Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach

Yinlam Chow
Stanford University
ychow@stanford.edu

Aviv Tamar
UC Berkeley
avivt@berkeley.edu

Shie Mannor
Technion
shie@ee.technion.ac.il

Marco Pavone
Stanford University
pavone@stanford.edu

Abstract

In this paper we address the problem of decision making within a Markov decision process (MDP) framework where risk and modeling errors are taken into account. Our approach is to minimize a risk-sensitive conditional-value-at-risk (CVaR) objective, as opposed to a standard risk-neutral expectation. We refer to such problem as CVaR MDP. Our first contribution is to show that a CVaR objective, besides capturing risk sensitivity, has an alternative interpretation as expected cost under worst-case modeling errors, for a given error budget. This result, which is of independent interest, motivates CVaR MDPs as a unifying framework for risk-sensitive and robust decision making. Our second contribution is to present an approximate value-iteration algorithm for CVaR MDPs and analyze its convergence rate. To our knowledge, this is the first solution algorithm for CVaR MDPs that enjoys error guarantees. Finally, we present results from numerical experiments that corroborate our theoretical findings and show the practicality of our approach.

1 Introduction

Decision making within the Markov decision process (MDP) framework typically involves the minimization of a risk-neutral performance objective, namely the *expected* total discounted cost [3]. This approach, while very popular, natural, and attractive from a computational standpoint, neither takes into account the *variability* of the cost (i.e., fluctuations around the mean), nor its *sensitivity* to modeling errors, which may significantly affect overall performance [12]. Risk-sensitive MDPs [9] address the first aspect by replacing the risk-neutral expectation with a *risk-measure* of the total discounted cost, such as variance, Value-at-Risk (VaR), or Conditional-VaR (CVaR). Robust MDPs [15], on the other hand, address the second aspect by defining a set of plausible MDP parameters, and optimize decision with respect to the expected cost under worst-case parameters.

In this work we consider risk-sensitive MDPs with a CVaR objective, referred to as CVaR MDPs. CVaR [1, 20] is a risk-measure that is rapidly gaining popularity in various engineering applications, e.g., finance, due to its favorable computational properties [1] and superior ability to safeguard a decision maker from the "outcomes that hurt the most" [22]. In this paper, by *relating risk to robustness*, we derive a novel result that further motivates the usage of a CVaR objective in a decision-making context. Specifically, we show that the CVaR of a discounted cost in an MDP is *equivalent* to the expected value of the same discounted cost in presence of worst-case perturbations of the MDP parameters (specifically, transition probabilities), provided that such perturbations are within a certain error budget. This result suggests CVaR MDP as a method for decision making under *both* cost variability *and* model uncertainty, motivating it as *unified framework* for planning under uncertainty.

Literature review: Risk-sensitive MDPs have been studied for over four decades, with earlier efforts focusing on exponential utility [9], mean-variance [24], and percentile risk criteria [7] . Recently, for the reasons explained above, several authors have investigated CVaR MDPs [20]. Specifically,

in [4], the authors propose a dynamic programming algorithm for finite-horizon risk-constrained MDPs where risk is measured according to CVaR. The algorithm is proven to asymptotically converge to an optimal risk-constrained policy. However, the algorithm involves computing integrals over continuous variables (Algorithm 1 in [4]) and, in general, its implementation appears particularly difficult. In [2], the authors investigate the structure of CVaR optimal policies and show that a Markov policy is optimal on an augmented state space, where the additional (continuous) state variable is represented by the running cost. In [8], the authors leverage such result to design an algorithm for CVaR MDPs that relies on discretizing occupation measures in the augmented-state MDP. This approach, however, involves solving a non-convex program via a sequence of linear-programming approximations, which can only shown to converge asymptotically. A different approach is taken by [5], [19] and [25], which consider a finite dimensional parameterization of control policies, and show that a CVaR MDP can be optimized to a *local* optimum using stochastic gradient descent (policy gradient). A recent result by Pflug and Pichler [17] showed that CVaR MDPs admit a dynamic programming formulation by using a state-augmentation procedure different from the one in [2]. The augmented state is also continuous, making the design of a solution algorithm challenging.

Contributions: The contribution of this paper is twofold. First, as discussed above, we provide a novel interpretation for CVaR MDPs in terms of robustness to modeling errors. This result is of independent interest and further motivates the usage of CVaR MDPs for decision making under uncertainty. Second, we provide a new optimization algorithm for CVaR MDPs, which leverages the state augmentation procedure introduced by Pflug and Pichler [17]. We overcome the aforementioned computational challenges (due to the continuous augmented state) by designing an algorithm that merges approximate value iteration [3] with linear interpolation. Remarkably, we are able to provide explicit error bounds and convergence rates based on contraction-style arguments. In contrast to the algorithms in [4, 8, 5, 25], given the explicit MDP model our approach leads to finite-time error guarantees, with respect to the *globally* optimal policy. In addition, our algorithm is significantly simpler than previous methods, and calculates the optimal policy *for all* CVaR confidence intervals and initial states simultaneously. The practicality of our approach is demonstrated in numerical experiments involving planning a path on a grid with thousand of states. To the best of our knowledge, this is the first algorithm to approximate globally-optimal policies for non-trivial CVaR MDPs whose error depends on the resolution of interpolation.

Organization: This paper is structured as follows. In Section 2 we provide background on CVaR and MDPs, we state the problem we wish to solve (i.e., CVaR MDPs), and motivate the CVaR MDP formulation by establishing a novel relation between CVaR and model perturbations. Section 3 provides the basis for our solution algorithm, based on a Bellman-style equation for the CVaR. Then, in Section 4 we present our algorithm and correctness analysis. In Section 5 we evaluate our approach via numerical experiments. Finally, in Section 6, we draw some conclusions and discuss directions for future work.

2 Preliminaries, Problem Formulation, and Motivation

2.1 Conditional Value-at-Risk

Let Z be a bounded-mean random variable, i.e., $\mathbb{E}[|Z|] < \infty$, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with cumulative distribution function $F(z) = \mathbb{P}(Z \leq z)$. In this paper we interpret Z as a cost. The *value-at-risk* (VaR) at confidence level $\alpha \in (0,1)$ is the $1 - \alpha$ quantile of Z, i.e., $\text{VaR}_\alpha(Z) = \min\{z \mid F(z) \geq 1 - \alpha\}$. The *conditional value-at-risk* (CVaR) at confidence level $\alpha \in (0,1)$ is defined as [20]:

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

where $(x)^+ = \max(x, 0)$ represents the positive part of x. If there is no probability atom at $\text{VaR}_\alpha(Z)$, it is well known from Theorem 6.2 in [23] that $\text{CVaR}_\alpha(Z) = \mathbb{E}\big[Z \mid Z \geq \text{VaR}_\alpha(Z)\big]$. Therefore, $\text{CVaR}_\alpha(Z)$ may be interpreted as the expected value of Z, conditioned on the α-portion of the tail distribution. It is well known that $\text{CVaR}_\alpha(Z)$ is decreasing in α, $\text{CVaR}_1(Z)$ equals to $\mathbb{E}(Z)$, and $\text{CVaR}_\alpha(Z)$ tends to $\max(Z)$ as $\alpha \downarrow 0$. During the last decade, the CVaR risk-measure has gained popularity in financial applications, among others. It is especially useful for controlling rare, but potentially disastrous events, which occur above the $1 - \alpha$ quantile, and are neglected by the VaR [22]. Furthermore, CVaR enjoys desirable axiomatic properties, such as coherence [1]. We refer to [26] for further motivation about CVaR and a comparison with other risk measures such as VaR.

A useful property of CVaR, which we exploit in this paper, is its alternative dual representation [1]:

$$\text{CVaR}_\alpha(Z) = \max_{\xi \in \mathcal{U}_{\text{CVaR}}(\alpha, \mathbb{P})} \mathbb{E}_\xi[Z], \tag{2}$$

where $\mathbb{E}_\xi[Z]$ denotes the ξ-weighted expectation of Z, and the *risk envelope* $\mathcal{U}_{\text{CVaR}}$ is given by $\mathcal{U}_{\text{CVaR}}(\alpha, \mathbb{P}) = \left\{ \xi : \xi(\omega) \in \left[0, \frac{1}{\alpha}\right], \int_{\omega \in \Omega} \xi(\omega)\mathbb{P}(\omega)d\omega = 1 \right\}$. Thus, the CVaR of a random variable Z may be interpreted as the worst-case expectation of Z, under a perturbed distribution $\xi\mathbb{P}$.

In this paper, we are interested in the CVaR of the total discounted cost in a sequential decision-making setting, as discussed next.

2.2 Markov Decision Processes

An MDP is a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, P, x_0, \gamma)$, where $\mathcal{X}$ and $\mathcal{A}$ are finite state and action spaces; $C(x, a) \in [-C_{\max}, C_{\max}]$ is a bounded deterministic cost; $P(\cdot|x, a)$ is the transition probability distribution; $\gamma \in [0, 1)$ is the discounting factor, and x_0 is the initial state. (Our results easily generalize to random initial states and random costs.)

Let the space of admissible histories up to time t be $H_t = H_{t-1} \times \mathcal{A} \times \mathcal{X}$, for $t \geq 1$, and $H_0 = \mathcal{X}$. A generic element $h_t \in H_t$ is of the form $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$. Let $\Pi_{H,t}$ be the set of all history-dependent policies with the property that at each time t the randomized control action is a function of h_t. In other words, $\Pi_{H,t} := \{\mu_0 : H_0 \to \mathbb{P}(\mathcal{A}), \mu_1 : H_1 \to \mathbb{P}(\mathcal{A}), \ldots, \mu_t : H_t \to \mathbb{P}(\mathcal{A})\}|\mu_j(h_j) \in \mathbb{P}(\mathcal{A})$ for all $h_j \in H_j, 1 \leq j \leq t\}$. We also let $\Pi_H = \lim_{t\to\infty} \Pi_{H,t}$ be the set of all history dependent policies.

2.3 Problem Formulation

Let $C(x_t, a_t)$ denote the stage-wise costs observed along a state/control trajectory in the MDP model, and let $\mathcal{C}_{0,T} = \sum_{t=0}^{T} \gamma^t C(x_t, a_t)$ denote the total discounted cost up to time T. The risk-sensitive discounted-cost problem we wish to address is as follows:

$$\min_{\mu \in \Pi_H} \quad \text{CVaR}_\alpha \left(\lim_{T\to\infty} \mathcal{C}_{0,T} \,\Big|\, x_0, \mu \right), \tag{3}$$

where $\mu = \{\mu_0, \mu_1, \ldots\}$ is the policy sequence with actions $a_t = \mu_t(h_t)$ for $t \in \{0, 1, \ldots\}$. We refer to problem (3) as CVaR MDP (One may also consider a related formulation combining mean and CVaR, the details of which are presented in the supplementary material).

The problem formulation in (3) directly addresses the aspect of risk sensitivity, as demonstrated by the numerous applications of CVaR optimization in finance (see, e.g., [21, 11, 6]) and the recent approaches for CVaR optimization in MDPs [4, 8, 5, 25]. In the following, we show a new result providing additional motivation for CVaR MDPs, from the point of view of *robustness to modeling errors*.

2.4 Motivation - Robustness to Modeling Errors

We show a new result relating the CVaR objective in (3) to the *expected* discounted-cost in presence of worst-case perturbations of the MDP parameters, where the perturbations are budgeted according to the "number of things that can go wrong". Thus, by minimizing CVaR, the decision maker also guarantees *robustness* of the policy.

Consider a trajectory $(x_0, a_0, \ldots, x_T)$ in a finite-horizon MDP problem with transitions $P_t(x_t|x_{t-1}, a_{t-1})$. We explicitly denote the time index of the transition matrices for reasons that will become clear shortly. The total probability of the trajectory is $P(x_0, a_0, \ldots, x_T) = P_0(x_0)P_1(x_1|x_0, a_0) \cdots P_T(x_T|x_{T-1}, a_{T-1})$, and we let $\mathcal{C}_{0,T}(x_0, a_0, \ldots, x_T)$ denote its discounted cost, as defined above.

We consider an adversarial setting, where an adversary is allowed to change the transition probabilities at each stage, under some budget constraints. We will show that, for a specific budget and perturbation structure, the expected cost under the worst-case perturbation is equivalent to the CVaR of the cost. Thus, we shall establish that, in this perspective, being risk sensitive is *equivalent* to being robust against model perturbations.

For each stage $1 \leq t \leq T$, consider a perturbed transition matrix $\hat{P}_t = P_t \circ \delta_t$, where $\delta_t \in \mathbb{R}^{\mathcal{X} \times \mathcal{A} \times \mathcal{X}}$ is a *multiplicative probability perturbation* and $\circ$ is the Hadamard product, under the condition that $\hat{P}_t$ is a stochastic matrix. Let Δ_t denote the set of perturbation matrices that satisfy this condition, and let $\Delta = \Delta_1 \times \cdots \times \Delta_T$ the set of all possible perturbations to the trajectory distribution.

We now impose a budget constraint on the perturbations as follows. For some budget $\eta \geq 1$, we consider the constraint

$$\delta_1(x_1|x_0, a_0)\delta_2(x_2|x_1, a_1) \cdots \delta_T(x_T|x_{T-1}, a_{T-1}) \leq \eta, \quad \forall x_0, \ldots, x_T \in \mathcal{X}, \forall a_0, \ldots, a_{T-1} \in \mathcal{A}. \tag{4}$$

Essentially, the product in Eq. (4) states that with small budget *the worst cannot happen at each time*. Instead, the perturbation budget has to be split (multiplicatively) along the trajectory. We note that Eq. (4) is in fact a constraint on the perturbation matrices, and we denote by $\Delta_\eta \subset \Delta$ the set of perturbations that satisfy this constraint with budget η. The following result shows an equivalence between the CVaR and the worst-case expected loss.

Proposition 1 (Interpretation of CVaR as a Robustness Measure) *It holds*

$$CVaR_{\frac{1}{\eta}}\left(\mathcal{C}_{0,T}(x_0, a_0, \ldots, x_T)\right) = \sup_{(\delta_1, \ldots, \delta_T) \in \Delta_\eta} \mathbb{E}_{\hat{P}}\left[\mathcal{C}_{0,T}(x_0, a_0, \ldots, x_T)\right], \tag{5}$$

where $\mathbb{E}_{\hat{P}}[\cdot]$ denotes expectation with respect to a Markov chain with transitions $\hat{P}_t$.

The proof of Proposition 1 is in the supplementary material. It is instructive to compare Proposition 1 with the dual representation of CVaR in (2) where both results convert the CVaR risk into a robustness measure. Note, in particular, that the perturbation budget in Proposition 1 has a *temporal* structure, which constrains the adversary from choosing the worst perturbation at each time step.

Remark 1 *An equivalence between robustness and risk-sensitivity was previously suggested by Osogami [16]. In that study, the* iterated *(dynamic) coherent risk was shown to be equivalent to a robust MDP [10] with a rectangular uncertainty set. The iterated risk (and, correspondingly, the rectangular uncertainty set) is very conservative [27], in the sense that* the worst can happen at each time step. *In contrast, the perturbations considered here are much less conservative. In general, solving robust MDPs without the rectangularity assumption is NP-hard. Nevertheless, Mannor et. al. [13] showed that, for cases where the number of perturbations to the parameters along a trajectory is upper bounded (budget-constrained perturbation), the corresponding robust MDP problem is tractable. Analogous to the constraint set (1) in [13], the perturbation set in Proposition 1 limits the total number of log-perturbations along a trajectory. Accordingly, we shall later see that optimizing problem (3) with perturbation structure (4) is indeed also tractable.*

Next section provides the fundamental theoretical ideas behind our approach to the solution of (3).

3 Bellman Equation for CVaR

In this section, by leveraging a recent result from [17], we present a dynamic programming (DP) formulation for the CVaR MDP problem in (3). As we shall see, the value function in this formulation depends on both the state and the CVaR confidence level α. We then establish important properties of such DP formulation, which will later enable us to derive an efficient DP-based approximate solution algorithm and provide correctness guarantees on the approximation error. All proofs are presented in the supplementary material.

Our starting point is a recursive decomposition of CVaR, whose proof is detailed in Theorem 10 of [17].

Theorem 2 (CVaR Decomposition, Theorem 21 in [17]) *For any $t \geq 0$, denote by $Z = (Z_{t+1}, Z_{t+2}, \ldots)$ the cost sequence from time $t + 1$ onwards. The conditional CVaR under policy μ, i.e., $CVaR_\alpha(Z \mid h_t, \mu)$, obeys the following decomposition:*

$$CVaR_\alpha(Z \mid h_t, \mu) = \max_{\xi \in \mathcal{U}_{CVaR}(\alpha, P(\cdot|x_t, a_t))} \mathbb{E}[\xi(x_{t+1}) \cdot CVaR_{\alpha\xi(x_{t+1})}(Z \mid h_{t+1}, \mu) \mid h_t, \mu],$$

where a_t is the action induced by policy $\mu_t(h_t)$, and the expectation is with respect to x_{t+1}.

Theorem 2 concerns a fixed policy μ; we now extend it to a general DP formulation. Note that in the recursive decomposition in Theorem 2 the right-hand side involves CVaR terms with different confidence levels than that in the left-hand side. Accordingly, we augment the state space $\mathcal{X}$ with an additional continuous state $\mathcal{Y} = (0, 1]$, which corresponds to the confidence level. For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the *value-function* $V(x, y)$ for the augmented state (x, y) is defined as:

$$V(x, y) = \min_{\mu \in \Pi_H} CVaR_y\left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0 = x, \mu\right).$$

4

Similar to standard DP, it is convenient to work with operators defined on the space of value functions [3]. In our case, Theorem 2 leads to the following definition of CVaR Bellman operator $\mathbf{T} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$:

$$\mathbf{T}[V](x,y) = \min_{a \in \mathcal{A}} \left[C(x,a) + \gamma \max_{\xi \in \mathcal{U}_{\mathrm{CVaR}}(y, P(\cdot|x,a))} \sum_{x' \in \mathcal{X}} \xi(x') V\left(x', y\xi(x')\right) P(x'|x,a) \right]. \quad (6)$$

We now establish several useful properties for the Bellman operator $\mathbf{T}[V]$.

Lemma 3 (Properties of CVaR Bellman Operator) *The Bellman operator $\mathbf{T}[V]$ has the following properties:*

1. *(Contraction.) $\|\mathbf{T}[V_1] - \mathbf{T}[V_2]\|_\infty \le \gamma \|V_1 - V_2\|_\infty$, where $\|f\|_\infty = \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} |f(x,y)|$.*

2. *(Concavity preserving in y.) For any $x \in \mathcal{X}$, suppose $yV(x,y)$ is concave in $y \in \mathcal{Y}$. Then the maximization problem in (6) is concave. Furthermore, $y\mathbf{T}[V](x,y)$ is concave in y.*

The first property in Lemma 3 is similar to standard DP [3], and is instrumental to the design of a converging value-iteration approach. The second property is nonstandard and specific to our approach. It will be used to show that the computation of value-iteration updates involves concave, and therefore *tractable* optimization problems. Furthermore, it will be used to show that a linear-interpolation of $V(x,y)$ in the augmented state y has a bounded error.

Equipped with the results in Theorem 2 and Lemma 3, we can now show that the fixed point solution of $\mathbf{T}[V](x,y) = V(x,y)$ is unique, and equals to the solution of the CVaR MDP problem (3) with $x_0 = x$ and $\alpha = y$.

Theorem 4 (Optimality Condition) *For any $x \in \mathcal{X}$ and $y \in (0,1]$, the solution to $\mathbf{T}[V](x,y) = V(x,y)$ is unique, and equals to $V^*(x,y) = \min_{\mu \in \Pi_H} CVaR_y \left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0 = x, \mu \right)$.*

Next, we show that the optimal value of the CVaR MDP problem (3) can be attained by a stationary Markov policy, defined as a greedy policy with respect to the value function $V^*(x,y)$. Thus, while the original problem is defined over the intractable space of history-dependent policies, a stationary Markov policy (over the augmented state space) is optimal, and can be readily derived from $V^*(x,y)$. Furthermore, an optimal history-dependent policy can be readily obtained from an (augmented) optimal Markov policy according to the following theorem.

Theorem 5 (Optimal Policies) *Let $\pi_H^* = \{\mu_0, \mu_1, \ldots\} \in \Pi_H$ be a history-dependent policy recursively defined as:*

$$\mu_k(h_k) = u^*(x_k, y_k), \ \forall k \ge 0, \quad (7)$$

with initial conditions x_0 and $y_0 = \alpha$, and state transitions

$$x_k \sim P(\cdot \mid x_{k-1}, u^*(x_{k-1}, y_{k-1})), \quad y_k = y_{k-1}\xi^*_{x_{k-1}, y_{k-1}, u^*}(x_k), \forall k \ge 1, \quad (8)$$

where the stationary Markovian policy $u^*(x,y)$ and risk factor $\xi^*_{x,y,u^*}(\cdot)$ are solution to the min-max optimization problem in the CVaR Bellman operator $\mathbf{T}[V^*](x,y)$. Then, π_H^* is an optimal policy for problem (3) with initial state x_0 and CVaR confidence level α.

Theorems 4 and 5 suggest that a value-iteration DP method [3] can be used to solve the CVaR MDP problem (3). Let an initial value-function guess $V_0 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be chosen arbitrarily. Value iteration proceeds recursively as follows:

$$V_{k+1}(x,y) = \mathbf{T}[V_k](x,y), \ \forall(x,y) \in \mathcal{X} \times \mathcal{Y}, \ k \in \{0, 1, \ldots, \}. \quad (9)$$

Specifically, by combining the contraction property in Lemma 3 and uniqueness result of fixed point solutions from Theorem 4, one concludes that $\lim_{k \to \infty} V_k(x,y) = V^*(x,y)$. By selecting $x = x_0$ and $y = \alpha$, one immediately obtains $V^*(x_0, \alpha) = \min_{\mu \in \Pi_H} CVaR_\alpha \left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu \right)$. Furthermore, an optimal policy may be derived from $V^*(x,y)$ according to the policy construction procedure in Theorem 5.

Unfortunately, while value iteration is conceptually appealing, its direct implementation in our setting is generally impractical since, e.g., the state y is continuous. In the following, we pursue an *approximation* to the value iteration algorithm (9), based on a linear interpolation scheme for y.

Algorithm 1 `CVaR Value Iteration with Linear Interpolation`

1: **Given:**
- $N(x)$ interpolation points $\mathbf{Y}(x) = \left\{y_1, \ldots, y_{N(x)}\right\} \in [0,1]^{N(x)}$ for every $x \in \mathcal{X}$ with $y_i < y_{i+1}, y_1 = 0$ and $y_{N(x)} = 1$.
- Initial value function $V_0(x,y)$ that satisfies Assumption 1.

2: For $t = 1, 2, \ldots$
- For each $x \in \mathcal{X}$ and each $y_i \in \mathbf{Y}(x)$, update the value function estimate as follows:
$$V_t(x, y_i) = \mathbf{T}_{\mathcal{I}}[V_{t-1}](x, y_i),$$

3: Set the converged value iteration estimate as $\widehat{V}^*(x, y_i)$, for any $x \in \mathcal{X}$, and $y_i \in \mathbf{Y}(x)$.
---

4   Value Iteration with Linear Interpolation

In this section we present an approximate DP algorithm for solving CVaR MDPs, based on the theoretical results of Section 3. The value iteration algorithm in Eq. (9) presents two main implementation challenges. The first is due to the fact that the augmented state y is continuous. We handle this challenge by using interpolation, and exploit the concavity of $yV(x,y)$ to bound the error introduced by this procedure. The second challenge stems from the the fact that applying $\mathbf{T}$ involves maximizing over ξ. Our strategy is to exploit the concavity of the maximization problem to guarantee that such optimization can indeed be performed effectively.

As discussed, our approach relies on the fact that the Bellman operator $\mathbf{T}$ preserves concavity as established in Lemma 3. Accordingly, we require the following assumption for the initial guess $V_0(x,y)$,

Assumption 1 *The guess for the initial value function $V_0(x,y)$ satisfies the following properties: 1) $yV_0(x,y)$ is concave in $y \in \mathcal{Y}$ and 2) $V_0(x,y)$ is continuous in $y \in \mathcal{Y}$ for any $x \in \mathcal{X}$.*

Assumption 1 may easily be satisfied, for example, by choosing $V_0(x,y) = \text{CVaR}_y(Z \mid x_0 = x)$, where Z is any arbitrary bounded random variable. As stated earlier, a key difficulty in applying value iteration (9) is that, for each state $x \in \mathcal{X}$, the Bellman operator has to be calculated for each $y \in \mathcal{Y}$, and $\mathcal{Y}$ is continuous. As an approximation, we propose to calculate the Bellman operator only for a finite set of values y, and interpolate the value function in between such interpolation points.

Formally, let $N(x)$ denote the number of interpolation points. For every $x \in \mathcal{X}$, denote by $\mathbf{Y}(x) = \left\{y_1, \ldots, y_{N(x)}\right\} \in [0,1]^{N(x)}$ the set of interpolation points. We denote by $\mathcal{I}_x[V](y)$ the linear interpolation of the function $yV(x,y)$ on these points, i.e.,

$$\mathcal{I}_x[V](y) = y_i V(x, y_i) + \frac{y_{i+1} V(x, y_{i+1}) - y_i V(x, y_i)}{y_{i+1} - y_i}(y - y_i),$$

where $y_i = \max\left\{y' \in \mathbf{Y}(x) : y' \leq y\right\}$ and y_{i+1} is the closest interpolation point such that $y \in [y_i, y_{i+1}]$, i.e., $y_{i+1} = \min\left\{y' \in \mathbf{Y}(x) : y' \geq y\right\}$. The interpolation of $yV(x,y)$ instead of $V(x,y)$ is key to our approach. The motivation is twofold: first, it can be shown [20] that for a discrete random variable Z, $y\text{CVaR}_y(Z)$ is piecewise linear in y. Second, one can show that the Lipschitzness of $yV(x,y)$ is preserved during value iteration, and exploit this fact to bound the linear interpolation error.

We now define the *interpolated* Bellman operator $\mathbf{T}_{\mathcal{I}}$ as follows:

$$\mathbf{T}_{\mathcal{I}}[V](x,y) = \min_{a \in \mathcal{A}}\left[C(x,a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x,a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V](y\xi(x'))}{y} P(x'|x,a)\right]. \qquad (10)$$

Remark 2 *Notice that by L'Hospital's rule one has $\lim_{y \to 0} \mathcal{I}_x[V](y\xi(x))/y = V(x,0)\xi(x)$. This implies that at $y = 0$ the interpolated Bellman operator is equivalent to the original Bellman operator, i.e., $\mathbf{T}[V](x,0) = \min_{a \in \mathcal{A}}\left\{C(x,a) + \gamma \max_{x' \in \mathcal{X}: P(x'|x,a) > 0} V(x', 0)\right\} = \mathbf{T}_{\mathcal{I}}[V](x,0)$.*

Algorithm 1 presents CVaR value iteration with linear interpolation. The only difference between this algorithm and standard value iteration (9) is the linear interpolation procedure described above. In the following, we show that Algorithm 1 converges, and bound the error due to interpolation. We begin by showing that the useful properties established in Lemma 3 for the Bellman operator $\mathbf{T}$ extend to the interpolated Bellman operator $\mathbf{T}_{\mathcal{I}}$.

Lemma 6 (Properties of Interpolated Bellman Operator) $\mathbf{T}_{\mathcal{I}}[V]$ *has the same properties of* $\mathbf{T}[V]$ *as in Lemma 3, namely 1) contraction and 2) concavity preservation.*

Lemma 6 implies several important consequences for Algorithm 1. The first one is that the maximization problem in (10) is concave, and thus may be solved efficiently at each step. This guarantees that the algorithm is *tractable*. Second, the contraction property in Lemma 6 guarantees that Algorithm 1 converges, i.e., there exists a value function $\widehat{V}^* \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$ such that $\lim_{n \to \infty} \mathbf{T}_{\mathcal{I}}^n[V_0](x, y_i) = \widehat{V}^*(x, y_i)$. In addition, the convergence rate is geometric and equals to γ.

The following theorem provides an error bound between approximate value iteration and exact value iteration (3) in terms of the interpolation resolution.

Theorem 7 (Convergence and Error Bound) *Suppose the initial value function $V_0(x, y)$ satisfies Assumption 1 and let $\epsilon > 0$ be an error tolerance parameter. For any state $x \in \mathcal{X}$ and step $t \geq 0$, choose $y_2 > 0$ such that $V_t(x, y_2) - V_t(x, 0) \geq -\epsilon$ and update the interpolation points according to the logarithmic rule: $y_{i+1} = \theta y_i$, $\forall i \geq 2$, with uniform constant $\theta \geq 1$. Then, Algorithm 1 has the following error bound:*

$$0 \geq \widehat{V}^*(x_0, \alpha) - \min_{\mu \in \Pi_H} CVaR_\alpha \left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu \right) \geq \frac{-\gamma}{1-\gamma} \mathbf{O}\left((\theta - 1) + \epsilon \right),$$

and the following finite time convergence error bound:

$$\left| \mathbf{T}_{\mathcal{I}}^n[V_0](x_0, \alpha) - \min_{\mu \in \Pi_H} CVaR_\alpha \left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu \right) \right| \leq \frac{\mathbf{O}\left((\theta - 1) + \epsilon \right) + \mathbf{O}(\gamma^n)}{1-\gamma}.$$

Theorem 7 shows that 1) the interpolation-based value function is a *conservative estimate* for the optimal solution to problem (3); 2) the interpolation procedure is *consistent*, i.e., when the number of interpolation points is arbitrarily large (specifically, $\epsilon \to 0$ and $y_{i+1}/y_i \to 1$), the approximation error tends to zero; and 3) the approximation error bound is $O((\theta - 1) + \epsilon)$, where $\log \theta$ is the *log-difference* of the interpolation points, i.e., $\log \theta = \log y_{i+1} - \log y_i$, $\forall i$.

For a pre-specified ϵ, the condition $V_t(x, y_2) - V_t(x, 0) \geq -\epsilon$ may be satisfied by a simple *adaptive procedure* for selecting the interpolation points $\mathbf{Y}(x)$. At each iteration $t > 0$, after calculating $V_t(x, y_i)$ in Algorithm 1, at each state x in which the condition does not hold, add a new interpolation point $y_2' = \frac{\epsilon y_2}{|V_t(x, y_2) - V_t(x, 0)|}$, and additional points between y_2' and y_2 such that the condition $\log \theta \geq \log y_{i+1} - \log y_i$ is maintained. Since all the additional points belong to the segment $[0, y_2]$, the linearly interpolated $V_t(x, y_i)$ remains unchanged, and Algorithm 1 proceeds as is. For bounded costs and $\epsilon > 0$, the number of additional points required is bounded.

The full proof of Theorem 7 is detailed in the supplementary material; we highlight the main ideas and challenges involved. In the first part of the proof we bound, for all $t > 0$, the Lipschitz constant of $yV_t(x, y)$ in y. The key to this result is to show that the Bellman operator $\mathbf{T}$ preserves the Lipschitz property for $yV_t(x, y)$. Using the Lipschitz bound and the concavity of $yV_t(x, y)$, we then bound the error $\frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y)$ for all y. The condition on y_2 is required for this bound to hold when $y \to 0$. Finally, we use this result to bound $\|\mathbf{T}_{\mathcal{I}}[V_t](x, y) - \mathbf{T}[V_t](x, y)\|_\infty$. The results of Theorem 7 follow from contraction arguments, similar to approximate dynamic programming [3].

5 Experiments

We validate Algorithm 1 on a rectangular grid world, where states represent grid points on a 2D terrain map. An agent (e.g., a robotic vehicle) starts in a safe region and its objective is to travel to a given destination. At each time step the agent can move to any of its four neighboring states. Due to sensing and control noise, however, with probability δ a move to a random neighboring state occurs. The stage-wise cost of each move until reaching the destination is 1, to account for fuel usage. In between the starting point and the destination there are a number of obstacles that the agent should avoid. Hitting an obstacle costs $M >> 1$ and terminates the mission. The objective is to compute a *safe* (i.e., obstacle-free) path that is *fuel efficient*.

For our experiments, we choose a 64×53 grid-world (see Figure 1), for a total of 3,312 states. The destination is at position $(60, 2)$, and there are 80 obstacles plotted in yellow. By leveraging Theorem 7, we use 21 log-spaced interpolation points for Algorithm 1 in order to achieve a small value function error. We choose $\delta = 0.05$, and a discount factor $\gamma = 0.95$ for an effective horizon of 200 steps. Furthermore, we set the penalty cost equal to $M = 2/(1-\gamma)$–such choice trades off high penalty for collisions and computational complexity (that increases as M increases). For the
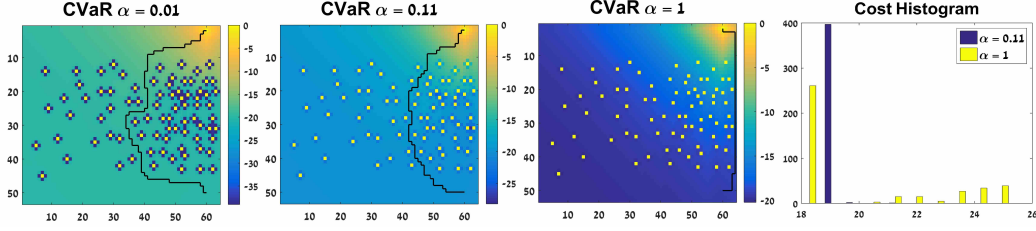
Figure 1: Grid-world simulation. Left three plots show the value functions and corresponding paths for different CVaR confidence levels. The rightmost plot shows a cost histogram (for 400 Monte Carlo trials) for a risk-neutral policy and a CVaR policy with confidence level $\alpha = 0.11$.

interpolation parameters discussed in Theorem 7, we set $\epsilon = 0.1$ and $\theta = 2.067$ (in order to have 21 logarithmically distributed grid points for the CVaR confidence parameter in $[0, 1]$).

In Figure 1 we plot the value function $V(x, y)$ for three different values of the CVaR confidence parameter α, and the corresponding paths starting from the initial position $(60, 50)$. The first three figures in Figure 1 show how by decreasing the confidence parameter α the average travel distance (and hence fuel consumption) slightly increases but the collision probability decreases, as expected. We next discuss robustness to modeling errors. We conducted simulations in which with probability 0.5 each obstacle position is perturbed in a random direction to one of the neighboring grid cells. This emulates, for example, measurement errors in the terrain map. We then trained both the risk-averse ($\alpha = 0.11$) and risk-neutral ($\alpha = 1$) policies on the nominal (i.e., unperturbed) terrain map, and evaluated them on 400 perturbed scenarios (20 perturbed maps with 20 Monte Carlo evaluations each). While the risk-neutral policy finds a shorter route (with average cost equal to 18.137 on successful runs), it is vulnerable to perturbations and fails more often (with over 120 failed runs). In contrast, the risk-averse policy chooses slightly longer routes (with average cost equal to 18.878 on successful runs), but is much more robust to model perturbations (with only 5 failed runs).

For the computation of Algorithm 1 we represented the concave piecewise linear maximization problem in (10) as a linear program, and concatenated several problems to reduce repeated overhead stemming from the initialization of the CPLEX linear programming solver. This resulted in a computation time on the order of two hours. We believe there is ample room for improvement, for example by leveraging parallelization and sampling-based methods. Overall, we believe our proposed approach is currently the most practical method available for solving CVaR MDPs (as a comparison, the recently proposed method in [8] involves infinite dimensional optimization). The Matlab code used for the experiments is provided in the supplementary material.

6   Conclusion

In this paper we presented an algorithm for CVaR MDPs, based on approximate value-iteration on an augmented state space. We established convergence of our algorithm, and derived finite-time error bounds. These bounds are useful to stop the algorithm at a desired error threshold.

In addition, we uncovered an interesting relationship between the CVaR of the total cost and the worst-case expected cost under adversarial model perturbations. In this formulation, the perturbations are correlated in time, and lead to a robustness framework significantly less conservative than the popular robust-MDP framework, where the uncertainty is temporally independent.

Collectively, our work suggests CVaR MDPs as a unifying and practical framework for computing control policies that are robust with respect to both stochasticity and model perturbations. Future work should address extensions to large state-spaces. We conjecture that a sampling-based approximate DP approach [3] should be feasible since, as proven in this paper, the CVaR Bellman equation is contracting (as required by approximate DP methods).

References

[1] P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3): 203–228, 1999.

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

[3] D. Bertsekas. *Dynamic programming and optimal control, Vol II*. Athena Scientific, 4th edition, 2012.

[4] V. Borkar and R. Jain. Risk-constrained Markov decision processes. *IEEE Transaction of Automatic Control*, 59(9):2574 – 2579, 2014.

[5] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems 27*, pages 3509–3517, 2014.

[6] K. Dowd. *Measuring market risk*. John Wiley & Sons, 2007.

[7] J. Filar, D. Krass, and K. Ross. Percentile performance criteria for limiting average Markov decision processes. *Automatic Control, IEEE Transactions on*, 40(1):2–10, 1995.

[8] W. Haskell and R. Jain. A convex analytic approach to risk-aware Markov decision processes. *SIAM Journal of Control and Optimization*, 2014.

[9] R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18 (7):356–369, 1972.

[10] G. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257280, 2005.

[11] G. Iyengar and A. Ma. Fast gradient descent method for mean-CVaR optimization. *Annals of Operations Research*, 205(1):203–212, 2013.

[12] S. Mannor, D. Simester, P. Sun, and J. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[13] S. Mannor, O. Mebel, and H. Xu. Lightning does not strike twice: Robust MDPs with coupled uncertainty. In *International Conference on Machine Learning*, pages 385–392, 2012.

[14] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[15] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[16] T. Osogami. Robustness and risk-sensitivity in markov decision processes. In *Advances in Neural Information Processing Systems*, pages 233–241, 2012.

[17] G. Pflug and A. Pichler. Time consistent decisions and temporal decomposition of coherent risk functionals. *Optimization online*, 2015.

[18] M. Phillips. *Interpolation and approximation by polynomials*, volume 14. Springer Science & Business Media, 2003.

[19] L. Prashanth. Policy gradients for cvar-constrained mdps. In *Algorithmic Learning Theory*, pages 155–169. Springer, 2014.

[20] R. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[21] R. Rockafellar, S. Uryasev, and M. Zabarankin. Master funds in portfolio analysis with general deviation measures. *Journal of Banking & Finance*, 30(2):743–778, 2006.

[22] G. Serraino and S. Uryasev. Conditional value-at-risk (CVaR). In *Encyclopedia of Operations Research and Management Science*, pages 258–266. Springer, 2013.

[23] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming*. SIAM, 2009.

[24] M. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

[25] A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via sampling. In *AAAI*, 2015.

[26] S. Uryasev, S. Sarykalin, G. Serraino, and K. Kalinchenko. VaR vs CVaR in risk management and optimization. In *CARISMA conference*, 2010.

[27] H. Xu and S. Mannor. The robustness-performance tradeoff in Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 1537–1544, 2006.

A Proofs of Theoretical Results

A.1 Proof of Proposition 1

By definition, we have that

$$\mathbb{E}_{\hat{P}}\left[C(x_0, a_0, \ldots, x_T)\right]$$

$$= \sum_{(x_0, a_0, \ldots, x_T)} P_0(x_0)\delta_1(x_1|x_0, a_0) \cdots P_T(x_T|x_{T-1}, a_{T-1})\delta_T(x_T|x_{T-1}, a_{T-1})C(x_0, a_0, \ldots, x_T)$$

$$= \sum_{(x_0, a_0, \ldots, x_T)} P(x_0, a_0, \ldots, x_T)\delta_1(x_1|x_0, a_0)\delta_2(x_2|x_1, a_1) \cdots \delta_T(x_T|x_{T-1}, a_{T-1})C(x_0, a_0, \ldots, x_T)$$

$$\doteq \sum_{(x_0, a_0, \ldots, x_T)} P(x_0, a_0, \ldots, x_T)\delta(x_0, a_0, \ldots, x_T)C(x_0, a_0, \ldots, x_T).$$

Note that by definition of the set Δ, for any $(\delta_1, \ldots, \delta_T) \in \Delta$ we have that $P(x_0, a_0, \ldots, x_T) > 0 \rightarrow \delta(x_0, a_0, \ldots, x_T) \geq 0$, and

$$\mathbb{E}\left[\delta(x_0, a_0, \ldots, x_T)\right] \doteq \sum_{(x_0, a_0, \ldots, x_T)} P(x_0, a_0, \ldots, x_T)\delta(x_0, a_0, \ldots, x_T) = 1.$$

Thus,

$$\sup_{(\delta_1, \ldots, \delta_T) \in \Delta_\eta} \mathbb{E}_{\hat{P}}\left[C(x_0, a_0, \ldots, x_T)\right]$$

$$= \sup_{\substack{0 \leq \delta(x_0, a_0, \ldots, x_T) \leq \eta, \\ \mathbb{E}[\delta(x_0, a_0, \ldots, x_T)] = 1}} \sum_{(x_0, a_0, \ldots, x_T)} P(x_0, a_0, \ldots, x_T)\delta(x_0, a_0, \ldots, x_T)C(x_0, a_0, \ldots, x_T) = \mathrm{CVaR}_{\frac{1}{\eta}}\left(C(x_0, a_0, \ldots, x_T)\right),$$

where the last equality is by the representation theorem for CVaR [23].

A.2 Proof of Lemma 3

The proof of monotonicity and constant shift properties follow directly from the definitions of the Bellman operator, by noting that $\xi(x')P(x'|x, a)$ is non-negative and $\sum_{x' \in \mathcal{X}} \xi(x')P(x'|x, a)] = 1$ for any $\xi \in \mathcal{U}_{\mathrm{CVaR}}(y, P(\cdot|x, a))$. For the contraction property, denote $c = \|V_1 - V_2\|_\infty$. Since

$$V_2(x, y) - \|V_1 - V_2\|_\infty \leq V_1(x, y) \leq V_2(x, y) + \|V_1 - V_2\|_\infty, \ \forall x \in \mathcal{X}, \ y \in \mathcal{Y},$$

by monotonicity and constant shift property,

$$\mathbf{T}[V_2](x, y) - \gamma\|V_1 - V_2\|_\infty \leq \mathbf{T}[V_1](x, y) \leq \mathbf{T}[V_2](x, y) + \gamma\|V_1 - V_2\|_\infty \ \forall x \in \mathcal{X}, \ y \in \mathcal{Y}.$$

This further implies that

$$|\mathbf{T}[V_1](x, y) - \mathbf{T}[V_2](x, y)| \leq \gamma\|V_1 - V_2\|_\infty \ \forall x \in \mathcal{X}, \ y \in \mathcal{Y}$$

and the contraction property follows.

Now, we prove the concavity preserving property. Assume that $yV(x, y)$ is concave in y for any $x \in \mathcal{X}$. Let $y_1, y_2 \in \mathcal{Y}$, and $\lambda \in [0, 1]$, and define $y_\lambda = (1 - \lambda)y_1 + \lambda y_2$. We have

$$(1 - \lambda)y_1 \mathbf{T}[V](x, y_1) + \lambda y_2 \mathbf{T}[V](x, y_2)$$

$$= (1 - \lambda)y_1 \min_{a_1 \in \mathcal{A}}\left[C(x, a_1) + \gamma \max_{\xi_1 \in \mathcal{U}_{\mathrm{CVaR}}(y_1, P(\cdot|x, a_1))} \sum_{x' \in \mathcal{X}} \xi_1(x')V\left(x', y_1\xi_1(x')\right)P(x'|x, a_1)\right]$$

$$+ \lambda y_2 \min_{a_2 \in \mathcal{A}}\left[C(x, a_2) + \gamma \max_{\xi_2 \in \mathcal{U}_{\mathrm{CVaR}}(y_2, P(\cdot|x, a_2))} \sum_{x' \in \mathcal{X}} \xi_2(x')V\left(x', y_2\xi_2(x')\right)P(x'|x, a_2)\right]$$

$$= \min_{a_1 \in \mathcal{A}}\left[(1 - \lambda)y_1 C(x, a_1) + \gamma \max_{\xi_1 \in \mathcal{U}_{\mathrm{CVaR}}(y_1, P(\cdot|x, a_1))} \sum_{x' \in \mathcal{X}} \xi_1(x')V\left(x', y_1\xi_1(x')\right)P(x'|x, a_1)(1 - \lambda)y_1\right]$$

$$+ \min_{a_2 \in \mathcal{A}}\left[\lambda y_2 C(x, a_2) + \gamma \max_{\xi_2 \in \mathcal{U}_{\mathrm{CVaR}}(y_2, P(\cdot|x, a_2))} \sum_{x' \in \mathcal{X}} \xi_2(x')V\left(x', y_2\xi_2(x')\right)P(x'|x, a_2)\lambda y_2\right]$$

10

$$\leq \min_{a\in\mathcal{A}} \left[y_\lambda C(x,a) + \gamma \max_{\substack{\xi_1\in\mathcal{U}_{\text{CVaR}}(y_1,P(\cdot|x,a))\\ \xi_2\in\mathcal{U}_{\text{CVaR}}(y_2,P(\cdot|x,a))}} \sum_{x'\in\mathcal{X}} P(x'|x,a)\left((1-\lambda)y_1\xi_1(x')V\left(x',y_1\xi_1(x')\right) + \lambda y_2\xi_2(x')V\left(x',y_2\xi_2(x')\right)\right) \right]$$

$$\leq \min_{a\in\mathcal{A}} \left[y_\lambda C(x,a) + \gamma \max_{\substack{\xi_1\in\mathcal{U}_{\text{CVaR}}(y_1,P(\cdot|x,a))\\ \xi_2\in\mathcal{U}_{\text{CVaR}}(y_2,P(\cdot|x,a))}} \sum_{x'\in\mathcal{X}} P(x'|x,a)\left((1-\lambda)y_1\xi_1(x') + \lambda y_2\xi_2(x')\right) V\left(x',\left((1-\lambda)y_1\xi_1(x') + \lambda y_2\xi_2(x')\right)\right) \right]$$

where the first inequality is by concavity of the $\min$, and the second is by the concavity assumption. Now, define $\xi = \frac{(1-\lambda)y_1\xi_1 + \lambda y_2\xi_2}{y_\lambda}$. When $\xi_1 \in \mathcal{U}_{\text{CVaR}}(y_1, P(\cdot|x,a))$ and $\xi_2 \in \mathcal{U}_{\text{CVaR}}(y_2, P(\cdot|x,a))$, we have that $\xi \in \left[0, \frac{1}{y_\lambda}\right]$ and $\sum_{x'\in\mathcal{X}} \xi(x')\mathbb{P}(x'|x,a) = 1$. We thus have

$$(1-\lambda)y_1 \mathbf{T}[V](x,y_1) + \lambda y_2 \mathbf{T}[V](x,y_2)$$

$$\leq \min_{a\in\mathcal{A}} \left[y_\lambda C(x,a) + \gamma \max_{\xi\in\mathcal{U}_{\text{CVaR}}(y_\lambda,P(\cdot|x,a))} \sum_{x'\in\mathcal{X}} P(x'|x,a)y_\lambda\xi(x')V\left(x',y_\lambda\xi(x')\right) \right]$$

$$= y_\lambda \min_{a\in\mathcal{A}} \left[C(x,a) + \gamma \max_{\xi\in\mathcal{U}_{\text{CVaR}}(y_\lambda,P(\cdot|x,a))} \sum_{x'\in\mathcal{X}} P(x'|x,a)\xi(x')V\left(x',y_\lambda\xi(x')\right) \right] = y_\lambda \mathbf{T}[V](x,y_\lambda).$$

Finally, to show that the inner problem in (6) is a concave maximization, we need to show that

$$\Lambda_{x,y,a}(z) := \begin{cases} zV(x',z)P(x'|x,a)/y & \text{if } y \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

is a concave function in $z \in \mathbb{R}$ for any given $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $a \in \mathcal{A}$. Suppose $zV(x,z)$ is a concave function in z. Immediately we can see that $\Lambda_{x,y,a}(z)$ is concave in z when $y = 0$. Also notice that when $y \in \mathcal{Y} \setminus \{0\}$, since the transition probability $P(x'|x,a)$ is non-negative, we have the result that $\Lambda_{x,y,a}(z)$ is concave in z. This further implies

$$\sum_{x'\in\mathcal{X}} \frac{P(x'|x,a)}{y}\Lambda_{x,y,a}(y\xi(x')) = \sum_{x'\in\mathcal{X}} \xi(x')V(x',y\xi(x'))P(x'|x,a)$$

is concave in ξ. Furthermore by combining the result with the fact that the feasible set of ξ is a polytope, we complete the proof of this claim.

A.3 Proof of Theorem 4

The first part of the proof is to show that for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$,
$$V_n(x,y) := \mathbf{T}^n[V_0](x,y) = \min_{\mu\in\Pi_M} \text{CVaR}_y\left(\mathcal{C}_{0,n} + \gamma^n V_0 \mid x_0 = x, \mu\right), \tag{11}$$

by induction, where the initial condition is $(x_0, y_0) = (x,y)$ and control action a_t is induced by $\mu(x_t, y_t)$. For $n = 1$, we have that $V_1(x,y) = \mathbf{T}[V_0](x,y) = \min_{\mu\in\Pi_M} C(x_0, a_0) + \gamma\text{CVaR}_y\left(C(x_1, a_1) + V_0(x_1) \mid x_0 = x, \mu\right)$ from definition. By induction hypothesis, assume the above expression holds at $n = k$. For $n = k+1$,

$$V_{k+1}(x,y) := \mathbf{T}^{k+1}[V_0](x,y) = \mathbf{T}[V_k](x,y)$$

$$= \min_{a\in\mathcal{A}} \left[C(x,a) + \gamma \max_{\xi\in\mathcal{U}_{\text{CVaR}}(y,P(\cdot|x,a))} \sum_{x'\in\mathcal{X}} \xi(x')V_k\left(x', \underbrace{y\xi(x')}_{y'}\right)P(x'|x,a) \right]$$

$$= \min_{a\in\mathcal{A}} \left[C(x,a) + \gamma \max_{\xi\in\mathcal{U}_{\text{CVaR}}(y,P(\cdot|x,a))} \sum_{x'\in\mathcal{X}} \xi(x')P(x'|x,a) \min_{\mu\in\Pi_M} \text{CVaR}_{y'}\left(\mathcal{C}_{0,k} + \gamma^k V_0 \mid x_0 = x', \mu\right) \right]$$

$$= \min_{a\in\mathcal{A}} \left[C(x,a) + \max_{\xi\in\mathcal{U}_{\text{CVaR}}(y,P(\cdot|x,a))} \mathbb{E}_\xi\left[\min_{\mu\in\Pi_M} \text{CVaR}_{y_1}\left(\mathcal{C}_{1,k+1} + \gamma^{k+1} V_0 \mid x_1, \mu\right) \right] \right]$$

$$= \min_{\mu\in\Pi_M} \text{CVaR}_y\left(\mathcal{C}_{0,k+1} + \gamma^{k+1} V_0 \mid x_0 = x, \mu\right),$$

$$\tag{12}$$

where the initial state condition is given by $(x_0, y_0) = (x, y)$. Thus, the equality in (11) is proved by induction.

The second part of the proof is to show that $V^*(x_0, y_0) = \min_{\mu \in \Pi_M} \text{CVaR}_{y_0} (\lim_{n \to \infty} \mathcal{C}_{0,n} \mid x_0, \mu)$. Recall $\mathbf{T}[V](x, y) = \min_{a \in \mathcal{A}} C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x,a))} \mathbb{E}_\xi[V \mid x, y, a]$. Since $\mathbf{T}$ is a contraction and V_0 is bounded, one obtains

$$V^*(x, y) = \mathbf{T}[V^*](x, y) = \lim_{n \to \infty} \mathbf{T}^n[V_0](x, y) = \lim_{n \to \infty} V_n(x, y)$$

for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The first and the second equality follow directly from Proposition 2.1 and Proposition 2.2 in [3] and the third equality follows from the definition of V_n. Furthermore since $V_0(x, y)$ is bounded for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the result in (12) implies

$$- \lim_{n \to \infty} \gamma^n \|V_0\|_\infty \le V^*(x_0, y_0) - \min_{\mu \in \Pi_M} \text{CVaR}_{y_0} \left(\lim_{n \to \infty} \mathcal{C}_{0,n} \mid x_0, \mu \right) \le \lim_{n \to \infty} \gamma^n \|V_0\|_\infty.$$

Therefore, by taking $n \to \infty$, we have just shown that for any $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, $V^*(x_0, y_0) = \min_{\mu \in \Pi_M} \text{CVaR}_{y_0} (\lim_{n \to \infty} \mathcal{C}_{0,n} \mid x_0, \mu)$.

The third part of the proof is to show that for the initial state x_0 and confidence interval y_0, we have that

$$V^*(x_0, y_0) = \min_{\mu \in \Pi_H} \text{CVaR}_{y_0} \left(\lim_{n \to \infty} \mathcal{C}_{0,n} \mid x_0, \mu \right).$$

At any $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, we first define the tth tail-subproblem of problem (3) as follows:

$$\mathbb{V}(x_t, y_t) = \min_{\mu \in \Pi_H} \text{CVaR}_{y_t} \left(\lim_{n \to \infty} \mathcal{C}_{t,n} \mid x_t, \mu \right)$$

where the tail policy sequence is equal to $\mu = \{\mu_t, \mu_{t+1}, \ldots\}$ and the action is given by $a_j = \mu_j(h_j)$ for $j \ge t$. For any history depend policy $\widetilde{\mu} \in \Pi_H$, we also define the $\widetilde{\mu}-$induced value function as $\text{CVaR}_{y_t} (\lim_{n \to \infty} \mathcal{C}_{t,n} \mid x_t, \widetilde{\mu})$ where $\widetilde{\mu} = \{\widetilde{\mu}_t, \widetilde{\mu}_{t+1}, \ldots\}$ and $a_j = \widetilde{\mu}_j(h_j)$ for $j \ge t$.

Now let μ^* be the optimal policy of the above t^{th} tail-subproblem. Clearly, the truncated policy $\widetilde{\mu} = \{\mu_{t+1}^*, \mu_{t+2}^*, \ldots\}$ is a feasible policy for the $(t+1)^{\text{th}}$ tail subproblem at any state x_{t+1} and confidence interval y_{t+1}:

$$\min_{\mu \in \Pi_H} \text{CVaR}_{y_{t+1}} \left(\lim_{n \to \infty} \mathcal{C}_{t+1,n} \mid x_{t+1}, \mu \right).$$

Collecting the above results, for any pair $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and with $a_t = \mu_t^*(x_t)$ we can write

$$\mathbb{V}(x_t, y_t) = C(x_t, a_t) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot|x_t, a_t))} \mathbb{E} \left[\xi(x_{t+1}) \cdot \underbrace{\text{CVaR}_{y_{t+1}} \left(\lim_{n \to \infty} \mathcal{C}_{t+1,n} \mid x_{t+1}, \widetilde{\mu} \right)}_{\mathbb{V}^{\widetilde{\mu}}(x_{t+1}, y_{t+1}), y_{t+1} = y_t \xi(x_{t+1})} \right]$$

$$\ge C(x_t, a_t) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot|x_t, a_t))} \mathbb{E}_\xi[\mathbb{V}(x_{t+1}, y_t \xi(x_{t+1})) \mid x_t, y_t, a_t] \ge \mathbf{T}[\mathbb{V}](x_t, y_t).$$

The first equality follows from the definition of $\mathbb{V}(x_t, y_t)$ and the decomposition of CVaRs (Theorem 2). The first inequality uses the inequality: $\mathbb{V}^{\widetilde{\mu}}(x, y) \ge \mathbb{V}(x, y), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$. The second inequality follows from the definition of Bellman operator $\mathbf{T}$.

On the other hand, starting at any state x_{t+1} and confidence interval y_{t+1}, let $\mu^* = \{\mu_{t+1}^*, \mu_{t+2}^*, \ldots\} \in \Pi_H$ be an optimal policy for the tail subproblem:

$$\min_{\mu \in \Pi_H} \text{CVaR}_{y_{t+1}} \left(\lim_{n \to \infty} \mathcal{C}_{t+1,n} \mid x_{t+1}, \mu \right).$$

For a given pair of $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, construct the "extended" policy $\widetilde{\mu} = \{\widetilde{\mu}_t, \widetilde{\mu}_{t+1}, \ldots\} \in \Pi_H$ as follows:

$$\widetilde{\mu}_t(x_t) = u^*(x_t, y_t), \text{ and } \widetilde{\mu}_j(h_j) = \mu_j^*(h_j) \text{ for } j \ge t+1,$$

where $u^*(x_t, y_t)$ is the minimizer of the fixed point equation

$$u^*(x_t, y_t) \in \operatorname*{argmin}_{a \in \mathcal{A}} C(x_t, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot|x_t, a))} \mathbb{E}_\xi[\mathbb{V}(x_{t+1}, y_t \xi(x_{t+1})) \mid x_t, y_t, a],$$

with y_t is the given confidence interval to the tth tail-subproblem and the transition from y_t to y_{t+1} is given by $y_{t+1} = y_t\xi^*(x_{t+1})$ where

$$\xi^* \in \arg\max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t,P(\cdot|x_t,a^*))} \mathbb{E}\left[\xi(x_{t+1}) \cdot \text{CVaR}_{y_t\xi(x_{t+1})}\left(\lim_{n\to\infty} \mathcal{C}_{t+1,n} \mid x_{t+1},\widetilde{\mu}\right)\right]$$

Since μ^* is an optimal, and a fortiori feasible policy for the tail subproblem (from time $t + 1$), the policy $\widetilde{\mu} \in \Pi_H$ is a feasible policy for the tail subproblem (from time t): $\min_{\mu \in \Pi_H} \text{CVaR}_{y_t}\left(\lim_{n\to\infty} \mathcal{C}_{t,n} \mid x_t, \mu\right)$. Hence, we can write

$$\mathbb{V}(x_t, y_t) \leq C(x_t, \widetilde{\mu}_t(x_t)) + \gamma\text{CVaR}_{y_t}\left(\lim_{n\to\infty} \mathcal{C}_{t+1,n} \mid x_t, \widetilde{\mu}\right).$$

Hence from the definition of μ^*, one easily obtains:

$\mathbb{V}(x_t, y_t)$

$\leq C(x_t, u^*(x_t, y_t)) + \gamma \max\limits_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t,P(\cdot|x_t,u^*(x_t,y_t)))} \mathbb{E}\left[\xi(x_{t+1}) \cdot \text{CVaR}_{y_t\xi(x_{t+1})}\left(\lim_{n\to\infty} \mathcal{C}_{t+1,n} \mid x_{t+1},\widetilde{\mu}\right) \mid x_t, y_t, u^*(x_t, y_t)\right]$

$= C(x_t, u^*(x_t, y_t)) + \gamma \max\limits_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t,P(\cdot|x_t,u^*(x_t,y_t)))} \mathbb{E}_\xi\left[\mathbb{V}(x_{t+1}, y_t\xi(x_{t+1})) \mid x_t, y_t, u^*(x_t, y_t)\right]$

$= \mathbf{T}[\mathbb{V}](x_t, y_t).$

Collecting the above results, we have shown that $\mathbb{V}$ is a fixed point solution to $V(x, y) = \mathbf{T}[V](x, y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Since the fixed point solution is unique, combining both of these arguments implies $V^*(x, y) = \mathbb{V}(x, y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Therefore, it follows that with initial state (x, y), we have $V^*(x, y) = \mathbb{V}(x, y) = \min_{\mu \in \Pi_H} \text{CVaR}_y\left(\lim_{T\to\infty} \mathcal{C}_{0,T} \mid x_0 = x, \mu\right)$.

Combining the above three parts of the proof, the claims of this theorem follows.

A.4   Proof of Theorem 5

Similar to the definition of the optimal Bellman operator $\mathbf{T}$, for any augmented stationary Markovin policy $u : \mathcal{X} \times \mathcal{Y} \to \mathcal{A}$, we define the policy induced Bellman operator $\mathbf{T}_u$ as

$$\mathbf{T}_u[V](x, y) = C(x, u(x, y)) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y,P(\cdot|x,u(x,y)))} \sum_{x' \in \mathcal{X}} \xi(x')V\left(x', y\xi(x')\right) P(x'|x, u(x, y)).$$

Analogous to Theorem 4, we can easily show that the fixed point solution to $\mathbf{T}_u[V](x, y) = V(x, y)$ is unique and the CVaR decomposition theorem (Theorem 2) further implies this solution equals to

$$\text{CVaR}_y\left(\lim_{T\to\infty} \mathcal{C}_{0,T} \mid x_0 = x, u_H\right),$$

where the history dependent policy $\pi_H = \{\mu_0, \mu_1, \ldots\}$ is given by $\mu_k(h_k) = u(x_k, y_k)$ for any $k \geq 0$, with initial states $x_0, y_0 = \alpha$, state transitions (8), but with augmented stationary Markovian policy u^* replaced by u.

To complete the proof of this theorem, we need to show that the augmented stationary Markovian policy u^* is optimal if and only if

$$\mathbf{T}[V^*](x, y) = \mathbf{T}_{u^*}[V^*](x, y), \ \forall x \in \mathcal{X}, \ y \in \mathcal{Y}, \tag{13}$$

where $V^*(x, y)$ is the unique fixed point solution of $\mathbf{T}[V](x, y) = V(x, y)$. Here an augmented stationary Markovian policy u^* is optimal if and only if the induced history dependent policy u_H^* in (7) is optimal to problem (3).

First suppose u^* is an optimal augmented stationary Markvoian policy. Then using the definition of u^* and the result from Theorem 4 that

$$V^*(x, y) = \min_{\mu \in \Pi_H} \text{CVaR}_y\left(\lim_{T\to\infty} \mathcal{C}_{0,T} \mid x_0 = x, \mu\right),$$

we immediately show that $V^*(x, y) = V_{u^*}(x, y)$. By the fixed point equation $\mathbf{T}[V^*](x, y) = V^*(x, y)$ and $\mathbf{T}_{u^*}[V_{u^*}](x, y) = V_{u^*}(x, y)$, this further implies (13) holds.

Second suppose u^* satisfies the equality in (13). Then by the fixed point equality $\mathbf{T}[V^*](x, y) = V^*(x, y)$, we immediately obtain the equation $V^*(x, y) = \mathbf{T}_{u^*}[V^*](x, y)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. since the fixed point solution to $\mathbf{T}_{u^*}[V](x, y) = V(x, y)$

is unique, we further show that $\mathbf{T}[V^*](x,y) = V^*(x,y) = V_{u^*}(x,y)$ and $V_{u^*}(x,y) = \min_{\mu \in \Pi_H} \mathrm{CVaR}_y \left(\lim_{T\to\infty} \mathcal{C}_{0,T} \mid x_0 = x, \mu \right)$ from Theorem 4. By using the policy construction formula in (7) to obtain the history dependent policy u_H^* and following the above arguments at which the augmented Markovian stationary policy u is replaced by u^*, this further implies

$$\min_{\mu \in \Pi_H} \mathrm{CVaR}_y \left(\lim_{T\to\infty} \mathcal{C}_{0,T} \mid x_0 = x, \mu \right) = \mathrm{CVaR}_y \left(\lim_{T\to\infty} \mathcal{C}_{0,T} \mid x_0 = x, u_H^* \right),$$

i.e., u^* is an optimal augmented stationary Markovian policy.

A.5 Proof of Lemma 6

We first proof the monotonicity property. Based on the definition of $\mathcal{I}_x[V](y)$, if $V_1(x,y) \geq V_2(x,y)$ $\forall x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have that

$$\mathcal{I}_x[V_1](y) = \frac{y_{i+1}V_1(x,y_{i+1})(y - y_i) + y_i V_1(x,y_i)(y_{i+1} - y)}{y_{i+1} - y_i}, \text{ if } y \in \mathbf{I}_i(x).$$

Since $y_i, y_{i+1} \in \mathcal{Y}$ and $(y_{i+1} - y), (y - y_i) \geq 0$ (because $y \in \mathbf{I}_i(x)$), we can easily see that $\mathcal{I}_x[V_1](y) \geq \mathcal{I}_x[V_2](y)$. As $y \in \mathcal{Y}$ and $\xi(\cdot)P(\cdot|x,a) \geq 0$ for any $\xi \in \mathcal{U}_{\mathrm{CVaR}}(y, P(\cdot|x,a))$, this further implies $\mathbf{T}_{\mathcal{I}}[V_1](x,y) \geq \mathbf{T}_{\mathcal{I}}[V_2](x,y)$.

Next we prove the constant shift property. Note from the definition of $\mathcal{I}_x[V](y)$ that

$$\mathcal{I}_x[V + K](y)$$
$$= y_i(V(x,y_i) + K) + \frac{y_{i+1}(V(x,y_{i+1}) + K) - y_i(V(x,y_i) + K)}{y_{i+1} - y_i}(y - y_i), \text{ if } y \in \mathbf{I}_i(x),$$
$$= yK + y_i V(x,y_i) + \frac{y_{i+1}V(x,y_{i+1}) - y_i V(x,y_i)}{y_{i+1} - y_i}(y - y_i), \text{ if } y \in \mathbf{I}_i(x)$$
$$= \mathcal{I}_x[V](y) + yK.$$

Therefore by definition of $\mathbf{T}_{\mathcal{I}}[V](x,y)$, the constant shift property: $\mathbf{T}_{\mathcal{I}}[V + K](x,y) = \mathbf{T}_{\mathcal{I}}[V](x,y) + \gamma K$ for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, follows directly from the above arguments.

Equipped with both properties in monotonicity and constant shift, the proof of contraction of $\mathbf{T}_{\mathcal{I}}$ directly follows from the analogous proof in Lemma 3.

Finally we prove the concavity preserving property. Assume $yV(x,y)$ is concave in $y \in \mathcal{Y}$ for any $x \in \mathcal{X}$. Then for $y_{i+2} > y_{i+1} > y_i$, $\forall i \in \{1, \ldots, N(x) - 2\}$ the following inequality immediately follows from the definition of a concave function:

$$\left. \frac{d\mathcal{I}_x[V](y)}{dy} \right|_{y \in \mathbf{I}_{i+1}(x)} = \frac{y_{i+1}V(x,y_{i+1}) - y_i V(x,y_i)}{y_{i+1} - y_i}$$
$$\geq \frac{y_{i+2}V(x,y_{i+2}) - y_{i+1}V(x,y_{i+1})}{y_{i+2} - y_{i+1}} = \left. \frac{d\mathcal{I}_x[V](y)}{dy} \right|_{y \in \mathbf{I}_{i+2}(x)}. \tag{14}$$

We then show that the following inequality in each of the following cases, whenever the slope exists:

$$\mathcal{I}_x[V](z_1) \leq \mathcal{I}_x[V](z_2) + \left. \frac{d\mathcal{I}_x[V](y)}{dy} \right|_{y=z_2} (z_1 - z_2), \ \forall z_1, z_2 \in \mathcal{Y} \setminus \{0\}.$$

(1) There exists $i \in \{1, \ldots, N(x) - 1\}$ such that $z_1, z_2 \in \mathbf{I}_{i+1}(x)$. In this case we have that

$$\left. \frac{d\mathcal{I}_x[V](y)}{dy} \right|_{y=z_1} = \left. \frac{d\mathcal{I}_x[V](y)}{dy} \right|_{y=z_2},$$

and this further implies

$$\mathcal{I}_x[V](z_1) = \mathcal{I}_x[V](z_2) + \left. \frac{d\mathcal{I}_x[V](y)}{dy} \right|_{y=z_2} (z_1 - z_2).$$

(2) There exists $i, j \in \{1, \ldots, N(x) - 2\}$, $i + 1 < j$ such that $z_1 \in \mathbf{I}_{i+1}(x)$ and $z_2 \in \mathbf{I}_j(x)$. In this case, without loss of generality we assume $j = i + 1$. The proof for case: $j > i + 2$ is omitted

14

for the sake of brevity, as it can be completed by iteratively applying the same arguments from case: $j = i + 2$. Since $z_1 \in \mathbf{I}_i(x)$, $z_2 \in \mathbf{I}_j(x)$, we have $z_2 - z_1 \geq 0$ and

$$\left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_1} \geq \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_2}.$$

Based on the definition of the linear interpolation function, we have that

$$\mathcal{I}_x[V](y_{i+1}) = y_{i+1} V(x, y_{i+1}) = \mathcal{I}_x[V](y_i) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y \in \mathbf{I}_{i+1}(x)} (y_{i+1} - y_i).$$

Furthermore, combining previous arguments with the definitions of $\mathcal{I}_x[V](z_1)$, $\mathcal{I}_x[V](z_2)$ implies that for $(z_2 - y_{i+1}) \geq 0$,

$$\begin{aligned}
\mathcal{I}_x[V](z_2) =& \mathcal{I}_x[V](y_{i+1}) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_2} (z_2 - y_{i+1}) \\
\leq& \mathcal{I}_x[V](y_{i+1}) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_1} (z_2 - y_{i+1}) \\
=& \mathcal{I}_x[V](y_i) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y \in \mathbf{I}_{i+1}(x)} (z_2 - y_i) \\
=& \mathcal{I}_x[V](z_1) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_1} (z_2 - z_1).
\end{aligned}$$

(3) There exists $i, j \in \{1, \ldots, N(x) - 2\}$, $i + 1 < j$ such that $z_2 \in \mathbf{I}_{i+1}(x)$ and $z_1 \in \mathbf{I}_j(x)$. In this case, without loss of generality we assume $j = i + 1$. The proof for case: $j > i + 2$ is omitted for the sake of brevity, as it can be completed by iteratively applying the same arguments from case: $j = i + 2$. Since $z_2 \in \mathbf{I}_{i+1}(x)$, $z_1 \in \mathbf{I}_j(x)$, we have $z_1 - z_2 \geq 0$ and

$$\left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_1} \leq \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_2}.$$

Similar to the analysis in the previous case, we have that

$$\mathcal{I}_x[V](y_i) = y_i V(x, y_i) = \mathcal{I}_x[V](y_{i+1}) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y \in \mathbf{I}_{i+1}(x)} (y_i - y_{i+1})$$

Furthermore, combining previous arguments with the definitions of $\mathcal{I}_x[V](z_1)$, $\mathcal{I}_x[V](z_2)$ implies that for $(z_2 - z_1) \leq 0$,

$$\begin{aligned}
\mathcal{I}_x[V](z_2) =& \mathcal{I}_x[V](y_i) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_2} (z_2 - y_i) \\
=& \mathcal{I}_x[V](y_{i+1}) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_2} (z_2 - y_{i+1}) \\
=& \mathcal{I}_x[V](z_1) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_2} (z_2 - z_1) \\
\leq& \mathcal{I}_x[V](z_1) + \left.\frac{d\mathcal{I}_x[V](y)}{dy}\right|_{y=z_1} (z_2 - z_1).
\end{aligned}$$

Thus we have just shown that the first order sufficient condition for concave functions, corresponding to $\mathcal{I}_x[V](y)$, holds, i.e., $\mathcal{I}_x[V](y)$ is concave in $y \in \mathcal{Y} \setminus \{0\}$ for any given $x \in \mathcal{X}$. Now since $\mathcal{I}_x[V](y)$ is a continuous piecewise linear function in $y \in \mathcal{Y}$ and a concave function when the domain is restricted to $\mathcal{Y} \setminus \{0\}$. By continuity this immediately implies that $\mathcal{I}_x[V](y)$ is concave in $y \in \mathcal{Y}$ as well. Then following the identical arguments in the proof of Lemma 3 for the concavity preserving property, we can thereby show that

$$y\mathbf{T}_{\mathcal{I}}[V](x, y) = \min_{a \in \mathcal{A}} \left\{ yC(x, a) + \max_{\xi \in \mathcal{U}_{\mathrm{CVaR}}(y, P(\cdot|x,a))} \sum_{x' \in \mathcal{X}} \mathcal{I}_{x'}[V](y\xi(x')) P(x'|x, a) \right\}$$

is concave in $y \in \mathcal{Y}$ for any given $x \in \mathcal{X}$.

A.6 Useful Intermediate Results

Lemma 8 *Let $f(y) : [0,1] \to R$ be a concave function, differentiable almost everywhere, with Lipschitz constant M. Then the linear interpolation $\mathcal{I}[f](y)$ is also concave, and with Lipschitz constant $M_I \leq M$.*

Proof For every segment $[y_j, y_{j+1}]$ in the linear interpolation, $f(y)$ is concave, and with Lipschitz constant M, and $\mathcal{I}[f](y)$ is linear. Also, $f(y_j) = \mathcal{I}[f](y_j)$, and $f(y_{j+1}) = \mathcal{I}[f](y_{j+1})$, by definition of the linear interpolation. Denote by c_j the magnitude of the slope of $\mathcal{I}[f](y)$ at $y \in [y_j, y_{j+1}]$.

Assume by contradiction that $c_j > \max_{y \in [y_j, y_{j+1}]} |f'(y)|$ whenever $f'(y)$ exists. Consider the case when $f(y_{j+1}) \geq f(y_j)$. This implies c_j is the slope of the interpolation function $\mathcal{I}[f](y)$ at $y \in [y_j, y_{j+1}]$. Then by the fundamental theorem of calculus, we have

$$f(y_{j+1}) - f(y_j) = \int_{y_j}^{y_{j+1}} f'(y)dy \leq \int_{y_j}^{y_{j+1}} |f'(y)|dy < \int_{y_j}^{y_{j+1}} c_j dy = (\mathcal{I}[f](y_{j+1}) - \mathcal{I}[f](y_j)),$$

contradicting $f(y_{j+1}) = \mathcal{I}[f](y_{j+1})$ and $f(y_j) = \mathcal{I}[f](y_j)$.

On the other hand, consider the case when $f(y_{j+1}) \leq f(y_j)$. This implies $-c_j$ is the slope of the interpolation function $\mathcal{I}[f](y)$ at $y \in [y_j, y_{j+1}]$. Again by fundamental theorem of calculus,

$$0 \leq f(y_{j+1}) - f(y_j) = \int_{y_j}^{y_{j+1}} f'(y)dy \geq \int_{y_j}^{y_{j+1}} -|f'(y)|dy > \int_{y_j}^{y_{j+1}} -c_j dy = \mathcal{I}[f](y_j) - \mathcal{I}[f](y_{j+1}).$$

Since $f(y_{j+1}) = \mathcal{I}[f](y_{j+1})$ and $f(y_j) = \mathcal{I}[f](y_j)$, which implies $\mathcal{I}[f](y_j) - \mathcal{I}[f](y_{j+1}) \geq 0$, the above expression clearly leads to a contradiction.

We finally have that $\max_{y \in [y_j, y_{j+1}]} |f'(y)| \geq c_j$ for segment $j \in \{1, \dots, N(x) - 1\}$. As this argument holds for each segment, by maximizing over j over $\{1, \dots, N(x) - 1\}$, we have that

$$M \geq \max_{j \in \{1, \dots, N(x)-1\}} \max_{y \in [y_j, y_{j+1}]} |f'(y)| \geq \max_{j \in \{1, \dots, N(x)-1\}} c_j = M_I.$$

The concavity property (thus differentiability almost everywhere) are well-known results of linear interpolation [18].

Lemma 9 *Let $yV(x,y)$ be Lipschitz with constant M, concave, and differentiable almost everywhere, for every $x \in \mathcal{X}$ and $y \in [0,1]$. Then $y\mathbf{T}[V](x,y)$ is also Lipschitz with constant $C_{\max} + \gamma M$.*

Proof For any given state-action pair $x \in \mathcal{X}$, and $a \in \mathcal{A}$, let $P(x') = P(x'|x,a)$ be the transition kernel. Consider the function

$$H(y) \doteq \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot))} \sum_{x' \in \mathcal{X}} y\xi(x')V(x', y\xi(x')) P(x').$$

Note that, by definition of $\mathcal{U}_{\text{CVaR}}$, and a change of variables $z(x') = y\xi(x')$, we can write $H(y)$ as follows:

$$H(y) = \max_{\substack{0 \leq z(x') \leq 1, \\ \sum_{x'} P(x')z(x') = y}} \sum_{x' \in \mathcal{X}} z(x')V(x', z(x')) P(x'). \tag{15}$$

The Lagrangian of the above maximization problem is

$$L(z, \lambda; y) = \sum_{x' \in \mathcal{X}} z(x')V(x', z(x')) P(x') - \lambda(\sum_{x'} P(x')z(x') - y).$$

Since $yV(x,y)$ is concave, the maximum is attained. By first order optimality condition the following expression holds:

$$\frac{\partial L(z, \lambda; y)}{\partial z(x')} = P(x')\frac{\partial \left[z(x')V(x', z(x'))\right]}{\partial z(x')} - \lambda P(x') = 0.$$

Summing the last expression over x', we obtain:

$$\sum_{x' \in \mathcal{X}} P(x')\frac{\partial \left[z(x')V(x', z(x'))\right]}{\partial z(x')} = \sum_{x' \in \mathcal{X}} \lambda P(x') = \lambda.$$

Now, from the Lipschitz property of $yV(x, y)$, we have

$$\left| \sum_{x' \in \mathcal{X}} \lambda P(x') \right| \leq \sum_{x' \in \mathcal{X}} P(x') \left| \frac{\partial \left[z(x')V\left(x', z(x')\right) \right]}{\partial z(x')} \right| \leq \sum_{x' \in \mathcal{X}} P(x')M = M.$$

Thus,

$$|\lambda| \leq \sum_{x' \in \mathcal{X}} P(x') \left| \frac{\partial \left[z(x')V\left(x', z(x')\right) \right]}{\partial z(x')} \right| \leq M.$$

Note that the objective in (15) does not depend on y. From the envelope theorem [14], it follows that

$$\frac{dH(y)}{dy} = \lambda,$$

therefore, $H(y)$ is Lipschitz, with constant M.

Now, by definition,

$$y\mathbf{T}[V](x, y) = \min_{a \in \mathcal{A}} \left[yC(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\mathrm{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} y\xi(x')V\left(x', y\xi(x')\right) P(x'|x, a) \right].$$

Using our Lipschitz result for $H(y)$, we have that for any $a \in \mathcal{A}$, the function

$$yC(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\mathrm{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} y\xi(x')V\left(x', y\xi(x')\right) P(x'|x, a)$$

is Lipschitz in y, with constant $C(x, a) + \gamma M$. Using again the envelope theorem [14], we obtain that $y\mathbf{T}[V](x, y)$ is Lipschitz, with constant $C_{\max} + \gamma M$.

Lemma 10 *Consider Algorithm 1. Assume that for any $x \in \mathcal{X}$, the initial value function satisfies that $yV_0(x, y)$ is Lipschitz (in y), with uniform constant M_0. We have that for any $t \in \{0, 1, \ldots, \}$, the function $yV_t(x, y)$ is Lipschitz in y for any $x \in \mathcal{X}$, with Lipschitz constant*

$$M_t = \frac{1 - \gamma^t}{1 - \gamma} C_{\max} + \gamma^t M_0 \leq \frac{C_{\max}}{1 - \gamma} + M_0, \ \forall t.$$

Proof Let $\mathbf{T}_{\mathcal{I}}[V]$ denote the application of the Bellman operator $\mathbf{T}$ to the linearly-interpolated version of $yV(x, y)$. We have, by definition, that

$$V_1(x, y) = \mathbf{T}_{\mathcal{I}}[V_0](x, y).$$

Using Lemma 8 and Lemma 9, we have that $V_1(x, y)$ is Lipschitz, with $M_1 \leq C_{\max} + \gamma M_0$.

Note now, that $V_2(x, y) = \mathbf{T}_{\mathcal{I}}[V_1](x, y)$. Thus, by induction, we have

$$M_t \leq \frac{1 - \gamma^t}{1 - \gamma} C_{\max} + \gamma^t M_0,$$

and the result follows.

A.7   Proof of Theorem 7

The proof of this theorem is split into three parts. In the first part, we bound the difference $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$ at each state $(x, y) \in \mathcal{X} \times \mathcal{Y}$ using the previous technical lemmas and Lipschitz property.

In the second part, we bound the difference of $\mathbf{T}_{\mathcal{I}}[V_t](x, y) - \mathbf{T}[V_t](x, y)$.

In the third part we bound the interpolation error using contraction properties of Bellman recursions.

First we analyze the bounds for $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$ in the following four cases. Notice that from Lemma 10, we have that $|d\mathcal{I}_x[V_t](y)/dy| \leq M := C_{\max}/(1 - \gamma) + M_0$.

(1) When $y = 0$ (for which $y \in \mathbf{I}_1(x)$).
Using previous analysis and L'Hospital's rule we have that $\lim_{y \to 0} \mathcal{I}_x[V_t](y)/y = V_t(x, 0)$. This further implies $\lim_{y \to 0} \mathcal{I}_x[V_t](y)/y - V_t(x, 0) = 0$.

(2) When $y \in \mathbf{I}_{i+1}(x)$, $2 \le i < N(x) - 1$.
Similar to the inequality in (14), by concavity of $yV_t(x, y)$ in $y \in \mathcal{Y}$, we have that

$$\frac{d\mathcal{I}_x[V_t](y)}{dy}\bigg|_{y \in \mathbf{I}_{i+1}(x)} = \frac{y_{i+1}V_t(x, y_{i+1}) - y_i V_t(x, y_i)}{y_{i+1} - y_i} \le \frac{yV_t(x, y) - y_i V_t(x, y_i)}{y - y_i},$$

and

$$\frac{d\mathcal{I}_x[V_t](y)}{dy}\bigg|_{y \in \mathbf{I}_{i+2}(x)} = \frac{y_{i+2}V_t(x, y_{i+2}) - y_{i+1}V_t(x, y_{i+1})}{y_{i+2} - y_{i+1}} \le \frac{y_{i+1}V_t(x, y_{i+1}) - yV_t(x, y)}{y_{i+1} - y}.$$

From the first inequality, for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we get,

$$\frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y) \le \frac{1}{y}\left(y_i V_t(x, y_i) + \frac{y_{i+1}V_t(x, y_{i+1}) - y_i V_t(x, y_i)}{y_{i+1} - y_i}(y - y_i) - yV_t(x, y)\right) \le 0. \tag{16}$$

On the other hand, rearranging the second inequality gives

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y))$$

$$\ge \frac{1}{y}\left(y_i V_t(x, y_i) + \frac{d\mathcal{I}_x[V_t](y)}{dy}\bigg|_{y \in \mathbf{I}_{i+1}(x)}(y - y_i) - y_{i+1}V_t(x, y_{i+1}) - \frac{d\mathcal{I}_x[V_t](y)}{dy}\bigg|_{y \in \mathbf{I}_{i+2}(x)}(y - y_{i+1})\right)$$

$$= \left(\frac{d\mathcal{I}_x[V_t](y)}{dy}\bigg|_{y \in \mathbf{I}_{i+1}(x)} - \frac{d\mathcal{I}_x[V_t](y)}{dy}\bigg|_{y \in \mathbf{I}_{i+2}(x)}\right)\frac{y - y_{i+1}}{y} \ge -2M\left(\frac{y_{i+1}}{y} - 1\right). \tag{17}$$

Furthermore by the Lipschitz property, we also have the following inequality as well:

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y))$$

$$= \frac{y_{i+1}V_t(x, y_{i+1})(y - y_i) + y_i V_t(x, y_i)(y_{i+1} - y)}{(y_{i+1} - y_i)y} - V_t(x, y)$$

$$\ge \frac{y_i V_t(x, y_i)(y - y_i) + y_i V_t(x, y_i)(y_{i+1} - y) - M(y_{i+1} - y_i)(y - y_i)}{(y_{i+1} - y_i)y} - V_t(x, y) \tag{18}$$

$$= \frac{y_i V_t(x, y_i) - M(y - y_i)}{y} - V_t(x, y) \ge -2M\left(1 - \frac{y_i}{y}\right).$$

Combining the inequalities (17) and (18), the following lower bound for $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$ holds:

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \ge \delta := -2M\min\left\{1 - \frac{y_i}{y}, \frac{y_{i+1}}{y} - 1\right\}, \ \forall y \in \mathbf{I}_{i+1}(x), \ i \ge 2.$$

From the above definition, when $y_i \le y \le (y_i + y_{i+1})/2$, the lower bound becomes $\delta = -2M(1 - y_i/y)$ and when $(y_i + y_{i+1})/2 \le y \le y_{i+1}$, the corresponding lower bound is $\delta = -2M(y_{i+1}/y - 1)$. In both cases, δ is minimized when $y = (y_i + y_{i+1})/2$. Therefore, the above analysis implies the following lower bound:

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \ge -2M\frac{y_{i+1} - y_i}{y_{i+1} + y_i}, \ \forall y \in \mathbf{I}_{i+1}(x), \ i \ge 2.$$

When $y_{i+1} = \theta y_i$ for $i \in \{2, \ldots, N(x) - 1\}$ for some constant $\theta \ge 1$, this further implies that

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \ge -2M\frac{\theta - 1}{\theta + 1} \ge -M(\theta - 1), \ \forall y \in \mathcal{Y} \setminus [0, \epsilon].$$

Then combining the results, here we get the following bound for $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$:

$$-M(\theta - 1) \le \frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y) \le 0, \ \forall y \in \mathbf{I}_{i+1}(x), \ i \ge 2.$$

18

(3) When $y \in \mathbf{I}_{N(x)}(x)$, i.e., $y \in (y_{N(x)-1}, 1]$.

Similar to the proof of case (2), we can show that for any $x \in \mathcal{X}$ and $y \in \mathbf{I}_{N(x)}(x)$, the same lines of arguments in inequality (16) and (18) hold, which implies

$$-2M\left(1 - y_{N(x)-1}\right) \leq -2M\left(1 - \frac{y_{N(x)-1}}{y}\right) \leq \frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x,y)) \leq 0.$$

When $y_{N(x)} = 1 = \theta y_{N(x)-1}$, this further shows that

$$-2My_{N(x)-1}(\theta - 1) = -2M\left(y_{N(x)} - y_{N(x)-1}\right) \leq \frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x,y)) \leq 0,$$

and

$$-2M(\theta - 1) \leq -\frac{2}{\theta}M(\theta - 1) \leq \frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x,y)) \leq 0.$$

(4) When $y \in \mathbf{I}_2(x)$, i.e., $y \in (0, y_2]$.

From inequality (16), the definition of $\mathcal{I}_x[V_t](y)$, we have that

$$0 \geq \frac{\mathcal{I}_x[V_t](y) - yV_t(x,y)}{y} = \frac{y(V_t(x,y_2) - V_t(x,y))}{y} = V_t(x,y_2) - V_t(x,y) \geq V_t(x,y_2) - V_t(x,0).$$

The first inequality is due to the fact that $yV_t(x,y)$ is concave in $y \in \mathcal{Y}$ for any $x \in \mathcal{X}$, thus the first order condition implies

$$\frac{y_2 V_n(x,y_2) - y_1 V_n(x,y_1)}{y_2 - y_1} \leq \frac{y V_n(x,y) - y_1 V_n(x,y_1)}{y - y_1}, \ \forall y \in \mathbf{I}_2(x),$$

and the last inequality is due to the similar fact that

$$V_t(x,w) = \frac{wV_t(x,w) - 0 \cdot V_t(x,0)}{w - 0} \leq \frac{zV_t(x,z) - 0 \cdot V_t(x,0)}{z - 0} = V_t(x,z), \ \forall z, w \in \mathcal{Y}, \ z \leq w.$$

Therefore the condition of this theorem implies

$$0 \geq \frac{\mathcal{I}_x[V_t](y) - yV_t(x,y)}{y} \geq -\epsilon, \ \forall t \geq 0, \ x \in \mathcal{X}, \ y \in \mathcal{Y}.$$

Combining the above four cases, we have that for each state $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$0 \geq \frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x,y) \geq -2M(\theta - 1) - \epsilon, \ \forall t.$$

Second, we bound the difference of $\mathbf{T}_{\mathcal{I}}[V_t](x,y) - \mathbf{T}[V_t](x,y)$. By recalling that $\xi(\cdot)P(\cdot|x,a)$ is a probability distribution for any $\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x,a))$, we then combine all previous arguments and show that at any $t \in \{0, 1, \ldots, \}$ and any $x \in \mathcal{X}, a \in \mathcal{A}, y \in \mathbf{Y}(x)$,

$$\max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x,a))} \sum_{x' \in \mathcal{X}, \xi(x') \neq 0} \left(\frac{\mathcal{I}_{x'}[V_t](y\xi(x'))}{y\xi(x')} - V_t(x', y\xi(x'))\right)\xi(x')P(x'|x,a) \geq -2M(\theta-1) - \epsilon.$$

This further implies

$$\mathbf{T}[V_t](x,y) - \gamma(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_t](x,y) \leq \mathbf{T}[V_t](x,y). \tag{19}$$

Third, we prove the error bound of interpolation based value iteration using the above properties. By putting $t = 0$ in (19), we have that

$$-\gamma(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_0](x,y) - \mathbf{T}[V_0](x,y) \leq 0.$$

Applying the Bellman operator $\mathbf{T}$ on all sides of the above inequality and noting that $\mathbf{T}$ is a translational invariant mapping, the above expression implies

$$\mathbf{T}^2[V_0](x,y) - \gamma^2(2M(\theta - 1) + \epsilon) \leq \mathbf{T}[\mathbf{T}_{\mathcal{I}}[V_0]](x,y) = \mathbf{T}[V_1](x,y) \leq \mathbf{T}^2[V_0](x,y).$$

By adding the inequality: $-\gamma(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_1](x,y) - \mathbf{T}[V_1](x,y) \leq 0$ to the above expression, this further implies the following expression:

$$\mathbf{T}^2[V_0](x,y) - \gamma(1 + \gamma)(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_1](x,y) = \mathbf{T}_{\mathcal{I}}^2[V_0](x,y) \leq \mathbf{T}^2[V_0](x,y).$$
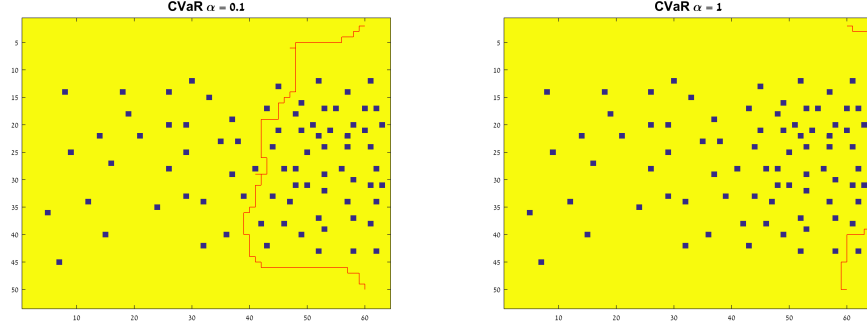
Figure 2: Grid-world - trajectory plots.

Then, by repeating this process, we can show that for any $n \in \mathbb{N}$, the following inequality holds:

$$\mathbf{T}^n[V_0](x,y) - \gamma \frac{1-\gamma^n}{1-\gamma}(2M(\theta-1)+\epsilon) \leq \mathbf{T}_{\mathcal{I}}^n[V_0](x,y) \leq \mathbf{T}^n[V_0](x,y).$$

Note that when $n \to \infty$, we have that γ^n converges to 0, $\mathbf{T}^n[V_0](x,y)$ converges to $\min_{\mu \in \Pi_H} \mathrm{CVaR}_y \left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x, \mu \right)$ (follow from Theorem 4) and $\mathbf{T}_{\mathcal{I}}^n[V_0](x,y)$ converges to $\widehat{V}^*(x,y)$ (follow from the contraction property in Lemma 6).

Furthermore, from Proposition 1.6.4 in [3], the contraction property of Bellman operator $\mathbf{T}$ implies that for any $x \in \mathcal{X}, y \in \mathcal{Y}$, the following expression holds:

$$|\mathbf{T}^n[V_0](x,y) - V^*(x,y)| \leq \frac{\gamma^n}{1-\gamma}(C_{\max} + \|Z\|_\infty)$$

where Z is the bounded random variable of the initial value function $V_0(x,y) = \mathrm{CVaR}_y(Z \mid x_0 = x)$ such that $\|V_0\|_\infty \leq \|Z\|_\infty$, and $V^*(x,y) = \min_{\mu \in \Pi_H} \mathrm{CVaR}_y \left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x, \mu \right)$. This further implies for any $x \in \mathcal{X}, y \in \mathcal{Y}$,

$$|\mathbf{T}_{\mathcal{I}}^n[V_0](x,y) - V^*(x,y)| \leq \gamma \frac{1-\gamma^n}{1-\gamma}(2M(\theta-1)+\epsilon) + \frac{\gamma^n}{1-\gamma}(C_{\max} + \|Z\|_\infty).$$

Then, by combining all the above arguments, we prove the claim of this theorem.

B Trajectory Plots

In Figure 2 we demonstrate simulated trajectories according to a policy that is greedy w.r.t. the value function, according to Theorem 5.

C Generalization to Mean-CVaR Optimization

In this section we extend our approach to MDPs with a mean-CVaR objective of the form:

$$\min_{\mu \in \Pi_H} \quad \lambda \mathbb{E}\left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu\right) + (1-\lambda)\mathrm{CVaR}_\alpha\left(\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu\right), \tag{20}$$

where $\lambda \in [0,1]$. Such an objective is common in practice [11], and is also useful for solving CVaR-constrained objectives using standard Lagrangian methods (see, e.g., [5]).

Now for any $\alpha_1, \alpha_2 \in [0,1]$, define

$$\rho_{\bar{\alpha}}(Z \mid H_t, \mu) = \lambda \mathrm{CVaR}_{\alpha_1}(Z \mid H_t, \mu) + (1-\lambda)\mathrm{CVaR}_{\alpha_2}(Z \mid H_t, \mu)$$

and notice that $\rho_{\bar{\alpha}}(Z \mid H_t, \mu) = \lambda \mathbb{E}(Z \mid H_t, \mu) + (1-\lambda)\mathrm{CVaR}_\alpha(Z \mid H_t, \mu)$ when the vector of CVaR confidence intervals is given by $\bar{\alpha} = (1, \alpha)$.

Theorem 11 *For any* $t \geq 0$, *denote by* $Z \doteq (Z_{t+1}, Z_{t+2}, \dots)$ *the cost sequence from time* $t + 1$ *onwards. The conditional mean-CVaR risk metric under policy* μ *obeys the following decomposition:*

$$\rho_{\bar{\alpha}}(Z \mid H_t, \mu) = \max_{\xi \in \mathcal{U}_{2CVaR}(\bar{\alpha}, P(\cdot \mid x_t, a_t))} \mathbb{E}[\mathbf{S}_\lambda(\xi(x_{t+1})) \cdot \rho_{\bar{\alpha}\mathbf{S}_\lambda(\xi(x_{t+1}))}(Z \mid H_{t+1}) \mid H_t]$$

where $\bar{\alpha} = (\alpha_1, \alpha_2)$ *is the vector of CVaR confidence intervals. The risk envelope is given by*

$$\mathcal{U}_{2CVaR}(\bar{\alpha}, P(\cdot \mid x_t, a_t)) = \left\{ \xi = (\xi_1, \xi_2) : \xi_i(x_{t+1}) \in \left[0, \frac{1}{\alpha_i}\right], \sum_{x_{t+1} \in \mathcal{X}} \xi_i(x_{t+1}) P(x_{t+1} \mid x_t, a_t) = 1, \forall i \right\},$$

$\mathbf{S}_\lambda(\xi) : \mathbb{R}^2 \mapsto \mathbb{R}$ *is a linear operator given by* $\lambda\xi_1 + (1 - \lambda)\xi_2$ *and* a_t *is the control input induced by policy* $\mu_t(h_t)$.

Now we extend the above analysis to Bellman recursion. With the generic state space $\mathcal{Y} = [0, 1]^2$, we now define the optimal Bellman operator at any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathbf{T}[V](x, y) = \min_{a \in \mathcal{A}} \left[C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{2CVaR}(y, P(\cdot \mid x, a))} \sum_{x' \in \mathcal{X}} \mathbf{S}_\lambda(\xi(x')) V(x', y\mathbf{S}_\lambda(\xi(x'))) P(x' \mid x, a) \right]. \tag{21}$$

Based on the decomposition result from Theorem 11, we now have the result on the convergence of Bellman recursion, analogous to Theorem 4 and 5, showing that the fixed point solution of $\mathbf{T}[V](x, y) = V(x, y)$ is unique and equals to the solution of (20) with $x_0 = x$ and $y_0 = (1, \alpha)$.

Theorem 12 *For any state* $x \in \mathcal{X}$ *and* $y = (y_1, y_2) \in [0, 1]^2$, *the fixed point solution of* $\mathbf{T}[V](x, y) = V(x, y)$ *is unique and is equal to* $V(x, y) := \min_{\mu \in \Pi_H} \lambda CVaR_{y_1} (\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu) + (1 - \lambda) CVaR_{y_2} (\lim_{T \to \infty} \mathcal{C}_{0,T} \mid x_0, \mu)$. *Furthermore, let* $\mu^* = \{\mu_0, \mu_1, \dots\} \in \Pi_H$ *be a policy recursively defined as in (7) with two-dimensional augmented state* $\{y_j\}$ *and initial condition* $y_0 = (1, \alpha)$. *Then* μ^* *is an optimal policy for the mean-CVaR problem (20) with initial condition* x_0 *and CVaR confidence level* α.

Extending the interpolation-based CVaR value iteration (Algorithm 1) for this case is straightforward, using a 2-D linear interpolation for $yV(x, y)$.