CVaR Optimization for MDPs: Existence and Computation of Optimal Policies

Rui Ding Eugene Feinberg
Applied Mathematics and Statistics Department
Stony Brook University Stony Brook, NY 11794
{rui.ding.1, eugene.feinberg}@stonybrook.edu

ABSTRACT

We study the problem of Conditional Value-at-Risk (CVaR) optimization for a finite-state Markov Decision Process (MDP) with total discounted costs and the reduction of this problem to a stochastic game with perfect information. The CVaR optimization problem for finite and infinite-horizon MDPs can be reformulated as a zero-sum stochastic game with a compact state space. This game has the following property: while the second player has perfect information including the knowledge of the decision chosen by the first player at the current time instance, the first player does not directly observe the augmented component of the state and does not know current and past decisions chosen by the second player. Using methods of convex analysis, we show optimal policies exist for this game and an optimal policy of the first player optimizes CVaR of the total discounted costs. In addition to proving existence of optimal policies, we provide algorithms for their computation.

1. INTRODUCTION

Decision making in the Markov decision process (MDP) framework is typically done in a risk-neutral setting, where the objective is either the expected total discounted cost/reward or average cos/reward per unit time. In recent works, CVaR objective has gained considerable interest in MDPs because it takes risk into an account. In particular [1] considered using the original CVaR functional on the discounted sum of the costs, while [4] suggested a nested reformulation of the CVaR functional. The reason for doing so is to obtain time consistency and decomposability of the resulting risk functional in the original state space. On the other hand, [1] wrote a CVaR minimax equation on the risk level augmented state space, using the time consistent CVaR decomposition theorem from [2].

Conditional Value-at-Risk (CVaR) is a coherent risk function that is widely used in engineering and finance. For a random variable Z defined on a probability space $(\Omega, \mathcal{F}, P)$, its CVaR for tail risk level $\alpha \in [0,1]$ is a conditional tail expectation: $CVaR_\alpha(Z) := E[Z|Z \geq VaR_\alpha(Z)]$, where the Value-at-Risk is $Var_\alpha(Z) := \min\{z : F_Z(z) \geq 1-\alpha\}$, and $F_Z(z) = P\{Z \leq z\}$ is the distribution function of Z. For a σ-algebra $\mathcal{G}$ on Ω such that $\mathcal{G} \subset \mathcal{F}$, the Pflug and Pichler CVaR decomposition theorem [2, lemma 22] states that, for a constant $\alpha \in [0,1]$, $CVaR_\alpha(Y) = \sup\{E[\xi \cdot CVaR_{\alpha\xi}(Y|\mathcal{G}) : \xi \text{ is } \mathcal{G}-\text{measurable}, \ 0 \leq \alpha\xi \leq 1, \text{ and } E[\xi] = 1\}$.

This paper deals with CVaR optimization for MDPs. An MDP is a tuple $(\mathbb{X}, \mathbb{A}, A(\cdot), c, p)$, where $\mathbb{X}$ is a set of states space, $\mathbb{A}$ is a set of actions space, and for each state $x \in \mathbb{X}$ there is a nonempty set of available actions $A(x) \subset \mathbb{A}$, and, c is a cost function, and p is a transition probability. The time $t = 0, 1, ...$ is discrete, and if at some time instance an action $a \in A(x)$ is selected at a state $x \in \mathbb{X}$, then the system moves to the next state according to $x' \sim p(\cdot|x,a)$ and the cost $c(x,a,x')$ is collected. Let $\beta \in [0,1]$ be a constant discount factor. An initial stat x and a policy π define a probability measure P_x^π on the space of trajectories $((\mathbb{X} \times \mathbb{A})^\infty, \mathcal{B}((\mathbb{X} \times \mathbb{A})^\infty))$, where $\mathcal{B}$ demotes a Borel σ-algebra. For a finite horizon N the total discounted cost is the random variable $Z_N = \sum_{t=0}^{N-1} \beta^t c(x_t, a_t) + \beta^N v_0(x_N)$ and for infinite-horizon it is $Z = \sum_{t=0}^{\infty} \beta^t c(x_t, a_t)$ where $\beta < 1$. In this work we study minimizing objective criteria $CVaR_\alpha(Z_N; x, \pi)$ for N-horizon problems and $CVaR_\alpha(Z; x, \pi)$ for infinite-horizon problems, where $\alpha \in [0,1]$, and $(\Omega, \mathcal{F}, P) = (\Omega, \mathcal{F}, P_x^\pi)$.

2. RISK-AUGMENTED STATE FORMULATION OF CVAR MDP

We consider CVaR optimization in MDP for finite sets $\mathbb{X}$ and $\mathbb{A}$ of states and actions respectively. Consider a finite N-horizon problem. The objective of CVaR MDP is to solve, for a given risk level $\alpha \in [0,1]$ and an initial state $x_0 \in \mathbb{X}$,

$$\min_{\pi \in \Pi_H} CVaR_\alpha(Z_N; x_0, \pi) \tag{1}$$

where Π_H is the class of nonrandomized history-dependent policies. Without loss of generality, we set $\mathbb{X} = \{1, 2, \ldots, M\}$. Following [2] and [3], define the CVaR risk envelope as: $\mathcal{U}_{CVaR}(\alpha, p(\cdot|x,a)) = \{b \in R^M : 0 \leq \alpha b_{x'} \leq 1, \ x' =$

$1, 2, \ldots, M$, $\sum_{x'=1}^{M} b_{x'} p(x'|x,a) = 1\}$, a convex polytope which can be replaced with the set: $B(x,\alpha,a) = \mathcal{U}_{CVaR}(\alpha, p(\cdot|x,a)) \cap \{b \in R^M : b_{x'} = 0 \text{ if } p(x'|x,a) = 0\}$.

Chow et al. [1] considered a state augmentation of the original CVaR MDP problem where the augmented state $(x,y) \in \mathbf{X}$ ($\mathbf{X} = \mathbb{X} \times [\mathbf{0},\mathbf{1}]$) consists of an original physical state $x \in \mathbb{X}$ and an assigned risk state $y \in [0,1]$ which indicates the tail risk level of CVaR at any particular subproblem. Hence for the initial stage problem we have the augmented state $(x_0, y_0 = \alpha)$. The state-augmented CVaR MDP is equivalent to a robust MDP on this augmented state space as a result of the CVaR decomposition theorem in [2]. Let us define $\mathcal{V}_N((x,y),\pi) = y \cdot CVaR_y(Z_N; x, \pi)$ the scaled value functions for a fixed policy π in this problem and let $\mathcal{V}_N(x,y)$ denote the scaled value functions for this problem which are attained by the optimal policy, $\forall (x,y) \in \mathbf{X}$. Results from [1] can be extended to show that the scaled value functions satisfy the following optimality equations, where we introduce the scaled Q-factor functions:

$$Q_{n+1}((x,y),a) = \max_{b \in B(x,y,a)} \sum_{x'=1}^{M} (y b_{x'} c(x,a,x') + \beta \mathcal{V}_n(x', y b_{x'})) p(x'|x,a), \forall (x,y) \in \mathbf{X}, a \in A(x), \quad (2)$$

and optimality equation for the finite horizon problem is:

$$\mathcal{V}_{n+1}(x,y) = \min_{a \in A(x)} \{Q_{n+1}((x,y),a)\}, \forall (x,y) \in \mathbf{X}, \quad (3)$$

for $n = 0, 1, \ldots$, with $\mathcal{V}_0(x,y) = y v_0(x), \forall y \in [0,1]$. The optimal action sets can be defined (when there are n stages left in a finite horizon problem where $n > 0$) as $A_n^*(x,y) = \{a \in A(x) | Q_n((x,y),a) = \mathcal{V}_n(x,y)\}, \forall x \in \mathbb{X}, y \in [0,1], n = 1, 2, \ldots$. Similarly, for the infinite horizon case, we have,

$Q((x,y),a) = \max_{b \in B(x,y,a)} \sum_{x'=1}^{M} (y b_{x'} c(x,a,x') + \beta \mathcal{V}(x', y b_{x'})) p(x'|x,a), \forall (x,y) \in \mathbf{X}, a \in A(x)$,

$\mathcal{V}(x,y) = \min_{a \in A(x)} \{Q((x,y),a)\}, \forall (x,y) \in \mathbf{X}$,

and infinite horizon optimal action sets are defined as $A^*(x,y) = \{a \in A(x) | Q((x,y),a) = \mathcal{V}(x,y)\}, \forall x \in \mathbb{X}, y \in [0,1]$.

Equations (2) and (3) are relevant to the robust MDP $(\mathbf{X}, \mathbb{A}, A(\cdot), \mathbb{B}, B(\cdot), \mathbf{c}, \mathbf{p})$, where the state space $\mathbf{X} = \mathbb{X} \times [\mathbf{0},\mathbf{1}]$, the action set $\mathbb{A}$ of player I is the set of actions in the original MDP, for which the sets nonempty finite sets $A(x)$ are defined. For the robust MDP, we define $A(\mathbf{x}) = A(x,y) := A(x)$ for all $\mathbf{x} = (x,y) \in \mathbb{X} \times [0,1]$. The space $\mathbb{B} = \mathbb{R}^M$. For each $\mathbf{x} = (x,y) \in \mathbf{X}$ the action set fot player II is $B(\mathbf{x},a) := B(x,y,a)$. If at state $(x,y) \in \mathbf{X}$ player I chooses an action $a \in A(x)$, and player II chooses an action $b \in B(x,y,a)$, then the next state will be $(x', y b'_x)$ and the cost incurred will be $\mathbf{c}((x,y,x'),a,b) = y c(x,a,x')$ with the transition probability $\mathbf{p}((x', y b'_x)|(x,y),a,b) := b_{x'} p(x'|x,a)$, where $x' \in \mathbb{X}$. Optimality equations (2), (3) recursively define the scaled value functions in this robust MDP, which is equal to the optimal tail risk level scaled CVaR value of the MDP problem in (1). The problem becomes a worst-case robust MDP if $y_0 = \alpha = 0$, and reduces to a risk neutral MDP if $y_0 = \alpha = 1$, where classic value iteration algorithm can be directly applied.

In the general case when $\alpha \in (0,1)$, the CVaR tail risk level in the augmented MDP problem is explicitly known only at the initial state. At the following steps it is changing and may not be directly observed. In the robust MDP, the player II controls risk, but there is no player II in the initial problem formulation. To resolve this issue, we introduce the following definition. A policy π for player I in the robust MDP $(\mathbf{X}, \mathbb{A}, A(\cdot), \mathbb{B}, B(\cdot), \mathbf{c}, \mathbf{p})$ is called autonomous if for each $t = 1, 2, \ldots$ for each two histories $h_t^{(i)} = x_0^{(i)}, y_0^{(i)}, a_0^{(i)}, b_0^{(i)}, \ldots, x_t^{(i)}, y_t^{(i)} \in H_t$, $i = 1, 2$, such that $y_0^{(1)} = y_0^{(2)}$, $x_n^{(1)} = x_n^{(2)}$ for $n = 0, \ldots, t$, and $a_n^{(1)} = a_n^{(2)}$ for $n = 0, \ldots, t-1$, then $\pi_t(\cdot|h_t^{(1)}) = \pi_t(\cdot|h_t^{(2)})$. The following theorem links the CVaR MDP with the robust MDP.

THEOREM 2.1. *There exists a nonrandomized autonomous optimal policy π^* minimizing CVaR of the total discounted costs for an MDP with a finite or infinite horizon over the class of nonrandomized policies. In addition, $y \cdot CVaR_y(Z_n; x, \pi^*) = \mathcal{V}_n(x,y), \forall n = 1, 2, \ldots, N$ and $y \cdot CVaR_y(Z; x, \pi^*) = \mathcal{V}(x,y), \forall (x,y) \in \mathbf{X}$.*

3. CVAR VALUE ITERATION: SLOPE CHARACTERIZATION

Lemmas below characterize the scaled value functions.

LEMMA 3.1. *$\mathcal{V}_n(x,y)$ and $Q_{n+1}((x,y),a)$ are concave and monotonically nondecreasing functions of $y \in [0,1]$ for every $n = 0, 1, \ldots$, state $x \in \mathbb{X}$ and action $a \in A(x)$. Moreover, if state space and action space are finite, then these functions are piecewise linear with finitely many linear segments.*

It is in fact sufficient to characterize the current assigned risk state in a finite horizon CVaR MDP via its slope on the optimal value functions. This gives efficient implementations of the CVaR value iteration algorithm that always produces an optimal policy even under nonuniqueness of the adversary(nature)'s optimal action in the robust MDP when it solves (2). For any risk level $y \in [0,1]$ of the optimal value function $\mathcal{V}_n(x,y)$ for the physical state is x and n stages left, define its left slope (left derivative in y) and right slope by functions $S^-(n,x,y)$, $S^+(n,x,y)$ with $S^-(n,x,y) \geq S^+(n,x,y)$. Define the linear segment on the function $\mathcal{V}_n(x,.)$ with slope $s > 0$ as: $D(n,x,s) = \{y \in [0,1] : S^+(n,x,y) \leq s \leq S^-(n,x,y)\}$.

LEMMA 3.2. *Let $A_n^*(x,y)$ denote the optimal action set for augmented state (x,y) with n stages left. If $S^-(n,x,y) = S^+(n,x,y)$, then $A_n^*(x,y) \subseteq A_n^*(x,z)$, $\forall z \in [0,1]$ such that $S^-(n,x,y) = S^-(n,x,z)$ or $S^-(n,x,y) = S^+(n,x,z)$.*

Lemma 3.2 implies that, $\forall x \in \mathbb{X}$ and $n = 0, 1, \ldots$, for each distinct slope s on the CVaR value functions $\mathcal{V}_n(x,.)$, we can define its optimal action set as $\tilde{A}_n^*(x,s)$, which is a set of actions that are optimal for all risk levels $y \in [0,1]$ such that

$S^-(n,x,y) = s$ or $S^+(n,x,y) = s$, i.e., $y \in D(n,x,s)$. As a consequence of Lemma 3.2, for all $x \in \mathbb{X}$ and $y \in (0,1)$,

$\tilde{A}_n^*(x, S^-(n,x,y)) \cup \tilde{A}_n^*(x, S^+(n,x,y)) \subset A_n^*(x,y)$

Based on the slope characterization we obtain the following lemma for the propagation of slope values in the CVaR Bellman equations (2), (3) corresponding to the robust MDP, where costs are dependent on the next state. Let $B_n^*(x,y,a)$ be the optimal response set at n stages left such that all vectors $\forall b \in B_n^*(x,y,a)$ are optimal solutions of the maximization problem in (2) for respective Q-factor function $Q_n((x,y),a)$.

LEMMA 3.3. *Given an augmented state $(x,y) \in \mathbf{X}$ at n stages to go, let $a \in \tilde{A}_n^*(x, S^-(n,x,y)) \cup \tilde{A}_n^*(x, S^+(n,x,y))$. If $a \in \tilde{A}_n^*(x, S^-(n,x,y))$, and at the next stage the physical state transitions to $x' \in \mathbb{X}$, then for all optimal response vector of the nature(adversary) $\forall b \in B_n^*(x,y,a)$ which determines the next assigned risk state by $y' = yb_{x'}$, we have:* $S^+(n-1,x',y') \leq \frac{S^-(n,x,y) - c(x,a,x')}{\beta} \leq S^-(n-1,x',y')$.

The similar result holds for $a \in \tilde{A}_n^*(x, S^+(n,x,y))$.

In the developed algorithm, the piecewise linear scaled value functions $\mathcal{V}_n(x,.)$ are represented by its slope sets $\mathcal{S}_{n,x}$ and break-point sets $\mathcal{Y}_{n,x}$, where $m^{(n,x)}$ is the number of unique slope segments: $\mathcal{S}_{n,x} = \{s_{n,x}^{(1)}, \ldots, s_{n,x}^{(m_{n,x})}\}, \mathcal{Y}_{n,x} = \{0 = y_{n,x}^{(0)}, y_{n,x}^{(1)}, \ldots, y_{n,x}^{(m_{n,x}-1)}, y_{n,x}^{(m_{n,x})} = 1\}$. Subroutine 1 constructs value functions recursively based on the minimax equations (2), (3), using a piecewise linear representation. Subroutine 2 characterizes propagation of slope information, which is sufficient for optimal decision making at subsequent stages.

Subroutine 1: CVaR Value Function Construction with Slope and Breakpoint Sets

Input: Slope sets $\mathcal{S}_{n-1,x}$ and break-point sets $\mathcal{Y}_{n-1,x}$, known cost and transition probabilities $c(x,a,x')$ and $p(x'|x,a), \forall x, x' \in \mathbb{X}, a \in A(x)$, and known discount factor β.

1. $\forall x \in \mathbb{X}, a \in A(x)$, compute the slope set and the break-point set of $Q_n((x,.),a)$, denoted as $\mathcal{S}_{n,x,a} = \{s_{n,x,a}^{(1)}, \ldots, s_{n,x,a}^{(m_{n,x,a})}\}, \mathcal{Y}_{n,x,a} = \{0 = y_{n,x,a}^{(0)}, y_{n,x,a}^{(1)}, \ldots, y_{n,x,a}^{(m_{n,x,a}-1)}, y_{n,x,a}^{(m_{n,x,a})} = 1\}$, where $m_{n,x,a}$ is the number of unique slope values on $Q_n((x,.),a)$. $\mathcal{S}_{n,x,a}$ is the unique ordered list of slope values: $\{c(x,a,x') + \beta s_{n-1,x'}^{(i)}\}_{x' \in \mathbb{X}, i \in \{1,\ldots,m_{n-1,x'}\}}$, sorted in decreasing order. Define the following index set $J_{n,x,a}^{(i),x'} = \{j \in \{1,\ldots,m_{n-1,x'}\} : c(x,a,x') + \beta s_{n-1,x'}^{(j)} \geq s_{n,x,a}^{(i)}\}$, $\forall x, x' \in \mathbb{X}, a \in A(x), i = 1,\ldots,m_{n,x,a}$. Then $\mathcal{Y}_{n,x,a}$ is computed as, $\forall i = 1,\ldots,m_{n,x,a}$ (with $y_{n,x,a}^{(0)} = 0$):

$y_{n,x,a}^{(i)} = \sum_{x' \in \mathbb{X}} \sum_{j \in J_{n,x,a}^{(i),x'}} p(x'|x,a)(y_{n-1,x'}^{(j)} - y_{n-1,x'}^{(j-1)})$.

2. Using $\mathcal{S}_{n,x,a}$ and $\mathcal{Y}_{n,x,a}, \forall x \in \mathbb{X}, a \in A(x)$, compute the slope set and break-point set representation of $\mathcal{V}_n(x,.), \forall x \in \mathbb{X}$, denoted as $\mathcal{S}_{n,x} = \{s_{n,x}^{(1)}, \ldots, s_{n,x}^{(m_{n,x})}\}, \mathcal{Y}_{n,x} = \{0 = y_{n,x}^{(0)}, y_{n,x}^{(1)}, \ldots, y_{n,x}^{(m_{n,x}-1)}, y_{n,x}^{(m_{n,x})} = 1\}$. This is done by taking $\mathcal{V}_n(x,y) = \min_{a \in A(x)} Q_n((x,y),a), \forall x \in \mathbb{X}, y \in [0,1]$. For all $i \in \{1,\ldots,m_{n,x}\}$, store the optimal action sets as $\tilde{A}_n^*(x, s_{n,x}^{(i)}) = \{a \in A(x) : Q_n((x,y),a) = \mathcal{V}_n(x,y) = \min_{a \in A(x)} Q_n((x,y),a), \forall y \in [y_{n,x}^{(i-1)}, y_{n,x}^{(i)}]\}$.

3. Output the slope set and break-point set characterization of $\mathcal{V}_n(x,.)$: $\mathcal{S}_{n,x}, \mathcal{Y}_{n,x}, \forall x \in \mathbb{X}$, and the optimal action sets for each slope segment at each state: $\tilde{A}_n^*(x,s), \forall x \in \mathbb{X}, s \in \mathcal{S}_{n,x}$.

Subroutine 2: Slope Information State Propagation

Input: Current state x, n stages to go, a slope value $s \in \mathcal{S}_{n,x}$ and an optimal action take from $a \in \tilde{A}_n^*(x,s)$, observed next state x' and known cost $c(x,a,x')$, characterization sets $\mathcal{S}_{n-1,x'}$ and $\mathcal{Y}_{n-1,x'}$ for the next stage value function at x'.

1. Compute the desired slope value $s' = \frac{s - c(x,a,x')}{\beta}$.

2. If $s' \in \mathcal{S}_{n-1,x'}$, set $s^- = s^+ = s'$, this slope suffices for making an optimal action in the next stage problem by taking an action from $\tilde{A}_{n-1}^*(x', s')$. Otherwise, set $s^- = \min\{s \in \mathcal{S}_{n-1,x'} : s \geq s'\}$ and $s^+ = \max\{s \in \mathcal{S}_{n-1,x'} : s \leq s'\}$.

3. Output the next slope information state: (x', s^-, s^+).

Algorithm 1: Finite Horizon CVaR Value Iteration with Slope Characterization

1. Set $n = 0$, $\mathcal{V}_0(x,y) = yv_0(x), \forall (x,y) \in \mathbf{X}$, which can be represented by slope set and break-point set characterization as $\mathcal{S}_{0,x} = \{v_0(x)\}, \mathcal{Y}_{0,x} = \{0,1\}, \forall x \in \mathbb{X}$.

2. For $n = 1,\ldots,N$: use subroutine 1 to compute $\mathcal{V}_n(x,y)$, $\forall (x,y) \in \mathbf{X}$, following (2) and (3), with input slope sets and break-point sets $\mathcal{S}_{n-1,x}, \mathcal{Y}_{n-1,x}, \forall x \in \mathbb{X}$. The output slope set and break-point set characterization of $\mathcal{V}_n(x,.)$ is stored as $\mathcal{S}_{n,x}, \mathcal{Y}_{n,x}, \forall x \in \mathbb{X}$. Store optimal action sets as $\tilde{A}_n^*(x,s), \forall x \in \mathbb{X}, s \in \mathcal{S}_{n,x}$.

3. For $t = 0,\ldots,N-1$: the current risk augmented state (x_t, y_t) with $N - t$ stages to go is characterized by the slope information state tuple (x_t, s_t^-, s_t^+), which is known exactly at the initial stage. Perform an optimal action $a_t \in \tilde{A}_{N-t}^*(x_t, s_t^*)$ where $s_t^* = s_t^-$ or s_t^+, observe the next physical state $x_{t+1} \in \mathbb{X}$ and incur cost $c(x_t, a_t, x_{t+1})$. Use subroutine 2 to identify the next augmented state $(x_{t+1}, s_{t+1}^-, s_{t+1}^+)$. (If $s_{t+1}^- > s_{t+1}^+$, the next assigned risk state y_{t+1} can be exactly identified.)

THEOREM 3.4. *Algorithm 1 always outputs an optimal non-randomized autonomous policy that minimizes the N-step CVaR value starting from any fixed $x_0 \in \mathbb{X}, y_0 = \alpha \in [0,1]$.*

Results similar to Lemma 3.2 and 3.3 can be generalized to the infinite horizon setting, where we can obtain an algorithm that computes an ϵ-optimal nonrandomized autonomous policy.

4. REFERENCES

[1] Chow, Y., et al., 2015. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. *NIPS 2015: 1522-1530*.

[2] Pflug, G., and A. Pichler, 2016. Time-Consistent Decisions and Temporal Decomposition of Coherent Risk Functionals. *Math. Oper. Res. 41: 682-699*.

[3] Rockafellar, R.T., and S. Uryasev, 2000. Optimization of Conditional Value-At-Risk. *Journal of Risk. 2: 21-42*.

[4] Shapiro, A., 2021. Tutorial on Risk Neutral, Distributionally Robust and Risk Averse Multistage Stochastic Programming. *Eur. J. Oper. Res. 288: 1-13*.