Risk-sensitive Markov decision processes with long-run CVaR criterion

Li Xia1 | Luyao Zhang1 | Peter W. Glynn2

1School of Business, Sun Yat-sen University, Guangzhou, China
2Department of Management Science and Engineering, Stanford University, Stanford, California, USA

Correspondence
Li Xia, School of Business, Sun Yat-sen University, Guangzhou, China.
Email: xiali5@sysu.edu.cn

Handling Editor: Stefanus Jasin

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2022YFA1004600; Guangdong Basic and Applied Basic Research Foundation, Grant/Award Numbers: 2023A1515012492, 2021A1515011984; Guangdong Province Key Laboratory of Computational Science, Grant/Award Number: 2020B1212000032; Fundamental Research Funds for the Central Universities, Grant/Award Number: 22wbj01; InnoHK: The Government of the HKSAR; Laboratory for AI-Powered Financial Technologies; National Natural Science Foundation of China, Grant/Award Numbers: 62073346, 72371253, 11931018, U1811462

Abstract
CVaR (Conditional value at risk) is a risk metric widely used in finance. However, dynamically optimizing CVaR is difficult, because it is not a standard Markov decision process (MDP) and the principle of dynamic programming fails. In this paper, we study the infinite-horizon discrete-time MDP with a long-run CVaR criterion, from the view of sensitivity-based optimization. By introducing a pseudo-CVaR metric, we reformulate the problem as a bilevel MDP model and derive a CVaR difference formula that quantifies the difference of long-run CVaR under any two policies. The optimality of deterministic policies is derived. We obtain a so-called Bellman local optimality equation for CVaR, which is a necessary and sufficient condition for locally optimal policies and only necessary for globally optimal policies. A CVaR derivative formula is also derived for providing more sensitivity information. Then we develop a policy iteration type algorithm to efficiently optimize CVaR, which is shown to converge to a local optimum in mixed policy space. Furthermore, based on the sensitivity analysis of our bilevel MDP formulation and critical points, we develop a globally optimal algorithm. The piecewise linearity and segment convexity of the optimal pseudo-CVaR function are also established. Our main results and algorithms are further extended to optimize the mean and CVaR simultaneously. Finally, we conduct numerical experiments relating to portfolio management to demonstrate the main results. Our work sheds light on dynamically optimizing CVaR from a sensitivity viewpoint.

KEYWORDS
Bellman local optimality equation, long-run CVaR, Markov decision process, risk-sensitive, sensitivity-based optimization

1 INTRODUCTION

The Markov decision process (MDP) is a fundamental mathematical model used to handle stochastic dynamic optimization problems (Feinberg & Shwartz, 2002; Puterman, 1994). The study of MDPs in operations research is multidisciplinary. It is deeply connected with reinforcement learning in computer science (Kaelbling et al., 1996; Sutton & Barto, 2018), optimal control in control science (Bertsekas, 2005; Lewis et al., 2012), dynamic discrete choice modeling in econometrics (Aguirregabiria & Mira, 2010; Rust, 1987), and so on. Traditional MDP theory focuses on the criteria of discounted or long-run average cost, where the principle of dynamic programming plays a key role. However, for other optimization criteria, such as risk metrics in finance, the corresponding optimization problems usually do not fit the standard MDP model and specific investigations are needed case by case.

Conditional value at risk (CVaR), also called average VaR (value at risk) or expected shortfall, is a widely used risk metric, built upon the VaR and variance related risk metrics (Rockafellar & Uryasev, 2000). Recently, it has also been used in other engineering fields, such as energy systems (Asensio & Contreras, 2016; Li et al., 2018), manufacturing and supply chains (Dixit et al., 2020; Li & Arreola-Risa, 2022; Xie et al., 2018), and so on. Suppose X is a continuous random variable representing a stochastic loss and its distribution function is denoted as $F_X(x)$. The CVaR corresponding...
to this stochastic loss at probability level $\alpha$ is defined as $\text{CVaR}_x(\alpha) := \mathbb{E}[X|X \geq F^{-1}_{X}(\alpha)] = \frac{1}{1-\alpha} \int \max\{0, x - F^{-1}_{X}(\alpha)\} \, dx$, which measures the conditional expectation of losses greater than a given quantile $F^{-1}_{X}(\alpha)$ (or called VaR$_x(\alpha)$), where $0 < \alpha < 1$. Compared with other risk metrics such as variance or VaR, CVaR can measure not only the down-side risk but also the expected value of large losses. Moreover, CVaR is a coherent risk measure (Artzner et al., 1999), which has desirable properties (monotonicity, translation equivariance, subadditivity, and positive homogeneity).

CVaR has been widely used to measure risk in static (single-stage) optimization problems (Alexander et al., 2006; Fu et al., 2009; Hong et al., 2014). However, it is difficult to handle the CVaR optimization problem in stochastic dynamic (multi-stage) scenarios because of time inconsistency (Boda & Filar, 2006; Pflug & Pichler, 2016). In the MDP terminology, we may define an instantaneous cost function for the long-run CVaR metric as $f(s, a) := F^{-1}_{X}(\alpha) + \frac{1}{1-\alpha}[c(s, a) - F^{-1}_{X}(\alpha)]^+$, where $(s, a)$ is the MDP’s one-step loss incurred in state $s$ with action $a$, $\cdot^+ := \max\{0, \cdot\}$, and $X$ is the random variable indicating the stochastic loss whose value is realized as $(s, a)$’s. We observe that $f(s, a)$ involves $F^{-1}_{X}(\alpha)$ that is affected by future losses and actions. Therefore, the CVaR cost function depends on future behaviors and is not additive or Markovian. The CVaR dynamic optimization problem does not fit a standard MDP model. Thus, the classical Bellman optimality equation does not hold and the principle of dynamic programming fails. However, Rockafellar and Uryasev (2002) discovered that the CVaR of the random variable $X$ is equivalent to $\min_{y \in \mathbb{R}} \{y + \frac{1}{1-\alpha} \mathbb{E}[X - y]^+\}$, where $y^+ = F^{-1}_{X}(\alpha)$ exactly achieves the minimum. With this equivalence, Bäuerle and Ott (2011) converted the CVaR minimization of discounted accumulative costs with finite or infinite horizon into a bilevel optimization problem, where the outer one is a static optimization problem with the auxiliary variable $y \in \mathbb{R}$ and the inner one is a standard MDP with the fixed $y$ and an augmented state space. Although the existence and properties of optimal policies were studied there, the efficient solution of such bilevel MDP problems was not presented. For different values of $y \in \mathbb{R}$, we have to solve different inner MDP problems, which are computationally intractable. Such techniques of equivalent MDP transformation and state augmentation were further extended to study the CVaR minimization in other general cases, such as semi-MDPs (Huang & Guo, 2016), continuous-time MDPs (Miller & Yang, 2017), unbounded costs (Uğurlu, 2017), just to name a few. On the other hand, Haskell and Jain (2015) utilized occupation measures to study the CVaR optimization of discounted cost infinite-horizon MDPs using mathematical programming, where the optimality of deterministic policies was also discussed. However, it is usually not efficient to solve a series of mathematical programs, especially when they are not convex or linear programs. Efficiently computing CVaR optimal policies is an important but not well-studied topic in the MDP literature.

Existing works primarily study the CVaR of discounted accumulated costs at a terminal stage $T$, say $\text{CVaR}_x(\tilde{X}_T)$, where $\tilde{X}_T := \sum_{t=0}^{T} y^t X_t$, $X_t$ is the cost at time $t$, and $y$ is the discount factor. However, the long-run average (per-step) cost (i.e., $\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[X_t]$) is the preferred criterion compared to the discounted criterion in many applications (Hernández-Lerma & Lasserre, 2012) such as in the analysis of communication networks or queueing systems (Ephremides & Verdu, 1989; Stidham & Weber, 1993; Tijms, 1986), where the objective $\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{CVaR}_x(\tilde{X}_t)$ makes sense when considering risks. Besides, there are many practical problems that concern about the risk of step-wise costs, such as the investment problem in finance and the optimal control in failure-prone manufacturing systems (Ahmadi-Javid & Malhamé, 2015). For example, in financial engineering, a risk-averse investor may be unable to tolerate high risk during the asset management process. Large downward fluctuations of the asset value over the time horizon $[0, T]$ may bring anxiety to the risk-averse investor and induce early withdrawal of the investment. In such a setting, the asset manager may wish to minimize not only the average cost $\mathbb{E}[X_t]$, but also the average value of $\text{CVaR}_x(\tilde{X}_t)$ over $t \in [0, T]$. When $T \to \infty$, this leads to the problem of finding an optimal policy that minimizes $\text{CVaR}_x(\tilde{X}_\infty)$ over the distribution of the steady-state random variable $X_\infty$. Similar steady-state risk measures have been studied in the literature, such as the steady-state variance in MDPs (Chung, 1994; Sobel, 1994; Xia, 2020). There is a recent paper by Bonetti et al. (2023) that also proposed the above steady-state CVaR in the setting of reinforcement learning, where they called such metric reward-based CVaR and a gradient-based TRPO algorithm was developed to find locally optimal solutions. However, steady-state CVaR optimization has not been studied in the setting of MDPs, and computationally efficient algorithms are particularly desired.

In this paper, we study a discrete-time undiscounted MDP with a long-run (steady-state) CVaR criterion. Different from the CVaR metrics for discounted costs, the long-run CVaR measures the conditional expectation of one-step costs when the MDP reaches steady-state. As the traditional approach of dynamic programming is not applicable for such non-standard MDP problems, we study this problem from the viewpoint of sensitivity-based optimization. By introducing a so-called pseudo-CVaR, we derive a bilevel MDP formulation with nested structure, where the inner one is a standard MDP for minimizing the pseudo-CVaR and the outer one is an optimization problem of a single parameter called pseudo-VaR. Then, we derive a closed-form difference formula to quantify the difference of long-run CVaR under any two policies. The CVaR difference formula has an elegant form that reduces the difficulty caused by the non-Markovian CVaR cost function. The optimality of deterministic policies is also shown from the nested formulation. With the CVaR difference formula, we further derive an optimality equation called the Bellman local optimality equation, which is necessary and sufficient for locally optimal policies in mixed policy.
space, while it is only necessary for globally optimal policies. A CVaR derivative formula is also obtained. With our CVaR sensitivity formulas and the Bellman local optimality equation, we develop an iterative algorithm that behaves similarly to the classical policy iteration algorithm and can efficiently converge to local optima. Based on the sensitivity analysis on the inner MDP, we derive a method to compute the so-called critical points and a globally optimal algorithm for CVaR MDPs is developed. The piecewise linearity and segment convexity of the optimal pseudo-CVaR function are also investigated. Our main results and algorithms are also extended to the scenario of optimizing mean and CVaR simultaneously. Finally, we study numerical examples relating to portfolio management to demonstrate the effectiveness of our approaches.

The main contributions of this paper are threefold. First, we study the discrete-time undiscounted MDP with the long-run CVaR criterion. To the best of our knowledge, our paper is the first to investigate MDP theory for minimizing long-run CVaR. Second, we derive the CVaR difference and derivative formulas, and the Bellman local optimality equation. These results work for the long-run CVaR MDPs with the aid of sensitivity analysis and optimization, which are different from the discounted MDPs using the tool of contraction operators. Third and also the most importantly, we develop a policy iteration type local algorithm and a sensitivity analysis based global algorithm, both are much more efficient than directly solving the bilevel MDP problem adopted in the literature (equivalent to solving a series of standard MDPs with the number of MDPs equal to the number of possible values of the outer-tier parameter \( \gamma \in \mathbb{R} \)). It seems that our work is the first to provide computationally efficient algorithms in the literature on CVaR MDPs.

The rest of the paper is organized as follows. In Section 2, we give the MDP formulation with the long-run CVaR criterion. In Section 3, we present the main results of this paper, including the CVaR difference formula, the Bellman local optimality equation, and the related theorems. In Section 4, we propose a policy iteration type algorithm and a sensitivity analysis based algorithm, which converges to a local and global optimum, respectively. In Section 5, we extend our approaches to mean-CVaR optimization. In Section 6, numerical experiments are conducted to demonstrate our main results. Finally, we conclude this paper and discuss future research topics in Section 7.

# 2 | PROBLEM FORMULATION

Consider a discrete-time MDP with tuple \( \mathcal{M} := (S, A, P, c) \), where the state space \( S \) and the action space \( A \) are both finite. A function \( P : S \times A \xrightarrow{D} S \) is the state transition probability kernel with element \( p(s'|s,a) \), where \( \leftrightarrow \) represents a mapping to the distribution on the successor \( S \). The element \( p(s'|s,a) \) indicates the transition probability to next state \( s' \) when action \( a \) is adopted in the current state \( s \). Obviously, we have \( \sum_{s' \in S} p(s'|s,a) = 1 \) for any \( (s,a) \in S \times A \). We denote \( c : S \times A \mapsto \mathbb{R} \) as the cost function and its element \( c(s,a) \) is the one-step (instantaneous) cost incurred in state \( s \) with action \( a \).

The minimum and the maximum of \( c \) are denoted as \( \underline{c} \) and \( \overline{c} \), respectively. A function \( d : S \times A \mapsto [0,1] \) describes a stationary randomized policy and its element \( d(s,a) \) indicates the probability of adopting action \( a \) in state \( s \), where \( \sum_{a \in A} d(s,a) = 1 \) for any \( s \in S \). If \( d \) is a stationary deterministic policy, we also use \( d(s) \in A \) to indicate the action adopted in state \( s \), with a slight abuse of notation. In this paper, we limit our discussion to stationary policies that make the associated Markov chains always ergodic (irreducible and aperiodic). A simple sufficient condition of ergodic MDPs is that there exists a policy under which the chain is irreducible and aperiodic, and that the support of the transition probabilities is independent of the action taken. Note that our main results can be extended to unichain MDPs, or even multichain cases with appropriate modifications. We denote \( D \) and \( D_0 \) as the spaces of stationary randomized and deterministic policies, respectively. The objective of an MDP is to find an optimal policy that optimizes a certain criterion of costs incurred over a time horizon.

Let \( X \) be a real-valued bounded random variable with cumulative distribution function \( F_X(x) = P(X \leq x) \). CVaR\(_{\alpha}(X) \) quantifies the expected value of \( X \) with its \((1 - \alpha)\)-tail distribution for a given probability level \( \alpha \in (0, 1) \), which is defined as

\[
\text{CVaR}_{\alpha}(X) := \frac{1}{1 - \alpha} \int_{\alpha}^{1} \text{VaR}_{\gamma}(X)d\gamma,
\]

where \( \text{VaR}_{\gamma}(X) \) denotes the \( \gamma \)-quantile of \( X \), also known as the value at risk, that is,

\[
\text{VaR}_{\gamma}(X) := \inf_{x \in \mathbb{R}} \{ x : F_X(x) \geq \gamma \} = F_X^{-1}(\gamma).
\]

When \( X \) has a continuous distribution, \( \text{CVaR}_{\alpha}(X) \) can also be rewritten as follows:

\[
\text{CVaR}_{\alpha}(X) = \mathbb{E}\{X | X \geq \text{VaR}_{\alpha}(X)\}.
\]

Rockafellar and Uryasev (2000, 2002) show that the CVaR of the random variable \( X \) is equivalent to solving the following convex optimization problem

\[
\text{CVaR}_{\alpha}(X) = \min_{y \in \mathbb{R}} \left\{ y + \frac{1}{1 - \alpha} \mathbb{E}[X - y]^+] \right\}, \tag{1}
\]

where \( y^* = \text{VaR}_{\alpha}(X) \) attains the minimum.

In this paper, we focus on the long-run CVaR criterion that measures the steady-state behavior of the MDP under a policy \( d \). Note that, in the rest of the paper, we omit the subscript \( \alpha \) of VaR and CVaR for notational simplicity, as \( \alpha \) is fixed unless otherwise declared. The definition of the long-run CVaR (or steady-state CVaR) is given as follows.
Definition 1. Let $C_t^d$ represent the random variable of the one-step cost at time $t$ under policy $d \in D$ with initial state $s_0 \in S$, then the long-run CVaR criterion is defined as

$$\text{CVaR}^d(s_0) := \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{CVaR}(C_t^d).$$

(2)

With the assumption of ergodicity of MDPs for each $d \in D$, the steady state-action distribution $\pi^d$ is well defined as

$$\pi^d(s, a) := \lim_{t \to \infty} \mathbb{P}^d(s_0, s, a) \quad \forall (s_0, s, a) \in S \times S \times A, d \in D,$$

(3)

where $\mathbb{P}^d(s_0, s, a)$ denotes the state-action visitation probability that $s_0 = s, a_t = a$ after $t$ time epochs we execute policy $d$ starting with initial state $s_0$. It is obvious that $\pi^d(s, a)$ is independent of initial state $s_0$. We also denote $\pi^d(s) := \sum_{s \in A} \pi^d(s, a)$ and $\pi^d(s, a) := \sum_{a \in A} \pi^d(s, a)\text{d}(s, a), s \in S$. The transition probability matrix of the MDP with policy $d$ is denoted as $P^d$ whose element is $P^d(s', s) := \sum_{a \in A} p^d(s, a)\text{d}(s, a), \forall s, s' \in S$.

Let $C^d$ be a random variable whose value is taken on $\{c(s, a) : (s, a) \in S \times A\}$ with corresponding probability distribution $\pi^d$. Below, we derive Lemma 1 to show that the long-run CVaR defined in Definition 1 equals the CVaR of $C^d$ for a given $d \in D$.

Lemma 1. For each stationary policy $d \in D$ and initial state $s_0 \in S$, it holds that

$$\text{CVaR}^d(s_0) = \text{CVaR}(C^d).$$

(4)

Proof. As both $C^d$ and $C_t^d$ are bounded variables with minimum $\underline{c}$ and maximum $\overline{c}$, the domain $y \in \mathbb{R}$ can be restricted to $[\underline{c}, \overline{c}]$ for computing $\text{CVaR}(C_t^d)$ and $\text{CVaR}(C^d)$ with (1)

$$\text{CVaR}(C^d) = \min_{y \in [\underline{c}, \overline{c}]} \left\{ y + \frac{1}{1-\alpha} \mathbb{E}[C^d - y]^+ \right\},$$

(5)

$$\text{CVaR}(C^d) = \min_{y \in [\underline{c}, \overline{c}]} \left\{ y + \frac{1}{1-\alpha} \mathbb{E}[C^d - y]^+ \right\}. $$

(6)

We apply (5) and (6) to estimate the gap between $\text{CVaR}(C_t^d)$ and $\text{CVaR}(C^d)$.

$$|\text{CVaR}(C_t^d) - \text{CVaR}(C^d)| = \left| \min_{y \in [\underline{c}, \overline{c}]} \left\{ y + \frac{1}{1-\alpha} \mathbb{E}[C^d - y]^+ \right\} - \min_{y \in [\underline{c}, \overline{c}]} \left\{ y + \frac{1}{1-\alpha} \mathbb{E}[C^d - y]^+ \right\} \right| \leq \frac{1}{1-\alpha} \max_{y \in [\underline{c}, \overline{c}]} \left| \mathbb{E}[C^d - y]^+ - \mathbb{E}[C^d - y]^+ \right|$$

where the first and third inequalities are ensured by $\mathbb{E}[c(s, a) - \min \{h_1(y_1), h_2(y_2)\}] \leq \max \{h_1(y_1), h_2(y_2)\}$ and $\mathbb{E}[c(s, a) - y]^+ \leq \overline{c} - \underline{c}, \forall y \in [\overline{c}, \underline{c}], (s, a) \in S \times A$, respectively, and the second inequality is ensured by absolute value inequalities. By applying (3) to (7), we derive

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{CVaR}(C_t^d) = \text{CVaR}(C^d) .$$

(8)

Thus, we substitute (8) into (2) and derive that the limit in (2) exists:

$$\text{CVaR}^d(s_0) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{CVaR}(C_t^d) = \text{CVaR}(C^d)$$

(9)

This completes the proof.

As the MDP is assumed always ergodic under any policy $d \in D$, the long-run CVaR of the MDP is independent of the initial state $s_0$ and further denoted as

$$\text{CVaR}^d := \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{CVaR}(C_t^d) = \lim_{T \to \infty} \text{CVaR}(C^d).$$

(10)

Similarly, we also have

$$\text{VaR}^d := \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{VaR}(C_t^d) = \lim_{T \to \infty} \text{VaR}(C^d).$$

(11)

The objective of the long-run CVaR MDP is to find an optimal policy $d^*$ that attains the minimal CVaR, that is,

$$d^* = \arg \min_{d \in D} \text{CVaR}^d ,$$

$$\text{CVaR}^* = \min_{d \in D} \text{CVaR}^d = \text{CVaR}^{d^*}. $$

(12)

To accommodate the CVaR metric, we introduce an auxiliary variable $y$ and define a new cost function as

$$\tilde{c}(y, s, a) := y + \frac{1}{1-\alpha}(c(s, a) - y)^+, $$

(13)

$$\forall (s, a) \in S \times A, \quad y \in \mathbb{R}. $$
applying (1) to (15), we derive a standard MDP model. Classical MDP theory, such as the long-run average optimization of MDP \( M(\text{VaR}^d) = \langle S, A, \mathcal{P}, \mathcal{C}, \tilde{c}(\cdot) \rangle \). However, the cost function \( \tilde{c}(\text{VaR}^d) \) depends on \( \text{VaR} \), the pseudo-CVaR equals the real CVaR of the long-run average optimization of MDP \( M(\text{VaR}^d) \). From (14) and (15), we see that the long-run CVaR optimization problem (12) is equivalent to a bilevel MDP problem (17) as the following mathematical program

\[
\min_{x \in \mathbb{R}^{S \times A}, \gamma \in \mathbb{R}} \left\{ \begin{array}{c}
y + \frac{1}{1 - \alpha} \sum_{(s, a) \in S \times A} x(s, a)[c(s, a) - y]^+ \\
\text{s.t.,} \\
\sum_{a \in A} x(s, a) = \sum_{(s', a') \in S \times A} x(s', a)p(s'|a, a) \\
\forall s \in S,
\end{array} \right.
\]

\[x(s, a) \geq 0, \quad \forall (s, a) \in S \times A. \tag{18}\]

Although the constraints of (18) are linear, the objective is quadratic and not convex. Therefore, (18) is not a convex optimization problem. In fact, as illustrated by the numerical experiments in Section 6, this problem may have multiple local minima, which is difficult to solve and establishes the nonconvexity.

Nevertheless, we find that it is not necessary to search \( y \) in the whole real number space \( \mathbb{R} \). With (1), we see that \( y^* \) in (16) equals \( \text{VaR}^d \). Thus, we define the set of all possible \( y^* \) as

\[\mathbb{Y} := \{ \text{VaR}^d : d \in D \}.\]

Enumerating every possible \( \text{VaR}^d \) to obtain \( \mathbb{Y} \) is still computationally intractable. Furthermore, we can use (11) to derive

\[\text{VaR}^d = \lim_{i \to \infty} \text{VaR}(C^d_i) = \lim_{i \to \infty} \inf\{\xi \in \mathbb{R} : P(c(s, a_i) \leq \xi) \geq \alpha\} \in \mathbb{Y}^o,\]

where \( \mathbb{Y}^o \) is defined as the set of all the possible values of \( c(s, a) \)'s, that is,

\[\mathbb{Y}^o := \{ c(s, a) : (s, a) \in S \times A \}. \tag{19}\]

Obviously, we have \( \mathbb{Y} \subseteq \mathbb{Y}^o \subseteq \mathbb{R} \) and \( |\mathbb{Y}^o| \leq |S||A| \). The number of standard MDPs in the bilevel problem (17) is significantly reduced and we directly derive the following lemma.

**Lemma 2.** The long-run CVaR minimization problem (12) is equivalent to the following bilevel MDP problem

\[\text{CVaR}^* = \min_{y \in \mathbb{Y}^o, d \in D} \text{CVaR}^d (y), \tag{20}\]
where the inner problem is a standard MDP of minimizing the pseudo-CVaR with a given $y$:

$$
\tilde{\text{CVaR}}^d (y) := \min_{d \in D} \text{CVaR}^d (y),
$$

$$
\tilde{d}^*(y) := \arg \min_{d \in D} \text{CVaR}^d (y). \tag{21}
$$

Lemma 2 indicates that a naive method for optimizing the long-run CVaR MDP is to solve a finite number of MDPs \{\mathcal{M}(y), y \in \mathbb{Y}\} in (21) and choose the one with the minimal pseudo-CVaR as $\min_{s \in \mathbb{Y}} \{\tilde{\text{CVaR}} (y)\}$. Most of literature work such as Bäuerle and Ott (2011) uses the above naive method to conduct numerical experiments. However, it is computationally intractable when the state and action spaces are large. How to efficiently solve the long-run CVaR MDP is still an open problem in the literature. In the rest of the paper, we use sensitivity analysis and optimization theory to study this problem in a new and efficient way. Different from traditional dynamic programming, sensitivity-based optimization theory studies the policy optimization of Markov systems by utilizing performance sensitivity information (Cao, 2007; Xia et al., 2014), based on the difference formula and the derivative formula.

With Lemma 2, we can directly derive the optimality of deterministic policies for the long-run CVaR MDP problem.

**Theorem 1.** The minimum CVaR $\tilde{\text{CVaR}}^d$ can be achieved by a deterministic policy.

**Proof.** We assume that $(y^*, d^*)$ is an optimal solution of (20). Obviously, the inner problem $\min_{d \in D} \text{CVaR}^d (y^*)$ is a standard MDP, and its minimum can be achieved by a deterministic policy $d_0$ (Puterman, 1994), that is, we have $\text{CVaR}^d = \text{CVaR}^{d_0} (y^*)$. Meanwhile, $\min_{s \in \mathbb{Y}} \text{CVaR}^d (y)$ must attain its minimum at $y^*$, otherwise it will conflict with the fact that $\text{CVaR}^d (y^*)$ is the minimum of (20). Thus, $y^*$ is the VaR of the MDP under policy $d_0$, and we have $\text{CVaR}^{d_0} (y^*) = \text{CVaR}^d = \text{CVaR}^d_0$. That is, the deterministic policy $d_0$ has the minimum CVaR. The theorem is proved. \qed

Therefore, we can focus only on the deterministic policy space $D_0$. From the viewpoint of sensitivity analysis, we compare the long-run average pseudo-CVaR difference of MDPs under any two policies $d$ and $d'$, where $d$ is the current policy and $d'$ is any other new policy. We can derive the following performance difference formula (refer to Chapter 4.1 of Cao, 2007)

$$
\tilde{\text{CVaR}}^{d'} (y) - \tilde{\text{CVaR}}^d (y) = \sum_{s \in S} \pi^{d'} (s) \left[ \sum_{s' \in S} [p(s'| s, d(s))] g^d (y, s') + \tilde{c}(y, s, d'(s)) - \tilde{c}(y, s, d(s)) \right], \tag{22}
$$

where $g^d (y)$ is a column vector called performance potentials whose element is defined as

$$
g^d (y, s) := \lim_{T \to \infty} \mathbb{E} \left\{ \sum_{t=0}^{T} [c(y, s, a_t) - \tilde{\text{CVaR}}^d (y)] s_0 = s \right\}, \quad s \in S. \tag{23}
$$

In the literature, $g^d (y)$ is also called the bias or relative value function (Puterman, 1994), which can be determined by the following Poisson equation

$$
g^d (y, s) = \tilde{c}(y, s, d(s)) - \tilde{\text{CVaR}}^d (y) + \sum_{s' \in S} p(s'| s, d(s)) g^d (y, s'), \quad s \in S. \tag{24}
$$

By setting $y = \text{VaR}^d$, with (15) and (22), we derive the CVaR difference formula

$$
\text{CVaR}^{d'} - \text{CVaR}^d = \sum_{s \in S} \pi^{d'} (s) \left[ \sum_{s' \in S} [p(s'| s, d'(s))] g^d (\text{VaR}^d, s') - p(s'| s, d(s)) g^d (\text{VaR}^d, s') + \tilde{c}(\text{VaR}^d, s, d'(s)) - \tilde{c}(\text{VaR}^d, s, d(s)) \right] + \text{CVaR}^{d'} - \tilde{\text{CVaR}}^{d'} (\text{VaR}^d), \quad \text{where } y = \text{VaR}^d. \tag{25}
$$

The last term in (25) quantifies the difference between the real CVaR and the pseudo-CVaR of policy $d'$, distorted by policy $d$. For notational simplicity, we define such distortion as

$$
\Delta_{\text{CVaR}} (d', d) := \text{CVaR}^{d'} - \tilde{\text{CVaR}}^{d'} (\text{VaR}^d) \leq 0, \quad \forall d, d' \in D_0, \tag{26}
$$

where the inequality directly follows (16). Therefore, we can rewrite (25) and derive the following lemma about the long-run CVaR difference formula.

**Lemma 3.** If the deterministic policy is changed from $d$ to a new policy $d'$, where $d, d' \in D_0$, then the difference of their long-run CVaR metrics is quantified by

$$
\text{CVaR}^{d'} - \text{CVaR}^d = \sum_{s \in S} \pi^{d'} (s) \left[ \sum_{s' \in S} [p(s'| s, d'(s)) - p(s'| s, d(s))] g^d (\text{VaR}^d, s') \right. \\
\left. + \tilde{c}(\text{VaR}^d, s, d'(s)) - \tilde{c}(\text{VaR}^d, s, d(s)) \right] + \Delta_{\text{CVaR}} (d', d). \tag{27}
$$

**Remark 2.** From the right-hand side of (27), we can see that the first part is the long-run average performance difference for a standard MDP model, where the cost function $\tilde{c}(y, s, a)$ has a constant $y = \text{VaR}^d$ estimated under the current policy $d$. The second part $\Delta_{\text{CVaR}} (d', d)$ is computationally demanding, but its value is always nonpositive. Therefore, we can develop an approach to generate an improved policy $d'$ based
on the above difference formula, which is described by the following theorem.

**Theorem 2.** For the current policy \( d \), if we find a new policy \( d' \in D_0 \) that satisfies

\[
\sum_{s' \in S} p(s' | s, d(s)) g^d(\text{VaR}^d, s') + \tilde{c}(\text{VaR}^d, s, d'(s)) \\
\leq \sum_{s' \in S} p(s' | s, d(s)) g^d(\text{VaR}^d, s') + \tilde{c}(\text{VaR}^d, s, d(s)), \forall s \in S,
\]

(28)

then we have CVaR\(^d\) \( \leq \) CVaR\(^d\). If the above inequality strictly holds for at least one state \( s \), then CVaR\(^d\) \( < \) CVaR\(^d\).

**Proof.** For the two policies \( d \) and \( d' \), we substitute (28) into the difference formula (27)

\[
\text{CVaR}^{d'} - \text{CVaR}^d \leq 0 + \Delta_{\text{CVaR}}(d', d) \leq 0,
\]

where the first inequality uses the fact that \( \pi^d(s) > 0 \) as the MDP is always ergodic, and the second inequality directly follows (26).

If the inequality of (28) strictly holds for at least one state \( s \), we directly have

\[
\text{CVaR}^{d'} - \text{CVaR}^d < 0 + \Delta_{\text{CVaR}}(d', d) \leq 0.
\]

Therefore, the theorem is proved. \( \square \)

Theorem 2 indicates an approach to generate improved policies: we only need to find new policies \( d' \) satisfying (28), where the values of vectors \( g^d(\text{VaR}^d) \) and \( \tilde{c}(\text{VaR}^d) \) are computable or estimatable based on the system sample path of the current policy \( d \). One example of generating improved policies is similar to policy improvement in classical policy iteration:

\[
d'(s) = \arg \min_{a \in A} \left\{ \tilde{c}(\text{VaR}^d, s, a) + \sum_{s' \in S} p(s' | s, a) g^d(\text{VaR}^d, s') \right\},
\]

(29)

\( s \in S \).

We further derive the following theorem that provides a necessary condition for optimal policies of the long-run CVaR MDP.

**Theorem 3.** An optimal deterministic policy \( d^* \) that achieves the minimal long-run CVaR must satisfy the so-called Bellman local optimality equations

\[
d^*(s) = \arg \min_{a \in A} \left\{ \tilde{c}(\text{VaR}^{d^*}, s, a) + \sum_{s' \in S} p(s' | s, a) g^{d^*}(\text{VaR}^{d^*}, s') \right\},
\]

(30)

\( s \in S \).

\[ g^{d^*}(\text{VaR}^{d^*}, s) + \text{CVaR}^{d^*} = \min_{a \in A} \left\{ \tilde{c}(\text{VaR}^{d^*}, s, a) + \sum_{s' \in S} p(s' | s, a) g^{d^*}(\text{VaR}^{d^*}, s') \right\}, \]

(31)

\( s \in S \).

**Proof.** The necessary condition (29) can be directly proved by (28) in Theorem 2. Below, we prove (30) based on the property of performance potentials. With the definition (23), we have

\[
\tilde{c}(\text{VaR}^{d^*}, s, d^*(s)) + \sum_{s' \in S} p(s' | s, d^*(s)) g^{d^*}(\text{VaR}^{d^*}, s').
\]

Substituting (29) into the above equation, we can directly derive

\[
\tilde{c}(\text{VaR}^{d^*}, s) + \text{CVaR}^{d^*} = \lim_{T \to \infty} \left\{ \sum_{t=0}^{T} \tilde{c}(\text{VaR}^{d^*}, s_t, a_t) - \text{CVaR}^{d^*} | s_0 = s \right\}.
\]

Similar to the Poisson equation (24), we can also derive

\[
\tilde{c}(\text{VaR}^{d^*}, s) + \text{CVaR}^{d^*} = \sum_{s' \in S} \tilde{c}(\text{VaR}^{d^*}, s, d^*(s)) + \sum_{s' \in S} p(s' | s, d^*(s)) g^{d^*}(\text{VaR}^{d^*}, s').
\]

We call (29) or (30) the Bellman local optimality equation for long-run CVaR MDPs, which is analogous to the classical Bellman optimality equation for long-run average MDPs, such as

\[
g^*(s) + \eta^* = \min_{a \in A} \left\{ c(s, a) + \sum_{s' \in S} p(s' | s, a) g^*(s') \right\}, \quad s \in S,
\]

(31)

where \( g^*(s) \) and \( \eta^* \) are the optimal value function and the optimal average cost, respectively. However, as opposed to the classical Bellman optimality equation (31) (which is necessary and sufficient for optimal policies), (30) is only necessary and not sufficient for a long-run CVaR optimal policy. This is because the term \( \tilde{c}(\text{VaR}^{d^*}, s, a) \) in (30) depends on the whole policy \( d^* \), while the cost function \( c(s, a) \) in (31) only depends on the current state and action. Another perspective is that the CVaR difference formula (27) has an extra term \( \Delta_{\text{CVaR}}(d', d) \) that does not arise in a standard MDP model.

**Remark 3.** We can use the Bellman local optimality equation (29) or (30) to find optimal policies. However, there may exist multiple fixed-point solutions of CVaR\(^{d^*}\) to (30), while the counterpart \( \eta^* \) for the classical Bellman optimality equation (31) is unique. These multiple solutions can be further recognized as local optima in a mixed policy space defined below.
Next, we discuss the performance derivative of mixed policies. For any two deterministic policies \(d, d' \in D_0\), we define \(d^{\delta,d'}\) as a mixed policy between \(d\) and \(d'\): adopt policy \(d'\) with probability \(\delta\) and adopt policy \(d\) with probability \(1 - \delta\), where \(\delta\) is a mixed probability and \(\delta \in [0, 1]\). Obviously, we have \(d^{0,d} = d\) and \(d^{1,d'} = d'\). By replacing \(d'\) with \(d^{\delta,d'}\) in (27), we compare the long-run CVaR difference of the MDP under policy \(d\) and \(d^{\delta,d'}\). For notational simplicity, we use the superscript “\(\delta\)” to identify the associated quantities of policy \(d^{\delta,d'}\). Thus, (27) can be rewritten as

\[
CVaR^\delta - CVaR^d = \sum_{s \in S} \pi^\delta(s) \delta
\]

\[
\times \left[ \sum_{s' \in S} \left[ p(s'|s,d'(s)) - p(s'|s,d(s)) \right] g^d(VaR^d, s') + \hat{c}(VaR^d, s, d'(s)) - \hat{c}(VaR^d, s, d(s)) \right] + \Delta CVaR(\delta, d).
\] (32)

When \(\delta \to 0\), we can validate that \(\Delta CVaR(\delta, d)\) goes to 0 with higher order of \(\delta\). Therefore, we can derive the following lemma about the long-run CVaR derivative formula, and the detailed proof can be found in the Appendix.

**Lemma 4.** For any two deterministic policies \(d, d' \in D_0\), the derivative of the long-run CVaR with respect to the mixed probability \(\delta\) is

\[
\frac{\partial CVaR^\delta}{\partial \delta} \bigg|_{\delta=0} = \sum_{s \in S} \pi^\delta(s) \left[ \sum_{s' \in S} \left[ p(s'|s,d'(s)) - p(s'|s,d(s)) \right] g^d(VaR^d, s') + \hat{c}(VaR^d, s, d'(s)) \right.
\]

\[
\left. - \hat{c}(VaR^d, s, d(s)) \right].
\] (33)

In the mixed policy space, we give the definition of local minimum of the long-run CVaR MDP problem.

**Definition 2.** A deterministic policy \(d \in D_0\) is locally optimal, if there exists \(\varepsilon > 0\) such that we always have \(CVaR^{d,\delta} \geq CVaR^d\) for any \(\delta \in [0, \varepsilon]\) and \(d' \in D_0\).

With Lemma 4 and Definition 2, we derive the following theorem of locally optimal policies.

**Theorem 4.** The policies \(d^*\) satisfying the optimality equation (29) or (30) are local optima of the long-run CVaR MDP problem in the mixed policy space.

**Proof.** Suppose the current policy \(d^*\) satisfies the optimality equation (29). We choose any new perturbed policy \(d' \in D_0\) and study the derivative of the long-run CVaR with respect to the mixed probability \(\delta\) along with the perturbation direction \((d', d')\):

\[
\frac{\partial CVaR^\delta}{\partial \delta} \bigg|_{\delta=0} = \sum_{s \in S} \pi^\delta(s) \left[ \sum_{s' \in S} \left[ p(s'|s,d'(s)) - p(s'|s,d^*(s)) \right] g^{d^*}(VaR^{d^*}, s') + \hat{c}(VaR^{d^*}, s, d'(s)) - \hat{c}(VaR^{d^*}, s, d^*(s)) \right].
\] (34)

As \(d^*\) satisfies (29), we have

\[
\hat{c}(VaR^{d^*}, s, d^*(s)) + \sum_{s' \in S} p(s'|s,d^*(s)) g^{d^*}(VaR^{d^*}, s') \leq \hat{c}(VaR^{d^*}, s, d'(s)) + \sum_{s' \in S} p(s'|s,d'(s)) g^{d^*}(VaR^{d^*}, s'), \forall a \in A.
\]

Substituting the above inequality into (34), we directly have

\[
\frac{\partial CVaR^\delta}{\partial \delta} \bigg|_{\delta=0} \geq 0,
\]

where we use the fact that \(\pi^\delta(s)\) is always positive. By using the first order of the Taylor expansion of \(CVaR^\delta\) with respect to \(\delta\), we observe that \(CVaR^\delta\) will always increase along with perturbation direction \((d^*, d')\), for any \(d' \in D_0\). Therefore, with Definition 2, we can see that the long-run CVaR at policy \(d^*\) is a local minimum in the mixed policy space.

The necessity of (29) or (30) for locally optimal policies can be directly shown in the same way as the proof of Theorem 3. Therefore, Theorems 3 and 4 indicate that the optimality equation (29) or (30) is necessary and sufficient for locally optimal policies, while only necessary for globally optimal policies. These results can be utilized to develop optimization algorithms for long-run CVaR MDPs.

## 4 | ALGORITHMS

In this section, we develop two algorithms based on the sensitivity analysis and optimization for long-run CVaR MDPs. One is a locally optimal algorithm that is based on the Bellman local optimality equation (29) and executed in a form of policy iteration. The other is a globally optimal algorithm that is based on the sensitivity analysis of the bilevel MDP problem (20).

### 4.1 | A locally optimal algorithm

With Theorems 2–4, we can develop an iterative algorithm to find a locally optimal policy, which satisfies the Bellman
A globally optimal algorithm is provided in Algorithm 1. The details are repeatedly generated based on Theorem 2. The main idea is to solve a standard MDP at the key quantities VaR\(^d\) and g\(^d\)\(\text{VaR}^d\) based on their definitions and (23), respectively.

Algorithm 1 is stated in the following theorem.

**Theorem 5.** Algorithm 1 converges to a local optimum of the long-run CVaR minimization problem.

**Proof.** First, we prove the convergence of Algorithm 1. Substituting (35) into (27), we have

\[
\text{CVaR}^{d^{(l+1)}} - \text{CVaR}^{d^{(l)}} = \sum_{s \in S} \pi^{d^{(l+1)}}(s) \times \left[ \sum_{s' \in S} [p(s'|s,d^{(l+1)}(s)) - p(s'|s,d^{(l)}(s))] g^{d^{(l)}}(\text{VaR}^{d^{(l)}}), d' \right] + \tilde{c}(\text{VaR}^{d^{(l)}}, d^{(l+1)}(s)) - \tilde{c}(\text{VaR}^{d^{(l)}}, s, d^{(l)}(s))
\]

+ \Delta_{\text{CVaR}}(d^{(l+1)}, d^{(l)})

< 0 + \Delta_{\text{CVaR}}(d^{(l+1)}, d^{(l)}) \leq 0.

Therefore, the long-run CVaR of newly generated policy by (35) is strictly reduced. As the deterministic policy space \(D_0\) is finite, Algorithm 1 will terminate after a finite number of iterations. Thus, the convergence is proved.

Next, we prove the local optimum of the convergence. When Algorithm 1 stops, from the stopping rule \(d^{(l)} = d^{(l-1)}\), we can see that the converged policy \(d^{(l)}\) must satisfy the optimality equation (29). With Theorem 4, the converged policy \(d^{(l)}\) is a local optimum in the mixed policy space. Thus, the theorem is proved.

It is known that policy iteration converges at least with a linear convergence rate and even at order 2 (quadratically) to the optimal value function under certain conditions, which partly accounts for the fast convergence of policy iteration in practice (Theorem 6.4.8 in Puterman, 1994, p. 181). As Algorithm 1 is of a form of policy iteration, it is reasonably expected to also have a fast convergence in practice, which is demonstrated by numerical experiments in Section 6. Another further interesting topic is to efficiently estimate the key quantities \(\text{VaR}^{d^{(l)}}\) and \(g^{d^{(l)}}(\text{VaR}^{d^{(l)}})\) based on system sample paths, which makes Algorithm 1 policy learning capable in the framework of reinforcement learning (Sutton & Barto, 2018; Zhou et al., 2022).

### 4.2 A globally optimal algorithm

In this subsection, we use the sensitivity analysis of the bilevel MDP to evaluate the effect of the outer parameter \(y\)'s perturbation on the optimal policy of the inner MDP, which leads to a globally optimal algorithm of the long-run CVaR MDP. The main idea is to solve a standard MDP at the
beginning and then determine the value range of parameter $y$ over which the optimal policy of the inner MDP $\mathcal{M}(y)$ remains unvaried. This is similar to the sensitivity analysis in LP, and we also call such value range the allowable range of $y$. First, we give the definition of critical points.

**Definition 3.** (Critical points) $y^c \in \mathbb{R}$ is called a critical point if there exist $y_1$ and $y_2$ with $y_1 < y^c < y_2$ such that the MDP $\mathcal{M}(y)$ has the same optimal policy for any $y \in [y_1, y_2]$, while this policy is not optimal for MDP $\mathcal{M}(y)$ with $y \in (y^c, y_2]$. We denote the set of critical points as $\mathcal{Y}^c$ and suppose $|\mathcal{Y}^c| = K$. Critical points play an important role in the following sensitivity analysis of the bilevel MDP. First, we characterize some important properties of the optimal pseudo-CVaR function $\overline{\text{CVaR}}^k(y)$, that is, the optimal values of inner MDPs, as stated in Theorem 6.

**Theorem 6.** $\overline{\text{CVaR}}^k(y)$ is a Lipschitz-continuous and piecewise linear function with respect to $y$, which can be further divided into convex segments by critical points in $\mathcal{Y}^c$.

**Proof.** We first prove the continuous property. For any $y_1, y_2 \in \mathbb{R}$, we have

$$
\left| \overline{\text{CVaR}}^k(y_2) - \overline{\text{CVaR}}^k(y_1) \right| = \left| \min_{d \in D_0} \sum_{(s,a) \in S \times A} \pi^d(s,a) \overline{c}(y_1, s, a) - \min_{d \in D_0} \sum_{(s,a) \in S \times A} \pi^d(s,a) \overline{c}(y_2, s, a) \right|
\leq \max_{d \in D_0} \sum_{(s,a) \in S \times A} \pi^d(s,a) \left| \overline{c}(y_1, s, a) - \overline{c}(y_2, s, a) \right|
\leq \max_{d \in D_0} \sum_{(s,a) \in S \times A} \pi^d(s,a) \left[ \left| y_1 - y_2 \right| + \frac{1}{1-\alpha} \left| c(s,a) - y_1 \right| + \left| c(s,a) - y_2 \right| \right]
\leq \left( \frac{1}{1-\alpha} + 1 \right) \left| y_1 - y_2 \right|,
$$

where the first inequality follows from $\min_d h_1(d) - \min_d h_2(d) \leq \max_d |h_1(d) - h_2(d)|$, the third inequality is ensured by $|b_1^+ - b_2^+| \leq |b_1 - b_2|$. When the probability level $\alpha$ is given, the coefficient $\frac{1}{1-\alpha} + 1$ is a constant. Thus, $\overline{\text{CVaR}}^k(y)$ is Lipschitz-continuous in $y$.

To further prove the latter two properties, it is worth noting that the MDPs $\{\mathcal{M}(y) : y \in [y^c, y^{c+1}]\}$ share a same optimal policy $d^*$ for any two adjacent critical points $y^c < y^c'$, that is,

$$
\overline{\text{CVaR}}^k(y) = \overline{\text{CVaR}}^{k'}(y) = y + \frac{1}{1-\alpha} \sum_{(s,a) \in S \times A} \pi^{d^*}(s,a)[c(s,a) - y] + \forall y \in [y^c, y^{c+1}].
$$

It is easy to validate that the term $y + \frac{1}{1-\alpha} \sum_{(s,a) \in S \times A} \pi^d(s,a)[c(s,a) - y] + \forall y \in [y^c, y^{c+1}]$ is piecewise linear and convex on $y \in [y^c, y^{c+1}]$ (Rockafellar & Uryasev, 2000, 2002). Thus, $\overline{\text{CVaR}}^k(y)$ is piecewise linear with respect to $y \in \mathbb{R}$ and can be further divided into convex segments by critical points in $\mathcal{Y}^c$.

Similar to the sensitivity analysis of LP, we develop a method for the sensitivity analysis of the inner MDP $\mathcal{M}(y)$. We aim to find the allowable range of $y$ under which the optimal policy of $\mathcal{M}(y)$ remains unvaried. We sort $\{c(s,a) : (s,a) \in S \times A\}$ in ascending order as $\{c_1, c_2, ..., c_{|S||A|}\}$. If $y \in [c_m, c_{m+1})$, then $1_{\{c(s,a) > y\}} = 1_{\{c(s,a) > c_m\}}$, $\forall (s,a) \in S \times A$, $m = 1, 2, ..., |S||A| - 1$. Therefore, we can rewrite $\overline{c}(y, s, a)$ in (13) as a linear function of $y$ as follows

$$
\overline{c}(y, s, a) = \rho_m(s,a)c(s,a) + y(1 - \rho_m(s,a)),
$$

where $\rho_m(s,a) := \frac{1}{1-\alpha} 1_{\{c(s,a) > c_m\}}$, $(s,a) \in S \times A$. (36)

where $\rho_m(s,a) := \frac{1}{1-\alpha} 1_{\{c(s,a) > c_m\}}$, is an $|S| \times |A|$ dimensional column vector of $1$‘s, and $\odot$ denotes the Hadamard product.

We denote the $k$th critical point as $y^c_k$, $k = 1, 2, ..., K$. For notational convenience, we also denote $y^c_{K+1} := +\infty$, although they are not critical points. For any given $y \in \{y^c_k, y^c_{k+1}\} \cap [c_m, c_{m+1})$, the optimal policy of the inner MDP $\mathcal{M}(y)$ remains unvaried and is denoted as $d^*_k$, that is,

$$
\begin{align*}
\widehat{d}^*_k(y) &\equiv \widehat{d}^*_k(y), \quad \forall y \in \{y^c_k, y^c_{k+1}\}, k = 0, 1, ..., K.
\end{align*}
$$

Substituting (36) into (14), we derive

$$
\overline{\text{CVaR}}^k(y)
= \sum_{(s,a) \in S \times A} \pi^*\overline{c}(y, s, a)
= \sum_{(s,a) \in S \times A} \pi^*\overline{c}(y, s, a)[c(s,a)c(s,a) + y(1 - \rho_m(s,a))].
$$

(39)
Thus, $\text{CVaR}^* (y)$ in (39) is a linear function of $y$ when $\mathbb{P} (Y = y) \in [y_k, y_{k+1}] \cap [c_m, c_{m+1})$. Moreover, $\text{CVaR}^* (y)$ is a convex and piecewise linear function of $y$ when $y \in [y_k, y_{k+1}]$, as also stated by Theorem 6. The number of convex segments is $K + 1$. To visualize these results, we give an illustration curve of $\text{CVaR}^* (y)$ in Figure 1, where $K = 3$, and it has four convex segments and two local optima. Obviously, $\text{CVaR}^* = \min \text{CVaR}^* (y)$ is not a convex optimization problem.

Next, from the viewpoint of sensitivity analysis, we derive a necessary and sufficient condition under which the optimal policy of MDP $\mathcal{M} (y)$ remains the same when $y$ is perturbed.

**Lemma 5.** Let $y$ belong to a given range $[c_m, c_{m+1})$, a policy $d^* \in D_0$ keeps optimal for the inner MDP $\mathcal{M} (y)$, if and only if it satisfies

$$
\begin{align*}
\sum_{y \in [y_1, y_{K+1}] \cap [c_m, c_{m+1})} \mu (y) \left[ (P_{m}^* - P_{m}^d) \cdot Z^{\text{ar}} \cdot P_{m}^{\text{ar}} \right] &\geq \rho_m^* \otimes c^{\text{ar}} - \rho_m^d \otimes c^{\text{ar}} + (P_{m}^{\text{ar}} - P_{m}^d) \\
\cdot Z^{\text{ar}} \cdot P_{m}^{\text{ar}} \otimes c^{\text{ar}} &\quad , \forall d^* \in D_0,
\end{align*}
$$

where $Z^{d} = (I - P^{d} + ee^T)^{-1}$ denotes the fundamental matrix under policy $d$ and $I$ is an identity matrix.

**Proof.** By the optimality condition of undiscounted MDPs (Cao, 2007, p. 187), a policy $d^* \in D_0$ is optimal for the MDP $\mathcal{M} (y)$, if and only if

$$
\begin{align*}
\tilde{c}^* (y) + P_{m}^{\text{ar}} \cdot Z^{\text{ar}} \cdot \tilde{c}^{\text{ar}} (y) &\leq \tilde{c}^d (y) + P_{m}^{\text{ar}} \cdot Z^{\text{ar}} \cdot \tilde{c}^{\text{ar}} (y), \\
\forall d^* \in D_0,
\end{align*}
$$

where $\tilde{c}^{\text{ar}} (y)$ is actually the performance potential $g^{\text{ar}} (y)$ defined in (23) (Xia & Glynn, 2016). Substituting (37) into (41), we can derive (40) by also utilizing the fact that both $P^{d}$ and $Z^{d}$ are stochastic matrices (Schweitzer, 1968).

Lemma 5 gives the necessary and sufficient condition (40) for identifying critical points. For notational simplicity, we rewrite (40) in a component form as below.

$$
y^{\text{T} \mathcal{V}_{1,m}} (s, a) \geq y^{\text{T} \mathcal{V}_{2,m}} (s, a), \quad \forall (s, a) \in S \times A,
$$

with the following definitions

$$
\begin{align*}
\mathcal{V}_{1,m} (s, a) &:= \rho_m (s, d^* (s)) - \rho_m (s, a) \\
&\quad + \sum_{s' \in S} \left[ p (s' | s, d^* (s)) - p (s' | s, a) \right] Z^{d^*} \cdot P^* (s, a), \\
\mathcal{V}_{2,m} (s, a) &:= \rho_m (s, d^* (s)) c (s, d^* (s)) - \rho_m (s, a) c (s, a) \\
&\quad + \sum_{s' \in S} \left[ p (s' | s, d^* (s)) - p (s' | s, a) \right] Z^{d^*} \cdot P^* (s, a) \otimes c^{d^*},
\end{align*}
$$

where $Z^{d^*}$ represents the $s'$th row of the fundamental matrix $Z^{d^*}$. Therefore, with (42), if $\mathcal{V}_{1,m} (s, a) < 0$, we can obtain an upper bound of $y$ as $y \leq \mathcal{V}_{2,m} (s, a) / \mathcal{V}_{1,m} (s, a)$; if $\mathcal{V}_{1,m} (s, a) > 0$, we can obtain a lower bound as $y \geq \mathcal{V}_{2,m} (s, a) / \mathcal{V}_{1,m} (s, a)$.

To obtain all the critical points $y^*$ in $[c_m, c_{m+1})$, we can iteratively conduct the above sensitivity analysis by starting at $c_m$, which is stated as Theorem 7.

![Figure 1: Illustration of piecewise-linear and segment-convex function $\text{CVaR}^* (y)$. Color figure can be viewed at wileyonlinelibrary.com](image-url)
Theorem 7. For any given $y \in [c_m, c_{m+1})$, assume $\tilde{d}^*$ is the optimal policy of the inner MDP $\mathcal{M}(y)$, then the upper bound of the allowable range of $y$ is determined by
\[
y^{(u)} = \begin{cases} 
\min_{(s,a) \in U_m} \left\{ \frac{T_{a,s}^y(s,a)}{U_m} \right\}, & U_m \neq \emptyset, \\
$c_{m+1}$, & U_m = \emptyset,
\end{cases}
\tag{43}
\]
where $U_m = \{(s,a) \in S \times A : \frac{T_{a,s}^y(s,a)}{T_{a,s}^y(s,a)} < 0\}$. Further, $y^{(u)} \in \mathbb{V}^c$ if and only if $y^{(u)} < c_{m+1}$.

Proof. By solving (42) at each state-action pair $(s,a)$ and taking the intersection of $y$ over $(s,a) \in S \times A$, we directly obtain (43). Combined with Definition 3, the theorem is proved.

If $y^{(o)} \in \mathbb{V}^c$, it indicates that the current optimal policy $\tilde{d}^*$ is not optimal anymore for the inner MDP $\mathcal{M}(y)$ if $y > y^{(o)}$. We need to compute a new optimal policy for proceeding the above sensitivity analysis. At point $y^{(o)}$, instead of solving a completely new MDP, we only need to mildly modify the current optimal policy $\tilde{d}^*$ to obtain the new optimal policy $(\tilde{d}^*)^l$.

\[
(\tilde{d}^*)^l(s) = \begin{cases} 
a^*, & s = s^*, \\
\tilde{d}^*(s), & s \neq s^*,
\end{cases}
\tag{44}
\]
where $(s^*, a^*) := \arg\min_{(s,a) \in U_m} \left\{ \frac{T_{a,s}^y(s,a)}{T_{a,s}^y(s,a)} \right\}$. Note that, the above operation can save a lot of computation, compared with solving a completely new MDP $\mathcal{M}(y)$.

Repeating the above sensitivity analysis, we can find all the critical points in $\mathbb{V}^c$. As indicated by Theorem 6 and (39), $\text{CVaR}(y)$ is divided into $K + 1$ convex segments. As the global optimum is achieved at $y \in \mathbb{V}^{(o)}$ by Lemma 2, the minimum of $\text{CVaR}(y)$ on the $k$th convex segment is easily obtained as $\min_{y \in \mathbb{V}^c \cap [c_m, c_{m+1}]} \text{CVaR}(y)$ by using the linear function (39). The global optimum can be obtained by choosing the minimum of $\text{CVaR}(y)$ among all convex segments, that is,
\[
\text{CVaR}^* = \min_{k=0,1,...,K} \min_{y \in \mathbb{V}^c \cap [c_m, c_{m+1}]} \left\{ \text{CVaR}(y) \right\}.
\tag{45}
\]
The CVaR globally optimal policy is $d^* = \tilde{d}^*_{k^*}$, where $k^*$ is the $k$th index that attains the minimum in (45).

Based on the above computing procedure, we develop an algorithm named CVaR Sensitivity Analysis (CVaR-SA), as described in Algorithm 2. The global convergence of this algorithm is guaranteed by Theorem 8.

Theorem 8. The CVaR-SA algorithm converges to the global optimum of the long-run CVaR MDP after a finite number of iterations.

Algorithm 2: CVaR-SA to find the global minimum of long-run CVaR MDPs.
1: Input: MDP parameters $(S, A, P, c)$
2: Output: A globally optimal policy $d^*$ and its minimum CVaR $V^*$
3: Initialization: set $m \leftarrow 1$, $y_1 \leftarrow c_1$, $V^* \leftarrow +\infty$
4: Compute the optimal policy $\tilde{d}^*$ of the inner MDP $\mathcal{M}(y_1)$
5: while $m < |S| \cup A$ do
6: Compute the value of $\text{CVaR}^*(y)$ using (39) and the current $\tilde{d}^*$
7: if $\text{CVaR}^*(y) < V^*$ then
8: $d^* \leftarrow \tilde{d}^*$, $V^* \leftarrow \text{CVaR}^*(y)$
9: end if
10: update $y^{(o)}$ based on (43)
11: while $y^{(o)} < c_{m+1}$ do
12: $y \leftarrow l + 1$, $y_1 \leftarrow c_m$, $\tilde{d}^* \leftarrow \tilde{d}^*_{l-1}$
13: end while
14: $m \leftarrow m + 1, l \leftarrow l + 1, y_1 \leftarrow c_m, \tilde{d}^* \leftarrow \tilde{d}^*_{l-1}$
15: end while
16: return $d^*$ and $V^*$

Proof. To prove the convergence, we only need to show that there is a finite number of critical points in any given range $[c_m, c_{m+1})$. Suppose $d^* \in D_0$ is the optimal policy for both MDPs $\mathcal{M}(y_1)$ and $\mathcal{M}(y_2)$ with $c_m \leq y_1 < y_2 < c_{m+1}$, we prove that $d^*$ is also optimal for MDPs $\{\mathcal{M}(y) : y_1 < y < y_2\}$. Using the optimality condition (40), for any $d' \in D_0$ we have
\[
y_1 \left[ \rho_{m}^{d^*} - \rho_{m}^{d'} + (P^{d^*} - P^{d'}) \cdot Z^{d^*} \cdot \rho_{m}^{d^*} \right] \\
\geq \rho_{m}^{d^*} \circ e^{d^*} - \rho_{m}^{d'} \circ e^{d^*} + (P^{d^*} - P^{d'}) \\
\cdot Z^{d^*} \cdot \rho_{m}^{d^*} \circ e^{d^*},
\]
\[
y_2 \left[ \rho_{m}^{d^*} - \rho_{m}^{d'} + (P^{d^*} - P^{d'}) \cdot Z^{d^*} \cdot \rho_{m}^{d^*} \right] \\
\geq \rho_{m}^{d^*} \circ e^{d^*} - \rho_{m}^{d'} \circ e^{d^*} + (P^{d^*} - P^{d'}) \\
\cdot Z^{d^*} \cdot \rho_{m}^{d^*} \circ e^{d^*}.
\]
For any $y = \lambda y_1 + (1 - \lambda)y_2$ with $0 < \lambda < 1$, by multiplying $\lambda$ and $(1 - \lambda)$ with the above two inequalities, respectively, and summing them, we derive that, $\forall d' \in D_0$,
\[
y[y] \left[ \rho_{m}^{d^*} - \rho_{m}^{d'} + (P^{d^*} - P^{d'}) \cdot Z^{d^*} \cdot \rho_{m}^{d^*} \right] \\
\geq \rho_{m}^{d^*} \circ e^{d^*} - \rho_{m}^{d'} \circ e^{d^*} + (P^{d^*} - P^{d'}) \\
\cdot Z^{d^*} \cdot \rho_{m}^{d^*} \circ e^{d^*},
\]
This completes the proof.
where we utilize the fact that \(1 \{c(s, a) > y\} = 1 \{c(s, a) > c_m\}\) for any \(y \in [c_m, c_{m+1})\). Thus, \(\hat{d}_s^*\) remains optimal for any MDP \(\mathcal{M}(y)\) when \(y_1 \leq y \leq y_2\). As the policy space is finite, the number of critical points among \([c_m, c_{m+1})\) is also finite. Therefore, we show that the number of total critical points among \([c_1, c_{|S|A}]\) is also finite and the CVaR-SA algorithm will stop after a finite number of iterations.

Moreover, we find that the global optimum of Algorithm 2 is directly guaranteed by (45), which is implemented in the algorithm.

Based on the analysis of Theorem 8, we have \(K = 0\) in the best case, and \(K = (|S||A| - 1) \times |A|^{|S|}\) in the worst case. Note that, the upper bound of \(K\) given here is a loose bound and it is difficult to further derive a tight upper bound of \(K\) due to the complicated relationship between the number of \(K\) and the MDPs’ parameters. We compare the CVaR-SA algorithm with a naive method that solves each MDP \(\mathcal{M}(y)\) with \(y \in \gamma^u\) in the following Remark 4, and further empirically show the superiority of the CVaR-SA algorithm over the naive method through Experiment 2 in Section 6.

Remark 4. Note that, the superiority of the CVaR-SA algorithm over the naive method is similar to the role of sensitivity analysis in linear programming. Specifically, the key computation load in the CVaR-SA algorithm is computing the fundamental matrix (inverse matrix) \(Z^{d}_1, Z^{d}_m\) and \(Z^{d}_{2m}\) of (43), which is computationally equivalent to do policy evaluation (obtain the performance potential \(Z^{d}_e\)) or the steady-state distribution \(Z^{d}_e\) (Xia & Glynn, 2016). Solving an MDP with policy iteration needs a number of iterations where the key computation load of each iteration is also the policy evaluation. The iteration complexity of policy iteration is still an open problem and it was observed that the iteration complexity of simple policy iteration (update policy only at one state per iteration) in the worst case is \(|A|^{|S|}\) (Littman et al., 1995). The computation complexity of the naive method in the literature for solving the CVaR MDPs is \(|\gamma^u|\) times of solving an MDP. As a comparison, the computation complexity of our CVaR-SA algorithm is solving one MDP plus \(|\gamma^u|\) times of policy evaluation, which is usually much more efficient than the naive method, as demonstrated by the experiments in Section 6.

5 EXTENSIONS TO MEAN-CVaR OPTIMIZATION

In practice, decision makers usually optimize the return and the risk together. Similar to the mean-variance optimization proposed by Markowitz (1952), we further study the mean-CVaR optimization in the scenario of long-run MDPs. All the results in Sections 3 and 4 can be extended to the combined metric of long-run mean and CVaR.

First, we define a combined cost function considering both the pseudo-CVaR and mean costs as below.

\[
f_\beta(y, s, a) := y + \frac{1}{1 - \alpha} \{c(s, a) - y\}^+ + \beta c(s, a),
\]

\((s, a) \in S \times A, y \in \mathbb{R}^{\|}, \) (46)

where \(\beta \geq 0\) is a coefficient to balance the weights between the pseudo-CVaR and mean costs. Therefore, the long-run average performance of the MDP under the cost function \(f_\beta(y)\) and the policy \(d\) is

\[
\eta^d(f_\beta(y)) = \sum_{(s, a)\in S \times A} \pi^d(s, a)f_\beta(y, s, a) = \text{CVaR}_d^\beta(y) + \beta \eta^d(c),
\]

where \(\text{CVaR}_d^\beta(y)\) is the pseudo-CVaR defined in (14) and \(\eta^d(c) := \sum_{(s, a)\in S \times A} \pi^d(s, a)c(s, a)\) is the mean cost of the MDP. Therefore, the long-run mean-CVaR optimization problem of MDPs is defined as below.

\[
d_\beta^* = \arg \min_{d\in D} \eta^d(f_\beta(y)) = \arg \min_{d\in D} \{ \text{CVaR}_d^\beta(c) + \beta \eta^d(c) \},
\]

\((s, a) \in S \times A, y \in \mathbb{R}^{\|}, \) (47)

which is equivalent to the following bilevel MDP problem

\[
\eta_\beta^* = \min_{d\in D} \min_{y\in \mathbb{R}^{\|}} \{ \eta^d(f_\beta(y)) \}_{y = \text{VaR}^d} = \min_{y\in \mathbb{R}^{\|}} \min_{d\in D} \{ \eta^d(f_\beta(y)) \}
\]

\(\) (48)

where the inner problem is a standard MDP \(\mathcal{M}_2(y) := \langle S, A, P, f_\beta(y) \rangle\) in lieu of \(\mathcal{M}(y)\).

Comparing the mean-CVaR optimization problem (48) and the CVaR minimization problem (17), these two problems are the same except that their cost functions \(f_\beta(y, s, a)\) and \(c(y, s, a)\) are different. Our main results and algorithms in the previous sections are also valid for this mean-CVaR optimization problem except that we have to replace \(c(y, s, a)\)’s with \(f_\beta(y, s, a)\)’s and also the associated quantities. More specifically, the potential function (23) in Section 3 should be replaced as

\[
g_\beta^d(y, s) := \lim_{t \to \infty} \mathbb{E} \left\{ \sum_{i=0}^T \left[ f_\beta(y, s, a_i) - \text{CVaR}_d^\beta(y) - \beta \eta^d(c) \right] s_0 = s \right\},
\]

\(s \in S.\) (49)

Critical points of mean-CVaR optimization can be also identified through the sensitivity analysis of MDPs and
TABLE 1 The iteration processes of Algorithm 1 under different initial policies.

| Iteration | \(d^0\) | VaR | CVaR | Iteration | \(d^0\) | VaR | CVaR |
|-----------|---------|-----|------|-----------|---------|-----|------|
| 0         | [3, 2, 3] | -7  | 4.36 | 0         | [2, 3, 3] | 4.5 | 4.96 |
| 1         | [3, 2, 1] | -7  | 2.41 | 1         | [2, 1, 1] | 1   | 3.22 |
| 2         | [3, 2, 1] | -7  | 2.41 | 2         | [2, 1, 1] | 1   | 3.22 |

Theorem 7

\[ y^{(a)} = \begin{cases} 
\min_{(s,a) \in U_m} \left[ \frac{T_{3,m}^{*}(s,a) + \beta T_{3,m}^{*}(s,a)}{T_{3,m}^{*}(s,a)} \right], & U_m \neq \emptyset, \\
\emptyset, & U_m = \emptyset,
\end{cases} \]

where \( T_{3,m}^{*}(s,a) = c(s, \tilde{d}^*(s)) - c(s, a) + \sum_{s' \in S} p(s'|s, \tilde{d}^*(s)) - p(s'|s, a) \cdot Z^{a}_{p} \cdot e^{a} \). The CVaR-SA Algorithm 2 can also find the global optimum of mean-CVaR optimization by replacing \( \mathcal{M}(y) \) with \( \mathcal{M}_2(y) \). Similarly, the policy iteration type Algorithm 1 can also find a local optimum of mean-CVaR optimization. In the next section, we conduct numerical experiments to demonstrate the effectiveness of our algorithms for mean-CVaR optimization.

6 | NUMERICAL EXPERIMENTS

In this section, we conduct three numerical experiments to validate the performance of our algorithms. Experiment 1 is used to verify that Algorithms 1 and 2 indeed converge to a local optimum and the global optimum, respectively. Experiment 2 illustrates the computational efficiency of these algorithms by comparing them with a naive method as the baseline. Finally, we apply our algorithms to a portfolio management problem in Experiment 3 that minimizes the mean and CVaR of investment losses simultaneously.

Experiment 1. Consider a discrete-time MDP with state space \( S := \{1, 2, 3\} \) and action space \( A := \{1, 2, 3\} \), where the values of transition probability \( P \) and cost function \( c \) are presented in Table A1 of the Appendix. The probability level is set as \( \alpha = 0.65 \).

We use Algorithms 1 and 2 to minimize the long-run CVaR. For the policy iteration Algorithm 1, arbitrarily choosing \( d \in D_0 \) as the initial policy, the algorithm always converges to one of the two locally optimal policies [3, 2, 1] and [2, 1, 1], with CVaR 2.41 and 3.22, respectively. We give two examples of iteration processes in Table 1. For the CVaR-SA Algorithm 2, we find that \( |Y^c| = 7 \) and \( |Y^g| = 3 \). The critical points of \( y^c \) divide CVaR\(^c\) (y) into four convex segments represented by different colored lines in Figure 2a. Each convex segment corresponds to a policy that remains optimal for the inner MDP \( \mathcal{M}(y) \) as y changes. Algorithm 2 starts with \( y = c = -9 \) and stops at \( y = \tilde{c} = 6 \). The minimum of CVaR\(^c\) (y) locates at the blue line, where the globally optimal policy is [3, 2, 1] with CVaR\(^c\) = 2.41, and the corresponding VaR is \( y^c = -7 \). Table A2 of the Appendix shows the iteration process of Algorithm 2.

Besides, we set \( \alpha = 0.95 \) and keep all of other parameters the same to test the influences of different probability levels on the computations. We find that Algorithm 1 always converges to the same optimal policy [2, 1, 1] with CVaR\(^c\) = 4, no matter which initial policy is chosen. It implies that this problem may only have a single local optimum. The results of Algorithm 2 are shown in Figure 2b and Table A2 of the Appendix. We can see that there is only one local optimum in Figure 2b, which is also the global optimum. The corresponding globally optimal policy is \( d^*_g = [2, 1, 1] \) and CVaR\(^c\) = 4 which is larger than CVaR\(^c\) = 2.41 when \( \alpha = 0.65 \). Figure 2b and Table A2 also illustrate that the values of \( \alpha \) do not influence the optimal policy in each convex segment and the values of critical points, which is supported by the fact that the \( \alpha \) in the computation processes of (21) and (43) can be canceled out.

Experiment 2. Consider five discrete-time MDPs with different scales, where the sizes of state and action spaces are set as \(|S_1| = 15, |A_1| = 5, |S_2| = 25, |A_2| = 10, |S_3| = 50, |A_3| = 30, |S_4| = 100, |A_4| = 30, |S_5| = 150, |A_5| = 100\), respectively. The values of transition probability of these MDPs are generated randomly, and the values of cost functions are generated randomly by a uniform distribution \( U(0, n) \) where \( n = 50, 100, 500, 1000, 4500 \) for these five MDPs, respectively. The probability level is set as \( \alpha = 0.95 \).

We adopt a naive method as baseline, which enumeratively solves MDP \( \mathcal{M}(y) \) for all \( y \in Y^c \) and chooses the minimum as CVaR\(^c\). We compare the CVaR performances and the computation times of Algorithms 1, 2, and the naive method through running each algorithm 100 replications with randomly selected initial policies for statistical analysis. These algorithms’ running times are compared by using a computer with an AMD CPU 1.5 GHz, Python 3.6.3, and Linux system. The results are shown in Figure 3 and Table 2, where \( T_{P_j}, T_{SA}, \) and \( T_{NM} \) represent the average computation time (seconds) of Algorithms 1, 2, and the naive method, respectively.

Figure 3 shows the CVaR performances of these three algorithms. We can see that both Algorithm 2 and the naive method reach the global minimum CVaR. Algorithm 1 may, respectively, converge to 12, 9, 8, 4, 3 different local optima for these five MDPs, which are shown by the black whiskers of standard deviations. When the size of MDP increases, Algorithm 1 has less local optima and the performance gaps between local optima and the global optimum become smaller. These phenomena may be understood from the conjecture that local optima will be more difficult to exist when the problem dimension becomes higher, as the direction of descent is more likely to be found when fixing a dimension. Thus, a local optimum is usually good enough in many practical problems, which is also often observed by
observe that |较好| algorithms. From Table 2, we can see that Algorithm 1 has viewed at wileyonlinelibrary.com

FIGURE 2 The curve of $\tilde{\text{CVaR}}^*$ ($y$) with $\tilde{d}^*_0 = [3, 1, 1], \tilde{d}^*_1 = [3, 2, 1], \tilde{d}^*_2 = [2, 2, 1], \tilde{d}^*_3 = [2, 1, 1]$. (a) $\alpha = 0.65$. (b) $\alpha = 0.95$. [Color figure can be viewed at wileyonlinelibrary.com]

FIGURE 3 CVaR-performance comparison of Algorithms 1, 2, and the naive method. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Computation time (seconds) comparison of Algorithms 1, 2, and the naive method.

| $|S|$ | $|A|$ | $|Y^a|$ | $|Y^r|$ | $T_{PT}$ | $T_{SA}$ | $T_{NM}$ |
|------|------|-------|-------|--------|--------|--------|
| 15   | 5    | 75    | 6     | 0.001  | 0.02   | 0.06   |
| 25   | 10   | 250   | 24    | 0.004  | 0.06   | 0.24   |
| 50   | 30   | 1500  | 95    | 0.27   | 5.37   | 41.40  |
| 100  | 30   | 3000  | 181   | 1.26   | 25.18  | 328.34 |
| 150  | 100  | 14,994| 336   | 10.46  | 141.21 | 7405.98|

practical applications of deep learning that also only find local optima.

Table 2 compares the computation efficiency of these algorithms. From Table 2, we can see that Algorithm 1 has the fastest computation speed. For Algorithm 2, we always observe that $|Y^r| \ll |Y^a|$. The computation time of Algorithm 2 is 3, 4, 8, 13, 52 times less than that of the naive method for these five MDPs, respectively. Such speed advan-

tage of Algorithm 2 becomes more significant as the problem size further increases. This experiment also demonstrates that Algorithm 2 usually converges fast in practice, although its theoretical complexity is not guaranteed. Finally, we use our algorithms to solve a portfolio management problem that minimizes both the mean and CVaR of investment losses.

Experiment 3. Consider a simplified version of portfolio management problem, where the investor’s initial wealth is $10,000. The investor needs to choose the wealth proportion allocated to a risk-free asset (asset 0) and two risky assets (stock 1, 2) at each investment time $t$. The state is defined as $s_t := (x_{1,t}, x_{2,t}, w_{1,t}, w_{2,t})$, where $x_{k,t} \in \{0, 1\} (k = 1, 2)$ represents the stock’s current economic state and $w_{k,t} \in \{0.2, 0.5, 0.8\} (k = 1, 2)$ is the holding proportion of stock $k$ at time $t$. The evolution of each stock’s economic state is supposed to be characterized by a Markov process, while the economic state transitions of the two stocks are independent of each other. The values of transition probabilities are described in Table A3 of the Appendix. The action space is $A := \{a_t = (a_{1,t}, a_{2,t}) : \sum_{k=1}^{2} a_{k,t} \leq 1, a_{k,t} \in \{0, 0.2, 0.5, 0.8\}\}$, where $a_{k,t}$ determines the wealth proportion allocated to stock $k$ at the next time $t+1$, that is, $w_{k,t+1} = a_{k,t} (k = 1, 2)$. The remaining wealth will be allocated to the risk-free asset, that is, $w_{0,t+1} = 1 - \sum_{k=1}^{2} w_{k,t+1}$. Hence, we have $|S| = 24$, $|A| = 6$ and the size of policy space $|D_0| = |A|^{|S|} = 6^{24}$ in this problem. The risk-free asset has a fixed return $r_f = 0.02\%$, while the returns of these two stocks $r_{k}(x_{k,t}) (k = 1, 2)$ are given specifically in Table A3 of the Appendix. The system instantaneous cost can be computed as

$$c(s_t, a_t, s_{t+1}) := -\left\{ r_f (1 - \sum_{k=1}^{2} a_{k,t}) + \sum_{k=1}^{2} r_k(x_{k,t+1}) a_{k,t} - b \sum_{k=1}^{2} |a_{k,t} - w_{k,t}| \right\},$$
where \( b = 0.05\% \) is the transaction cost per unit wealth. The investor aims to find an optimal investment policy \( d^* \) in (47) to minimize both the mean and CVaR of investment losses, and we set \( \beta = 6, \alpha = 0.85 \).

First, we use Algorithm 1 to solve this problem. Through traversing \( d \in D_B \) as the initial policy of Algorithm 1, we obtain three different locally optimal policies, and we give three associated iteration processes for instance in Table 3. Note that, the local optimum obtained by the third iteration process in Table 3 is also the global optimum, the first iteration process is exactly the convergence result when solely optimizing the mean. We also adopt Algorithm 2 as a comparison. Through computation, we find that \( |Y^\alpha| = 97 \) and \( |Y^\beta| = 48 \). The minimal mean-CVaR combined metric is \( \eta^* = -228.65 \) and the optimal policy \( d^*_G \) is shown in Table A4 of the Appendix.

We also compare the distributions of the system cost under the globally optimal policy \( d^*_G \), the locally optimal policy \( d^*_L \) obtained by the second iteration process in Table 3, the mean-optimal policy \( d^*_M \) that optimizes the mean only and a naive investment policy (chooses actions randomly at different states) \( d_N \), whose Gauss fitted curves are illustrated in Figure 4. The performance of \( d^*_G, d^*_L, d^*_M, d_N \) and the corresponding specific policies are shown in Table 4 as follows and Table A4 of the Appendix, respectively.

Figure 4 demonstrates that the right tail of the system cost distribution under the globally optimal policy is thinner than that under the naive investment policy and the mean-optimal policy. The system cost distribution under the mean-optimal policy has the heaviest right tail, which indicates that optimizing the mean merely may cause extreme losses. Moreover, as shown in Table 4, the globally optimal policy has the minimal mean-CVaR weighted value, which also demonstrates the effectiveness of our algorithms for mean-CVaR optimization.

### 7 DISCUSSION AND CONCLUSION

In this paper, we study long-run CVaR optimization in the framework of MDPs. Because of the non-Markovian CVaR cost function, this dynamic optimization problem is not a standard MDP and traditional dynamic programming is not applicable. We study this problem by using sensitivity-based optimization. The long-run CVaR difference formula and derivative formula are both derived. Moreover, a so-called Bellman local optimality equation and other properties of optimal policies are obtained. With the CVaR sensitivity formulas, we further develop a policy iteration type algorithm to minimize long-run CVaR, and its local convergence is proved. With the sensitivity analysis of the bilevel MDP formulation, we further develop a globally optimal algorithm based on the computation of critical points. The property of piecewise linear and segment convex of the optimal pseudo-CVaR function is also established. Our main results and algorithms are extended to mean-CVaR optimization.

Along with this research direction, it is valuable to further study other research topics. One topic is to extend the single objective of CVaR to multiple objectives, such as CVaR, mean, or their ratio metrics. Another topic is to extend our CVaR policy iteration type algorithm to other forms, such as value iteration or policy gradient type algorithms for CVaR, which are widely used in reinforcement learning for discounted criteria. Combining the techniques of sampling efficiency and neural network approximation for model-free scenarios is also a promising direction. The Bellman local optimality equation and the CVaR sensitivity formulas may play important roles in the study of these topics.

### ACKNOWLEDGMENTS

The authors are grateful to the valuable comments raised by the reviewers and the editors, which greatly improved the paper. This work was supported in part by the National Key Research and Development Program of China (2022YFA1004600), the National Natural Science Foundation of China (72371253, 62073346, 11931018, U1811462),

#### TABLE 3 The iteration processes of Algorithm 1 with different initial policies.

| Iteration # | \( \eta^\beta \) | Iteration # | \( \eta^\beta \) | Iteration # | \( \eta^\beta \) |
|-------------|----------------|-------------|----------------|-------------|----------------|
| 0           | 47.88          | 0           | -217.31        | 0           | -50.01         |
| 1           | -189.34        | 1           | -217.31        | 1           | -228.65        |
| 2           | -189.34        | -           | -             | 2           | -228.65        |

#### TABLE 4 The performance of different policies.

| Metric | \( d^*_G \) | \( d^*_L \) | \( d^*_M \) | \( d_N \) |
|--------|-------------|-------------|-------------|-----------|
| \( \eta^\beta \) | -228.65     | -217.31     | -189.34     | 46.76     |
| Mean   | -78.99      | -59.55      | -98.43      | -49.17    |
| CVaR   | 245.29      | 139.99      | 401.24      | 341.78    |
| VaR    | 178         | 140         | 376         | 259       |
the Guangdong Basic and Applied Basic Research Foundation (2023A1515012492, 2021A1515011984), the Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2020B1212060032), the Fundamental Research Funds for the Central Universities at the Sun Yat-sen University (22wklj01). The work described in this paper was also partially supported by InnoHK initiative, The Government of the HKSAR, and Laboratory for AI-Powered Financial Technologies.

**REFERENCES**

Aguirregabiria, V., & Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1), 38–67.

Ahmadi-Javid, A., & Malhamé, R. (2015). Optimal control of a multistate failure-prone manufacturing system under a conditional value-at-risk cost criterion. *Journal of Optimization Theory and Applications*, 167, 716–732.

Alexander, S., Coleman, T. F., & Li, Y. (2006). Minimizing CVaR and VaR for portfolio of derivatives. *Journal of Banking Finance*, 30, 583–605.

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9, 203–228.

Asensio, M., & Contreras, J. (2016). Stochastic unit commitment in isolated systems with renewable penetration under CVaR assessment. *IEEE Transactions on Smart Grid*, 7(3), 1356–1367.

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

Bertsekas, D. P. (2005). *Dynamic programming and optimal control* (Vol. I). Athena Scientific.

Boda, K., & Filar, J. (2006). Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63, 169–186.

Bonetti, M., Bisi, L., & Restelli, M. (2023). Risk-averse optimization of reward-based coherent risk measures. *Artificial Intelligence*, 316, 103845.

Cao, X. R. (2007). Stochastic learning and optimization: A sensitivity-based approach. Springer.

Chung, K. J. (1994). Mean-variance tradeoffs in an undiscounted MDP: The unichain case. *Operations Research*, 42, 184–188.

Dixit, V., Verma, P., & Tiwari, M. K. (2020). Assessment of pre and post-disaster supply chain resilience based on network structural parameters with CVaR as a risk measure. *International Journal of Production Economics*, 227, 107655.

Ephremides, A., & Verdu, S. (1989). Control and optimization methods in communication network problems. *IEEE Transactions on Automatic Control*, 34(9), 930–942.

Feinberg, E. A., & Shwartz, A. (Eds.). (2002). *Handbook of Markov decision processes: Methods and applications*. Kluwer.

Fu, M. C., Hong, L. J., & Hu, J. Q. (2009). Conditional Monte Carlo estimation of quantile sensitivities. *Management Science*, 55(12), 2019–2027.

Haskell, W. B., & Jain, R. (2015). A convex analytic approach to risk-aware Markov decision processes. *SIAM Journal on Control and Optimization*, 53(3), 1569–1598.

Hernández-Lerma, O., & Lasserre, J. B. (2012). *Discrete-time Markov control processes: Basic optimality criteria*. Springer Science & Business Media.

Hong, L. J., Hu, Z., & Liu, G. (2014). Monte Carlo methods for value-at-risk and conditional value-at-risk: A review. *ACM Transactions on Modeling and Computer Simulation*, 24, 1–37.

Huang, Y., & Guo, X. (2016). Minimum average value-at-risk for finite horizon semi-Markov decision processes in continuous time. *SIAM Journal on Optimization*, 26(1), 1–28.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237–285.

Lewis, F. L., Vrabie, D. L., & Syrmos, V. L. (2012). *Optimal control* (3rd ed.). John Wiley & Sons.

Li, B., & ArreolaRisa, A. (2022). Minimizing conditional value-at-risk under a modified basestock policy. *Production and Operations Management*, 31(4), 1822–1838.

Li, P., Yu, D., Yang, M., & Wang, J. (2018). Flexible look-ahead dispatch realized by robust optimization considering CVaR of wind power. *IEEE Transactions on Power Systems*, 33(5), 5330–5340.

Littman, M. L., Dean, T. L., & Kaelbling, L. P. (1995). On the complexity of solving Markov decision problems. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., Burlington, MA.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7, 77–91.

Miller, C. W., & Yang, I. (2017). Optimal control of conditional value-at-risk in continuous time. *SIAM Journal on Control and Optimization*, 55(2), 856–884.

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

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 3(3), 21–42.

Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking Finance*, 26, 1443–1471.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55, 999–1033.

Schweitzer P. J. (1968). Perturbation theory and finite Markov chains. *Journal of Applied Probability*, 5(2), 401–413.

Sobel, M. J. (1994). Mean-variance tradeoffs in an undiscounted MDP. *Operations Research*, 42, 175–183.

Stidham, S., & Weber, R. (1993). A survey of Markov decision models for control of networks of queues. *Queueing Systems*, 13, 291–314.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Tijms, H. C. (1986). *Stochastic modelling and analysis: A computational approach*. John Wiley & Sons, Inc.

Xia, L. (2020). A tutorial on event-based optimization: A new optimization framework. *Discrete Event Dynamic Systems: Theory and Applications*, 24(2), 103–132.

Xia, L., & Glynn, P. W. (2016). A generalized fundamental matrix for computing fundamental quantities of Markov systems. arXiv:1604.04343.

Xie, Y., Wang, H., & Lu, H. (2018). Coordination of supply chains with a retailer under the mean-CVaR criterion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(7), 1039–1053.

Ügürlu, K. (2017). Controlled Markov decision processes with AVaR criteria for unbounded costs. *Journal of Computational and Applied Mathematics*, 319, 24–37.

Zhou, Z., Athey, S., & Wager, S. (2022). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1), 148–183.

**APPENDIX**

**Proof of Lemma 4.** With the CVaR difference formula (32) for deterministic policy $\delta$ and mixed policy $\delta^d,\delta^d$, we can see
that the derivative is written as below.

\[
\frac{\partial \text{CVaR}^\delta}{\partial \delta} \bigg|_{\delta = 0} = \lim_{\delta \to 0} \frac{\text{CVaR}^\delta - \text{CVaR}^d}{\delta}
\]

\[
= \lim_{\delta \to 0} \sum_{s \in S} \pi^\delta(s) \left[ \sum_{s' \in S} [p(s'|s, d(s)) - p(s'|s,d(s)))]g^d(\text{VaR}^d, s') + \tilde{c}(\text{VaR}^d, s, d(s))
\]

\[
- \tilde{c}(\text{VaR}^d, s, d(s)) \right] + \lim_{\delta \to 0} \frac{\Delta_{\text{CVaR}}(\delta, d)}{\delta}
\]

\[
= \sum_{s \in S} \pi^d(s) \left[ \sum_{s' \in S} [p(s'|s, d(s))]
\]

\[
- p(s'|s,d(s))]g^d(\text{VaR}^d, s') + \tilde{c}(\text{VaR}^d, s, d(s))
\]

\[
- \tilde{c}(\text{VaR}^d, s, d(s)) \right] + \lim_{\delta \to 0} \frac{\Delta_{\text{CVaR}}(\delta, d)}{\delta}, \quad (A1)
\]

where we use the fact that \(\lim_{\delta \to 0} \pi^\delta(s) = \pi^d(s)\) because \(d_{\delta,d}^{\delta,d} = d\). Below, we study the last term in the above equation. From the definition (26), we have

\[
\Delta_{\text{CVaR}}(\delta, d) = \text{CVaR}^\delta - \text{CVaR}^d (\text{VaR}^d)
\]

Substituting (13) into the above equation, we have

\[
\Delta_{\text{CVaR}}(\delta, d) = \text{VaR}^\delta - \text{VaR}^d + \frac{1}{1 - \alpha} \sum_{s \in S} \pi^\delta(s, a)
\]

\[
\times [c(s, a) - \text{VaR}^\delta]^+ - [c(s, a) - \text{VaR}^d]^+.
\]

(A2)

Without loss of generality, we assume \(\text{VaR}^\delta > \text{VaR}^d\). We further define the following sets of state-action pairs

\[
H^+ := \{(s,a) \in S \times A : c(s,a) \geq \text{VaR}^\delta\},
\]

\[
H^- := \{(s,a) \in S \times A : c(s,a) \leq \text{VaR}^d\},
\]

\[
H^+ := \{(s,a) \in S \times A : c(s,a) > \text{VaR}^\delta\}.
\]

(A3)
Substituting \((A3)\) into \((A2)\), we have
\[
\Delta_{\text{CVaR}}(\delta, d) = \text{VaR}^\delta - \text{VaR}^d + \frac{1}{1 - \alpha} \times \left\{ \sum_{s, a \in H^+} \pi^\delta(s, a)[c(s, a) - \text{VaR}^\delta - c(s, a) + \text{VaR}^d] \right. \\
+ \left. \sum_{s, a \in H^-} \pi^\delta(s, a)[\text{VaR}^\delta - c(s, a) - c(s, a)] + \text{VaR}^d \right\} \\
+ \left\{ \sum_{s, a \in H^+} \pi^\delta(s, a)[\text{VaR}^\delta - 2c(s, a) + \text{VaR}^d] \right\} \\
= \text{VaR}^\delta - \text{VaR}^d - \frac{1}{1 - \alpha} \sum_{s, a \in H^+} \pi^\delta(s, a) \\
+ \frac{1}{1 - \alpha} \sum_{s, a \in H^-} \pi^\delta(s, a)[\text{VaR}^\delta - 2c(s, a) + \text{VaR}^d] \\
= \frac{1}{1 - \alpha} \sum_{s, a \in H^+} \pi^\delta(s, a)[\text{VaR}^\delta - 2c(s, a) + \text{VaR}^d],
\]
where the last equality utilizes the fact that \(\sum_{s, a \in H^+} \pi^\delta(s, a) = 1 - \alpha\). From \((A3)\), we can see that
\[H^+ \rightarrow \Phi, \quad \text{when} \ \delta \rightarrow 0. \quad (A5)\]

With the Mean Value Theorem, we also have
\[
\text{VaR}^\delta - 2c(s, a) + \text{VaR}^d \rightarrow 0, \quad \text{when} \ \delta \rightarrow 0, (s, a) \in H^+. \quad (A6)
\]

Substituting \((A5)\) and \((A6)\) into \((A4)\), we have
\[
\frac{\partial \Delta_{\text{CVaR}}(\delta, d)}{\partial \delta} \bigg|_{\delta=0} = \frac{1}{1 - \alpha} \left\{ \frac{\partial \text{VaR}^\delta}{\partial \delta} \sum_{s, a \in H^+} \pi^\delta(s, a) \bigg|_{\delta=0} \\
+ \sum_{s, a \in H^+} \frac{\partial \pi^\delta(s, a)}{\partial \delta}[\text{VaR}^\delta - 2c(s, a) + \text{VaR}^d] \bigg|_{\delta=0} \right\} \\
= \frac{1}{1 - \alpha} \left\{ 0 + 0 \right\} = 0. \quad (A7)
\]

Therefore, substituting the above Equation \((A7)\) into \((A1)\), we directly have
\[
\frac{\partial \text{CVaR}^\delta}{\partial \delta} \bigg|_{\delta=0} = \sum_{s \in S} \pi^\delta(s) \left[ \sum_{s' \in S} [p(s' | s, d(s)) - p(s' | s, d(s))] \\
\times g^d(\text{VaR}^d, s') + \tilde{c}(\text{VaR}^d, s, d(s)) - \tilde{c}(\text{VaR}^d, s, d(s))] \right].
\]

Then Lemma 4 is proved. \(\square\)