eta}(\tilde{y})\|\boldsymbol{q}(\tilde{y})) + \mathrm{KL}(\boldsymbol{\eta}(\tilde{x}|\tilde{y})\|\boldsymbol{q}(\tilde{x}|\tilde{y})), \tag{23}$$

where the conditional relative entropy is defined as

$$\mathrm{KL}(\boldsymbol{\eta}(\tilde{x}|\tilde{y})\|\boldsymbol{q}(\tilde{x}|\tilde{y})) = \mathbb{E}^{\boldsymbol{\eta}}\left[\log\frac{\boldsymbol{\eta}(\tilde{x}|\tilde{y})}{\boldsymbol{q}(\tilde{x}|\tilde{y})}\right] .$$

with $\mathbb{E}^{\boldsymbol{\eta}}[f(\tilde{x},\tilde{y})]$ as a shorthand notation to indicate that $(\tilde{x},\tilde{y}) \sim \boldsymbol{\eta}$.We can now decompose EVaR from its definition in (3) as

$$
\begin{aligned}
\mathrm{EVaR}_\alpha[\tilde{x}] &= \inf_{\boldsymbol{\eta}\in\Delta_{m\cdot N}:\boldsymbol{\eta}\ll\boldsymbol{q}} \left\{\mathbb{E}^{\boldsymbol{\eta}}[\tilde{x}] \mid \mathrm{KL}(\boldsymbol{\eta} \mid \boldsymbol{q}) \le -\log\alpha\right\} \\
&\overset{\text{(a)}}{=} \inf_{\boldsymbol{\eta}\in\Delta_{m\cdot N}:\boldsymbol{\eta}\ll\boldsymbol{q}} \left\{\mathbb{E}^{\boldsymbol{\eta}}[\tilde{x}] \mid \mathrm{KL}(\boldsymbol{\eta}(\tilde{y})\|\boldsymbol{q}(\tilde{y})) + \mathrm{KL}(\boldsymbol{\eta}(\tilde{x}|\tilde{y})\|\boldsymbol{q}(\tilde{x}|\tilde{y})) \le -\log\alpha\right\} \\
&= \inf_{\boldsymbol{\eta}\in\Delta_{m\cdot N}:\boldsymbol{\eta}\ll\boldsymbol{q}} \left\{\mathbb{E}^{\boldsymbol{\eta}}[\tilde{x}] \mid \mathrm{KL}(\boldsymbol{\eta}(\tilde{y})\|\boldsymbol{q}(\tilde{y})) + \mathbb{E}^{\boldsymbol{\eta}}\left[\mathbb{E}^{\boldsymbol{\eta}}\left[\log\frac{\boldsymbol{\eta}(\tilde{x}|\tilde{y})}{\boldsymbol{q}(\tilde{x}|\tilde{y})}\right] \mid \tilde{y}\right] \le -\log\alpha\right\} \\
&\overset{\text{(b)}}{=} \inf_{\boldsymbol{\eta}\in\Delta_{m\cdot N},\boldsymbol{\zeta}\in(0,1]^N:\boldsymbol{\eta}\ll\boldsymbol{q}} \left\{\mathbb{E}^{\boldsymbol{\eta}}[\mathbb{E}^{\boldsymbol{\eta}}[\tilde{x} \mid \tilde{y}]] \mid \begin{array}{c}\mathrm{KL}(\boldsymbol{\eta}(\tilde{y})\|\boldsymbol{q}(\tilde{y})) + \mathbb{E}^{\boldsymbol{\eta}}[-\log(\zeta_{\tilde{y}})] \le -\log\alpha \\ \mathbb{P}_{\boldsymbol{\eta}}\left[\mathbb{E}^{\boldsymbol{\eta}}\left[\log(\boldsymbol{\eta}(\tilde{x}|\tilde{y})/\boldsymbol{q}(\tilde{x}|\tilde{y})) \mid \tilde{y}\right] \le -\log(\zeta_{\tilde{y}})\right] = 1\end{array}\right\} \\
&\overset{\text{(c)}}{=} \inf_{\boldsymbol{\xi}\in\Delta_N,\boldsymbol{\zeta}\in(0,1]^N:\boldsymbol{\xi}\ll\hat{\boldsymbol{p}}} \left\{\mathbb{E}^{\boldsymbol{\xi}}[\mathrm{EVaR}_{\zeta_{\tilde{y}}}[\tilde{x}|\tilde{y}]] \mid \mathrm{KL}(\boldsymbol{\xi}\|\hat{\boldsymbol{p}}) + \mathbb{E}^{\boldsymbol{\xi}}[-\log(\zeta_{\tilde{y}})] \le -\log\alpha\right\} \\
&= \inf_{\boldsymbol{\xi}\in\Delta_N,\boldsymbol{\zeta}\in(0,1]^N:\boldsymbol{\xi}\ll\hat{\boldsymbol{p}}} \left\{\sum_i \xi_i\,\mathrm{EVaR}_{\zeta_i}[\tilde{x} \mid \tilde{y}=i] \mid \sum_{i=1}^N \xi_i\log(\xi_i/\hat{p}_i) - \sum_{i=1}^N \xi_i\log(\zeta_i) \le -\log\alpha\right\}.
\end{aligned}
$$

Here, we decompose the relative entropy of $\boldsymbol{\eta}$ and $\boldsymbol{q}$ using (23) in step (a) and then use the tower property of the expectation operator in the next step. In step (b), we introduce a variable ζ_i for each realization of $\tilde{y}=i$ with $i\in\mathcal{N}$ to decouple the influence of $\boldsymbol{\eta}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}})$, under each $\tilde{y}$, in the inequality constraint. Finally, we replace the conditional EVaR definition by solving for $\boldsymbol{\eta}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}})$ for a given $\boldsymbol{\zeta}$ in step (c), and representing $\boldsymbol{\eta}(\tilde{\boldsymbol{y}})$ using $\boldsymbol{\xi}$. ☐

The first part of our theorem follows directly from Proposition A.1. Suppose that $\alpha > 0$; the result follows for $\alpha = 0$ because $\mathrm{EVaR}_0[\cdot]$ reduces to $\operatorname{ess\,inf}$. Then, the second part of the corollary holds as

$$
\begin{aligned}
\mathrm{EVaR}_\alpha[r(\tilde{s},\tilde{a},\tilde{s}')] &= \inf_{\boldsymbol{\zeta}\in(0,1]^N,\,\boldsymbol{\xi}\in\Delta_N} \left\{\sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\zeta_s}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \mid \sum_{s\in\mathcal{S}} \xi_s\log\frac{\xi_s}{\zeta_s\hat{p}_s} \le -\log\alpha\right\} \\
&\le \inf_{\boldsymbol{\zeta}\in(0,1]^N,\,\boldsymbol{\xi}\in\Delta_N} \left\{\sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\zeta_s}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \mid \sum_{s\in\mathcal{S}} \xi_s\log\frac{\xi_s}{\zeta_s\hat{p}_s} \le -\log\alpha,\ \boldsymbol{\xi}\le\alpha^{-1}\hat{\boldsymbol{p}}\right\} \\
&\le \inf_{\boldsymbol{\xi}\in\Delta_N} \left\{\sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \mid \boldsymbol{\xi}\le\alpha^{-1}\hat{\boldsymbol{p}}\right\}.
\end{aligned}
$$

The first inequality follows from adding a constraint on the pairs on the $\boldsymbol{\xi}$ considered by the infimum. The second inequality follows by fixing $\zeta_s = \hat{\zeta}_s$ with $\hat{\zeta}_s = \alpha\xi_s\hat{p}_s^{-1}$ for each $s\in\mathcal{S}$. This is an upper bound because $\hat{\zeta}_s$ is feasible in the infimum:

$$
\sum_{s\in\mathcal{S}} \xi_s\log\frac{\xi_s}{\hat{\zeta}_s\hat{p}_s} = -\log\alpha \le -\log\alpha\,.
$$

The value $\hat{\zeta}_s$ is well-defined since $\hat{p}_s > 0$ and the constraint $\boldsymbol{\xi}\le\alpha^{-1}\hat{\boldsymbol{p}}$ ensures that $\hat{\zeta}_s \le 1$. Also, we can relax the constraint $\zeta_s > 0 \Rightarrow \xi_s > 0$ to $\xi_s \ge 0$ because $\mathrm{EVaR}_0[\tilde{x}] = \lim_{\alpha\to 0}\mathrm{EVaR}_\alpha[\tilde{x}]$, and, therefore, the infimum is not affected. Finally, the inequality in the corollary follows immediately by further upper bounding the decomposition above by adding a constraint. ☐

A.4 Proof of Theorem 5.2

The equality develops from Theorem 5.1 as

$$
\begin{aligned}
\max_{\pi\in\Pi}\ \mathrm{VaR}_\alpha^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] &= \max_{\pi\in\Pi}\ \sup_{\boldsymbol{\zeta}\in\Delta_S:\alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}}\ \min_{s\in\mathcal{S}} \left(\mathrm{VaR}^{\tilde{a}\sim\boldsymbol{\pi}(s)}_{\alpha\zeta_s\hat{p}_s^{-1}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s]\right) \\
&= \sup_{\boldsymbol{\zeta}\in\Delta_S:\alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}}\ \max_{\pi\in\Pi}\ \min_{s\in\mathcal{S}} \left(\mathrm{VaR}^{\tilde{a}\sim\boldsymbol{\pi}(s)}_{\alpha\zeta_s\hat{p}_s^{-1}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s]\right) \\
&= \sup_{\boldsymbol{\zeta}\in\Delta_S:\alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}}\ \min_{s\in\mathcal{S}} \left(\max_{\boldsymbol{d}\in\Delta_A}\mathrm{VaR}^{\tilde{a}\sim\boldsymbol{d}}_{\alpha\zeta_s\hat{p}_s^{-1}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s]\right),
\end{aligned}
$$

where we first change the order of maximum and supremum, followed by changing the order of $\max_\pi \min_s$ with $\min_s \max_\pi$. The latter is a direct consequence of the interchangeability property of the maximum operation (Shapiro, 2017, Proposition 2.2). $\square$

A.5 Proof of Proposition 5.3

The proof mainly relies on correcting the decomposition of lower quantile proposed in (Li et al., 2022).

Proposition A.2. *Given an $\tilde{x} \in \mathbb{X}$, suppose that a random variable $\tilde{y}\colon \Omega \to \mathcal{N} = \{1, \ldots, N\}$ is distributed as $\hat{\boldsymbol{p}} = (\hat{p}_i)_{i=1}^N$ with $\hat{p}_i > 0$. Then:*

$$Q_\alpha(\tilde{x}) \quad = \quad \sup_{\boldsymbol{\zeta} \in [0,1]^N} \left\{ \min_{i \in \mathcal{N} : \zeta_i < 1} Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i) \mid \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha \right\}, \tag{24}$$

where we interpret the supremum to be minus infinity if its feasible set is empty, which only occurs if $\alpha = 0$.

Proof. First, we decompose lower quantile using its definition as

$$\begin{aligned}
Q_\alpha(\tilde{x}) = \sup\left\{ z \mid \mathbb{P}\left[\tilde{x} < z\right] < \alpha \right\} &\overset{\text{(a)}}{=} \sup_{z \in \mathbb{R}} \left\{ z \mid \sum_{i=1}^N \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] \hat{p}_i < \alpha \right\} \\
&\overset{\text{(b)}}{=} \sup_{z \in \mathbb{R}, \boldsymbol{\zeta} \in [0,1]^N} \left\{ z \mid \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha,\ \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] < \zeta_i, \forall i \in \mathcal{N} : \zeta_i < 1 \right\} \\
&\overset{\text{(c)}}{=} \sup_{z \in \mathbb{R}, \boldsymbol{\zeta} \in [0,1]^N} \left\{ z \mid z < Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i), \forall i \in \mathcal{N} : \zeta_i < 1,\ \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha \right\} \\
&\overset{\text{(d)}}{=} \sup_{\boldsymbol{\zeta} \in [0,1]^N} \left\{ \sup_{z \in \mathbb{R}} \left\{ z \mid z < Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i), \forall i \in \mathcal{N} : \zeta_i < 1 \right\} \mid \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha \right\} \\
&\overset{\text{(e)}}{=} \sup_{\boldsymbol{\zeta} \in [0,1]^N} \left\{ \min_{i \in \mathcal{N} : \zeta_i < 1} Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i) \mid \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha \right\}.
\end{aligned}$$

We decompose the probability $\mathbb{P}\left[\tilde{x} < z\right]$ in terms of the conditional probabilities $\mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right]$ in step (a) and then lower-bound them by an auxiliary variable ζ_i in step (b). In step (c), we exploits the following equivalence:

$$\mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] < \zeta_i \quad \Leftrightarrow \quad z < Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i)$$

The direction $\Leftarrow$ in the equivalence follows from the definition of $Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i)$:

$$z < Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i) = \inf\left\{ z \mid \mathbb{P}\left[\tilde{x} \le z \mid \tilde{y} = i\right] \ge \zeta_i \right\} \ \Rightarrow\ \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] < \zeta_i.$$

The direction $\Rightarrow$ follows from the definition of VaR (see equation (1)), which implies that VaR upper-bounds any z that satisfies the left-hand condition:

$$\mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] < \zeta_i \ \Rightarrow\ Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i) = \sup\ \left\{ z \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] < \zeta_i \right\} \ge z,$$

yet $Q_{\zeta_i}(\tilde{x} \mid \tilde{y} = i) \neq z$ otherwise since $\mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right]$ is right continuous, there must exist some $\epsilon > 0$ for which $\mathbb{P}\left[\tilde{x} < z + \epsilon \mid \tilde{y} = i\right] < \zeta_i$ hence:

$$z = \sup\ \left\{ z \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x} < z \mid \tilde{y} = i\right] < \zeta_i \right\} \ge z + \epsilon > z,$$

which leads to a contradiction. In step (e), we solve for z. Finally, we obtain the form in (24). $\square$

The decomposition proposed in Proposition A.2 can now be used, exactly as was done for the case of VaR, to obtain a decomposition for the risk averse MDP:

$$
\begin{aligned}
\max_{\pi\in\Pi} Q_{\alpha}^{\tilde{a}\sim\boldsymbol{\pi}(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] \;&=\; \max_{\pi\in\Pi} \sup_{\boldsymbol{\zeta}\in[0,1]^S} \left\{ \min_{s\in\mathcal{S}:\zeta_s<1} \; Q_{\zeta_s}^{\tilde{a}\sim\pi(s)}(r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s) \;\middle|\; \sum_{s=1}^{S} \zeta_s \hat{p}_s < \alpha \right\} \\
&= \sup_{\boldsymbol{\zeta}\in[0,1]^S} \max_{\pi\in\Pi} \left\{ \min_{s\in\mathcal{S}:\zeta_s<1} \; Q_{\zeta_s}^{\tilde{a}\sim\pi(s)}(r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s) \;\middle|\; \sum_{s=1}^{S} \zeta_s \hat{p}_s < \alpha \right\} \\
&= \sup_{\boldsymbol{\zeta}\in[0,1]^S} \left\{ \min_{s\in\mathcal{S}:\zeta_s<1} \; \left(\max_{\boldsymbol{d}\in\Delta_A} Q_{\zeta_s}^{\tilde{a}\sim\boldsymbol{d}}(r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s) \;\middle|\; \sum_{s=1}^{S} \zeta_s \hat{p}_s < \alpha \right) \right\}.
\end{aligned}
$$