Les Cahiers du GERAD ISSN: 0711–2440

On dynamic program decompositions of static risk measures

J.L. Hau, E. Delage, M. Ghavamzadeh, M. Petrik

G–2023–13

April 2023

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2023
– Bibliothèque et Archives Canada, 2023

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2023
– Library and Archives Canada, 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

On dynamic program decompositions of static risk measures

Jia Lin Hau [a]

Erick Delage [b]

Mohammad Ghavamzadeh [c]

Marek Petrik [a]

[a] *Department of Computer Science, University of New Hampshire, Durham, NH, USA*

[b] *Google Research, Mountain View, Sunnyvale, San Francisco, CA, USA*

[c] *Department of Decision Sciences, HEC Montréal & GERAD, Montréal (Qc), Canada*

erick.delage@hec.ca

April 2023
Les Cahiers du GERAD
G–2023–13

Abstract : Optimizing static risk-averse objectives in Markov decision processes is challenging because they do not readily admit dynamic programming decompositions. Prior work has proposed to use a dynamic decomposition of risk measures that help to formulate dynamic programs on an augmented state space. This paper shows that several existing decompositions are inherently inexact, contradicting several claims in the literature. In particular, we give examples that show that popular decompositions for CVaR and EVaR risk measures are strict overestimates of the true risk values. However, an exact decomposition is possible for VaR, and we give a simple proof that illustrates the fundamental difference between VaR and CVaR dynamic programming properties.

Acknowledgements: We thank Yinlam Chow for mentioning that there may be a duality gap in the CVaR decomposition.

1 Introduction

Optimizing static risk-averse objectives in Markov decision processes (MDPs) is challenging because they do not readily admit dynamic programming decompositions. Prior work has proposed to use a dynamic decomposition of risk measures that help to formulate dynamic programs on an augmented state space. This paper shows that several existing decompositions are inherently inexact, which contradicts several claims in the literature [4, 10, 11]. In particular, we give examples that show that several popular decompositions for the *conditional value at risk* (CVaR) and *entropic value at risk* (EVaR) strictly overestimate the true values when used in the context of MDPs.

As our first contribution, we construct an example that shows that the CVaR dynamic program in Chow et al. [4] significantly overestimates the optimal value function. This result contradicts the optimality claims in Chow et al. [4] and, likely, the CVaR optimality claims in Li et al. [10]. Because our CVaR counterexample exploits a duality gap, it only applies to *policy optimization* and does not contradict the *policy evaluation* decomposition in Pflug and Pichler [12]. This counterexample is important because the decomposition has been employed widely in the reinforcement learning literature.[1]

As our second contribution, we construct an example that shows that the EVaR dynamic program in Ni and Lai [11] significantly overestimates the value function of a fixed policy. The example contradicts the optimality claims in Ni and Lai [11] and applies both to policy evaluation and optimization.

As our third contribution, we derive a decomposition for VaR. This derivation mirrors the derivation of the quantile MDP decomposition in Li et al. [10], but with several technical differences. In particular, a quantile of an atomic distribution is an interval rather than a unique value. The quantile MDP decomposition in Li et al. [10] is defined for the lower bound of the quantile interval, while VaR is typically defined as the upper bound of the quantile interval [8]. Also, our derivation of the VaR decomposition demonstrates why the decomposition ideas work for VaR but fail for CVaR and EVaR.

Our counterexamples do not affect the correctness of several other risk decomposition techniques. The decomposition proposed for quantile, or *value at risk* (VaR), objective in Li et al. [10] uses a different approach and remains unaffected. The *parametric dynamic programs* proposed for CVaR and EVaR [2, 3, 5, 9] also remain unaffected. The parametric dynamic programs use the *primal* risk measure representation and do not suffer from the same duality gap issues as the augmentation methods that use the *dual* representation of the risk measures, such as [4].

2 Framework

A *monetary risk measure* $\psi\colon \mathbb{X} \to \mathbb{R}$ is a monotone mapping that assigns a real value to each random variable from the set $\mathbb{X}\colon \Omega \to \mathbb{R}$ of real-valued random variables defined over a *finite* probability space Ω with $O = |\Omega|$. Random variables in this paper are adorned with a tilde, such as $\tilde{x} \in \mathbb{X}$. All risk measures in this paper are defined for a random variable $\tilde{x}$ that represents *rewards*.

Because we restrict our attention to finite probability spaces, we can represent any random variable $\tilde{x} \in \mathbb{X}$ as a vector $\boldsymbol{x} \in \mathbb{R}^O$. We also use $\boldsymbol{q} \in \Delta(O)$ to represent a probability distribution over Ω where $\Delta(O)$ represents the O-dimensional simplex. Using this notation, we have that

$$\mathbb{E}[\tilde{x}] \;=\; \boldsymbol{q}^\top \boldsymbol{x}\,.$$

As the decision model, we consider a *finite-horizon MDP* with a horizon $T = 1$. Although we only consider the finite-horizon objective, our results also generalize to discounted infinite-horizon problems. The MDP has two states $\mathcal{S} = \{s_1, s_2\} = \{1, 2\}$ and two actions $\mathcal{A} = \{a_1, a_2\} = \{1, 2\}$. Let $S = S$

[1] 280 citations on Google Scholar at the time of writing

and $A = |\mathcal{A}|$. The initial distribution is $\hat{\boldsymbol{p}} \in \Delta(S)$. We differentiate vectors from scalars using a bold notation.

The transition probability function is $p\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ such that $p(s, a, s')$ represents the transition probability from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ after taking $a \in \mathcal{A}$. 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 reward associated with the transition to s' from s after taking an action a. To avoid divisions by 0 that are not central to our claims, we assume that $\hat{p}_s > 0$ and $p(s, a, s') > 0$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Finally, we use $\boldsymbol{p}_{s,a} = p(s, a, \cdot) \in \Delta(S)$ to denote the vector of transition probabilities.

The solution to an MDP is a policy π from the set $\Pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ of all *history-dependent randomized* policies. Given that the horizon is $T = 1$, all history-dependent policies are also *stationary* (or state-dependent). We use $\pi(s, a)$ to denote the probability of an action a in a state s, and $\boldsymbol{\pi}(s) = \pi(s, \cdot) \in \Delta(A)$ to denote the A-dimensional vector of action probabilities in a state s.

The objective of computing an optimal risk-averse policy in the MDP with a horizon $T = 1$ can be formalized as

$$\max_{\pi \in \Pi} \psi \left[r(\tilde{s}, \tilde{a}, \tilde{s}') \mid \tilde{s} \sim \hat{\boldsymbol{p}}, \tilde{a} \sim \boldsymbol{\pi}(\tilde{s}), \tilde{s}' \sim \boldsymbol{p}_{\tilde{s},\tilde{a}} \right]. \tag{1}$$

We use the term *policy evaluation* to refer to the problem of computing the objective value in (1) for a fixed policy π. Alternatively, we use the term *policy optimization* to refer to solving for the optimal π in (1).

3 CVaR policy optimization gap

This section shows that a common CVaR decomposition proposed in Chow et al. [4] and used to optimize risk-averse policies is inherently sub-optimal regardless of how closely one discretizes the state space.

The *conditional value at risk* (CVaR) for a risk level $\alpha \in [0, 1]$ and a random variable $\tilde{x} \in \mathbb{X}$ distributed as $\tilde{x} \sim \boldsymbol{q}$ is defined as (e.g., [8, definition 11.8])

$$\mathrm{CVaR}_\alpha \left[\tilde{x}\right] \;=\; \inf_{\boldsymbol{\xi} \in \Delta(O)} \left\{ \boldsymbol{\xi}^\top \boldsymbol{x} \mid \alpha \cdot \boldsymbol{\xi} \le \boldsymbol{q} \right\}. \tag{2}$$

CVaR is also known as the average value at risk, and sometimes styled as CV@R or AV@R and can be defined for $\alpha \in (0, 1]$ as (e.g., [14, eq. (6.23)])

$$\mathrm{CVaR}_\alpha \left[\tilde{x}\right] \;=\; \sup_{z \in \mathbb{R}} \left(z - \alpha^{-1} \mathbb{E} \left[z - \tilde{x} \right]_+ \right).$$

CVaR, as we define it, applies to $\tilde{x}$ that represents rewards and assumes that a higher value of the risk measure is preferable to a lower value. Also, our CVaR becomes less risk-averse when the risk level α increases. Other definitions common in the literature include CVaR applied to costs, CVaR applied to negative rewards, or CVaR with a risk level $1 - \alpha$.

The following theorem represents one of the key results used to decompose the risk measure in multi-stage decision-making.

Theorem 1 (lemma 22 in [12]). *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,*

$$\mathrm{CVaR}_\alpha \left[r(\tilde{s}, \tilde{a}, \tilde{s}') \right] \quad = \quad \min_{\boldsymbol{\zeta} \in \mathcal{Z}_{\mathrm{C}}} \sum_{s \in \mathcal{S}} \zeta_s \, \mathrm{CVaR}_{\alpha \zeta_s \hat{p}_s^{-1}} \left[r(s, \tilde{a}, \tilde{s}') \mid \tilde{s} = s \right], \tag{3}$$

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

$$\mathcal{Z}_{\mathrm{C}} \;=\; \left\{ \boldsymbol{\zeta} \in \Delta(S) \mid \alpha \cdot \boldsymbol{\zeta} \le \hat{\boldsymbol{p}} \right\}. \tag{4}$$

The notation in Theorem 1 differs superficially from lemma 22 in [12]. Specifically, our CVaR is defined for rewards rather than costs, the meaning of our α corresponds to $1-\alpha$ in [12], and we use $\xi_s = z_s \hat{p}_s$ as the optimization variable.

We include a simple proof of Theorem 1 below for completeness.

Proof. 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 $\boldsymbol{q} \in \Delta(O)$ 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 $\boldsymbol{\xi} \in \mathbb{R}^O$ 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 vector $\boldsymbol{x}_s = \boldsymbol{x}_{s,\cdot,\cdot} \in \mathbb{R}^{S \cdot A}$ is defined analogously to $\boldsymbol{\xi}_s$.

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

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

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 $\boldsymbol{\xi}_s$ are independent given $\boldsymbol{\zeta}$, we get that

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

The step (a) follows from the interchangeability principle [14, 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 $\tilde{x} = r(\tilde{s}, \tilde{a}, \tilde{s}')$ is conditionally distributed on $\tilde{s} = s$ according to $q_{s,a,s'} \hat{p}_s^{-1}$ and the equality follows from the definition of CVaR in (2). □

The decomposition in Theorem 1 is important because it shows that the CVaR evaluation can be formulated as a dynamic program. The theorem shows that CVaR at the time $t = 0$ decomposes into a convex combination of CVaR values at the 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 $\boldsymbol{\xi}$, 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.

Chow et al. [4] propose to adapt the decomposition from Theorem 1 to the policy optimization setting as

$$\begin{aligned}\max_{\pi\in\Pi}\mathrm{CVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\mid \tilde{a}\sim\boldsymbol{\pi}(\tilde{s})\right] &= \max_{\pi\in\Pi}\min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\left(\mathrm{CVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}')\mid \tilde{a}\sim\boldsymbol{\pi}(s),\tilde{s}=s\right]\right)\\ &\overset{??}{=} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_{\mathrm{C}}}\sum_{s\in\mathcal{S}}\zeta_s\left(\max_{\boldsymbol{d}\in\Delta(\mathcal{A})}\mathrm{CVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}')\mid \tilde{a}\sim\boldsymbol{d},\tilde{s}=s\right]\right).\end{aligned} \tag{5}$$

Chow et al. [4] use the decomposition in (5) to formulate a dynamic program with the current risk level as an additional state variable.

The following theorem shows that the equality in (5) marked with a question mark is false in general.

Theorem 2. *There exists an MDP and a risk level $\alpha\in(0,1)$ such that*

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

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

Before proving Theorem 2, we discuss its implications. First, Theorem 2 contradicts theorems 5 and 7 in Chow et al. [4] and shows that their algorithm is inherently suboptimal regardless of the discretization resolution. Theorem 2 also contradicts the optimality of the accelerated dynamic program proposed in Stanko and Macek [15].

Second, Li et al. [10, section 5.1] recently proposed to optimize the CVaR objective in MDPs using a quantile representation. Their approach closely resembles the accelerated algorithm of Stanko and Macek [15]. Our additional analysis and numerical simulations indicate that the decomposition in Li et al. [10] is identical to Chow et al. [4], in which case, Theorem 2 also contradicts theorem 4 in Li et al. [10].

Finally, it is important to emphasize that Theorem 2 only applies to the policy optimization setting and does not contradict Theorem 1, which holds for the evaluation of policies that assign the same action distribution to each history of states and action (i.e., are independent of the hypothesized values of $\boldsymbol{\zeta}$).

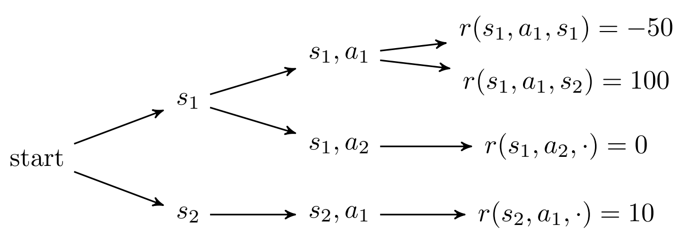

Figure 1: Rewards of MDP M_{C} used in the proof of Theorem 2. The dot indicates that the rewards are independent of the next state.

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

$$r(s_1,a_1,s_1) = -50, \qquad r(s_1,a_1,s_2) = 100,$$
$$r(s_1,a_2,s_1) = r(s_1,a_2,s_2) = 0, \qquad r(s_2,a_1,s_1) = r(s_2,a_1,s_2) = 10\,.$$

The transition probabilities in M_{C} are

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

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

To simplify the notation, 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

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

Because CVaR is convex in the distribution [7] and any distribution for $r(\tilde{s},\tilde{a},\tilde{s}')$ obtained from a $\pi \in \Pi$ 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, we can reformulate the *left-hand side* of (6) in terms of θ_π as

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

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 (6) can be expressed using the convexity of CVaR in the distribution by algebraic manipulation as

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

Using the notation introduced above and the sufficiency of optimizing over deterministic policies only, the inequality in (6) becomes

$$\max_{\pi\in\{\pi_1,\pi_2\}} \min_{\boldsymbol{\zeta}\in\mathcal{Z}_\mathrm{C}} \theta_\pi(\boldsymbol{\zeta}) \;<\; \min_{\boldsymbol{\zeta}\in\mathcal{Z}_\mathrm{C}} \max_{\pi\in\{\pi_1,\pi_2\}} \theta_\pi(\boldsymbol{\zeta}) . \tag{7}$$

Figure 2 demonstrates the inequality in (7) numerically, with a rectangle representing the left-hand maximum and a pentagon representing the right-hand 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 (7) 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 cash invariant. The function $\theta_{\pi_1}(\cdot)$ is piecewise-linear and convex, and its slope can be computed using the subgradient that satisfies for each $s\in\mathcal{S}$ and $\hat{\boldsymbol{\zeta}}\in\mathcal{Z}_\mathrm{C}$ [4]:

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

Simple algebraic manipulation then shows that

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

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 a rectangle and a pentagon in Figure 2, respectively. □

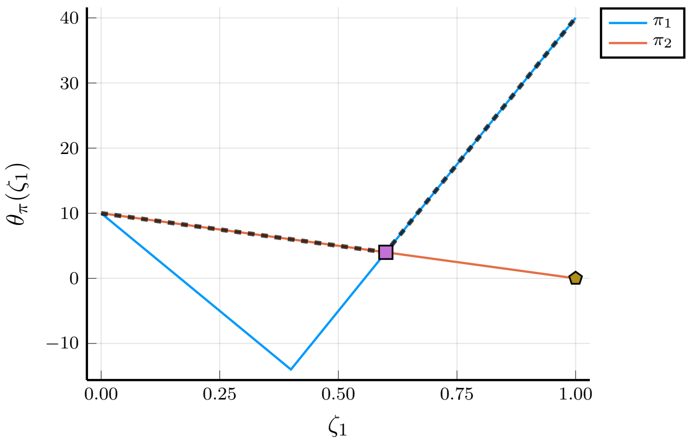

Figure 2: Functions in the CVaR counterexample in the proof of Theorem 2. The dashed line shows the function $\zeta_{s_1} \mapsto \max_{\pi \in \{\pi_1, \pi_2\}} \theta_\pi([\zeta_{s_1}, 1 - \zeta_{s_1}])$.

In summary, the decomposition in Theorem 1 cannot be exploited in policy optimization because the inequality in the derivation above may not be tight:

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

where the last equality follows from the interchangeability property of optimization and expected value [14, theorem 7.92].

4 EVaR policy evaluation gap

This section shows that a decomposition for EVaR proposed in Ni and Lai [11] is inexact even when considering the policy evaluation setting and an arbitrarily-fine discretization.

The *entropic value at risk* (EVaR) is defined as [1]

$$\begin{aligned} \mathrm{EVaR}_\alpha[\tilde{x}] &= \sup_{\beta > 0} \beta^{-1} \left(-\log \mathbb{E}\left[\exp\left(-\beta \tilde{x} \right) \right] + \log \alpha \right) \\ &= \inf_{\boldsymbol{\xi} \ll \boldsymbol{q}} \left\{ \mathbb{E}_{\boldsymbol{\xi}}[\tilde{x}] \mid \mathrm{KL}(\boldsymbol{\xi} \| \boldsymbol{q}) \le -\log \alpha \right\}. \end{aligned} \tag{8}$$

Here, KL is the standard KL-divergence defined for each $\boldsymbol{x}, \boldsymbol{y} \in \Delta(\Omega)$ as

$$\mathrm{KL}(\boldsymbol{x} \| \boldsymbol{y}) = \sum_{\omega \in \Omega} x_\omega \log \left(\frac{x_\omega}{y_\omega} \right).$$

the KL-divergence is defined only when $\boldsymbol{x}$ is aboslutely continuous with respect to $\boldsymbol{y}$, denoted as $\boldsymbol{x} \ll \boldsymbol{y}$, and defined as $y_\omega = 0 \Rightarrow x_\omega = 0$ for each $\omega \in \Omega$.

Ni and Lai [11] 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

$$\mathrm{EVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\right] \overset{??}{=} \min_{\boldsymbol{\xi}\in\mathcal{Z}_\mathrm{E}} \sum_{s\in\mathcal{S}} \xi_s \, \mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right] , \tag{9}$$

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

$$\mathcal{Z}_\mathrm{E} = \left\{ \boldsymbol{\xi} \in \Delta(\mathcal{S}) \mid \sum_{s\in\mathcal{S}} \xi_s \log(\xi_s/\hat{p}_s) \le -\log\alpha \,,\ \underbrace{\boldsymbol{\xi} \le \alpha^{-1}\hat{\boldsymbol{p}}}_{\text{implicit in [11]}} \right\} . \tag{10}$$

Note that we use variables $\xi_s = z_s\hat{p}_s$ in comparison with z_s in Ni and Lai [11].

The constraint $\boldsymbol{\xi} \le \alpha^{-1}\hat{\boldsymbol{p}}$ in (10) is not stated explicitly in Ni and Lai [11] but is necessary because $\mathrm{EVaR}_{\alpha'}[\cdot]$ is defined only for $\alpha' \in (0,1)$ (and extended to $[0,1]$). When $\alpha' = \alpha\xi_s\hat{p}_s^{-1}$ in (9) it must also satisfy for each $s \in \mathcal{S}$ that

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

This additional constraint on $\boldsymbol{\xi}$ implies that $\mathcal{Z}_\mathrm{E} \subseteq \mathcal{Z}_\mathrm{C}$.

The following theorem shows that the equality in (9) does not hold even in the policy evaluation setting.

Theorem 3. *There exists an MDP with a single action and $\alpha \in (0,1)$ such that*

$$\mathrm{EVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right] < \min_{\boldsymbol{\xi}\in\mathcal{Z}_\mathrm{E}} \sum_{s\in\mathcal{S}} \xi_s \, \mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,a_1,\tilde{s}') \mid \tilde{s}=s\right] , \tag{11}$$

where $\tilde{s} \sim \hat{\boldsymbol{p}}$, $\tilde{s}' \sim p(\tilde{s},a_1,\cdot)$, and set $\mathcal{Z}_\mathrm{E}$ defined by (10).

Theorem 3 demonstrates a stronger failure mode than Theorem 2, since it applies to both policy evaluation and policy optimization settings.

Proof. 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 $\hat{p}_{s_1} = \hat{p}_{s_2} = 0.5$.

Because $\mathcal{Z}_\mathrm{E} \subseteq \mathcal{Z}_\mathrm{C}$, the right-hand side of (11) can be lower-bounded by CVaR as

$$\begin{aligned} \min_{\boldsymbol{\xi}\in\mathcal{Z}_\mathrm{E}} \sum_{s\in\mathcal{S}} \xi_s \, \mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,a_1,\tilde{s}')\right] &= \min_{\boldsymbol{\xi}\in\mathcal{Z}_\mathrm{E}} \sum_{s\in\mathcal{S}} \xi_s r(s) \\ &\ge \min_{\boldsymbol{\xi}\in\mathcal{Z}_\mathrm{C}} \sum_{s\in\mathcal{S}} \xi_s r(s) = \mathrm{CVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right] . \end{aligned} \tag{12}$$

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 [8].

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 [1, proposition 3.2]), we can further lower-bound (12) as

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

Therefore, (11) holds with an inequality.

To prove by contradiction that the inequality in (11) is strict, suppose that

$$\mathrm{EVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right] \;=\; \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s\, \mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,a_1,\tilde{s}')\right] . \tag{14}$$

Equalities (14) and (13) imply that $\mathrm{EVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right] = \mathrm{CVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right]$ which is false in general [1].

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

$$\mathrm{CVaR}_\alpha\left[r(\tilde{s},a_1,\tilde{s}')\right] = \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{C}}} \sum_{s\in\mathcal{S}} \xi_s r(s) = \max\left\{0,\, \frac{\hat{p}_{s_1}+\alpha-1}{\alpha}\right\} . \tag{15}$$

Choose α such that $1-\alpha < \hat{p}_{s_1}$, then the optimal $\boldsymbol{\xi}^\star$ in (15) 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 (8), and, therefore, there exists an $\epsilon > 0$ such that

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

which proves the desired inequality. □

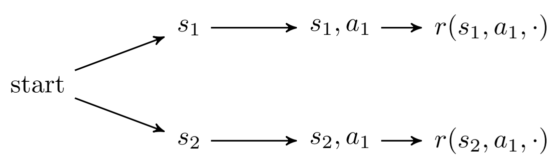

Figure 3: Rewards of the MDP M_{E} used in the proof of Theorem 3. The dot indicates that the rewards are independent of the next state.

We propose a correct decomposition of EVaR in the following theorem and employ it to establish that the decomposition in (9) overestimates the actual value of EVaR.

Theorem 4. *Given a random variable $\tilde{x} \in \mathbb{X}$ and a discrete variable $\tilde{y}\colon \Omega \to \mathcal{N} = \{1,\ldots,N\}$, with probabilities denoted as $\{\hat{p}_i\}_{i=1}^N$, we have that*

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

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 [6, theorem 2.5.3]), which states for any probability distributions $\boldsymbol{\eta}, \boldsymbol{q} \in \Delta(\Omega)$ that

$$\mathrm{KL}(\boldsymbol{\eta}\|\boldsymbol{q}) = \mathrm{KL}(\boldsymbol{\eta}(\tilde{y})\|\boldsymbol{q}(\tilde{x})) + \mathrm{KL}(\boldsymbol{\eta}(\tilde{x}|\tilde{y})\|\boldsymbol{q}(\tilde{x}|\tilde{y})), \tag{16}$$

where the conditional relative entropy is defined as

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

We can now decompose EVaR from its definition in (8) as

$$
\begin{aligned}
\mathrm{EVaR}_\alpha\left[\tilde{x}\right] &= \inf_{\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}\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}\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{\zeta}\in(0,1]^N,\boldsymbol{\eta}\ll\boldsymbol{q}} \left\{\mathbb{E}_{\boldsymbol{\eta}}[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{\zeta}\in(0,1]^N,\boldsymbol{\xi}\ll\boldsymbol{q}} \left\{\mathbb{E}_{\boldsymbol{\xi}}[\mathrm{EVaR}_{\zeta_{\tilde{y}}}\left[\tilde{x}|\tilde{y}\right]] \mid \mathrm{KL}(\boldsymbol{\xi}(\tilde{y})\|\boldsymbol{q}(\tilde{y})) + \mathbb{E}_{\boldsymbol{\xi}}[-\log(\zeta_{\tilde{y}})] \le -\log\alpha\right\} \\
&= \inf_{\boldsymbol{\zeta}\in(0,1]^N,\,\boldsymbol{\xi}\in\Delta(N)} \left\{\sum_i \xi_i\,\mathrm{EVaR}_{\zeta_i}\left[\tilde{x} \mid \tilde{y}=i\right] \mid \boldsymbol{\xi}\ll\hat{\boldsymbol{p}}, \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 (16) 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 $\boldsymbol{\eta}$ for a given ξ in step (c), and then get the final expression by algebraic manipulation. □

Corollary 1. *Given any finite MDP with horizon $T = 1$, we have that*

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

where

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

Moreover, the EVaR can be upper-bounded as

$$
\mathrm{EVaR}_\alpha\left[r(\tilde{s},\tilde{a},\tilde{s}')\right] \quad \le \quad \min_{\boldsymbol{\xi}\in\mathcal{Z}_{\mathrm{E}}} \sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\alpha\xi_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right]. \tag{17}
$$

Proof. The first part of the corollary follows directly from Theorem 4. Suppose that $\alpha > 0$; the result follows for $\alpha = 0$ because $\mathrm{EVaR}_0\left[\cdot\right]$ reduces to ess inf. Then, the second part of the corollary holds as

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

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 $\mathrm{EVaR}_0\left[\tilde{x}\right] = \lim_{\alpha\to 0}\mathrm{EVaR}_\alpha\left[\tilde{x}\right]$, 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. □

5 VaR decomposition

This section discusses a dynamic program decomposition for value-at-risk (VaR) whose decomposition resembles the CVaR and EVaR decompositions described above. We provide a new proof of the VaR decomposition to elucidate the differences that make it optimal in contrast with CVaR and EVaR decompositions. Our VaR decomposition closely resembles the quantile MDP approach in Li et al. [10] with a few technical modifications that can significantly impact the computed value.

Value at risk (VaR) in modern risk management literature (e.g., [8, 14]) is typically defined for a risk level $\alpha \in [0,1]$ and a random variable $\tilde{x} \in \mathbb{X}$ as

$$\mathrm{VaR}_\alpha[\tilde{x}] \;=\; \sup\;\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} < z] \le \alpha\} \;=\; \inf\;\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} \le z] > \alpha\}\,. \tag{18}$$

Note that $\mathrm{VaR}_1[\tilde{x}] = \infty$. For the equality between the two definitions above, see, for example, Follmer and Schied [8, remark A.20].

To contrast the typical definition of VaR with the quantile definition in Li et al. [10], it is helpful to summarize how VaR is related to the quantile of a random variable. Recall that $q \in \mathbb{R}$ is an α-*quantile* of $\tilde{x} \in \mathbb{X}$ when

$$\mathbb{P}[\tilde{x} \le q] \ge \alpha \quad \text{and} \quad \mathbb{P}[\tilde{x} < q] \le \alpha. \tag{19}$$

In general, the set of quantiles is an interval $[q^-_{\tilde{x}}(\alpha), q^+_{\tilde{x}}(\alpha)]$ with the bounds computed as [8, appendix A.3]

$$\begin{aligned} q^-_{\tilde{x}}(\alpha) \;&=\; \sup\{z \mid \mathbb{P}[\tilde{x} < z] < \alpha\} = \inf\{z \mid \mathbb{P}[\tilde{x} \le z] \ge \alpha\} \\ q^+_{\tilde{x}}(\alpha) \;&=\; \inf\{z \mid \mathbb{P}[\tilde{x} \le z] > \alpha\} = \sup\{z \mid \mathbb{P}[\tilde{x} < z] \le \alpha\}\,. \end{aligned}$$

When the distribution of $\tilde{x}$ is absolutely continuous (atomless), then $q^+_{\tilde{x}}(\alpha) = q^-_{\tilde{x}}(\alpha)$ 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

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

The objective in Li et al. [10] is to maximize the quantile operator $Q_\alpha \colon \mathbb{X} \to \mathbb{R}$ defined for rewards $\tilde{x} \in \mathbb{X}$ and a risk level $\alpha \in [0,1]$ as

$$Q_\alpha(\tilde{x}) \;=\; \inf_{z\in\mathbb{R}}\;\{z \mid \mathbb{P}[\tilde{x} \le z] \ge \alpha\}\;. \tag{20}$$

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

$$Q_\alpha(\tilde{x}) \;=\; q^-_{\tilde{x}}(\alpha), \qquad \text{but} \qquad \mathrm{VaR}_\alpha[\tilde{x}] \;=\; q^+_{\tilde{x}}(\alpha)\,. \tag{21}$$

As a result, the quantile MDP objective in (20) coincides with the VaR value only when the quantile is unique. Example 1 demonstrates that this may not always be the case.

Theorem 5. *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:*

$$\mathrm{VaR}_\alpha[\tilde{x}] \quad = \quad \sup_{\boldsymbol{\zeta}\in\Delta(N)} \left\{ \min_i\; \mathrm{VaR}_{\alpha\zeta_i\hat{p}_i^{-1}}[\tilde{x} \mid \tilde{y} = i] \mid \alpha \cdot \boldsymbol{\zeta} \le \hat{\boldsymbol{p}} \right\}, \tag{22}$$

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

Proof. First, we decompose VaR using the definition in (18) as

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

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 equivalence follows immediately from the fact that $\mathrm{VaR}_{\zeta_i}[\tilde{x} \mid \tilde{y} = i]$ is a ζ_i-quantile and satisfies (19), namely:

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

The direction $\Rightarrow$ follows from the definition of VaR (see equation (18)), 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 (22) 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 is amenable to policy optimization. The main difference between the VaR decomposition and CVaR is that VaR can be expressed as a maximization or a supremum.

Theorem 6. *Given any finite MDP with horizon $T = 1$, we have that:*

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

Proof. The equality develops from Theorem 6 as

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

where we first change the order of maximum and supremum, followed by changing the order of $\max_\pi \min_s$ with $\min_s \max_\pi$, which follows based on the interchangeability property of the maximum operation [13, proposition 2.2]. □

For completeness, we finally present the valid decompositions for the lower quantile and lower quantile MDP.

Theorem 7. *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:*

$$q_\alpha^-(\tilde{x}) \quad = \quad \sup_{\boldsymbol{\zeta}\in[0,1]^N} \left\{ \min_{i:\zeta_i<1} \; q_{\zeta_i}^-(\tilde{x} \mid \tilde{y} = i) \mid \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha \right\}, \tag{23}$$

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

We note that the difference with the result presented in [10] resides in the constraint imposed on $\boldsymbol{\zeta}$ replacing the weak inequality with a strong one.

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

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

We decompose the probability $\mathbb{P}[\tilde{x} < z]$ 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 exploits the following equivalence:

$$\mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] < \zeta_i \quad \Leftrightarrow \quad z < q_{\zeta_i}^-(\tilde{x} \mid \tilde{y} = i)$$

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

$$z < q_{\zeta_i}^-(\tilde{x} \mid \tilde{y} = i) = \inf\{z \mid \mathbb{P}[\tilde{x} \le z \mid \tilde{y} = i] \ge \zeta_i\} \;\Rightarrow\; \mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] < \zeta_i.$$

The direction $\Rightarrow$ follows from the definition of VaR (see equation (18)), which implies that VaR upper-bounds any z that satisfies the left-hand condition:

$$\mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] < \zeta_i \;\Rightarrow\; q_{\zeta_i}^-(\tilde{x} \mid \tilde{y} = i) = \sup\ \{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} < z \mid \tilde{y} = i] < \zeta_i\} \ge z,$$

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

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

which leads to a contradiction. In step (e), we solve for z. Finally, we obtain the form in (23). □

Corollary 2. *Given any finite MDP with horizon $T = 1$, we have that:*

$$\max_{\pi\in\Pi} q_\alpha^-(r(\tilde{s},\tilde{a},\tilde{s}') \mid \tilde{a} \sim \boldsymbol{\pi}(\tilde{s})) = \sup_{\boldsymbol{\zeta}\in[0,1]^N} \left\{ \min_{i:\zeta_i<1} \; \max_{\boldsymbol{d}\in\Delta(A)} q_{\zeta_s}^-(r(s,\tilde{a},\tilde{s}') \mid \tilde{a} \sim \boldsymbol{d}) \mid \sum_{i=1}^N \zeta_i \hat{p}_i < \alpha \right\}.$$

6 Conclusion

Our examples show that several popular approaches to solving MDPs with static CVaR and EVaR risk measures are inherently suboptimal. We also give a new decomposition bound for EVaR and adapt an existing quantile MDP decomposition to the VaR objective.

References

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

[2] N. Bäuerle and J. Ott. Markov decision processes with average-value-at-risk criteria. Mathematical Methods of Operations Research, 74(3):361–379, 2011.

[3] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In Neural Information Processing Systems (NIPS), 2014.

[4] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: A CVaR optimization approach. In Neural Information Processing Systems (NIPS), 2015.

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

[6] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, 2nd edition, July 2006.

[7] E. Delage, D. Kuhn, and W. Wiesemann. "Dice"-sion-making under uncertainty: When can a random decision reduce risk? Management Science, 65(7):3282–3301, 2019.

[8] H. Follmer and A. Schied. Stochastic Finance: Introduction in Discrete Time. De Gruyter Graduate, fourth edition, 2016.

[9] J. L. Hau, M. Petrik, and M. Ghavamzadeh. Entropic risk optimization in discounted MDPs. In Artificial Intelligence and Statistics (AISTATS), 2023.

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

[11] X. Ni and L. Lai. EVaR optimization for risk-sensitive reinforcement learning, Feb. 2022. URL https://faculty.engineering.ucdavis.edu/lai/wp-content/uploads/sites/38/2022/02/paper_v13.pdf.

[12] G. C. Pflug and A. Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. Mathematics of Operations Research, 41(2):682–699, 2016.

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

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

[15] S. Stanko and K. Macek. Risk-averse distributional reinforcement learning: A CVaR optimization approach. In International Joint Conference on Computational Intelligence, pages 412–423, 2019.