Abstract—We consider the stochastic shortest path planning problem in MDPs, i.e., the problem of designing policies that ensure reaching a goal state from a given initial state with minimum accrued cost. In order to account for rare but important realizations of the system, we consider a nested dynamic coherent risk total cost functional rather than the conventional risk-neutral total expected cost. Under some assumptions, we show that optimal, stationary, Markovian policies exist and can be found via a special Bellman’s equation. We propose a computational technique based on difference convex programs (DCPs) to find the associated value functions and therefore the risk-averse policies. A rover navigation MDP is used to illustrate the proposed methodology with conditional-value-at-risk (CVaR) and entropic-value-at-risk (EVaR) coherent risk measures.

I. INTRODUCTION

Shortest path problems [10], i.e., the problem of reaching a goal state form an initial state with minimum total cost, arise in several real-world applications, such as driving directions on web mapping websites like MapQuest or Google Maps [41] and robotic path planning [18]. In a shortest path problem, if transitions from one system state to another is subject to stochastic uncertainty, the problem is referred to as a stochastic shortest path (SSP) problem [13], [44]. In this case, we are interested in designing policies such that the total expected cost is minimized. Such planning under uncertainty problems are indeed equivalent to an undiscounted total cost Markov decision processes (MDPs) [37] and can be solved efficiently via the dynamic programming method [13], [12].

However, emerging applications in path planning, such as autonomous navigation in extreme environments, e.g., subterranean [24] and extraterrestrial environments [2], not only require reaching a goal region, but also risk-awareness for mission success. Nonetheless, the conventional total expected cost is only meaningful if the law of large numbers can be invoked and it ignores important but rare system realizations. In addition, robust planning solutions may give rise to behavior that is extremely conservative.

Risk can be quantified in numerous ways. For example, mission risks can be mathematically characterized in terms of chance constraints [34], [33], utility functions [23], and distributional robustness [48]. Chance constraints often account for Boolean events (such as collision with an obstacle or reaching a goal set) and do not take into consideration the tail of the cost distribution. To account for the latter, risk measures have been advocated for planning and decision making tasks in robotic systems [32]. The preference of one risk measure over another depends on factors such as sensitivity to rare events, ease of estimation from data, and computational tractability. Artzner et al. [8] characterized a set of natural properties that are desirable for a risk measure, called a coherent risk measure, and have henceforth obtained widespread acceptance in finance and operations research, among others. Coherent risk measures can be interpreted as a special form of distributional robustness, which will be leveraged later in this paper.

Conditional value-at-risk (CVaR) is an important coherent risk measure that has received significant attention in decision making problems, such as MDPs [20], [19], [36], [9]. General coherent risk measures for MDPs were studied in [39], [3], wherein it was further assumed the risk measure is time consistent, akin to the dynamic programming property. Following the footsteps of [39], [46] proposed a sampling-based algorithm for MDPs with static and dynamic coherent risk measures using policy gradient and actor-critic methods, respectively (also, see a model predictive control technique for linear dynamical systems with coherent risk objectives [45]).

A method based on stochastic reachability analysis was proposed in [17] to estimate a CVaR-safe set of initial conditions via the solution to an MDP. A worst-case CVaR SSP planning method was proposed and solved via dynamic programming in [27]. Also, total cost undiscounted MDPs with static CVaR measures were studied in [16] and solved via a surrogate MDP, whose solution approximates the optimal policy with arbitrary accuracy.

In this paper, we propose a method for designing policies for SSP planning problems, such that the total accrued cost in terms of dynamic, coherent risk measures is minimized (a generalization of the problems considered in [27] and [16]) to dynamic, coherent risk measures). We begin by showing that, under the assumption that the goal region is reachable in finite time with non-zero probability, the total accumulated risk cost is always bounded. We further show that, if the coherent risk measures satisfy a Markovian property, we can find optimal, stationary, Markovian risk-averse policies via solving a special Bellman’s equation. We also propose a computational method based on difference convex programming to solve the Bellman’s equation and therefore design risk-averse policies. We elucidate the proposed method via numerical examples involving a rover navigation MDP and CVaR and entropic-value-at-risk (EVaR) measures.

The rest of the paper is organized as follows. In the next section, we review some definitions and properties used in the sequel. In Section III, we present the problem under study and show its well-posedness under an assumption. In Section IV, we present the main result of the paper, i.e., a
Consider a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), a filtration \(\mathcal{F}_t \subset \cdots \subset \mathcal{F}_T \subset \mathcal{F}\), and an adapted sequence of random variables (stage-wise costs) \(c_t, t = 0, \ldots, T\), where \(T \in \mathbb{N} \cup \{\infty\}\). For \(t = 0, \ldots, T\), we further define the spaces \(C_t = L_p(\Omega, \mathcal{F}_t, \mathbb{P}), p \in [1, \infty]\), \(C_{t,T} = C_t \times \cdots \times C_T\), and \(\mathcal{C} = C_0 \times C_1 \times \cdots\). In order to describe how one can evaluate the risk of sub-sequence \(c_1, \ldots, c_T\) from the perspective of stage \(t\), we require the following definitions.

**Definition 2 (Conditional Risk Measure):** A mapping \(\rho_{t,T} : C_{t,T} \rightarrow C_t\), where \(0 \leq t \leq N\), is called a conditional risk measure, if it has the following monotonicity property:

\[
\rho_{t,T}(c) \leq \rho_{t,T}(c'), \quad \forall c, \forall c' \in C_{t,T} \text{ such that } c \preceq c'.
\]

**Definition 3 (Dynamic Risk Measure):** A dynamic risk measure is a sequence of conditional risk measures \(\rho_{t,T} : C_{t,T} \rightarrow C_t, t = 0, \ldots, T\).

One fundamental property of dynamic risk measures is their consistency over time [39, Definition 3]. If a risk measure is time-consistent, we can define the one-step conditional risk measure \(\rho_t : C_{t+1} \rightarrow C_t, t = 0, \ldots, T-1\) as follows:

\[
\rho_t(c_{t+1}) = \rho_{t,t+1}(0, c_{t+1}),
\]

and for all \(t = 1, \ldots, T\), we obtain:

\[
\rho_{t,T}(c_t, \ldots, c_T) = \rho_t(c_t + \rho_{t+1}(c_{t+1} + c_{t+2} + \cdots + \rho_{T-1}(c_{T-1} + \rho_T(c_T))) \cdots)
\]

Note that the time-consistent risk measure is completely defined by one-step conditional risk measures \(\rho_t, t = 0, \ldots, T-1\) and, in particular, for \(t = 0\), it defines a risk measure of the entire sequence \(c \in C_{0:T}\).

At this point, we are ready to define a coherent risk measure.

**Definition 4 (Coherent Risk Measure):** We call the one-step conditional risk measures \(\rho_t : C_{t+1} \rightarrow C_t, t = 1, \ldots, N-1\) as in [2] a coherent risk measure, if it satisfies the following conditions:

- **Convexity:** \(\rho_t(\lambda c + (1-\lambda)c') \leq \lambda \rho_t(c) + (1-\lambda)\rho_t(c')\), for all \(\lambda \in (0,1)\) and all \(c, c' \in C_{t+1}\);
- **Monotonicity:** If \(c \preceq c'\), then \(\rho_t(c) \leq \rho_t(c')\) for all \(c, c' \in C_{t+1}\);
- **Translational Invariance:** \(\rho_t(c' + c) = \rho_t(c') + c\) for all \(c \in C_t\) and \(c' \in C_{t+1}\);
- **Positive Homogeneity:** \(\rho_t(\beta c) = \beta \rho_t(c)\) for all \(c \in C_{t+1}\) and \(\beta \geq 0\).

In fact, we can show that there exists a dual (or distributionally robust) representation for any coherent risk measure. Let \(m, n \in [1, \infty]\) such that \(1/m + 1/n = 1\) and

\[
\mathcal{P} = \left\{ q \in L_n(S, 2^S, \mathbb{P}) \mid \sum_{s \in S} q(s)p(s)' = 1, \forall q \geq 0 \right\}.
\]

**Proposition 1 (Proposition 4.14 in [26]):** Let \(\mathcal{Q}\) be a closed convex subset of \(\mathcal{P}\). The one-step conditional risk measure \(\rho_t : C_{t+1} \rightarrow C_t, t = 1, \ldots, N-1\) is a coherent risk measure if and only if

\[
\rho_t(c) = \sup_{q \in \mathcal{Q}} \langle c, q \rangle_{\mathcal{Q}}, \quad \forall c \in C_{t+1},
\]
where \( \langle \cdot, \cdot \rangle_\mathcal{Q} \) denotes the inner product in \( \mathcal{Q} \).

Hereafter, all risk measures are assumed to be coherent.

### III. Problem Formulation

Next, we formally describe the risk-averse SSP problem. We also demonstrate that, if the goal state is reachable in finite time, the risk-averse SSP problem is well-posed. Let \( \pi = \{ \pi_0, \pi_1, \ldots \} \) be an admissible policy.

**Problem 1:** Consider MDP \( \mathcal{M} \) as described in Definition 1. Given an initial state \( s_0 \neq s^g \), we are interested in solving the following problem

\[
\pi^* \in \arg \min_\pi J(s_0, \pi), \quad (4)
\]

where

\[
J(s_0, \pi) = \lim_{T \to \infty} \rho_{sT} \left( c(s_0, \pi_0), \ldots, c(s_T, \pi_T) \right), \quad (5)
\]

is the total risk functional for the admissible policy \( \pi \).

In fact, we are interested in reaching the goal state \( s^g \) such that the total risk cost is minimized.\(^1\) Note that the risk-averse deterministic shortest problem can be obtained as a special case when the transitions are deterministic. We define the optimal risk value function as

\[
J^*(s) = \min_\pi J(s, \pi), \quad \forall s \in \mathcal{S},
\]

and call a stationary policy \( \pi = \{ \mu, \mu, \ldots \} \) (denoted \( \mu \)) optimal if \( J(s, \mu) = J^*(s) = \min_\pi J(s, \pi), \quad \forall s \in \mathcal{S} \).

We posit the following assumption, which implies that the goal state is reachable eventually under all policies.

**Assumption 1 (Goal is Reachable in Finite Time):** Regardless of the policy used and the initial state, there exists an integer \( \tau \) such that there is a positive probability that the goal state \( s^g \) is visited after no more than \( \tau \) stages.\(^2\)

We then have the following observation with respect to Problem 1:\(^3\)

**Proposition 2:** Let Assumption 1 hold. Then, the risk-averse SSP problem, i.e., Problem 1, is well-posed and \( J(s_0, \pi) \) is bounded for all policies \( \pi \).

**Proof:** Assumption 1 implies that for each admissible policy \( \pi \), we have \( p_{\pi} = \max_{s \in \mathcal{S}} \mathbb{P}(s_{\pi} \neq s^g | s_0 = s, \pi) < 1 \). That is, given a policy \( \pi \), the probability \( p_{\pi} \) of not visiting the goal state \( s^g \) is less than one. Let \( p = \max_\pi p_{\pi} \). Remark

\(^1\)An important class of SSP planning problems are concerned with minimum-time reachability. Indeed, our formulation also encapsulates minimum-time problems, in which for MDP \( \mathcal{M} \), we have \( c(s) = 1 \), for all \( s \in \mathcal{S} \).

\(^2\)If instead of one goal state \( s^g \), we were interested in a set of goal states \( \mathcal{G} \subset \mathcal{S} \), it suffices to define \( \tau = \inf \{ t \mid \mathbb{P}(s_t \in \mathcal{G} | s_0 = s_0, \pi) > 0 \} \). Then, all the paper’s derivations can be applied. For the sake of simplicity of the presentation, we present the results for a single goal state.

\(^3\)Note that Problem 1 is ill-posed in general. For example, if the induced Markov chain for an admissible policy is periodic, then the limit in (5) may not exist. This is in contrast to risk-averse discounted infinite-horizon MDPs [3], for which we only require non-negativity and boundedness of immediate costs.

that \( p_{\pi} \) is dependent solely on \( \{ \tau_1, \tau_2, \ldots, \tau_\pi \} \). Moreover, since \( \mathcal{A} \) is finite, the number of \( \tau \)-stage policies is also finite, which implies finiteness of \( p_{\pi} \). Hence, \( p < 1 \) as well. Therefore, for any policy \( \pi \) and initial state \( s \), we obtain

\[
\mathbb{P}(s_{2\tau} \neq s^g | s_0 = s, \pi) = \mathbb{P}(s_{2\tau} \neq s^g | s_\tau \neq s^g, s_0 = s, \pi) \times \mathbb{P}(s_\tau \neq s^g | s_0 = s, \pi) \leq p^2. \quad \text{(6)}
\]

Then, by induction, we can show that, for any admissible SSP policy \( \pi \), we have

\[
\mathbb{P}(s_{k\tau} \neq s^g | s_0 = s, \pi) \leq p^k, \quad \forall s \in \mathcal{S},
\]

and \( k = 1, 2, \ldots. \) Indeed, we can show that the risk-averse cost incurred in the \( \tau \) periods between \( \tau k \) and \( \tau(k + 1) - 1 \) is bounded as follows

\[
\rho_0(\cdots \rho_{\tau(k+1)−1}(c_{\tau k} + \cdots + c_{\tau(k+1)−1}) \cdots) \quad (7a)
\]

\[
\leq \rho(c_{\tau} + \cdots + c_{\tau k−1}) \quad (7b)
\]

\[
= \sup_{q \in \mathcal{Q}} \langle \tau \tilde{c}, q \rangle_\mathcal{Q} \quad (7c)
\]

\[
\leq \langle \tau \tilde{c}, q^* \rangle_\mathcal{Q} \quad (7d)
\]

\[
= \tau \tilde{c} \sum_{s \in \mathcal{S}} \mathbb{P}(s_{\tau k} \neq s^g | s_0 = s, \pi) q^*(s, \pi) \quad (7e)
\]

\[
\leq \tau \tilde{c} \times \sup_{q \in \mathcal{Q}} \mathbb{P}(s_{k \tau} \neq s^g | s_0 = s, \pi) \sum_{s \in \mathcal{S}} |q^*(s, \pi)| \quad (7f)
\]

\[
\leq \tau \tilde{c} \sum_{s \in \mathcal{S}} \mathbb{P}(s_{k \tau} \neq s^g | s_0 = s, \pi) \sum_{s \in \mathcal{S}} |q^*(s, \pi)| \quad (7g)
\]

where in (7b) we used the translational invariance property of coherent risk measures and defined \( \rho = \rho_0 \circ \cdots \circ \rho_{\tau(k+1)−1} \). Since any finite compositions of the coherent risk measures is a risk measure [42], we have that \( \rho \) is also a coherent risk measure. Moreover, since the immediate cost function \( c \) is upper-bounded, from the monotonicity property of the coherent risk measure \( \rho \), we obtain (7e). Equality (7d) is derived from Proposition 1. Inequality (7e) is obtained via defining \( q^* = \text{argmax}_{q \in \mathcal{Q}} \langle \tau \tilde{c}, q \rangle_\mathcal{Q} \). In inequality (7f), we used Hölder inequality and finally we used (5) to obtain the last inequality. Thus, the risk-averse total cost \( J(s, \pi), s \in \mathcal{S} \), exists and is finite, because given Assumption 1 we have

\[
|J(s_0, \pi)| = \lim_{T \to \infty} \rho_0 \cdots \rho_{\tau−1} \circ \cdots \circ \rho_0(c_0 + \cdots + c_{\tau}) \quad (8)
\]

\[
\leq \sum_{k=0}^{\infty} \rho_0(\cdots \rho_{\tau(k+1)−1}(c_{\tau k} + \cdots + c_{\tau(k+1)−1}) \cdots) \quad (8)
\]

\[
\leq \sum_{k=0}^{\infty} \tau \tilde{c} \sum_{s \in \mathcal{S}} \mathbb{P}(s_{\tau k} \neq s^g | s_0 = s, \pi) \sum_{s \in \mathcal{S}} |q^*(s, \pi)| \quad (8)
\]

where in the first equality above we used the translational invariance property and in the first inequality we used the sub-additivity property of coherent risk measures. Hence, \( J(s_0, \pi) \) is bounded for all \( \pi \).

### IV. Risk-Averse SSP Planning

This section presents the paper’s main result, which includes a special Bellman’s equation for finding the risk value functions for Problem 1. Furthermore, assuming that the coherent risk measures satisfy a Markovian property, we show that the optimal risk-averse policies are stationary and Markovian.
To begin with, note that at any time $t$, the value of $\rho_t$ is $F_t$-measurable and is allowed to depend on the entire history of the process $\{s_0, s_1, \ldots\}$ and we cannot expect to obtain a Markov optimal policy [35]. In order to obtain Markov optimal policies for Problem 1, we need the following property [39, Section 4] of risk measures.

**Definition 5** (Markov Risk Measure [25]): A one-step conditional risk measure $\rho$ is defined in (2), at time index $\pi$, value functions $J^k$ converge to the optimal risk value function $J^*$ is optimal if and only if $\mu$ attains the minimum in Bellman’s equation (11).

**Proof:** For every positive integer $M$, an initial state $s_0$, and policy $\pi$, we can split the nested risk cost (2), where $\rho_t$ is continuous [40, Corollary 3.1] and hence the limiting process and the measure $\rho_t$ commute. The limit term is indeed the total risk cost starting at $s_{\pi}$, i.e., $J(s_{\pi}, \pi)$. Next, we show that under Assumption 1, this term remains bounded. From (5) in the proof of Proposition 2, we have

$$J(s_{\pi}, \pi) = \lim_{T \to \infty} \rho_{t+1}(c_{t+1} \cdots + \rho_T(c_T) \cdots)$$

where

$$\rho_t(v(s_{t+1})) = \sigma_t(v(s_{t+1}), s_t, T(s_{t+1}|s_t, \alpha_t)). \quad (9)$$

In fact, if $\rho_t$ is a coherent risk measure, $\sigma_t$ also satisfies the properties of a coherent risk measure (Definition 4).

**Assumption 2:** The one-step coherent risk measure $\rho_t$ is a Markov risk measure.

We can now present the main result in the paper, a form of Bellman’s equations for solving the risk-averse SSP problem.

**Theorem 1:** Consider MDP $P$ as described in Definition 1 and let Assumptions 1 and 2 hold. Then, the following statements are true for the risk-averse SSP problem:

(i) Given (non-negative) initial condition $J^0(s), s \in S$, the sequence generated by the recursive formula (dynamic programming)

$$J^{k+1}(s) = \min_{a \in A_t} \left( c(s, \alpha) + \sigma \{ J^k(s'), s, T(s'|s, \alpha) \} \right), \quad \forall s \in S; \quad (10)$$

converges to the optimal risk value function $J^*(s), s \in S$.

(ii) The optimal risk value functions $J^*(s), s \in S$ are the unique solution to the Bellman’s equation

$$J^*(s) = \min_{a \in A_t} \left( c(s, \alpha) + \sigma \{ J^*(s'), s, T(s'|s, \alpha) \} \right), \quad \forall s \in S; \quad (11)$$

(iii) For any stationary Markovian policy $\mu$, the risk averse value functions $J(s, \mu(s)), s \in S$ are the unique solutions to

$$J(s, \mu) = c(s, \mu(s)) + \sigma \{ J(s, \mu), s, T(s'|s, \alpha) \}, \quad \forall s \in S; \quad (12)$$

(iv) A stationary Markovian policy $\mu$ is optimal if and only if $\mu$ attains the minimum in Bellman’s equation (11).

Proof: For every positive integer $M$, an initial state $s_0$, and policy $\pi$, we can split the nested risk cost (5), where $\rho_t$ is defined in (2), at time index $\pi$ and obtain

$$J(s_0, \pi) = \rho_0(c_0 + \cdots + \rho_{\pi-1}(c_{\pi-1} - 1) + \rho_T(c_T) \cdots)$$

where we used the fact that the one-step coherent risk measures $\rho_t$ are continuous [40, Corollary 3.1] and hence the
Remark that the middle term in the inequality above is the \(\tau M\)-stage risk-averse cost of the policy \(\pi\) with the terminal cost \(J^0(s,m)\). Given Assumption 2, from [39, Theorem 2], the minimum of this cost is generated by the dynamic programming recursion (10) after \(\tau M\) iterations. Taking the minimum over the policy \(\pi\) on every side of (19) yields:

\[
-p^M \max_{s \in S} |J^0(s)| + J^*(s_0) - \frac{\tau \bar{c} p^M}{1 - p} \\
\leq J^\tau M(s_0) \\
\leq p^M \max_{s \in S} |J^0(s)| + J^*(s_0) + \frac{\tau \bar{c} p^M}{1 - p},
\]

for all \(s_0\) and \(M\). Finally, let \(k = \tau M\). Since the above inequality holds for all \(M\), taking the limit of \(M \to \infty\) gives:

\[
\lim_{M \to \infty} J^\tau M(s_0) = \lim_{k \to \infty} J^k(s_0) = J^*(s_0), \quad \forall s_0 \in S. \quad (21)
\]

(ii) Taking the limit, \(k \to \infty\), of both sides of (10) yields:

\[
\lim_{k \to \infty} J^{k+1}(s) = \lim_{k \to \infty} \min_{\alpha \in \text{Act}} \left( c(s, \alpha) + \sigma \left\{ J^k(s'), s, T(s'|s, \alpha) \right\} \right), \quad \forall s \in S. \quad (21)
\]

Since the limit and the minimization commute over a finite number of alternatives, we have:

\[
J^*(s) = \lim_{k \to \infty} \min_{\alpha \in \text{Act}} \left( c(s, \alpha) + \sigma \left\{ J^k(s'), s, T(s'|s, \alpha) \right\} \right), \quad \forall s \in S.
\]

Finally, because \(\sigma\) is continuous [40, Corollary 3.1], the limit and \(\sigma\) commute as well and from (21), we obtain:

\[
J^*(s) = \min_{\alpha \in \text{Act}} \left( c(s, \alpha) + \sigma \left\{ J^*(s'), s, T(s'|s, \alpha) \right\} \right), \quad \forall s \in S.
\]

To show uniqueness, note that for any \(J(s), s \in S\) satisfying the above equation, the dynamic programming recursion (10) starting at \(J(s), s \in S\) replicates \(J(s), s \in S\) and from Part (i) we infer \(J(s) = J^*(s)\) for all \(s \in S\).

(iii) Given a stationary Markovian policy \(\mu\), at every state \(s\), we have \(\alpha = \mu(s)\), hence from Item (i), we have:

\[
J^{k+1}(s, \mu) = \min_{\alpha \in \mu(s)} \left( c(s, \alpha) + \sigma \left\{ J^k(s', \mu), s, T(s'|s, \alpha) \right\} \right), \quad \forall s \in S.
\]

Since the minimum is only over one element, we have:

\[
J^{k+1}(s, \mu) = c(s, \mu(s)) + \sigma \left\{ J^k(s', \mu), s, T(s'|s, \alpha) \right\}, \quad \forall s \in S, \quad \text{with which} \quad k \to \infty \text{converges uniquely (see Item (ii)) to} \quad J(s, \mu).
\]

(iv) The stationary policy attains its minimum in (11), if:

\[
J^*(s) = \min_{\alpha \in \mu(s)} \left( c(s, \alpha) + \sigma \left\{ J^*(s'), s, T(s'|s, \alpha) \right\} \right)
\]

\[
= c(s, \mu(s)) + \sigma \left\{ J^*(s'), s, T(s'|s, \alpha) \right\}, \quad \forall s \in S.
\]

Then, Part (iii) and the above equation imply \(J(s, \mu) = J^*(s)\) for all \(s \in S\). Conversely, if \(J(s, \mu) = J^*(s)\) for all \(s \in S\), then Items (ii) and (iii) imply that \(\mu\) is optimal.

At this point, we should highlight that, in the case of conditional expectation as the coherent risk measure \(\rho\) and the risk value functions, we have the above equation, the dynamic programming recursion (10) simplifies to [11, Proposition 5.2.1] for the risk-neutral SSP problem. In fact, Theorem 1 is a generalization of [11, Proposition 5.2.1] to the risk-averse case.

Recursion (dynamic programming) (10) represents the Value Iteration (VI) for finding the risk value functions. In general, value iteration converges with infinite number of iterations \((k \to \infty)\), but it can be shown, for a stationary policy \(\mu\) resulting in an acyclic induced Markov chain, the VI algorithm converges in \(|S|\) of steps (see the derivation for the risk-neutral SSP problem in [14]).

Alternatively, one can design risk-averse policies using Policy Iteration (PI). That is, starting with an initial policy \(\mu_0\), we can carry out policy evaluation via (12) followed by a policy improvement step, which calculates an improved policy \(\mu_0^{k+1}\), as

\[
\mu_0^{k+1}(s) = \arg \min_{\alpha \in \text{Act}} \left( c(s, \alpha) + \sigma \left\{ J^{k+1}(s, \alpha), s, T(s'|s, \alpha) \right\} \right), \quad \forall s \in S.
\]

This process is repeated until no further improvement is found in terms of the risk value functions: \(J^{k+1}(s) = J^k(s)\) for all \(s \in S\).

However, we do not pursue VI or PI approaches further in this work. The main obstacle for using VI and PI is that equations (10)-(12) are nonlinear (and non-smooth) in the risk value functions may require significant computational burden (see the specialized non-smooth Newton Method in [39] for solving similar nonlinear VIs). Instead, the next section presents a computational method based on difference convex programs (DCPs).

V. A DCP Computational Approach

In this section, we propose a computational method based on DCPs to find the risk value functions and subsequently policies that minimize the accrued dynamic risk in the SSP planning. Before stating the DCP formulation, we show that the Bellman operator in (10) is non-decreasing. Let

\[
\mathcal{D}_\pi J = c(s, \pi(s)) + \sigma \left\{ J(s'), s, T(s'|s, \pi(s)) \right\}, \quad \forall s \in S,
\]

\[
\mathcal{D}_J : = \min_{\alpha \in \text{Act}} \left( c(s, \alpha) + \sigma \left\{ J(s', s, T(s'|s, \alpha)) \right\} \right), \quad \forall s \in S.
\]

Lemma 1: Let Assumptions 1 and 2 hold. For all \(v, w \in \mathcal{L}_m(S, 2^S, \mathbb{P})\), if \(v \leq w\), then \(\mathcal{D}_\pi v \leq \mathcal{D}_\pi w\) and \(\mathcal{D} v \leq \mathcal{D} w\).

Proof: Since \(\rho\) is a Markov risk measure, we have \(\rho(v) = \sigma(v, T)\) for all \(v \in \mathcal{L}_m(S, 2^S, \mathbb{P})\). Furthermore, since \(\rho\) is a coherent risk measure from Proposition 1, we know that (3) holds. Inner producting both sides of \(v \leq w\) with the probability measure \(q \in \mathcal{Q} \subset \mathbb{P}\) from right and taking the supremum over \(\mathcal{Q}\) yields

\[
\sup_{q \in \mathcal{Q}} \langle v, q \rangle_\mathcal{Q} \leq \sup_{q \in \mathcal{Q}} \langle w, q \rangle_\mathcal{Q}.
\]

From Proposition 1, we have \(\sigma(v, s, T) = \sigma(v, s, T)\).
techniques of nonlinear optimization [14] more efficiently.

Yet, a locally optimal solution can be obtained based on globally [28], as feature selection in machine learning [30] and inverse DCP [28]. Many applications require solving DCPs, such as the rover navigation MDP (also used for risk-averse SSP planning for CVaR and EVaR risk measures used in our numerical experiments in the next section. Note that for the risk-neutral conditional expectation measure, optimization (23) becomes a linear program, since \( \sigma \{ J'(s'), s, T(s'|s, \alpha) \} = \sum_{s' \in S} T(s'|s, \alpha) J(s') \) is linear in the decision variables \( J \).

VI. Numerical Experiments

In this section, we evaluate the proposed method for risk-averse SSP planning with a rover navigation MDP (also used in [2, 3]). We consider the traditional total expectation as well as CVaR and EVaR. The experiments were carried out on a MacBook Pro with 2.8 GHz Quad-Core Intel Core i5 and 16 GB of RAM. The resultant linear programs and DCPs were solved using CVX [21] with DCCP [43] add-on.

An agent (e.g., a rover) must autonomously navigate a 2-dimensional terrain map (e.g., Mars surface) represented by an \( M \times N \) grid with 0.25\( MN \) obstacles. Thus, the state space is given by \( S = \{ s'| i = x + y, x \in \{1, \ldots, M\}, y \in \{1, \ldots, N\} \} \) with \( x = 1, y = 0 \) being the leftmost bottom grid. Since the rover can move from cell to cell, its action set is \( \text{Act} = \{E, W, N, S\} \). The actions move the robot from its current cell to a neighboring cell, with some uncertainty. The state transition probabilities for various cell types are shown for actions \( E \) (East) and \( N \) (North) in Figure 2. Other actions lead to similar transitions. Hitting an obstacle incurs the immediate cost of 5, while the goal grid region has zero immediate cost. Any other grid has a cost of 1 to represent fuel consumption.

Once a solution \( J^* \) to optimization problem (22) is found, we can find a corresponding stationary Markovian policy as

\[
\mu^*(s) = \arg \min_{\alpha \in \text{Act}} \left( c(s, \alpha) + \sigma \{ J^*(s, \alpha), s, T(s'|s, \alpha) \} \right), \forall s \in S.
\]

A. DCPs for Risk-Averse SSP Planning

Assumption 1 implies that each \( \rho \) is a coherent, Markov risk measure. Hence, the mapping \( v \mapsto \sigma(v, \cdot, \cdot) \) is convex (because \( \sigma \) is also a coherent risk measure). We next show that optimization problem (22) is in fact a DCP.

Let \( f_0 = 0, g_0(J) = \sum_{s \in S} J(s), f_1(J) = J(s), g_1(s, \alpha) = c(s, \alpha), \) and \( g_2(J) = \sigma(J, \cdot, \cdot) \). Note that \( f_0 \) and \( g_1 \) and \( g_2 \) are convex (constant) functions and \( g_0, f_1, \) and \( g_2 \) are convex functions in \( J \). Then, (22) can be expressed as the minimization

\[
\inf_J f_0 - g_0(J)
\]

subject to

\[
f_1(J) - g_1(s, \alpha) - g_2(J) \leq 0, \forall s, \alpha.
\]
simulate the policy by first solving an exponential cone program. Note that these calculations can be carried out offline for policy synthesis and then the policy can be applied for risk-averse robot path planning.

The table also outlines the failure ratios of each risk measure. In this case, EVaR outperformed both CVaR and total expectation in terms of robustness, which is consistent with the fact that EVaR is a more conservative risk measure. Lower failure/collision rates were observed for $\varepsilon = 0.3$, which correspond to more risk-averse policies. In addition, these results suggest that, although total expectation can be used as a measure of performance in high number of Monte Carlo simulations, it may not be practical to use it for real-world planning under uncertainty scenarios. CVaR and EVaR seem to be a more efficient metric for performance in shortest path planning under uncertainty.

### VII. Conclusions

We proposed a method based on dynamic programming for designing risk-averse policies for the SSP problem. We presented a computational approach in terms of difference convex programs for finding the associated risk value functions and hence the risk-averse policies. Future research will extend to risk-averse MDPs with average costs and risk-averse MDPs with linear temporal logic specifications, where the former problem is cast as a special case of the risk-averse SSP problem. In this work, we assumed the states are fully observable, we will study the SSP problems with partial state observation [4], [2] in the future, as well.

### References

[1] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[2] M. Ahmadi, M. Ono, M. D. Ingham, R. M. Murray, and A. D. Ames. Risk-averse planning under uncertainty. In *2020 American Control Conference (ACC)*, pages 3305–3312. IEEE, 2020.

[3] M. Ahmadi, U. Rosolia, M. D. Ingham, R. M. Murray, and A. D. Ames. Constrained risk-averse Markov decision processes. In *The 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[4] M. Ahmadi, R. Sharan, and J. W. Burdick. Stochastic finite state control of POMDPs with LTL specifications. *arXiv preprint arXiv:2001.07679*, 2020.

[5] A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *J. Optimization Theory and Applications*, 155(3):1105–1123, 2012.

[6] A. Ahmadi-Javid and M. Fallah-Tafiri. Portfolio optimization with entropic value-at-risk. *Euro. J. Operational Res.*, 279(1):225–241, 2019.

[7] A. Ahmadi-Javid and A. Pichler. An analytical study of norms and banach spaces induced by the entropic value-at-risk. *Mathematics and Financial Economics*, 11(4):527–550, 2017.

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

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

[10] R. Bellman. On a routing problem. *Quarterly of applied mathematics*, 16(1):87–90, 1958.

[11] D. Bertsekas. Dynamic programming and optimal control. *Athena Scientific*, Nashua, NH, USA, vol. 1, 2017.

[12] D. Bertsekas and H. Yu. Stochastic shortest path problems under weak conditions. *Lab. for Information and Decision Systems Report LIDS-P-2909*, MIT, 2013.

[13] D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Math. Operations Research*, 16(3):580–595, 1991.

[14] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[15] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[16] S. Carpin, Y. Chow, and M. Pavone. Risk aversion in finite markov decision processes using total cost criteria and average value at risk. In *IEEE Int. Conf. Robotics and Automation*, pages 335–342, 2016.

| $(M \times N)_\rho$ | $J^*(s_0)$ | Total Time [s] | U.O. | F.R. |
|------------------|---------|--------------|-----|-----|
| $(4 \times 5)_{CVaR_{0.7}}$ | 18.76 | 0.58 | 2 | 14% |
| $(10 \times 10)_{CVaR_{0.7}}$ | 55.72 | 1.12 | 4 | 19% |
| $(10 \times 20)_{CVaR_{0.7}}$ | 47.36 | 1.36 | 8 | 21% |
| $(4 \times 5)_{EVaR_{0.3}}$ | 25.69 | 0.57 | 2 | 10% |
| $(10 \times 10)_{EVaR_{0.3}}$ | 43.86 | 1.16 | 4 | 13% |
| $(10 \times 20)_{EVaR_{0.3}}$ | 49.03 | 1.34 | 8 | 15% |

### Table 1

**Comparison between total expectation, CVaR, and EVaR risk measures.** $(M \times N)_\rho$ denotes the grid-world of size $M \times N$ and one-step coherent risk measure $\rho$. Total Time denotes the time taken by the CVX solver to solve the associated linear programs or DCPs. U.O. denotes the number of single grid uncertain obstacles used for robustness test. F.R. denotes the failure rate out of 100 Monte Carlo simulations.
In this appendix, we present the specific DCPs for finding the risk value functions for two coherent risk measures studied in our numerical experiments, namely, CVaR and EVaR.

For a given confidence level $\varepsilon \in (0, 1)$, value-at-risk (VaR$_c$) denotes the $(1 - \varepsilon)$-quantile of the value cost variable. CVaR$_c$ is the expected loss in the $(1 - \varepsilon)$-tail given that the particular threshold VaR$_c$ has been crossed. CVaR$_c$ is given by

$$
\rho_{\varepsilon}(c_{t+1}) = \inf_{\zeta \in \mathbb{R}} \left( \frac{1}{\varepsilon} \mathbb{E} \left[ \left( c_{t+1} - \zeta \right)^+ \right] + \frac{1}{\varepsilon} \mathbb{E} \left[ \mathbf{1}_{\zeta > c_{t+1}} \mathcal{F}_t \right] \right),
$$

where $(\cdot)^+$ is the positive part of $(\cdot)$. A value of $\varepsilon \approx 1$ corresponds to a risk-neutral policy; whereas, a value of $\varepsilon \to 0$ is rather a risk-averse policy.

In fact, Theorem 1 can applied to CVaR since it is a coherent risk measure. For MDP $\mathcal{M}$, the risk value functions can be computed by DCP (23), where $g_{2}(J) = \inf_{c \in \mathbb{R}} \left\{ \zeta + \frac{1}{\varepsilon} \sum_{s' \in S} (J(s') - \zeta + \mathbb{E} [T(s' | s, a)]) \right\}$, where the infimum on the right hand side of the above equation can be absorbed into the overall infimum problem, i.e., $\inf_{f : \mathcal{M}} \zeta$. Note that $g_{2}(J)$ above is convex in $\zeta$ (Theorem 1).

Unfortunately, CVaR ignores the losses below the VaR threshold (since it is only concerned with the average of VaR at the $(1 - \varepsilon)$-tail of the cost distribution). EVaR is the tightest upper bound in the sense of Chernoff inequality for the value at risk (VaR) and CVaR and its dual representation is associated with the relative entropy. In fact, it was shown in [7] that EVaR$_c$ and CVaR$_c$ are equal only if there are no losses ($c \to 0$) below the VaR$_c$ threshold. In addition, EVaR is a strictly monotone risk measure; whereas, CVaR is only monotone [6] (see Definition 4). EVaR$_c$ is given by

$$
\rho_{\varepsilon}(c_{t+1}) = \inf_{\zeta > 0} \left( \frac{\mathbb{E} [\zeta^{c_{t+1}} | \mathcal{F}_t]}{\varepsilon} \right)^{1/\varepsilon}.
$$

Similar to CVaR$_c$, for EVaR$_c$, $\varepsilon \to 1$ is a risk-neutral case; whereas, $\varepsilon \to 0$ corresponds to a risk-averse case. In fact, it was demonstrated in [5, Proposition 3.2] that $\lim_{\varepsilon \to 0} \text{EVaR}_c(c) = \varepsilon \text{ ess sup} \{c\} = \bar{c}$ (worst-case cost).

Since EVaR$_c$ is a coherent risk measure, the conditions of Theorem 1 hold. Since $\zeta > 0$, using the change of variables, $J \equiv \zeta J$ (note that this change of variables is monotone increasing in $\zeta$ [1]), we can compute EVaR value functions by solving (23), where $f_0 = 0, f_1(J) = \tilde{J}, g_0(J) = \sum_{s \in S} J(s), g_1(c) = \zeta c$, and $g_2(J) = \log \left( \sum_{s' \in S} \mathbb{E} [J(s') | s, a] \right)$.

Similar to the CVaR case, the infimum over $\zeta$ can be lumped into the overall infimum problem, i.e., $\inf_{f : \mathcal{M}} \zeta > 0$. Note that $g_{2}(\tilde{J})$ is convex in $\tilde{J}$, since the logarithm of sums of exponentials is convex [15, p. 72].