View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Caltech Authors - Main

Risk-Averse Planning Under Uncertainty

Mohamadreza Ahmadi, Masahiro Ono, Michel D. Ingham,
Richard M. Murray, and Aaron D. Ames

arXiv:1909.12499v1 [cs.RO] 27 Sep 2019

***Abstract*— We consider the problem of designing policies for partially observable Markov decision processes (POMDPs) with dynamic coherent risk objectives. Synthesizing risk-averse *optimal* policies for POMDPs requires infinite memory and thus undecidable. To overcome this difficulty, we propose a method based on bounded policy iteration for designing stochastic but finite state (memory) controllers, which takes advantage of standard convex optimization methods. Given a memory budget and optimality criterion, the proposed method modifies the stochastic finite state controller leading to sub-optimal solutions with lower coherent risk.**

I. Introduction

With the rise of autonomous systems being deployed in real-world settings, the associated risk that stems from unknown and unforeseen circumstances is correspondingly on the rise. In particular, in safety-critical scenarios, such as aerospace applications, decision making should account for risk. For example, spacecraft control technology relies heavily on a relatively large and highly skilled mission operations team that generates detailed time-ordered and event-driven sequences of commands. This approach will not be viable in the future with increasing number of missions and a desire to limit the operations team and Deep Space Network (DSN) costs. Future spaceflight missions will be at large distances and light- time delays from Earth, requiring novel capabilities for astronaut crews and ground operators to manage spacecraft consumables such as power, water, propellant, and life support systems to prevent mission failure. In order to maximize the science returns under these conditions, the ability to deal with emergencies and safely explore remote regions are becoming more and more important [18]. Even in Mars rover navigation problems, finding planning policies that minimize risk is of utmost importance due to the uncertainties present in Mars surface data [20] as illustrated in Figure 1.

Risk can be quantified in numerous ways. For example, mission risks can be mathematically characterized in terms of chance constraints [21]–[23]. 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.* [3] 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.

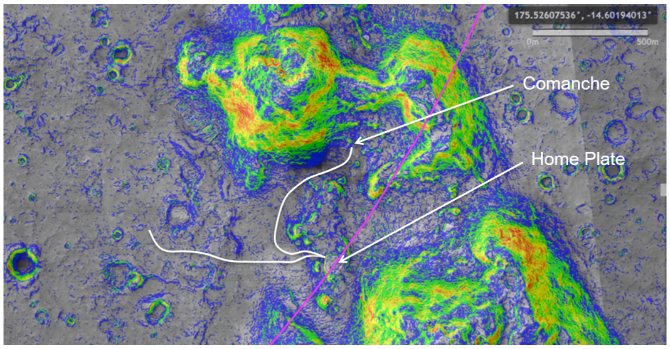

Fig. 1. Mars surface slope uncertainty for Mars rover navigation: regions with slopes ranging any values within (blue) $5° - 10°$, (green) $10° - 15°$, (yellow) $15° - 20°$, (orange) $20° - 25°$, (red) $\geqslant 25°$, and (the rest) $< 5°$ or no data.

An important example of a coherent risk measure is the conditional value-at-risk (CVaR) that has received significant attention in decision making problems such as Markov decision processes (MDPs) [5], [8], [9], [27]. General coherent risk measures for MDPs were studied in [29], wherein it was further assumed the risk measure is *time consistent*, similar to the dynamic programming property. Following the footsteps of the latter contribution, [32] 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 [31]).

However, in many aerospace applications, sensing constraints does not allow for full-state observation and decision making involves partial observation [2], [19]. These problems can be represented as a partially observable Markov decision process (POMDP), where decision making is subject to uncertainty stemming from stochastic outcomes as well as partial observation [16]. In this paper, we propose a method based on bounded policy iteration to design sub-optimal risk-averse policies for POMDPs. To this end, we first discuss that the problem of designing risk-averse optimal policies is undecidable in general. Then, we show that a stochastic but finite-memory controller can be synthesized to upper-bound the dynamic risk. Given a memory budget, we propose a policy iteration method to synthesize these finite-state controllers that can increase the number of memory states to improve risk-aversity. We illustrate our proposed method with a numerical example of path planning under uncertainty.

The rest of the paper is organized as follows. The next section reviews some preliminary notions and definitions used in the sequel. In Section III, we discuss POMDPs with coherent risk

M. Ahmadi, R. M. Murray, and A. D. Ames are with the California Institute of Technology, 1200 E. California Blvd., MC 104-44, Pasadena, CA 91125, e-mail: ({mrahmadi, murray, ames}@caltech.edu). M. Ono and M. D. Ingham are with the NASA Jet Propulsion Laboratory, 4800 Oak Grove Dr, Pasadena, CA 91109, e-mail: ({masahiro.ono, michel.d.ingham}@jpl.nasa.gov)

measures. In Section IV, we propose sub-optimal stochastic finite state controllers that minimize the upper-bound on the coherent risk. In Section V, a bounded policy iteration algorithm is formulated to design risk-averse stochastic finite controllers. In Section VI, we elucidate our results with a numerical example. Finally, in Section VII, we conclude the paper and give directions for future research.

II. Preliminaries

In this section, we briefly review some notions and definitions used throughout the paper.

A. Markov Chains

A Markov chain $\mathcal{M}$ is composed of the state space $\mathcal{S}$, the transition probability defined as the conditional distribution $T(.|s) : \mathcal{S} \to [0,1]$ such that $\sum_{s' \in \mathcal{S}} T(s'|s) = 1,\ \forall s \in \mathcal{S}$, and the initial distribution ι_{init} such that $\sum_{s \in \mathcal{S}} \iota_{\text{init}}(s) = 1$. An infinite path, denoted by the superscript ω, of the Markov chain $\mathcal{M}$ is a sequence of states $\pi = s_0 s_1 \cdots \in \mathcal{S}^\omega$ such that $T(s_{t+1}|s_t) > 0$ for all t and $\iota_{\text{init}}(s_0) > 0$. The probability space over such paths is the defined as follows. The sample space Ω is the set of infinite paths with initial state $s \in \mathcal{S}$, *i.e.*, $\Omega = \text{Paths}(s)$. $\Sigma_{\text{Paths}(s)}$ is the least σ-algebra on $\text{Paths}(s)$ containing $\text{Cyl}(\omega)$, where $\text{Cyl}(\omega) = \{\omega' \in \text{Paths}(s) \mid \omega \text{ is a prefix of } \omega'\}$ is the cylinder set. Finally, in order to specify the probability measure over all sets of events in $\Sigma_{\text{Paths}(s)}$, it is sufficient to provide the probability of each cylinder set, which can be computed as $\text{Pr}_{\mathcal{M}}\left[\text{Cyl}(s_0 \ldots s_n)\right] = \iota_{\text{init}}(s_0) \prod_{0 \leqslant t \leqslant n} T(s_{t+1} \mid s_t)$. Once the probability measure is defined over the cylinder sets, the expectation operator $\mathbb{E}_{\mathcal{M}}$ is also uniquely defined. In the sequel, we remove the subscript whenever the Markov chain is clear from the context.

B. Partially Observable Markov Decision Process

***Definition** 1 (POMDP):* A *POMDP*, $\mathcal{PM}$, consists of:

- States $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ of the autonomous agent(s) and world model,
- Actions $Act = \{\alpha_1, \ldots, \alpha_{|Act|}\}$ available to the robot,
- Observations $\mathcal{O} = \{o_1, \ldots, o_{|\mathcal{O}|}\}$,
- A Transition function $T(s_j|s_i, \alpha)$,
- A cost, $c(s_i, \alpha_i) \geqslant 0$, for each state $s_i \in \mathcal{S}$ and action $\alpha_i \in Act$.

This paper considers *finite* POMDPs where $\mathcal{S}$, Act, and $\mathcal{O}$ are finite sets. For each action the probability of making a transition from state $s_i \in \mathcal{S}$ to state $s_j \in \mathcal{S}$ under action $\alpha \in Act$ is given by $T(s_j|s_i, \alpha)$. For each state s_i, an observation $o \in \mathcal{O}$ is generated independently with probability $O(o|s_i)$. The starting world state is given by the distribution $\iota_{\text{init}}(s_i)$. The probabilistic components of a POMDP model must satisfy the following:

$$\begin{cases} \sum_{s \in \mathcal{S}} T(s|s_i, \alpha) = 1, & \forall s_i \in \mathcal{S}, \alpha \in Act \\ \sum_{o \in \mathcal{O}} O(o|s) = 1, & \forall s \in \mathcal{S} \\ \sum_{s \in \mathcal{S}} \iota_{\text{init}}(s) = 1. & \end{cases}$$

Given a POMDP, we can define beliefs or distributions over states at each time step to keep track of sufficient statistics with finite description [4]. The beliefs $b \in \Delta(\mathcal{S})$, with $\Delta(\mathcal{S})$ being the set of probability distributions over $\mathcal{S}$, for all $s \in \mathcal{S}$ can be computed using the Bayes' law as follows:

$$b_0(s) = \frac{\iota_{\text{init}}(s) O(o_0 \mid s)}{\sum_{o \in O} \iota_{\text{init}}(s) O(o \mid s)}, \tag{1}$$

$$b_t(s) = \frac{O(o_t \mid s, \alpha_t) \sum_{s' \in \mathcal{S}} T(s \mid s', \alpha_t) b_{N-1}(s')}{\sum_{s \in \mathcal{S}} O(o_t \mid s, \alpha_t) \sum_{s' \in \mathcal{S}} T(s \mid s', \alpha_t) b_{N-1}(s')}, \tag{2}$$

for all $t \geqslant 1$. It is also worth mentioning that (2) is referred to as the *belief update equation*.

C. Stochastic Finite State Control of POMDPs

It is well established that designing optimal policies for POMDPs based on the (continuous) belief states require uncountably infinite memory or internal states [7], [15], [17]. This paper focuses on a particular class of POMDP controllers, namely, *stochastic finite state controllers*. These controllers lead to a finite state space Markov chain for the closed loop controlled system.

***Definition** 2 (Stochastic Finite State Controller):* Let $\mathcal{PM}$ be a POMDP with observations $\mathcal{O}$, actions Act, and initial distribution ι_{init}. A *stochastic finite state controller* for $\mathcal{PM}$ is given by the tuple $\mathcal{G} = (G, \omega, \kappa)$ where

- $G = \{g_1, g_2, \ldots, g_{|G|}\}$ is a finite set of internal states (I-states).
- $\omega : G \times \mathcal{O} \to \Delta(G \times Act)$ is a function of internal stochastic finite state controller states g_k and observation o, such that $\omega(g_k, o)$ is a probability distribution over $G \times Act$. The next internal state and action pair (g_l, α) is chosen by independent sampling of $\omega(g_k, o)$. By abuse of notation, $\omega(g_l, \alpha|g_k, o)$ will denote the probability of transitioning to internal stochastic finite state controller state g_l and taking action α, when the current internal state is g_k and observation o is received.
- $\kappa : \Delta(\mathcal{S}) \to \Delta(G)$ chooses the starting internal FSC state g_0, by independent sampling of $\kappa(\iota_{\text{init}})$, given initial distribution ι_{init} of $\mathcal{PM}$. $\kappa(g|\iota_{\text{init}})$ will denote the probability of starting the FSC in internal state g when the initial POMDP distribution is ι_{init}.

Closing the loop around a POMDP with a stochastic finite state controller yields the following transition system.

***Definition** 3 (Global Markov Chain):* Let POMDP $\mathcal{PM}$ have state space $\mathcal{S}$ and let G be the I-states of stochastic finite state controller $\mathcal{G}$. The global Markov chain $\mathcal{M}_{\mathcal{S} \times G}^{\mathcal{PM}, \mathcal{G}}$ (or simply $\mathcal{M}$, where the stochastic finite state controller and the POMDP are clear from the context) with execution $\sigma = \{[s_0, g_0], [s_1, g_1], \ldots\}$, $[s_t,\ g_t] \in \mathcal{S} \times G$ evolves as follows:

- The probability of initial global state $[s_0, g_0]$ is

$$\iota_{\text{init}}^{\mathcal{M}}\left([s_0, g_0]\right) = \iota_{\text{init}}(s_0) \kappa(g_0|\iota_{\text{init}}).$$

- The state transition probability, $T^{\mathcal{M}}$, is given by
$$T^{\mathcal{M}}\left([s_{t+1}, g_{t+1}] \,|[s_t, g_t]\right) = \sum_{o \in \mathcal{O}} \sum_{\alpha \in Act} O(o|s_t)\omega(g_{t+1}, \alpha|g_t, o)T(s_{t+1}|s_t, \alpha).$$

Note that the global Markov chain arising from a finite state space POMDP also has a finite state space.

D. Coherent Risk Measures

Consider a probability space $(\Omega, \mathcal{F}, P)$, a filteration $\mathcal{F}_0 \subset \cdots \mathcal{F}_N \subset \mathcal{F}$, and an adapted sequence of random variables (stage-wise costs) $c_t,\ t = 0, \ldots, N$, where $N \in \mathbb{N}_{\geqslant 0} \cup \{\infty\}$. We further define the spaces $\mathcal{C}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$, $p \in [0, \infty)$, $t = 0, \ldots, N$ and let $\mathcal{C}_{t:N} = \mathcal{C}_t \times \cdots \times \mathcal{C}_N$ and $\mathcal{C} = \mathcal{C}_0 \times \mathcal{C}_1 \times \cdots$. We further assume that the sequence $c \in \mathcal{C}$ is almost surely bounded, *i.e.*,
$$\max_t \operatorname{essup} |c_t(\omega)| < \infty.$$

In order to describe how one can evaluate the risk of subsequence $c_t, \ldots, c_N$ from the perspective of stage t, we require the following definitions.

***Definition** 4 (Conditional Risk Measure):* A mapping $\rho_{t:N} : \mathcal{C}_{t:N} \to \mathcal{C}_t$, where $0 \leqslant t \leqslant N$, is called a *conditional risk measure*, if it has the following monoticity property:
$$\rho_{t:N}(c) \leqslant \rho_{t:N}(c'), \quad \forall c, \forall c' \in \mathcal{C}_{t:N} \text{ such that } c \leqslant c', \quad (3)$$
where the inequalities should be understood componentwise.

***Definition** 5 (Dynamic Risk Measure):* A *dynamic risk measure* is a sequence of conditional risk measures $\rho_{t:N} : \mathcal{C}_{t:N} \to \mathcal{C}_t$, $t = 0, \ldots, N$.

One fundamental property of dynamic risk measures is their consistency over time. That is, if c will be as good as c' from the perspective of some future time θ, and they are identical between time τ and θ, then c should not be worse than c' from the current time's perspective.

***Definition** 6 (Time-Consistent Risk Measure):* A dynamic risk measure $\{\rho_{t:N}\}_{t=0}^{T}$ is called *time-consistent* if for all $0 \leqslant t \leqslant \tau < \theta \leqslant T$ and all sequences $Z, W \in \mathcal{C}_{t:N}$ the conditions
$$c_t = c'_t,\ t = \tau, \ldots, \theta - 1, \quad \text{and} \quad \rho_{\theta,T}(Z_\theta, \ldots, c_t) \leqslant \rho_{\theta,T}(W_\theta, \ldots, c'_t)$$
imply
$$\rho_{\tau,N}(c_\tau, \ldots, c_t) \leqslant \rho_{\tau,N}(c'_\tau, \ldots, c'_t). \quad (4)$$

If a risk measure is time-consistent, we can define the one-step conditional risk measure $\rho_t : \mathcal{C}_{t+1} \to \mathcal{C}_t$, $t = 0, \ldots, N-1$ as follows:
$$\rho_t(c_{t+1}) = \rho_{t,t+1}(0, c_{t+1}), \quad (5)$$
and for all $t = 1, \ldots, N$, we obtain:
$$\rho_{t,N}(c_t, \ldots, c_N) = \rho_t\big(c_t + \rho_{t+1}(c_{t+1} + \rho_{t+2}(c_{t+2} + \cdots + \rho_{N-1}\left(c_{N-1} + \rho_N(c_N)\right) \cdots))\big). \quad (6)$$

Note that the time-consistent risk measure is completely defined by one-step conditional risk measures ρ_t, $t = 0, \ldots, N-1$ and, in particular, for $t = 0$, (6) define a risk measure of the entire sequence $c \in \mathcal{C}_{0:N}$.

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

***Definition** 7 (Coherent Risk Measure):* We call the one-step conditional risk measures $\rho_t : \mathcal{C}_{t+1} \to \mathcal{C}_t$, $t = 1, \ldots, N-1$ as in (6) a *coherent risk measure* if it satisfies the following conditions

- **Convexity:** $\rho_t(\lambda c + (1-\lambda)c') \leqslant \lambda\rho_t(c) + (1-\lambda)\rho_t(c')$, for all $\lambda \in (0,1)$ and all $c, c' \in \mathcal{C}_{t+1}$;
- **Monotonicity:** If $c \leqslant c'$ then $\rho_t(c) \leqslant \rho_t(c')$ for all $c, c' \in \mathcal{C}_{t+1}$;
- **Time Consistency:** $\rho_t(c + c') = c + \rho_t(c')$ for all $c \in \mathcal{C}_t$ and $c' \in \mathcal{C}_{t+1}$;
- **Positive Homogeneity:** $\rho_t(\beta c) = \beta\rho_t(c)$ for all $c \in \mathcal{C}_{t+1}$ and $\beta \geqslant 0$.

Henceforth, all the risk measures considered are assumed to be coherent. In this paper, we are interested in the discounted infinite horizon problems. Let $\gamma \in (0,1)$ be a given discount factor. For $N = 0, 1, \ldots$, we define the functionals
$$\begin{aligned}\rho^{\gamma}_{0,N}(c_0, \ldots, c_N) &= \rho_{0,N}(c_0, \gamma c_1, \ldots, \gamma^N c_N) \\ &= \rho_0\Bigg(c_0 + \rho_1\big(\gamma c_1 + \rho_2(\gamma^2 c_2 + \cdots \\ &\quad + \rho_{N-1}\left(\gamma^{N-1} c_{N-1} + \rho_N(\gamma^N c_N)\right) \cdots)\big)\Bigg),\end{aligned}$$
which are the same as (6) for $t = 0$, but with discounting γ^t applied to each c_t. Finally, we have total discounted risk functional $\xi_\gamma : \mathcal{C} \to \mathbb{R}$ defined as
$$\xi_\gamma(Z) = \lim_{N \to \infty} \rho^{\gamma}_{0,N}(c_0, \ldots, c_N). \quad (7)$$

From [29, Theorem 3], we have that ξ_γ is convex, monotone, and positive homogenoeus.

III. Risk-Averse POMDPs

Notions of coherent risk and dynamic risk measures discussed in the previous section have been developed and applied in microeconomics and mathematical finance fields in the past two decades [33]. Generally speaking, risk-averse decision making is concerned with the behavior of agents, e.g. consumers and investors, who, when exposed to uncertainty, attempt to lower that uncertainty. The agent averts to agree to a situation with an unknown payoff rather than another situation with a more predictable payoff but possibly lower expected payoff. In a Markov decision making setting, the main idea in risk-averse control is to replace the conventional conditional expectation of the cumulative reward or cost objectives with more general risk measures.

Consider a stationary (policies, transition probabilities, and cost functions do not depend explicitly on time) controlled Markov process $\{s_t\}$, $t = 0, 1, \ldots$. Each policy $\pi = \{\pi_t\}_{t=0}^{\infty}$

leads to a cost sequence $c_t = c(s_t, \alpha_t)$, $t = 0, 1, \ldots$. We define the dynamic risk of evaluating the γ-discounted cost of the policy π as

$$V_\gamma(\pi, s_0) = \xi_\gamma\big(c(s_0, \alpha_0), c(s_1, \alpha_1), \ldots\big), \tag{8}$$

where ξ_γ is defined in (7). In this work, we are interested in addressing the following problem:

***Problem** 1: For a given POMDP $\mathcal{PM}$, a discount factor $\gamma \in (0,1)$, and a total risk functional V_γ as in* (8) *with $\{\rho_t\}_{t=0}^{\infty}$ being coherent risk measures, compute*

$$\pi^* \in \operatorname{argmin}_\pi V_\gamma(\pi, b_0).$$

We refer to a controlled Markov process with the "nested" objective (8) a *risk-averse* Markov process. Many applications such as portfolio allocation problems [10] and organ transplant decisions [14] require a risk-averse Markov model. It was also previously demonstrated in [9], [24] that coherent risk measure objectives can account for modeling errors and parametric uncertainty in MDPs.

The main challenge is that at any time t, the value of ρ_t is $\mathcal{F}_t$-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 [25].

In order to obtain Markov optimal policies, we need to make the following assumption (see [29, Section 4] for more details):

***Assumption** 1: For any function $\phi(s_t, a_t, s_{t+1})$, we have*

$$\rho_t(\phi(s_t, a_t, s_{t+1})) = \mathrm{R}\left\{\phi(s_t, a_t, \cdot), s_t, p(s_t, a_t)\right\}, \tag{9}$$

where $a_t = \pi(s_t)$. The function R *is called a* Markov risk transition mapping.

Note that the Markov risk transition mapping depends on the function ϕ, the states s, and probability vector $p(s, a)$. The dot in $\phi(s_t, a_t, \cdot)$ on the right hand side of (9) represents a dummy variable that is integrated/summed out with respect to the s_t-th row of the transition probability matrix $p(s_t, a_t)$. The simplest case of the Markov risk transition mapping is the conditional expectation $\mathbb{E}\{\cdot \mid s_t, a_t\}$, *i.e.*,

$$\begin{aligned}\mathrm{R}\left\{\phi(s_t, a_t, \cdot), s_t, p(s_t, a_t)\right\} &= \mathbb{E}\{\phi(s_t, a_t, s_{t+1}) \mid s_t, a_t\}\\ &= \sum_{s_{t+1}} \phi(s_t, a_t, s_{t+1}) T(s_{t+1} \mid s_t, a_t).\end{aligned}$$

If R is a coherent risk measure as described in Definition 7, then the Markov policies are sufficient to ensure optimality [29]. In particular, for the CVaR risk measure, the Markov risk transition mapping is given by

$$\begin{aligned}&\mathrm{R}\{\phi, s, p(s,a)\}\\ &\quad= \inf_{z \in \mathbb{R}} \left\{ z + \frac{1}{\alpha} \sum_{s'} \big(\phi(s, a, s') - z\big)_+ T(s' \mid s, a) \right\}.\end{aligned} \tag{10}$$

The risk-averse formulation can be extended to POMDPs as follows.

***Theorem** 1: Consider the POMDP $\mathcal{PM}$ with the nested risk objective* (8) *and $\gamma \in (0,1)$. Let Assumption 1 hold, let ρ_t, $t = 0, 1, \ldots$ be coherent risk measures as described in Definition 7, and suppose $c(\cdot, \cdot)$ be non-negative and upper-bounded. Then, the stationary optimal policy $\pi^* = \{\pi_t^*\}_{t=0}^{\infty} = \{\pi^*\}_{t=0}^{\infty}$ is the solution of the following Bellman's equations*

$$V_\gamma(b) = \min_{\alpha \in Act} \Big(c(b, \alpha) + \gamma \mathrm{R}\left\{ V_\gamma(b'), b, p(b' \mid b, \alpha) \right\} \Big), \tag{11a}$$

$$\pi^*(b) \in \operatorname{argmin}_{\alpha \in Act} \Big(c(b, \alpha) + \gamma \mathrm{R}\left\{ V_\gamma(b'), b, p(b' \mid b, \alpha) \right\} \Big), \tag{11b}$$

where $c(b, \alpha) = \sum_{s \in \mathcal{S}} c(s, \alpha) b(s)$.

Proof: Note that a POMDP can be represented as an MDP over the belief states (2). Hence, a POMDP is a controlled Markov process with states $b \in \Delta(\mathcal{S})$, where the controlled belief transition probability is described as

$$\begin{aligned}p(b' \mid b, \alpha) &= \sum_{o \in \mathcal{O}} p(b' \mid b, o, \alpha)\ p(o \mid b, \alpha)\\ &= \sum_{o \in \mathcal{O}} \delta\left(b' - \frac{O(o \mid s, \alpha) \sum_{s' \in \mathcal{S}} T(s \mid s', \alpha) b(s')}{\sum_{s \in \mathcal{S}} O(o \mid s, \alpha) \sum_{s' \in \mathcal{S}} T(s \mid s', \alpha) b(s')} \right)\\ &\qquad \times \sum_{s \in \mathcal{S}} O(o \mid s, \alpha) \sum_{s'' \in \mathcal{S}} T(s \mid s'', \alpha) b(s''),\end{aligned} \tag{12}$$

with

$$\delta(a) = \begin{cases} 1 & a = 0, \\ 0 & \text{otherwise}. \end{cases}$$

Then, given that $c(\cdot, \cdot)$ is non-negative and upper-bounded, from [16, Theorem 8.6.2] and [29, Theorem 4], we infer that from the Bellman equations (11) we can obtain the optimal policies. ■

We can use a method based on policy iteration to solve the dynamic programming equations (11) to design risk-averse optimal policies. To this end, for $k = 0, 1, \ldots$, given a stationary Markov policy π^k, we calculate the corresponding value function as

$$V_\gamma^k(b) = c\left(b, \pi^k(b)\right) + \gamma \mathrm{R}\left\{ V_\gamma^k(b), b, p(b' \mid b, \pi^k(b) \right\}. \tag{13a}$$

Then, we compute the next policy π^{k+1} as

$$\pi^{k+1}(b) \in \operatorname{argmin}_\pi \Big(c(b, \pi(b)) + \gamma \mathrm{R}\big\{ V_\gamma^k(b), b, p(b' \mid b, \pi(b) \big\} \Big). \tag{13b}$$

Unfortunately, the problem of designing risk-averse optimal Markovian policies for POMDPs is undecidable in general. This follows from [17, Theorem 4.4] by noting that $\inf_\pi V_\gamma = \sup_\pi (-V_\gamma)$.

In the subsequent section, we demonstrate that, if instead of considering policies with infinite-memory we search over finite-memory policies, then we can minimize upper-bounds on the total risk cost functional (8).

IV. Risk-Averse Stochastic Finite State Controllers

Under a stochastic finite state controller, the POMDP is transformed into a Markov chain $\mathcal{M}_{\mathcal{S}\times\mathcal{G}}^{\mathcal{PM}\times\mathcal{G}}$ with design probability distributions ω and κ. We define the total risk functional of this parametric Markov chain as

$$V_\gamma(\mathcal{G},\iota_{\text{init}}) = \xi_\gamma\big(c([s_1,g_1],\alpha_1), c([s_2,g_2],\alpha_2),\ldots\big), \quad (14)$$

where α_ts and g_ts are drawn from the probability distribution $\omega(g_{t+1},\alpha_t \mid g_t, o_t)$. In this setting, Problem 1 can be expressed as

***Problem** 2: For a given POMDP $\mathcal{PM}$, a stochastic finite state controller $\mathcal{G}$, a discount factor $\gamma \in (0,1)$, and a total risk functional V_γ as in* (14) *with $\{\rho_t\}_{t=1}^{\infty}$ being coherent risk measures, compute*

$$(\omega^*,\kappa^*) \in \operatorname{argmin}_{\omega,\kappa} V_\gamma(\mathcal{G},\iota_{\text{init}}).$$

The optimal value of Problem 2 provides an upper-bound to that of Problem 1, since a stochastic finite state controller only contains finite memory states and it can be at best as good as the belief-based optimal policy (with infinite memory). The latter claim can also be shown using [12, Theorem 1], which indicates that any improvement in the parameters of a stochastic finite state controller (in the sense of optimizing the value functions) is at most as good as the belief value function.

For POMDPs controlled by stochastic finite state controllers, the dynamic program is developed in the global state space $\mathcal{S}\times G$. The value function is defined over this global state space, and policy iteration techniques must also be carried out in the global state space. For a given stochastic finite state controller, $\mathcal{G}$, and the POMDP $\mathcal{PM}$, the value function $V_{\gamma,\mathcal{M}}([s,g])$ is the discounted dynamic risk measure under $\mathcal{G}$, and can be computed by solving a set of equations:

$$\begin{aligned} &V_{\gamma,\mathcal{M}}([s,g]) \\ &\quad= \sum_{\alpha\in Act}\sum_{g'\in\mathcal{G},o\in\mathcal{O}} \omega(g',\alpha \mid g,o)O(o|g')c([s,g],\alpha) \\ &\quad+ \gamma\mathrm{R}\Big\{V_{\gamma,\mathcal{M}}([s',g']),[s,g],T^{\mathcal{M}}\left([s',g']\,|[s,g]\right)\Big\}, \quad (15)\end{aligned}$$

where

$$p(\alpha \mid g) = \sum_{g'\in\mathcal{G},o\in\mathcal{O}} \omega(g',\alpha \mid g,o)O(o|g').$$

Then, for each s, the optimal value function over the induced Markov Chain $\mathcal{M}$ can be computed by taking the minimum of the above equation over all I-states g

$$V_{\gamma,\mathcal{M}}^*(s) := \min_{g\in\mathcal{G}} V_{\gamma,\mathcal{M}}([s,g]). \quad (16)$$

Since $v \mapsto \mathrm{R}(v,\cdot,\cdot)$ is convex (because R is a coherent risk measure), (15) can be solved by a convex optimization.

We end this section by demonstrating that the optimal values obtained using the stochastic finite state controllers upper-bound those of the belief-based (infinite-memory) policy.

***Proposition** 1: Consider the POMDP $\mathcal{PM}$ and the Markov chain $\mathcal{M}$ induced by the stochastic finite state controller $\mathcal{G}$. Then, for all $s\in\mathcal{S}$, we have $V_\gamma^*(b(s)) \leqslant V_{\gamma,\mathcal{M}}^*(s)$.*

Proof: The value function of the induced Markov chain $\mathcal{M}$ satisfies (15) for all $[s,g]\in\mathcal{S}\times\mathcal{G}$. For each I-state g, the value function in beliefs can be computed as

$$V_\gamma([b,g]) := \sum_{s\in\mathcal{S}} b(s)V_{\gamma,\mathcal{M}}([s,g]),$$

and the optimal value function given by

$$V_\gamma^*(b) = \min_{g\in\mathcal{G}} \sum_{s\in\mathcal{S}} b(s)V_{\gamma,\mathcal{M}}([s,g]).$$

Applying Hölder inequality to the right-hand side of the above equality, we obtain

$$\begin{aligned} V_\gamma^*(b) &\leqslant \min_{g\in G}\left(\sup|\sum_{s\in\mathcal{S}} b(s)|\right)|V_{\gamma,\mathcal{M}}([s,g])| \\ &= \min_{g\in\mathcal{G}} V_{\gamma,\mathcal{M}}([s,g]),\end{aligned}$$

where in the last inequality we used the fact that $\sum_s b(s) = 1$ since $b\in\Delta(\mathcal{S})$ and the fact that $V_{\gamma,\mathcal{M}}([s,g])$ is non-negative (since c is non-negative). From (16), we infer $V_\gamma^* \leqslant V_{\gamma,\mathcal{M}}^*$. ■

V. A Bounded Policy Iteration Algorithm for Risk-Averse stochastic finite state controllers

So far, we showed that synthesizing an infinite memory controller for POMDPs with coherent risk objectives is undecidable. On the other hand, a stochastic finite state controller can upper-bound the coherent risk for a POMDP. In this section, we provide a computational method based on bounded policy iteration to design risk-averse stochastic finite state controllers. Furthermore, we propose techniques for minimizing the upper bound on the total coherent risk by adding I-states to the algorithm in order to escape local minima.

Policy iteration incrementally improves a controller by alternating between two steps: Policy Evaluation and Policy Improvement, until convergence to an optimal policy [6]. During policy improvement, a dynamic programming update using the so called *dynamic programming backup equation (DP Backup)* is used. For a risk-averse POMDP, the DP Backup is given by

$$V_\gamma(b) \;=\; \min_{\alpha\in Act}\Big(c(b,\alpha) \;+\; \gamma\mathrm{R}\big\{V_\gamma(b),b,p(b' \;\mid\; b,\alpha)\big\}\Big),$$

The r.h.s. of the DP Backup can be applied to any risk value function. The effect is a risk reduction (if possible) at every belief state. However, DP Backup is difficult to use directly as it must be computed at each belief state in the belief space, which is uncountably infinite.

In [13], [26], a methodology called the Bounded Policy Iteration is proposed for stochastic finite state controllers, which allows stochastic finite state controllers with fewer I-states to have comparable performance in comparison with deterministic finite state controllers, while allowing the stochastic finite state controller to grow in a bounded fashion – only one (or a few) I-state(s) need to be added at a time to escape a local minima.

Before presenting our proposed bounded policy iteration method for risk-averse stochastic finite state controllers, we recall the following important definition.

***Definition** 8 (Tangent Belief State):* A belief state b is called a *tangent belief state*, if $V_\gamma(b)$ touches the DP Backup of $V_\gamma(b)$ from above. Since $V_\gamma(b)$ must equal V_g^β for some g, we also say that the I-state g is tangent to the backed up value function V_γ at b.

Equipped with this definition, the two steps involved in our algorithm is described next.

A. I-States Improvement via Convex Optimization

Let $\vec{V}_{\gamma,\mathcal{M}}(g) \in \mathbb{R}^{|S|}$ denote the vectorized $V_{\mathcal{M}}([s,g])$ in s. We say that an I-state g is *improved*, if the tunable stochastic finite state controller parameters associated with that I-state can be adjusted so that $\vec{V}^*_{\gamma,\mathcal{M}}(g)$ decreases.

As a first step, we point out that the search over κ can be dropped. This is simply because the initial I-state is chosen by computing the best valued I-state for the given initial belief, *i.e.*, $\kappa(g_{\text{init}}) = 1$, where

$$g_{\text{init}} = \underset{g}{\text{argmin}} \ \left(\vec{\iota}^{\mathcal{M}}_{\text{init}}\right)^T \vec{V}_{\gamma,\mathcal{M}}(g).$$

After initialization, we pose the improvement as a convex optimization as follows:

I-state Improvement Convex Optimization: For the I-state g, the following convex optimization is constructed over the variables ϵ, $\omega(g',\alpha|g,o)$, $\forall g',\alpha,o$

$$\begin{gathered}
\max_{\epsilon>0,\omega(g',\alpha|g,o)} \quad \epsilon \\
\text{subject to} \\
\text{Improvement Constraint:} \\
V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leqslant \text{r.h.s. of (15)}, \quad \forall s \in \mathcal{S}, \\
\text{Probability Constraints:} \\
\sum_{(g',\alpha)\in G\times Act} \omega(g',\alpha \mid g,o) = 1, \quad \forall o \in \mathcal{O}, \\
\omega(g',\alpha \mid g,o) \geqslant 0, \quad \forall g' \in G, \alpha \in Act, o \in \mathcal{O}. \qquad (17)
\end{gathered}$$

The above convex optimization searches for ω values that improve the I-state value vector $\vec{V}^*_{\gamma,\mathcal{M}}(g)$ by maximizing the decision variable ϵ. If an improvement is found, *i.e.*, $\epsilon > 0$, the parameters of the I-state are updated by the corresponding minimizing ω.

Algorithm 1 Bounded Policy Iteration For Synthesizing Risk-Averse Stochastic Finite State Controllers

```
Input: (a) An initial feasible stochastic finite state controller,
    $\mathcal{G}$. (b) Maximum size of stochastic finite state controller
    $N_{max}$. (c) $N_{new} \leqslant N_{max}$ number of I-states
 1: $improved \leftarrow True$
 2: Compute the value vectors, $\vec{V}_{\gamma,\mathcal{M}}$ based on the convex
    optimization (18).
 3: while $|G| \leqslant N_{max}$ and $improved = True$ do
 4:   $improved \leftarrow False$
 5:   for all I-states $g \in G$ do
 6:     Set up the I-State Improvement Convex Optimiza-
        tion (17).
 7:     if I-State Improvement Convex Optimization results
        in optimal $\epsilon > 0$ then
 8:       Replace the parameters for I-state $g$
 9:       $improved \leftarrow True$
10:       Compute the value vectors, $\vec{V}_{\gamma,\mathcal{M}}$ based on the
          convex optimization (18).
11:   if $improved = False$ and $|G| < N_{max}$ then
12:     $n_{added} \leftarrow 0$
13:     $N'_{new} \leftarrow \min(N_{new}, N_{max} - |G|)$
14:     Try to add $N'_{new}$ I-state(s) to $\mathcal{G}$ via Algorithm 2 in
        Section V-B.
15:     $n_{added} \leftarrow$ actual number of I-states added in previ-
        ous step.
16:     if $n_{added} > 0$ then
17:       $improved \leftarrow True$
18:       Compute the value vectors, $\vec{V}_{\gamma,\mathcal{M}}$ based on the
          convex optimization (18).
Output: $\mathcal{G}$
```

Algorithm 1 outlines the main steps in the bounded policy iteration for risk-averse stochastic finite state controllers. The algorithm has two distinct parts. First, for fixed parameters of the stochastic finite state controller (ω), policy evaluation is carried out, in which $V_{\gamma,\mathcal{M}}([s,g])$ is computed using the following convex optimization (Steps 2, 10 and 18): For each I-state g, we have the following:

$$\begin{gathered}
\min_{\epsilon_1>0,\epsilon_2>0,V_{\gamma,\mathcal{M}}} \quad \epsilon_1 - \epsilon_2 \\
\text{subject to} \\
V_{\gamma,\mathcal{M}}([s,g]) - \big(\text{r.h.s. of (15)}\big) \leqslant \epsilon_1, \quad \forall s \in \mathcal{S}, \\
V_{\gamma,\mathcal{M}}([s,g]) - \big(\text{r.h.s. of (15)}\big) \geqslant \epsilon_2, \quad \forall s \in \mathcal{S}. \qquad (18)
\end{gathered}$$

In fact, the above optimization solves (15) for $V_{\gamma,\mathcal{M}}$. Second, after evaluating the current coherent risk function, an improvement is carried out either by changing the parameters of existing I-states, or if no new parameters can improve any I-state, then a fixed number of I-states are added to escape the local minima (Steps 14-17). This is described in Section V-B.

B. Escaping Local Minima by Adding I-States

At some point of running the algorithm, no I-state may be improved with further iterations, *i.e.*, $\forall g \in G$, the corresponding convex optimization (17) yields an optimal value of $\epsilon = 0$. Then, policy iteration has reached a local minimum if and only if $\vec{V}_{\gamma,\mathcal{M}}(g)$ is tangent to the backed up value function for all $g \in G$ [26]. The dual variables corresponding to the Improvement Constraints in (17) provide those belief states that are tangent to the risk function. The process for adding

Algorithm 2 Adding I-states to Escape Local Minima

```
Input: (a) Set $B$ of tangent beliefs for each I-state. (b) A function $node : B \to G$ identifying the I-state which yields each tangent belief. (c) $N_{new}$ the maximum number of I-states to add. (d) $\vec{V}_{\gamma,\mathcal{M}}(g)$ the computed risk value functions at each node $g \in G$.
1:  $N_{added} \leftarrow 0$.
2:  repeat
3:    Pick $b \in B$, $B \leftarrow B \backslash \{b\}, g \leftarrow node(b)$.
4:    $Fwd = \varnothing$
5:    for all $(\alpha, o) \in (Act \times \mathcal{O})$ do
6:      if $Pr(o|b) = \sum_{s \in \mathcal{S}} b(s)O(o|s) > 0$ then
7:        Look ahead one step to compute forwarded beliefs $b_{o,\alpha}(s') = \sum_s T(s'|s,\alpha)\frac{O(o|s)b(s)}{\sum_{o' \in \mathcal{O}} O(o'|s)b(s)}$.
8:        $Fwd \leftarrow Fwd \cup \{b_{o,\alpha}\}$.
9:    for all $b \in Fwd$ do
10:     Apply a dynamic programming backup step
          $$V_\gamma^{BU}(b) = \min_{\alpha \in Act} \Big( c(b,\alpha) + \gamma \mathrm{R}\big\{V_\gamma(b), b, p(b' \mid b, \alpha)\big\}\Big),$$
        where $V_\gamma(b(s)) = \min_{g \in G} b_{o,\alpha}(s) V_{\gamma,\mathcal{M}}([s,g])$ and $b_{o,\alpha}$ is computed for each product state $s' \in \mathcal{S}$ as follows $b_{o,\alpha}(s') = \sum_s T(s'|s,\alpha)\frac{O(o|s)b(s)}{\sum_{o' \in \mathcal{O}} O(o'|s)b(s)}$.
11:     Note the minimizing action $\alpha^*$ and I-state $g^*$.
12:   if $V_\gamma^{BU}(b) < V_\gamma(b)$ for $b \in Fwd$ then
13:     Add new deterministic I-state $g_{new}$ such that $\omega(g_{new}|g^*, \alpha^*, o) = 1$, $\forall o \in \mathcal{O}$.
14:     $N_{added} \leftarrow N_{added} + 1$.
15:   if $N_{added} \geqslant N_{new}$ then
16:     return
17: until $B = \varnothing$.
```

I-states involves forwarding the tangent beliefs one step and then checking if the value of those forwarded beliefs can be improved. The procedure for adding I-states is provided in Algorithm 2.

Algorithm 2 can be understood as follows. Assume that a tangent belief b exists for some I-state g. Instead of directly improving the value of the tangent belief, the algorithm tries to improve the value of forwarded beliefs reachable in one step from the tangent beliefs. First, the forwarded beliefs are computed (Step 4-8). Then, the corresponding risk value functions are applied to a DP Backup (Steps 9-11). If some action α^* and successor I-state g^* can in fact reduce the risk value (Step 12), then a new I-state is added which deterministically leads to this action and successor I-state (Steps 13-14). Note that at the end of the algorithm, the newly added I-states, g_{new} have no incoming edges, *i.e.*, no pre-existing I-states transition to g_{new}. However, when the other I-states are improved in subsequent policy improvement steps, they generate transitions to any g_{new} added. This new I-state then improves the value of the original tangent belief.

VI. Numerical Example

An agent (e.g. a robot) has to autonomously navigate a two dimensional terrain map (e.g. Mars surface) represented by a 10×10 grid world (100 states) with 15 obstacles of different shapes. At each time step the agent can move to any of its eight neighboring states (diagonal moves are allowed). 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 incurs the cost of 10 leading to termination, while the goal grid region has reward 80. The discount factor is $\gamma = 0.95$. After a move is chosen, the observation of the agent is assumed to be binary, *i.e.*, either an obstacle is detected in the next cell that the robot is moving to or not. Similar to [9], in our simulations, we included an obstacle and target position perturbation in a random direction to one of the neighboring grid cells with probability 0.2 to represent uncertainty in the terrain map (recall the uncertainty in Mars terrain maps as shown in Figure 1).

The objective is to compute a safe (*i.e.*, obstacle-free) path that is fuel efficient. To this end, we consider CVaR as the coherent risk measure. CVaR is given by

$$\rho_t(c_{t+1}) = \inf_{z \in \mathbb{R}} \left\{ z + \frac{1}{\alpha}\mathbb{E}\left[(c_{t+1} - z)_+ \mid \mathcal{F}_t\right]\right\}, \qquad (19)$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ and the infimum should be understood point-wise. In general, the confidence level α may be $\mathcal{F}_t$-measurable function with values in the interval $(0,1)$. Here, we assume $\alpha \in (0,1)$. A value of $\alpha \simeq 1$ corresponds to a risk-neutral policy; whereas, a value of $\alpha \simeq 0$ is rather a risk-averse policy. For CVaR risk measure, (15) can be computed as

$$\begin{aligned} V_{\gamma,\mathcal{M}}([s,g]) = &\sum_{\alpha, g', o} \omega(g', \alpha \mid g, o) O(o|g') c([s,g],\alpha) \\ &+ \gamma \inf_{z \in \mathbb{R}} \Bigg\{ z + \frac{1}{\alpha} \sum_{g', s', o, \alpha} \big(V\left([s',g']\right) - z\big)_+ \\ &\times O(o \mid s)\omega(g', \alpha \mid g, o) T(s' \mid s, \alpha)\Bigg\}, \end{aligned}$$

where the infimum on the right hand side of the above equation can either be solved by line search techniques or by representation in terms of an elementary linear programming problem since it is convex in z [28, Theorem 1] (the function $(\cdot)_+$ is increasing and convex [25, Lemma A.1., p. 117]).

Figure 2 depicts the policies and the value functions computed for the grid world based on the bounded policy iteration technique in Section V. For these experiments, we used 2 internal states for the stochastic finite state controller and the corresponding convex optimizations were solved using CVX toolbox [11] in MATLAB.

As it can be observed from Figure 2, the risk-neutral policy leads to shorter paths from different cells to the target. However, on 100 perturbed scenarios, it performed poorly with 43 failures. On the other hand, the risk-averse policy

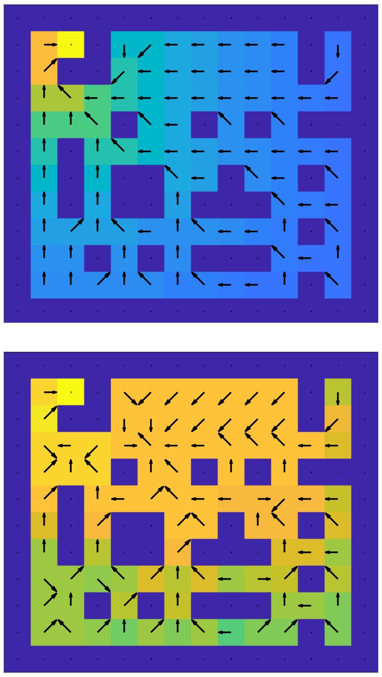

Fig. 2. Numerical results obtained based on the proposed risk-averse control method for two different confidence levels of (top) $\alpha = 0.9$ and (bottom) $\alpha = 0.1$. The yellow square at $(1, 2)$ denotes the goal region. The arrows represent the actions (or the moves) with the highest probability.

leads to longer routes from cells to the target chooses, but it resulted only in 3 failed scenarios. These results parallel those obtained in [9], wherein risk-averse policies in terms of CVaR for MDPs were studied.

VII. Conclusions

We proposed a method based on bounded policy iteration and convex optimization to design risk-averse stochastic finite state controllers for POMDPs. Future research will explore risk-averse polices for POMDPs that maximize the satisfaction probability of a set of high-level mission specifications in terms of temporal logic formulae [1], [30]. Furthermore, the risk-averse policy synthesis technique will be applied for designing risk-averse planning policies for traversing on uncertain Mars surface (as depicted in Figure 1).

References

[1] M. Ahmadi, R. Sharan, and J. Burdick. Stochastic finite state control of pomdps with ltl specifications. *In Preparation*, 2019.

[2] M. Ahmadi, A. Singletary, J. W. Burdick, and A. D. Ames. Safe Policy Synthesis in Multi-Agent POMDPs via Discrete-Time Barrier Functions. In *58th Conference on Decision and Control*, Nice, France, 2019.

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

[4] K. J. Astrom. Optimal control of Markov decision processes with incomplete state estimation. *J. Mathematical Anal. and Appl.,*, (10):174–205, 1965.

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

[6] Dimitri P Bertsekas. *Dynamic programming and stochastic control*. Number 10. Academic Press, 1976.

[7] Anthony R. Cassandra, Leslie Pack Kaelbling, and Michael L. Littman. Acting optimally in partially observable stochastic domains. In *AAAI*, pages 1023–1028, 1994.

[8] Y. Chow and M. Ghavamzadeh. Algorithms for cvar optimization in mdps. In *Advances in neural information processing systems*, pages 3509–3517, 2014.

[9] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems*, pages 1522–1530, 2015.

[10] J. Gonzalo and J. Olmo. Differences between short-and long-term risk aversion: An optimal asset allocation perspective. *Oxford Bulletin of Economics and Statistics*, 81(1):42–61, 2019.

[11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[12] Eric A Hansen. Solving POMDPs by searching in policy space. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 211–219. Morgan Kaufmann Publishers Inc., 1998.

[13] Eric A. Hansen. Sparse stochastic finite-state controllers for pomdps. 2008.

[14] R. L. Heilman, E. P. Green, K. S. Reddy, A. Moss, and B. Kaplan. Potential impact of risk and loss aversion on the process of accepting kidneys for transplantation. *Transplantation*, 101(7):1514–1517, 2017.

[15] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998.

[16] V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, 2016.

[17] O. Madani, S. Hanks, and A. Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1):5 – 34, 2003.

[18] C. L. McGhan, T. Vaquero, A. R. Subrahmanya, O. Arslan, R. Murray, M. D. Ingham, M. Ono, T. Estlin, B. Williams, and M. Elaasar. The resilient spacecraft executive: An architecture for risk-aware operations in uncertain environments. In *Aiaa Space 2016*, page 5541. 2016.

[19] P. Nilsson, S. Haesaert, R. Thakker, K. Otsu, C. I. Vasile, A. Agha-Mohammadi, R. M. Murray, and A. D. Ames. Toward specification-guided active mars exploration for cooperative robot teams. In *Robotics: Science and Systems*, 2018.

[20] M. Ono, M. Heverly, B. Rothrock, E. Almeida, F. Calef, T. Soliman, N. Williams, H. Gengl, T. Ishimatsu, A. Nicholas, et al. Mars 2020 site-specific mission performance analysis: Part 2. surface traversability. In *2018 AIAA SPACE and Astronautics Forum and Exposition*, page 5419, 2018.

[21] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

[22] M. Ono, B. C. Williams, and L. Blackmore. Probabilistic planning for continuous dynamic systems under bounded risk. *Journal of Artificial Intelligence Research*, 46:511–577, 2013.

[23] Masahiro Ono. Closed-loop chance-constrained mpc with probabilistic resolvability. In *2012 IEEE 51st IEEE conference on decision and control (CDC)*, pages 2611–2618. IEEE, 2012.

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

[25] Jonathan Theodor Ott. *A Markov decision model for a surveillance application and risk-sensitive Markov decision processes*. 2010.

[26] Pascal Poupart and Craig Boutilier. Bounded finite state controllers. In *NIPS*, 2003.

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

[28] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[29] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

[30] R. Sharan and J. Burdick. Finite state control of pomdps with ltl specifications. In *American Control Conference*, 2014.

[31] S. Singh, Y. Chow, A. Majumdar, and M. Pavone. A framework for time-consistent, risk-sensitive model predictive control: Theory and algorithms. *IEEE Transactions on Automatic Control*, 2018.

[32] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, 2016.

[33] D. Vose. *Risk analysis: a quantitative guide*. John Wiley & Sons, 2008.