Risk-Averse Decision Making Under Uncertainty

Mohamadreza Ahmadi, Ugo Rosolia, Michel D. Ingham, Richard M. Murray, *Fellow, IEEE*, and Aaron D. Ames, *Fellow, IEEE*

***Abstract*—A large class of decision making under uncertainty problems can be described via Markov decision processes (MDPs) or partially observable MDPs (POMDPs), with application to artificial intelligence and operations research, among others. In this article, we consider the problem of designing policies for MDPs and POMDPs with objectives and constraints in terms of dynamic coherent risk measures rather than the traditional total expectation, which we refer to as the *constrained risk-averse problem*. Our contributions can be described as follows: first, for MDPs, under some mild assumptions, we propose an optimization-based method to synthesize Markovian policies. We then demonstrate that such policies can be found by solving difference convex programs (DCPs). We show that our formulation generalize linear programs for constrained MDPs with total discounted expected costs and constraints; second, for POMDPs, we show that, if the coherent risk measures can be defined as a Markov risk transition mapping, an infinite-dimensional optimization can be used to design Markovian belief-based policies. For POMDPs with stochastic finite-state controllers (FSCs), we show that the latter optimization simplifies to a (finite dimensional) DCP. We incorporate these DCPs in a policy iteration algorithm to design risk-averse FSCs for POMDPs. We demonstrate the efficacy of the proposed method with numerical experiments involving conditional-value-at-risk and entropic-value-at-risk risk measures.**

***Index Terms*—Markov processes, stochastic systems, uncertain systems.**

I. Introduction

AUTONOMOUS systems are being increasingly deployed in real-world settings. Hence, the associated risk that stems from unknown and unforeseen circumstances is correspondingly on the rise. This demands for autonomous systems that can make appropriately conservative decisions when faced with uncertainty in their environment and behavior. Mathematically speaking, risk can be quantified in numerous ways, such as chance constraints [1] and distributional robustness [2]. However, applications in autonomy and robotics require more "nuanced assessments of risk" [3], motivating the need for risk-averse safety analysis [4] and synthesis [5] for autonomous systems (see an application to bipedal robots [6]).

Artzner et al. [7] characterized a set of natural properties that are desirable for a risk measure, called a coherent risk measure, and have obtained widespread acceptance in finance and operations research, among other fields.

A popular model for representing sequential decision making under uncertainty is a Markov decision processes (MDP) [8]. MDPs with coherent risk objectives were studied in [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], and [19]. In [9] and [11], the authors proposed a sampling-based algorithm for finding saddle point solutions using policy gradient methods. However, Tamar et al. [11] required the risk envelope appearing in the dual representation of the coherent risk measure to be known with an explicit canonical convex programming formulation. While this may be the case for CVaR, mean-semideviation, and spectral risk measures [12], such explicit form is not known for general coherent risk measures, such as EVaR. Furthermore, it is not clear whether the saddle point solutions are a lower bound or upper bound to the optimal value. Saddle-point problems are solved also in [13] to compute stochastic approximations to risk-aware MDPs. Also, policy-gradient-based methods require calculating the gradient of the coherent risk measure, which is not available in explicit form in general. For the CVaR measure, MDPs with risk constraints and total expected costs were studied in [14] and [15] and locally optimal solutions were found via policy gradients, as well. However, this method also leads to saddle point solutions (which cannot be shown to be upper bounds or lower bounds of the optimal value) and cannot be applied to general coherent risk measures. In addition, because the objective and the constraints are in terms of different coherent risk measures, the authors assume there exists a policy that satisfies the CVaR constraint (feasibility assumption), which may not be the case in general. Following the footsteps of [16], a promising approach based on approximate value iteration was proposed for MDPs with CVaR objectives in [17]. A policy iteration algorithm for finding policies that minimize total coherent risk measures for MDPs was studied in [18] and a computational nonsmooth Newton method was proposed in [18]. Similarly an offline iterative algorithm was proposed also in [19].

When the states of the agent and/or the environment are not directly observable, a partially observable MDP (POMDP) can

Manuscript received 24 July 2022; accepted 18 March 2023. Date of publication 3 April 2023; date of current version 29 December 2023. Recommended by Associate Editor Z. Shu. *(Corresponding author: Mohamadreza Ahmadi.)*

Mohamadreza Ahmadi, Ugo Rosolia, Richard M. Murray, and Aaron D. Ames are with the Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: mrahmadi@caltech.edu; urosolia@caltech.edu; murray@caltech.edu; ames@caltech.edu).

Michel D. Ingham is with the NASA Jet Propulsion Laboratory, Pasadena, CA 91109 USA (e-mail: michel.d.ingham@jpl.nasa.gov).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2023.3264178.

Digital Object Identifier 10.1109/TAC.2023.3264178

be used to study decision making under uncertainty introduced by the partial state observability [20]. POMDPs with coherent risk measure objectives were studied in [21] and [22]. Despite the elegance of the theory, no computational method was proposed to design policies for general coherent risk measures. In [23], we proposed a method for finding finite-state controllers (FSCs) for POMDPs with objectives defined in terms of coherent risk measures, which takes advantage of convex optimization techniques. However, the method can only be used if the risk transition mapping is affine in the policy.

Summary of contributions: In this article, we consider MDPs and POMDPs with both objectives and constraints in terms of coherent risk measures. Our contributions are fourfold.

1) For MDPs, we use the Lagrangian framework and reformulate the problem into a inf-sup problem. For Markov risk transition mappings, we propose an optimization-based method to design Markovian policies that lower bound the constrained risk-averse problem.
2) For MDPs, we evince that the optimization problems are in the special form of DCPs and can be solved by the disciplined convex–concave programming (DCCP) method. We also demonstrate that these results generalize linear programs for constrained MDPs with total discounted expected costs and constraints.
3) For POMDPs, we demonstrate that, if the coherent risk measures can be defined as a Markov risk transition mapping, an infinite-dimensional optimization can be used to design Markovian belief-based policies, which in theory requires infinite memory to synthesize (in accordance with classical POMDP complexity results).
4) For POMDPs with stochastic FSCs, we show that the latter optimization converts to a (finite dimensional) DCP and can be solved by the DCCP framework. We incorporate these DCPs in a policy iteration algorithm to design risk-averse FSCs for POMDPs.

We assess the efficacy of the proposed method with numerical experiments involving conditional-value-at-risk (CVaR) and entropic-value-at-risk (EVaR) risk measures.

Preliminary results on risk-averse MDPs were presented in [24]. This article, in addition to providing detailed proofs and new numerical analysis in the MDP case, generalizes [24] to partially observable systems (POMDPs) with dynamic coherent risk objectives and constraints.

The rest of this article is organized as follows. In the following section, we briefly review some notions used in this article. In Section III, we formulate the problem under study. In Section IV, we present the optimization-based method for designing risk-averse policies for MDPs. In Section V, we describe a policy iteration method for designing finite-memory controllers for risk-averse POMDPs. In Section VI, we illustrate the proposed methodology via numerical experiments. Finally, Section VII concludes this article.

Notation: We denote by $\mathbb{R}^n$ the n-dimensional Euclidean space and $\mathbb{N}_{\geq 0}$ the set of nonnegative integers. Throughout this article, we use bold font to denote a vector and $(\cdot)^\top$ for its transpose, e.g., $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$, with $n \in \{1, 2, \ldots\}$. For a vector $\boldsymbol{a}$, we use $\boldsymbol{a} \succeq (\preceq) \mathbf{0}$ to denote element-wise nonnegativity (nonpositivity) and $\boldsymbol{a} \equiv \mathbf{0}$ to show all elements of $\boldsymbol{a}$ are zero. For two vectors $a, b \in \mathbb{R}^n$, we denote their inner product by $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$, i.e., $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^\top \boldsymbol{b}$. For a finite set $\mathcal{A}$, we denote its power set by $2^{\mathcal{A}}$, i.e., the set of all subsets of $\mathcal{A}$. For a probability space $(\omega_T, \mathcal{F}, \mathbb{P})$ and a constant $p \in [1, \infty)$, $\mathcal{L}_p(\omega_T, \mathcal{F}, \mathbb{P})$ denotes the vector space of real valued random variables c for which $\mathbb{E}|c|^p < \infty$.

II. Preliminaries

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

A. Markov Decision Processes

An MDP is a tuple $\mathcal{M} = (\mathcal{S}, \text{Act}, T, \kappa_0)$ consisting of a set of states $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ of the autonomous agent(s) and world model, actions $\text{Act} = \{\alpha_1, \ldots, \alpha_{|\text{Act}|}\}$ available to the agent, a transition function $T(s_j | s_i, \alpha)$, and κ_0 describing the initial distribution over the states.

This article considers *finite* MDPs, where $\mathcal{S}$ and Act 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 $\alpha \in \text{Act}$ is given by $T(s_j | s_i, \alpha)$. The probabilistic components of an MDP must satisfy the following:

$$\begin{cases} \sum_{s \in \mathcal{S}} T(s | s_i, \alpha) = 1 & \forall s_i \in \mathcal{S} \quad \forall \alpha \in \text{Act} \\ \sum_{s \in \mathcal{S}} \kappa_0(s) = 1. \end{cases}$$

B. Partially Observable MDPs

A *POMDP* is a tuple $\mathcal{PM} = (\mathcal{M}, \mathcal{O}, O)$ consisting of an MDP $\mathcal{M}$, observations $\mathcal{O} = \{o_1, \ldots, o_{|\mathcal{O}|}\}$, and an observation model $O(o \mid s)$. We consider *finite* POMDPs, where $\mathcal{O}$ is a finite set. Then, for each state s_i, an observation $o \in \mathcal{O}$ is generated independently with probability $O(o | s_i)$, which satisfies

$$\sum_{o \in \mathcal{O}} O(o | s) = 1 \quad \forall s \in \mathcal{S}.$$

In POMDPs, the states $s \in \mathcal{S}$ are not directly observable. The beliefs $b \in \Delta(\mathcal{S})$, i.e., the probability of being in different states, 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:

$$b_0(s) = \frac{\kappa_0(s) O(o_0 \mid s)}{\sum_{o \in O} \kappa_0(s) O(o \mid s)} \tag{1}$$

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

for all $t \geq 1$.

C. Finite-State Control of POMDPs

It is well established that designing optimal policies for POMDPs based on the (continuous) belief states requires uncountably infinite memory or internal states [25], [26]. This article focuses on a particular class of POMDP controllers, namely, FSCs.

A *stochastic FSC* for $\mathcal{PM}$ is given by the tuple $\mathcal{G} = (G, \omega_T, \kappa)$, where $G = \{g_1, g_2, \ldots, g_{|G|}\}$ is a finite set of internal states (I-states), $\omega_T : G \times \mathcal{O} \to \Delta(G \times \text{Act})$ is a function of internal stochastic FSC states g_k and observation o, such that

$\omega_T(g_k, o)$ is a probability distribution over $G \times \text{Act}$. The next internal state and action pair (g_l, α) is chosen by independent sampling of $\omega_T(g_k, o)$. By abuse of notation, $\omega_T(g_l, \alpha | g_k, o)$ will denote the probability of transitioning to internal stochastic FSC 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 $\kappa(\kappa_0)$, given initial distribution κ_0 of $\mathcal{PM}$, and $\kappa(g|\kappa_0)$ will denote the probability of starting the FSC in internal state g when the initial POMDP distribution is κ_0.

D. Coherent Risk Measures

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a filtration $\mathcal{F}_0 \subset \cdots \mathcal{F}_N \subset \mathcal{F}$, and an adapted sequence of random variables (stage-wise costs) c_t, $t = 0, \ldots, N$, where $N \in \mathbb{N}_{\geq 0} \cup \{\infty\}$. For $t = 0, \ldots, N$, we further define the spaces $\mathcal{C}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, \mathbb{P}), p \in [1, \infty), \mathcal{C}_{t:N} = \mathcal{C}_t \times \cdots \times \mathcal{C}_N$, and $\mathcal{C} = \mathcal{C}_0 \times \mathcal{C}_1 \times \cdots$. We assume that the sequence $\boldsymbol{c} \in \mathcal{C}$ is almost surely bounded (with exceptions having probability zero), i.e., $\max_t \text{ess sup} |c_t(\omega)| < \infty$.

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 1 (Conditional Risk Measure): A mapping $\rho_{t:N} : \mathcal{C}_{t:N} \to \mathcal{C}_t$, where $0 \leq t \leq N$, is called a *conditional risk measure*, if it has the following monoticity property:

$$\rho_{t:N}(\boldsymbol{c}) \leq \rho_{t:N}(\boldsymbol{c}') \qquad \forall \boldsymbol{c}, \forall \boldsymbol{c}' \in \mathcal{C}_{t:N} \text{ such that } \boldsymbol{c} \preceq \boldsymbol{c}'.$$

Definition 2 (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 [18, Definition 3]. 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 perspective at time τ.

In this article, we focus on time consistent, coherent risk measures, which satisfy four nice mathematical properties, as defined below [12, p. 298].

Definition 3 (Coherent Risk Measure): We call the one-step conditional risk measures $\rho_t : \mathcal{C}_{t+1} \to \mathcal{C}_t$, $t = 1, \ldots, N-1$ a *coherent risk measure* if it satisfies the following conditions.

1) 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 $c, c' \in \mathcal{C}_{t+1}$.
2) Monotonicity: If $c \leq c'$ then $\rho_t(c) \leq \rho_t(c')$ for all $c, c' \in \mathcal{C}_{t+1}$.
3) Translational invariance: $\rho_t(c + c') = c + \rho_t(c')$ for all $c \in \mathcal{C}_t$ and $c' \in \mathcal{C}_{t+1}$.
4) Positive homogeneity: $\rho_t(\beta c) = \beta \rho_t(c)$ for all $c \in \mathcal{C}_{t+1}$ and $\beta \geq 0$.

We are interested in the discounted infinite horizon problems. Let $\gamma \in (0,1)$ be a given discount factor. For $t = 0, 1, \ldots$, we define the functional

$$\rho^{\gamma}_{0,t}(c_0, \ldots, c_t) = \rho_{0,t}\left(c_0, \gamma c_1, \ldots, \gamma^t c_t\right)$$
$$\rho_0\big(c_0 + \rho_1\big(\gamma c_1 + \rho_2\big(\gamma^2 c_2 + \cdots + \rho_{t-1}\left(\gamma^{t-1} c_{t-1} + \rho_t\left(\gamma^t c_t\right)\right)\cdots\big)\big)\big).$$

Finally, we have total discounted risk functional $\rho^{\gamma} : \mathcal{C} \to \mathbb{R}$ defined as

$$\rho^{\gamma}(\boldsymbol{c}) = \lim_{t \to \infty} \rho^{\gamma}_{0,t}(c_0, \ldots, c_t). \tag{3}$$

From [18, Th. 3], we have that ρ^{γ} is convex, monotone, and positive homogeneous.

For details regarding different examples of coherent of risk measures we consider in this article, i.e., CVaR and EVaR, please refer to the Appendix.

III. Problem Formulation

Consider a stationary controlled Markov process $\{q_t\}$, $t = 0, 1, \ldots$ (an MDP or a POMDP) with initial probability distribution κ_0, wherein policies, transition probabilities, and cost functions do not depend explicitly on time. Each policy $\pi = \{\pi_t\}_{t=0}^{\infty}$ leads to cost sequences $\boldsymbol{c}_t = c(q_t, \alpha_t)$, $t = 0, 1, \ldots$ and $\boldsymbol{d}^i_t = d^i(q_t, \alpha_t), t = 0, 1, \ldots, i = 1, 2, \ldots, n_c$. We define the dynamic risk of evaluating the γ-discounted cost of a policy π as

$$J_{\gamma}(\kappa_0, \pi) = \rho^{\gamma}\left(c(q_0, \alpha_0), c(q_1, \alpha_1), \ldots\right) \tag{4}$$

and the γ-discounted dynamic risk constraints of executing policy π as

$$D^i_{\gamma}(\kappa_0, \pi) = \rho^{\gamma}\left(d^i(q_0, \alpha_0), d^i(q_1, \alpha_1), \ldots\right) \leq \beta^i \qquad i = 1, 2, \ldots, n_c \tag{5}$$

where ρ^{γ} is defined in (3), $q_0 \sim \kappa_0$, and $\beta^i > 0$, $i = 1, 2, \ldots, n_c$, are given constants. We assume that $c(\cdot, \cdot)$ and $d^i(\cdot, \cdot)$, $i = 1, 2, \ldots, n_c$, are nonnegative and upper bounded. For a discount factor $\gamma \in (0,1)$, an initial condition κ_0, and a policy π, we infer from [18, Th. 3] that both $J_{\gamma}(\kappa_0, \pi)$ and $D^i_{\gamma}(\kappa_0, \pi)$ are well-defined (bounded), if c and d are bounded.

In this work, we are interested in addressing the following problem.

Problem 1: For a controlled Markov decision process (an MDP or a POMDP), a discount factor $\gamma \in (0,1)$, and a total risk functional $J_{\gamma}(\kappa_0, \pi)$ as in (4) and total cost constraints (5), where $\{\rho_t\}_{t=0}^{\infty}$ are coherent risk measures, compute

$$\pi^* \in \text{argmin}_{\pi} \; J_{\gamma}(\kappa_0, \pi)$$
$$\textit{subject to} \quad \boldsymbol{D}_{\gamma}(\kappa_0, \pi) \preceq \boldsymbol{\beta}. \tag{6}$$

We call a controlled Markov process with the "nested" objective (4) and constraints (5) a *constrained risk-averse Markov process*.

For MDPs, the authors in [17] and [27] showed that such coherent risk measure objectives can account for modeling errors and parametric uncertainties. We can also interpret Problem 1 as designing policies that minimize the accrued costs in a risk-averse sense[1] and at the same time ensuring that the system constraints, e.g., fuel constraints, are not violated even in the rare but costly scenarios.

[1] With the exception of conditional expectation as the coherent risk measure.

Note that in Problem 1 both the objective function and the constraints are in general nondifferentiable and nonconvex in policy π (with the exception of total expected cost as the coherent risk measure ρ^γ [28]). Therefore, finding optimal policies in general may be hopeless. Instead, we find suboptimal polices by taking advantage of a Lagrangian formulation and then using an optimization form of Bellman's equations.

Next, we show that the constrained risk-averse problem is equivalent to a nonconstrained inf-sup risk-averse problem thanks to the Lagrangian method.

Proposition 1: Let $J_\gamma(\kappa_0)$ be the value of Problem 1 for a given initial distribution κ_0 and discount factor γ. Then, (i) the value function satisfies

$$J_\gamma(\kappa_0) = \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi, \boldsymbol{\lambda}) \tag{7}$$

where $\boldsymbol{\lambda}$ is the vector of the Lagrange multipliers and

$$L_\gamma(\pi, \boldsymbol{\lambda}) = J_\gamma(\kappa_0, \pi) + \langle \boldsymbol{\lambda}, (\boldsymbol{D}_\gamma(\kappa_0, \pi) - \boldsymbol{\beta}) \rangle \tag{8}$$

is the Lagrangian function.

(ii) Furthermore, a policy π^* is optimal for Problem 1, if and only if $J_\gamma(\kappa_0) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi^*, \boldsymbol{\lambda})$.

Proof: (i) If for some π Problem 1 is not feasible, then $\sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi, \boldsymbol{\lambda}) = \infty$. In fact, if the ith constraint is not satisfied, i.e., $D^i_\gamma > \beta^i$, we can achieve the latter supremum by choosing $\lambda_i \to \infty$, while keeping the rest of λ^is constant or zero. If Problem 1 is feasible for some π, then the supremum is achieved by setting $\boldsymbol{\lambda} = \mathbf{0}$. Hence, $L_\gamma(\boldsymbol{\lambda}, \pi) = J_\gamma(\kappa_0, \pi)$ and

$$\inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi, \boldsymbol{\lambda}) = \inf_{\pi : \boldsymbol{D}_\gamma(\kappa_0, \pi) \leq \boldsymbol{\beta}} J_\gamma(\kappa_0, \pi)$$

which implies (i).

(ii) If π is optimal, then, from (7), we have

$$J_\gamma(\kappa_0) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi^*, \boldsymbol{\lambda}).$$

Conversely, if $J_\gamma(\kappa_0) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi, ' \boldsymbol{\lambda})$ for some π', then from (7), we have $\inf_\pi \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L_\gamma(\pi, ' \boldsymbol{\lambda})$. Hence, π' is the optimal policy. ■

IV. Constrained Risk-Averse MDPs

At any time t, the value of ρ_t is $\mathcal{F}_t$-measurable and is allowed us to depend on the entire history of the process $\{s_0, s_1, \ldots\}$ and we cannot expect to obtain a Markov optimal policy [29], [30]. In order to obtain Markov policies, we need the following property [18].

Definition 4 (Markov Risk Measure): Let $m, n \in [1, \infty)$ such that $1/m + 1/n = 1$ and $\mathcal{P} = \{p \in \mathcal{L}_n(\mathcal{S}, 2^{\mathcal{S}}, \mathbb{P}) \mid \sum_{s' \in \mathcal{S}} p(s') \mathbb{P}(s') = 1, p \geq 0\}$. A one-step conditional risk measure $\rho_t : \mathcal{C}_{t+1} \to \mathcal{C}_t$ is a Markov risk measure with respect to the controlled Markov process $\{s_t\}, t = 0, 1, \ldots$, if there exists a risk transition mapping $\sigma_t : \mathcal{L}_m(\mathcal{S}, 2^{\mathcal{S}}, \mathbb{P}) \times \mathcal{S} \times \mathcal{P} \to \mathbb{R}$ such that for all $v \in \mathcal{L}_m(\mathcal{S}, 2^{\mathcal{S}}, \mathbb{P})$ and $\alpha_t \in \pi(s_t)$, we have

$$\rho_t(v(s_{t+1})) = \sigma_t(v(s_{t+1}), s_t, p(s_{t+1} | s_t, \alpha_t)) \tag{9}$$

where $p : \mathcal{S} \times Act \to \mathcal{P}$ is called the controlled kernel.

In fact, if ρ_t is a coherent risk measure, σ_t also satisfies the properties of a coherent risk measure (Definition 3). In this article, since we are concerned with MDPs, the controlled kernel is simply the transition function T.

Assumption 1: The one-step coherent risk measure ρ_t is a Markov risk measure.

The simplest case of the risk transition mapping is in the conditional expectation case $\rho_t(v(s_{t+1})) = \mathbb{E}\{v(s_{t+1}) \mid s_t, \alpha_t\}$, i.e.,

$$\begin{aligned} &\sigma\{v(s_{t+1}), s_t, p(s_{t+1} | s_t, \alpha_t)\} \\ &= \mathbb{E}\{v(s_{t+1}) \mid s_t, \alpha_t\} \\ &= \sum_{s_{t+1} \in \mathcal{S}} v(s_{t+1}) T(s_{t+1} \mid s_t, \alpha_t). \end{aligned} \tag{10}$$

Note that in the total discounted expectation case σ is a linear function in v rather than a convex function, which is the case for a general coherent risk measures. For example, for the CVaR risk measure, the Markov risk transition mapping is given by

$$\begin{aligned} &\sigma\{v(s_{t+1}), s_t, p(s_{t+1} | s_t, \alpha_t)\} \\ &= \inf_{\zeta \in \mathbb{R}} \left\{ \zeta + \frac{1}{\varepsilon} \sum_{s_{t+1} \in \mathcal{S}} (v(s_{t+1}) - \zeta)_+ T(s_{t+1} \mid s_t, \alpha_t) \right\} \end{aligned}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ is a convex function in v.

If σ is a coherent Markov risk measure, then the Markov policies are sufficient to ensure optimality [18].

In the next result, we show that we can find a lower bound to the solution to Problem 1 via solving an optimization problem. We later show that this optimization problem has some nice properties that can be used to synthesize risk-averse policies.

Theorem 1: Consider an MDP $\mathcal{M}$ with the nested risk objective (4), constraints (5), and discount factor $\gamma \in (0, 1)$. Let Assumption 1 hold and ρ_t, $t = 0, 1, \ldots$ be coherent risk measures as described in Definition 3. Then, the solution $(\boldsymbol{V}^*_\gamma, \boldsymbol{\lambda}^*)$ to the following optimization problem (Bellman's equation):

$$\sup_{\boldsymbol{V}_\gamma, \boldsymbol{\lambda} \succeq \mathbf{0}} \langle \boldsymbol{\kappa_0}, \boldsymbol{V}_\gamma \rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\beta} \rangle$$

subject to

$$\begin{aligned} V_\gamma(s) \leq\ & c(s, \alpha) + \langle \boldsymbol{\lambda}, \boldsymbol{d}(s, \alpha) \rangle \\ & + \gamma \sigma\{V_\gamma(s'), s, p(s' | s, \alpha)\} \quad \forall s \in \mathcal{S} \quad \forall \alpha \in Act \end{aligned} \tag{11}$$

satisfies

$$J_\gamma(\kappa_0) \geq \langle \boldsymbol{\kappa_0}, \boldsymbol{V}^*_\gamma \rangle - \langle \boldsymbol{\lambda}^*, \boldsymbol{\beta} \rangle. \tag{12}$$

Proof: From Proposition 1, we have known that (7) holds. Hence, we have

$$\begin{aligned} J_\gamma(\kappa_0) &= \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} (J_\gamma(\kappa_0, \pi) + \langle \boldsymbol{\lambda}, (\boldsymbol{D}_\gamma(\kappa_0, \pi) - \boldsymbol{\beta}) \rangle) \\ &= \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} (J_\gamma(\kappa_0, \pi) + \langle \boldsymbol{\lambda}, \boldsymbol{D}_\gamma(\kappa_0, \pi) \rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\beta} \rangle) \\ &= \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} (\rho^\gamma(c) + \langle \boldsymbol{\lambda}, \rho^\gamma(d) \rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\beta} \rangle) \\ &= \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} (\rho^\gamma(c) + \rho^\gamma(\langle \boldsymbol{\lambda}, d \rangle) - \langle \boldsymbol{\lambda}, \boldsymbol{\beta} \rangle) \end{aligned}$$

$$\geq \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left(\rho^{\gamma}(c + \langle \boldsymbol{\lambda}, d \rangle) - \langle \boldsymbol{\lambda}, \beta \rangle\right)$$

$$\geq \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \inf_{\pi} \left(\rho^{\gamma}(c + \langle \boldsymbol{\lambda}, d \rangle) - \langle \boldsymbol{\lambda}, \beta \rangle\right) \tag{13}$$

wherein the fourth, fifth, and sixth inequalities above, we used the positive homogeneity property of ρ^{γ}, subadditivity property of ρ^{γ} [7, Proposition 2.1], and the minimax inequality, respectively. Since $\langle \boldsymbol{\lambda}, \beta \rangle$ does not depend on π, to find the solution the infimum it suffices to find the solution to

$$\inf_{\pi} \rho^{\gamma}(\tilde{c})$$

where $\tilde{c} = c + \boldsymbol{\lambda}' d$. The value to the abovementioned optimization can be obtained by solving the following Bellman equation [18, Th. 4]:

$$V_{\gamma}(s) = \inf_{\alpha \in \text{Act}} \left(\tilde{c}(s, \alpha) + \gamma \sigma\{V_{\gamma}(s'), s, p(s'|s, \alpha)\}\right).$$

Next, we show that the solution to the abovementioned Bellman equation can be alternatively obtained by solving the convex optimization

$$\sup_{V_{\gamma}} \langle \kappa_0, V_{\gamma} \rangle$$

subject to

$$V_{\gamma}(s) \leq \tilde{c}(s, \alpha) + \gamma \sigma\{V_{\gamma}(s'), s, p(s'|s, \alpha)\} \quad \forall s, \alpha. \tag{14}$$

Define

$$\mathfrak{D}_{\pi} v := \tilde{c}(s, \pi(s)) + \gamma \sigma\{v(s'), s, p(s'|s, \pi(s))\} \qquad \forall s \in \mathcal{S}$$

and $\mathfrak{D} v := \min_{\alpha \in \text{Act}} (\tilde{c}(s, \alpha) + \gamma \sigma\{v(s'), s, p(s'|s, \alpha)\})$ for all $s \in \mathcal{S}$. From [18, Lemma 1], we infer that $\mathfrak{D}_{\pi}$ and $\mathfrak{D}$ are nondecreasing; i.e., for $v \leq w$, we have $\mathfrak{D}_{\pi} v \leq \mathfrak{D}_{\pi} w$ and $\mathfrak{D} v \leq \mathfrak{D} w$. Therefore, if $V_{\gamma} \leq \mathfrak{D}_{\pi} V_{\gamma}$, then $\mathfrak{D}_{\pi} V_{\gamma} \leq \mathfrak{D}_{\pi}(\mathfrak{D}_{\pi} V_{\gamma})$, which implies that $V_{\gamma} \leq \mathfrak{D}_{\pi} V_{\gamma} \leq \mathfrak{D}_{\pi}(\mathfrak{D}_{\pi} V_{\gamma}) = \mathfrak{D}_{\pi}^2 V_{\gamma}$. By repeated application of $\mathfrak{D}_{\pi}$, we obtain

$$V_{\gamma} \leq \mathfrak{D}_{\pi} V_{\gamma} \leq \mathfrak{D}_{\pi}^2 V_{\gamma} \leq \mathfrak{D}_{\pi}^{\infty} V_{\gamma} = V_{\gamma}^*.$$

Note that by definition of $\mathfrak{D}_{\pi}(\cdot)$, any feasible solution to (14) must satisfy $V_{\gamma} \leq \mathfrak{D}_{\pi} V_{\gamma}$ and, hence, must satisfy $V_{\gamma} \leq V_{\gamma}^*$. Thus, given that all entries of κ_0 are positive, V_{γ}^* is the optimal solution to (14). Substituting (14) back in the last inequality in (13) yields the result. ■

Once the values of $\boldsymbol{\lambda}^*$ and $\boldsymbol{V}_{\gamma}^*$ are found by solving optimization problem (11), we can find the policy as

$$\pi^*(s) \in \text{argmin}_{\alpha \in \text{Act}} \Big(c(s, \alpha) + \langle \boldsymbol{\lambda}^*, \boldsymbol{d}(s, \alpha) \rangle + \gamma \sigma\{V_{\gamma}^*(s'), s, p(s'|s, \alpha)\}\Big). \tag{15}$$

One interesting observation is that if the coherent risk measure ρ^t is the total discounted expectation, Theorem 1 can be simplified to the following Corollary, which was formulated in [28] for constrained MDPs using properties of Markov processes.

Corollary 1: Let the assumptions of Theorem 1 hold and let $\rho_t(\cdot) = \mathbb{E}(\cdot|s_t, \alpha_t)$, $t = 1, 2, \ldots$. Then, the solution $(\boldsymbol{V}_{\gamma}^*, \boldsymbol{\lambda}^*)$ to optimization (11) satisfies

$$J_{\gamma}(\kappa_0) = \langle \boldsymbol{\kappa_0}, \boldsymbol{V}_{\gamma}^* \rangle - \langle \boldsymbol{\lambda}^*, \beta \rangle.$$

Furthermore, with $\rho_t(\cdot) = \mathbb{E}(\cdot|s_t, \alpha_t)$, $t = 1, 2, \ldots$, optimization (11) becomes a linear program.

Proof: From the derivation in (13), we observe the two inequalities are from the application of (a) the subadditivity property of ρ^{γ} and (b) the max–min inequality. Next, we show that in the case of total expectation both of these properties lead to an equality.

a) Subadditivity property of ρ^{γ}: for total expectation, we have

$$\sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t c_t + \sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t \langle \boldsymbol{\lambda}, \boldsymbol{d}_t \rangle = \sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t (c_t + \langle \boldsymbol{\lambda}, \boldsymbol{d}_t \rangle).$$

Thus, equality holds.

b) Max–min inequality: in the $\rho_{\kappa_0}^{\gamma}(\cdot) = \sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t(\cdot)$ case, both the objective function and the constraints are linear in the decision variables π and $\boldsymbol{\lambda}$. Therefore, the sixth line in (13) reads as

$$\inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left(\rho^{\gamma}(c + \langle \boldsymbol{\lambda}, \boldsymbol{d} \rangle) - \langle \boldsymbol{\lambda}, \beta \rangle\right)$$

$$= \inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left(\sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t (c_t + \langle \boldsymbol{\lambda}, \boldsymbol{d}_t \rangle) - \langle \boldsymbol{\lambda}, \beta \rangle\right). \tag{16}$$

Since the expression inside parentheses above is convex in π ($\mathbb{E}_{\kappa_0}^{\pi}$ is linear in the policy) and concave (linear) in $\boldsymbol{\lambda}$. From Minimax Theorem [31], we have that the following equality holds:

$$\inf_{\pi} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left(\sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t (c_t + \langle \boldsymbol{\lambda}, \boldsymbol{d}_t \rangle) - \langle \boldsymbol{\lambda}, \beta \rangle\right)$$

$$= \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \inf_{\pi} \left(\sum_t \mathbb{E}_{\kappa_0}^{\pi} \gamma^t (c_t + \langle \boldsymbol{\lambda}, \boldsymbol{d}_t \rangle) - \langle \boldsymbol{\lambda}, \beta \rangle\right).$$

Furthermore, from (10), we see that σ is linear in v for total expectation. Therefore, the constraint in (11) is linear in V_{γ} and $\boldsymbol{\lambda}$. Since $\langle \boldsymbol{\kappa_0}, \boldsymbol{V}_{\gamma} \rangle - \langle \boldsymbol{\lambda}, \beta \rangle$ is also linear in V_{γ}s and $\boldsymbol{\lambda}$s, optimization (11) becomes a linear program in the case of total expectation coherent risk measure. ■

In [24], we presented a method based on difference convex programs to solve (11), wherein ρ^{γ} is an arbitrary coherent risk measure and we described the specific structure of the optimization problem for conditional expectation, CVaR, and EVaR. In fact, it was shown that (11) can be written in a standard DCP [32] format as

$$\inf_{\boldsymbol{V}_{\gamma}, \boldsymbol{\lambda} \succeq \mathbf{0}} \; f_0(\boldsymbol{\lambda}) - g_0(\boldsymbol{V}_{\gamma})$$

subject to

$$f_1(V_{\gamma}) - g_1(\boldsymbol{\lambda}) - g_2(V_{\gamma}) \leq 0 \quad \forall s, \alpha. \tag{17}$$

DCPs arise in many applications, such as feature selection in machine learning [33] and inverse covariance estimation in statistics [34]. Although DCPs can be solved globally [32], e.g., using branch and bound algorithms [35], a locally optimal solution can be obtained based on techniques of nonlinear optimization [36] more efficiently. In particular, in this work, we use a variant of the convex–concave procedure [37], [38], wherein the concave terms are replaced by a convex upper

bound and solved. In fact, the DCCP [38] technique linearizes DCP problems into a (disciplined) convex program (carried out automatically via the DCCP Python package [38]), which is then converted into an equivalent cone program by replacing each function with its graph implementation. Then, the cone program can be solved readily by available convex programming solvers, such as CVXPY [39].

We end this section by pointing out that solving (11) using the DCCP method only finds the (local) saddle points to optimization problem (11). Nevertheless, every saddle point to (11) satisfies (12) (from Theorem 1). In fact, every saddle point is a lower bound of the optimal value of Problem 1.

V. Constrained Risk-Averse POMDPs

So far, we considered MDPs where information about the states are directly observable. In this section, we propose a policy synthesis methodology for MDPs with partial state information. In fact, for POMDPs, we can find a lower bound to the solution to Problem 1 via solving an infinite-dimensional optimization problem. Note that a POMDP is equivalent to a belief MDP $\{b_t\}$, $t = 1, 2, \ldots$, where b_t is defined in (2).

Theorem 2: Consider a POMDP $\mathcal{PM}$ with the nested risk objective (4) and constraint (5) with $\gamma \in (0, 1)$. Let Assumption 1 hold, let ρ_t, $t = 0, 1, \ldots$ be coherent risk measures, and suppose $c(\cdot, \cdot)$ and $\{d^i(\cdot, \cdot)\}_{i=1}^{n_c}$ be nonnegative and upper bounded. Then, the solution (λ^*, V_γ^*) to the following Bellman's equation:

$$\sup_{\boldsymbol{V}_\gamma, \boldsymbol{\lambda} \succeq 0} \langle \boldsymbol{b}_0, \boldsymbol{V}_\gamma \rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\beta} \rangle$$

subject to

$$V_\gamma(b) \leq c(b, \alpha) + \langle \boldsymbol{\lambda}, \boldsymbol{d}(b, \alpha) \rangle + \gamma \sigma\{V_\gamma(b'), b, p(b'|b, \alpha)\} \quad \forall b \in \Delta(\mathcal{S}) \quad \forall \alpha \in \text{Act} \tag{18}$$

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

$$J_\gamma(b_0) \geq \langle \boldsymbol{b}_0, \boldsymbol{V}_\gamma^* \rangle - \langle \boldsymbol{\lambda}^*, \boldsymbol{\beta} \rangle. \tag{19}$$

Proof: Note that a POMDP can be represented as an MDP over the belief states (2) with initial distribution (1). Hence, a POMDP is a controlled Markov process with states $b \in \Delta(\mathcal{S})$, where the controlled belief transition probability is described as

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

with

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

The rest of the proof follows the same footsteps on Theorem 1 over the belief MDP with $p(b'|b, \alpha)$ as defined above. ■

Unfortunately, since $b \in \Delta(\mathcal{S})$ and hence $V_\gamma : \Delta(\mathcal{S}) \to \mathbb{R}$, optimization (18) is infinite-dimensional and we cannot solve it efficiently.

If the one-step coherent risk measure ρ_t is the total discounted expectation, we can show that optimization problem (18) simplifies to an infinite-dimensional linear program and equality holds in (19). This can be proved following the same lines as the proof of Corollary 1 but for the belief MDP. Hence, Theorem 2 also provides an optimization based solution to the constrained POMDP problem.

A. Risk-Averse FSC Synthesis Via Policy Iteration

In order to synthesize risk-averse FSCs, we employ a policy iteration algorithm. Policy iteration incrementally improves a controller by alternating between two steps: Policy evaluation (computing value functions by fixing the policy) and policy improvement (computing the policy by fixing the value functions), until convergence to a satisfactory policy [40]. For a risk-averse POMDP, policy evaluation can be carried out by solving (18). However, as mentioned earlier, (18) is difficult to use directly as it must be computed at each (continuous) belief state in the belief space, which is uncountably infinite.

In the following, we show that if instead of considering policies with infinite-memory, we search over finite-memory policies, then we can find suboptimal solutions to Problem 1 that lower bound $J_\gamma(\kappa_0)$. To this end, we consider stochastic but finite-memory controllers as described in Section II-C.

Closing the loop around a POMDP with an FSC $\mathcal{G}$ induces a Markov chain. The global Markov chain $\mathcal{MC}_{\mathcal{S} \times G}^{\mathcal{PM}, \mathcal{G}}$ (or simply $\mathcal{MC}$, where the stochastic FSC and the POMDP are clear from the context) with execution $\{[s_0, g_0], [s_1, g_1], \ldots\}$, $[s_t,\ g_t] \in \mathcal{S} \times G$. The probability of initial global state $[s_0, g_0]$ is

$$\iota_{\text{init}}([s_0, g_0]) = \kappa_0(s_0) \kappa(g_0 | \kappa_0).$$

The state transition probability, $T^{\mathcal{M}}$, is given by

$$\begin{aligned} &T^{\mathcal{M}}([s_{t+1}, g_{t+1}] \,|[s_t, g_t]) \\ &= \sum_{o \in \mathcal{O}} \sum_{\alpha \in \text{Act}} O(o|s_t) \omega_T(g_{t+1}, \alpha | g_t, o) T(s_{t+1} | s_t, \alpha). \end{aligned}$$

B. Risk Value Function Computation

Under an FSC, 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_T and κ. The closed-loop Markov chain $\mathcal{M}_{\mathcal{S} \times \mathcal{G}}^{\mathcal{PM} \times \mathcal{G}}$ is a controlled Markov process with $\{q_t\} = \{[s_t, g_t]\}$, $t = 1, 2, \ldots$. In this setting, the total risk functional (4) becomes a function of ι_{init} and FSC $\mathcal{G}$, i.e.,

$$J_\gamma(\iota_{\text{init}}, \mathcal{G}) = \rho^\gamma\left(c([s_0, g_0], \alpha_0), c([s_1, g_1], \alpha_1), \ldots\right) \quad s_0 \sim \kappa_0,\ g_0 \sim \kappa \tag{20}$$

where α_ts and g_ts are drawn from the probability distribution $\omega_T(g_{t+1}, \alpha_t \mid g_t, o_t)$. The constraint functionals $D_\gamma^i(\iota_{\text{init}}, \mathcal{G})$, $i = 1, 2, \ldots, n_c$ can also be defined similarly.

Let $J_\gamma(\iota_{\text{init}})$ be the value of Problem 1 under an FSC $\mathcal{G}$. Then, it is evident that $J_\gamma(b_0) \geq J_\gamma(\iota_{\text{init}})$, since FSCs restrict the search space of the policy π. That is, they can only be as good as the (infinite-dimensional) belief-based policy $\pi(b)$ as $|G| \to \infty$ (infinite-memory).

Risk value function optimization: For POMDPs controlled by stochastic FSCs, the dynamic program is developed in the global state space $\mathcal{S} \times G$. From Theorem 1, we see that for a given FSC, $\mathcal{G}$, and POMDP $\mathcal{PM}$, the value function $V_{\gamma,\mathcal{M}}([s,g])$ can be computed by solving the following finite dimensional optimization:

$$\sup_{\boldsymbol{V}_{\gamma,\mathcal{M}},\boldsymbol{\lambda}\succeq\mathbf{0}} \ \langle \iota_{\text{init}}, \boldsymbol{V}_{\gamma,\mathcal{M}}\rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\beta}\rangle$$

subject to

$$V_{\gamma,\mathcal{M}}([s,g]) \leq \sum_{\alpha\in\text{Act}} p(\alpha \mid g)\tilde{c}([s,g],\alpha) + \gamma\sigma\left\{V_{\gamma,\mathcal{M}}([s,'g']),[s,g],T^{\mathcal{M}}([s,'g']\,|[s,g])\right\} \quad \forall s\in\mathcal{S} \quad \forall g\in G \tag{21}$$

where $p(\alpha \mid g) = \sum_{g'\in\mathcal{G},o\in\mathcal{O}} \omega_T(g,'\alpha \mid g,o)O(o|g')$, and $\tilde{c}([s,g],\alpha) = c([s,g],\alpha) + \langle \boldsymbol{\lambda}, \boldsymbol{d}([s,g],\alpha)\rangle$. Then, the solution $(\boldsymbol{V}^*_{\gamma,\mathcal{M}}, \boldsymbol{\lambda}^*)$ satisfies

$$J_\gamma(\iota_{\text{init}}) \geq \langle \iota_{\text{init}}, \boldsymbol{V}^*_{\gamma,\mathcal{M}}\rangle - \langle \boldsymbol{\lambda}^*, \boldsymbol{\beta}\rangle. \tag{22}$$

Note that since ρ^γ is a coherent, Markov risk measure (Assumption 1), $v \mapsto \sigma(v,\cdot,\cdot)$ is convex (because σ is also a coherent risk measure). In fact, optimization problem (21) is indeed a DCP in the form of (17), where we should replace V_γ with $V_{\gamma,\mathcal{M}}$ and set $f_0(\boldsymbol{\lambda}) = \langle \boldsymbol{\lambda}, \boldsymbol{\beta}\rangle$, $g_0(\boldsymbol{V}_{\gamma,\mathcal{M}}) = \langle \iota_{\text{init}}, \boldsymbol{V}_{\gamma,\mathcal{M}}\rangle$, $f_1(V_{\gamma,\mathcal{M}}) = V_{\gamma,\mathcal{M}}$, $g_1(\boldsymbol{\lambda}) = \sum_{\alpha\in\text{Act}} p(\alpha \mid g)\tilde{c}([s,g],\alpha)$, and $g_2(V_{\gamma,\mathcal{M}}) = \gamma\sigma(V_{\gamma,\mathcal{M}},\cdot,\cdot)$.

The abovementioned optimization is in standard DCP form because f_0 and g_1 are convex (linear) functions of $\boldsymbol{\lambda}$ and g_0, f_1, and g_2 are convex functions in $V_{\gamma,\mathcal{M}}$.

Solving (17) gives a set of value functions $V_{\gamma,\mathcal{M}}$. In the following section, we discuss how to use the solutions from this DCP in our proposed policy iteration algorithm to sequentially improve the FSC parameters ω_T.

C. I-States Improvement

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

To begin with, we compute the initial I-state by finding the best valued I-state for a given initial belief, i.e., $\kappa(g_{\text{init}}) = 1$, where

$$g_{\text{init}} = \operatorname*{argmax}_{g\in G} \left\langle \iota_{\text{init}}, \vec{V}_{\gamma,\mathcal{M}}(g)\right\rangle.$$

After this initialization, we search for FSC parameters ω_T that result in an improvement.

I-state improvement optimization: Given value functions $V_{\gamma,\mathcal{M}}([s,g])$ for all $s\in\mathcal{S}$ and $g\in G$ and Lagrangian parameters $\boldsymbol{\lambda}$, for every I-state g, we can find FSC parameters ω_T that result in an improvement by solving the following optimization:

$$\max_{\epsilon>0,\omega_T(g,'\alpha|g,o)} \ \epsilon$$

subject to

Improvement Constraint:

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq \text{r.h.s. of (21)} \quad \forall s\in\mathcal{S}$$

Probability Constraints:

$$\sum_{(g,'\alpha)\in G\times\text{Act}} \omega_T(g,'\alpha \mid g,o) = 1 \quad \forall o\in\mathcal{O}$$

$$\omega_T(g,'\alpha \mid g,o) \geq 0 \quad \forall g'\in G, \alpha\in\text{Act}, o\in\mathcal{O}. \tag{23}$$

Note that the abovementioned optimization searches for ω_T values that improve the I-state value vector $\vec{V}^*_{\gamma,\mathcal{M}}(g)$ by maximizing the auxiliary decision variable ϵ.

Optimization problem (23) is in general nonconvex. This can be inferred from the fact that, although the first term in the r.h.s. of (21) is linear in ω_T, its convexity or concavity is not clear in the σ term for a general coherent risk measure. Fortunately, we can prove the following result, where in we show that for several examples of coherent risk measures (23) either becomes a linear program or a convex optimization problem.

Proposition 2: Let $\boldsymbol{V}_{\gamma,\mathcal{M}}$ and $\boldsymbol{\lambda}$ be given. Then, the I-state improvement optimization (23) is a linear program for conditional expectation and CVaR risk measures. Furthermore, (23) is a convex optimization for EVaR risk measure.

Proof: Please refer to the Appendix. ■

If no improvement is achieved by optimization (23), i.e., $\epsilon = 0$, for fixed number of internal states $|G|$, we can increase $|G|$ by one following the footsteps of the bounded policy iteration method proposed in [23, Sec. V.B].

D. Policy Iteration Algorithm

Algorithm 1 outlines the main steps in the proposed policy iteration method for the constrained risk-averse FSC synthesis. The algorithm has two distinct parts. First, for fixed parameters of the FSC (ω_T), policy evaluation is carried out, in which $V_{\gamma,\mathcal{M}}([s,g])$ and $\boldsymbol{\lambda}$ are computed using DCP (21) (Steps 2, 10, and 18). Second, after evaluating the current value functions and the Lagrange multipliers, an improvement is carried out either by changing the parameters of existing I-states via optimization (23), 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) based on the method proposed in [23, Sec. V.B].

VI. Numerical Experiments

In this section, we evaluate the proposed methodology with numerical experiments. In addition to the traditional total expectation, we consider two other coherent risk measures, namely,

Algorithm 1: Policy Iteration For Synthesizing Constrained Risk-Averse FSC.

Input: (a) An initial feasible FSC, $\mathcal{G}$. (b) Maximum size of FSC N_{max}. (c) $N_{new} \leq N_{max}$ number of I-states

$improved \leftarrow True$
Compute the value vectors, $\vec{V}_{\gamma,\mathcal{M}}$ and Lagrange multipliers $\boldsymbol{\lambda}$, based on DCP (21).
while $|G| \leq N_{max}$ **and** $improved = True$ **do**
  $improved \leftarrow False$
  for all I-states $g \in G$ **do**
    Solve the I-State Improvement Optimization (23).
    if I-State Improvement Optimization results in $\epsilon > 0$ **then**
      Replace the parameters ω_T for I-state g
      $improved \leftarrow True$
      Compute the value vectors, $\vec{V}_{\gamma,\mathcal{M}}$ and Lagrange multipliers $\boldsymbol{\lambda}$, based on optimization (21).
  if $improved = False$ **and** $|G| < N_{max}$ **then**
    $n_{added} \leftarrow 0$
    $N'_{new} \leftarrow \min(N_{new}, N_{max} - |G|)$
    Try to add N'_{new} I-state(s) to $\mathcal{G}$.
    $n_{added} \leftarrow$ actual number of I-states added in previous step.
    if $n_{added} > 0$ **then**
    $improved \leftarrow True$
      Compute the value vectors, $\vec{V}_{\gamma,\mathcal{M}}$ and Lagrange multipliers $\boldsymbol{\lambda}$, based on optimization (21).

Output: $\mathcal{G}$

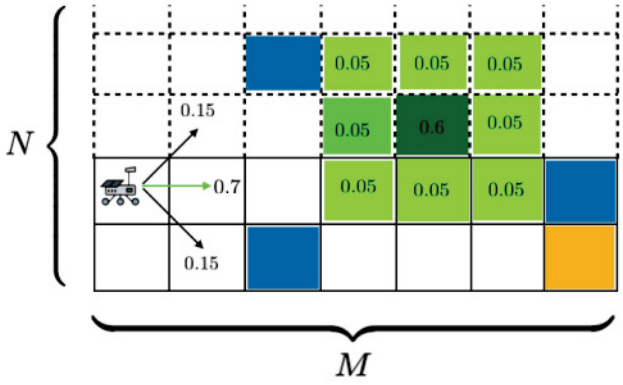


Fig. 1. Grid-world illustration for the rover navigation example. Blue cells denote the obstacles and the yellow cell denotes the goal.

CVaR and EVaR. All 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 CVXPY [39] with DCCP [38] add-on in Python.

A. Rover MDP Example Set up

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 $0.25MN$ obstacles as shown in Fig. 1. The state space is given by $\mathcal{S} = \{s_i | i = x + My, x \in \{1, \ldots, M\}, y \in \{1, \ldots, N\}\}$. The action set available to the robot is Act $= \{E, W, N, S, \text{NE}, \text{NW}, \text{SE}, \text{SW}\}$. The state transition probabilities for various cell types are shown for actions E in Fig. 2, i.e., the agent moves to the grid implied by the action with 0.7 probability but can also move to any adjacent ones with 0.3 probability. Partial observability arises because the rover cannot determine obstacle cell location from measurements directly. The observation space is $\mathcal{O} = \{o_i | i = x + My, x \in \{1, \ldots, M\}, y \in \{1, \ldots, N\}\}$. Once at an adjacent cell to an obstacle, the rover can identify an actual obstacle position (dark green) with probability 0.6, and a distribution over the nearby cells (light green).

Hitting an obstacle incurs the immediate cost of 10, while the goal grid region has zero immediate cost. Any other grid has a cost of 2 to represent fuel consumption. The discount factor is set to $\gamma = 0.95$.

The objective is to compute a safe path that is fuel efficient, i.e., solving Problem 1. To this end, we consider total expectation, CVaR, and EVaR as the coherent risk measures.

Once a policy is calculated, as a robustness test, inspired by [17], we included a set of single grid obstacles that are perturbed in a random direction to one of the neighboring grid cells with probability 0.3 to represent uncertainty in the terrain map. For each risk measure, we run 100 Monte Carlo simulations with the calculated policies and count failure rates, i.e., the number of times a collision has occurred during a run.

B. MDP Results

To evaluate the technique discussed in Section IV, we assume that there is no partial observation. In our experiments, we consider four grid-world sizes of 10×10, 15×15, 20×20, and 30×30 corresponding to 100, 225, 400, and 900 states, respectively. For each grid-world, we randomly allocate 25% of the grids to obstacles, including 3, 6, 9, and 12 uncertain (single-cell) obstacles for the 10×10, 15×15, 20×20, and 30×30 grids, respectively. In each case, we solve DCP (11) (linear program in the case of total expectation) with $|\mathcal{S}||\text{Act}| = MN \times 8 = 8MN$ constraints and $MN + 2$ variables (the risk value functions V_γs, Langrangian coefficient λ, and ζ for CVaR and EVaR). In these experiments, we set $\varepsilon = 0.2$ for CVaR and EVaR coherent risk measures to represent risk-averse policies. The fuel budget (constraint bound β) was set to 50, 10, 200, and 600 for the 10×10, 15×15, 20×20, and 30×30 grid-worlds, respectively. The initial condition was chosen as $\kappa_0(s_M) = M - 1$, i.e., the agent starts at the second left most grid at the bottom.

A summary of our numerical experiments is provided in Table I. Note the computed values of Problem 1 satisfy $\mathbb{E}(c) \leq \text{CVaR}_\varepsilon(c) \leq \text{EVaR}_\varepsilon(c)$, which is consistent with the fact that EVaR is a more conservative coherent risk measure than CVaR [41].

For total expectation coherent risk measure, the calculations took significantly less time, since they are the result of solving a set of linear programs. For CVaR and EVaR, a set of DCPs were solved. CVaR calculation was the most computationally involved. This observation is consistent with [42], where it was

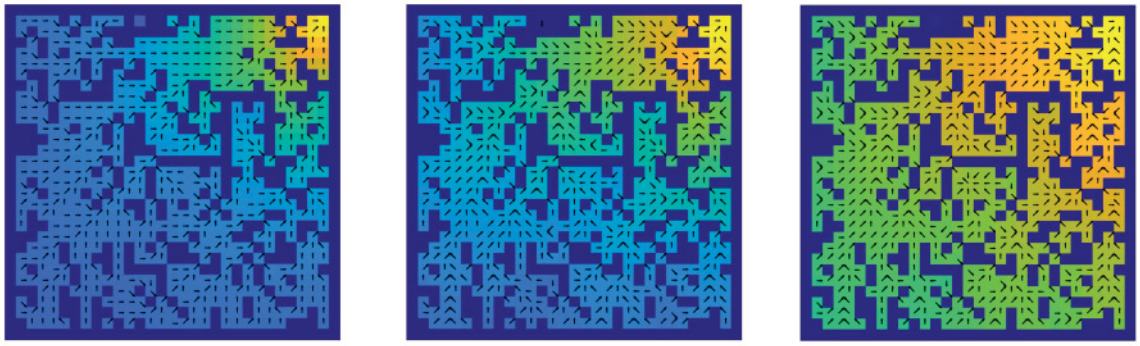


Fig. 2. Results for the MDP example with total expectation (left), CVaR (middle), and EVaR (right) coherent risk measures. The goal is located at the yellow cell. Notice the 9 single cell obstacles used for robustness test.

TABLE I
COMPARISON BETWEEN TOTAL EXPECTATION, CVAR, AND EVAR COHERENT RISK MEASURES

$(M \times N)_{\rho_t}$	$J_\gamma(\kappa_0)$	Total Time [s]	# U.O.	F.R.
$(10 \times 10)_{\mathrm{E}}$	9.12	0.8	3	11%
$(15 \times 15)_{\mathrm{E}}$	12.53	0.9	6	23%
$(20 \times 20)_{\mathrm{E}}$	19.93	1.7	9	33%
$(30 \times 30)_{\mathrm{E}}$	27.30	2.4	12	41%
$(10 \times 10)_{\mathrm{CVaR}_{0.7}}$	$\geqslant$12.04	5.8	3	8%
$(15 \times 15)_{\mathrm{CVaR}_{0.7}}$	$\geqslant$14.83	9.3	6	18%
$(20 \times 20)_{\mathrm{CVaR}_{0.7}}$	$\geqslant$20.19	10.34	9	19%
$(30 \times 30)_{\mathrm{CVaR}_{0.7}}$	$\geqslant$34.95	14.2	12	32%
$(10 \times 10)_{\mathrm{CVaR}_{0.2}}$	$\geqslant$14.45	6.2	3	3%
$(15 \times 15)_{\mathrm{CVaR}_{0.2}}$	$\geqslant$17.82	9.0	6	5%
$(20 \times 20)_{\mathrm{CVaR}_{0.2}}$	$\geqslant$25.63	11.1	9	13%
$(30 \times 30)_{\mathrm{CVaR}_{0.2}}$	$\geqslant$44.83	15.25	12	22%
$(10 \times 10)_{\mathrm{EVaR}_{0.7}}$	$\geqslant$14.53	4.8	3	4%
$(15 \times 15)_{\mathrm{EVaR}_{0.7}}$	$\geqslant$16.36	8.8	6	11%
$(20 \times 20)_{\mathrm{EVaR}_{0.7}}$	$\geqslant$29.89	10.5	9	15%
$(30 \times 30)_{\mathrm{EVaR}_{0.7}}$	$\geqslant$54.13	14.99	12	12%
$(10 \times 10)_{\mathrm{EVaR}_{0.2}}$	$\geqslant$18.03	5.8	3	1%
$(15 \times 15)_{\mathrm{EVaR}_{0.2}}$	$\geqslant$21.10	8.7	6	3%
$(20 \times 20)_{\mathrm{EVaR}_{0.2}}$	$\geqslant$24.08	10.2	9	7%
$(30 \times 30)_{\mathrm{EVaR}_{0.2}}$	$\geqslant$63.04	14.25	12	10%

$(M \times N)_{\rho t}$ denotes experiments with grid-world of size $M \times N$ and one-step coherent risk measure ρt. $J_\gamma(\kappa_0)$ is the valued of the constrained risk-averse problem (Problem 1). Total Time denotes the time taken by the CVXPY solver to solve the associated linear programs or DCPs in seconds. # 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 with the computed policy.

discussed that EVaR calculation is much more efficient than CVaR. 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 more conservative. In addition, these results imply that, although discounted total expectation is a measure of performance in high number of Monte Carlo simulations, it may not be practical to use it for mission-critical decision making under uncertainty scenarios. CVaR and especially EVaR seem to be a more reliable metric for performance in planning under uncertainty.

For the sake of illustrating the computed policies, Fig. 3 depicts the results obtained from solving DCP (11) for a 30×30 grid-world. The arrows on grids depict the (sub)optimal actions and the heat map indicates the values of Problem 1 for each grid state. Note that the values for EVaR are greater than those for CVaR and the values for CVaR are greater from those of total expectation. This is in accordance with the theory that $\mathbb{E}(c) \leq \mathrm{CVaR}_\varepsilon(c) \leq \mathrm{EVaR}_\varepsilon(c)$ [41]. In addition, by inspecting the computed actions in obstacle dense areas of the grid-world (for example, the middle right area), we infer that the actions in more risk-averse cases (especially, for EVaR) have a higher tendency to steer the agent away from the obstacles given the diagonal transition uncertainty as depicted in Fig. 2; whereas, for total expectation, the actions are merely concerned about reaching the goal.

C. POMDP Results

In our experiments, we consider two grid-world sizes of 10×10 and 20×20 corresponding to 100 and 400 states, respectively. For each grid-world, we allocate 25% of the grid to obstacles, including 8, and 16 uncertain (single-cell) obstacles for the 10×10 and 20×20 grids, respectively. In each case, we run Algorithm 1 for risk-averse FSC synthesis with $N_{\max} = 6$ and a maximum number of 100 iterations were considered.

In these experiments, we set the confidence level $\varepsilon = 0.15$ for CVaR and EVaR coherent risk measures. The fuel budget (constraint bound β) was set to 50 and 200 for the 10×10 and 20×20 grid-worlds, respectively. The initial condition was chosen as $\kappa_0(s_M) = 1$, i.e., the agent starts at the right most grid at the bottom.

A summary of our numerical experiments is provided in Table II. Note the computed values of Problem 1 satisfy $\mathbb{E}(c) \leq \mathrm{CVaR}_\varepsilon(c) \leq \mathrm{EVaR}_\varepsilon(c)$ [41].

For total expectation coherent risk measure, the calculations took significantly less time, since they are the result of solving a set of linear programs. For CVaR and EVaR, a set of DCPs were solved in the risk value function computation step. In

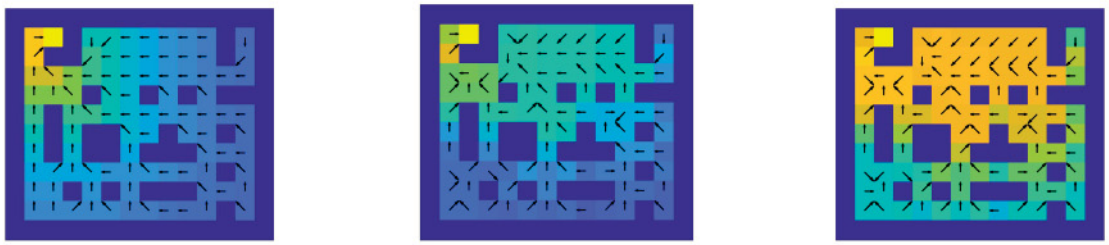


Fig. 3. Results for the POMDP example with total expectation (left), CVaR (middle), and EVaR (right) coherent risk measures. The goal is located at the yellow cell. Notice the nine single cell obstacles used for robustness test.

TABLE II
COMPARISON BETWEEN TOTAL EXPECTATION, CVAR, AND EVAR COHERENT RISK MEASURES

$(M \times N)_{\rho_t}$	$J_\gamma(\iota_{\text{init}})$	AIT [s]	# U.O.	F.R.
$(10 \times 10)_{\mathbb{E}}$	10.53	0.2	3	15%
$(20 \times 20)_{\mathbb{E}}$	19.98	0.3	9	37%
$(10 \times 10)_{\text{CVaR}_{0.7}}$	$\geqslant$11.02	2.9	3	9%
$(20 \times 20)_{\text{CVaR}_{0.7}}$	$\geqslant$20.19	7.5	9	22%
$(10 \times 10)_{\text{CVaR}_{0.2}}$	$\geqslant$16.53	3.1	3	4%
$(20 \times 20)_{\text{CVaR}_{0.2}}$	$\geqslant$24.92	7.6	9	16%
$(10 \times 10)_{\text{EVaR}_{0.7}}$	$\geqslant$15.02	3.3	3	5%
$(20 \times 20)_{\text{EVaR}_{0.7}}$	$\geqslant$23.42	9.9	9	11%
$(10 \times 10)_{\text{EVaR}_{0.2}}$	$\geqslant$19.62	3.9	3	2%
$(20 \times 20)_{\text{EVaR}_{0.2}}$	$\geqslant$29.36	9.7	9	6%

$(M \times N)_{\rho_t}$ denotes experiments with grid-world of size $M \times N$ and one-step coherent risk measure ρ_t. $J_\gamma(\iota_{\text{init}})$ is the valued of the constrained risk-averse POMDP problem (Problem 1). AIT denotes the average time spent for each iteration of Algorithm 1. # 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 with the computed policy.

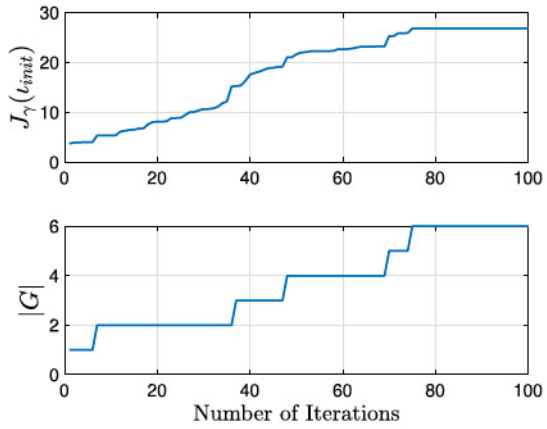


Fig. 4. Evolution of the lower bound and the number of i-states with respect to the number of iterations of Algorithm 1 for the 20 $\times$ 20 grid-world and EVaR coherent risk measure.

the I-state improvement step, a set of linear programs were solved for CVaR and convex optimizations for EVaR. Hence, EVaR calculation was the most computationally involved in this case.

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, tallying with the fact that EVaR is conservative. In addition, these results suggest that, although discounted total expectation is 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 especially EVaR seem to be a more reliable metric for performance in planning under uncertainty.

For the sake of illustrating the computed policies, Fig. 3 depicts the results obtained from solving (21) for a 20 $\times$ 20 grid-world. The arrows on grids depict the (sub)optimal actions and the heat map indicates the values of (21) for each grid state. Note that the values for EVaR are greater than those for CVaR and the values for CVaR are greater from those of total expectation. This is in accordance with the theory that $\mathbb{E}(c) \leq \text{CVaR}_\varepsilon(c) \leq \text{EVaR}_\varepsilon(c)$ [41].

Moreover, for the 20 $\times$ 20 grid-world with EVaR coherent risk measure, Fig. 4 depicts the evolution of the number of FSC I-states $|G|$ and the lower bound on the optimal value of Problem 1, $J_\gamma(\iota_{\text{init}})$, with respect to the iteration number of Algorithm 1. We can see that as the number of I-states increase, the lower bound is improved.

VII. CONCLUSION

We proposed an optimization-based method for designing policies for MDPs and POMDPs with coherent risk measure objectives and constraints. We showed that such value function optimizations are in the form of DCPs. In the case of POMDPs, we proposed a policy iteration method for finding suboptimal FSCs that lower bound the constrained risk-averse problem and we demonstrated that dependent on the coherent risk measure of interest the policy search can be carried out via a linear program or a convex optimization. Numerical experiments were provided to show the efficacy of our approach. In particular, we showed that considering coherent risk measures lead to significantly lower collision rates in Monte Carlo simulations in navigation problems.

In this work, we focused on discounted infinite horizon risk-averse problems. Future work will explore other cost criteria [43]. The interested reader is referred to our preliminary results on total cost risk-averse MDPs [44], where in Bellman's equations for the risk-averse stochastic shortest path problem are derived. Expanding on the latter work, we will also explore high-level mission specifications in terms of temporal logic formulas for risk-averse MDPs and POMDPs [45], [46]. Another area for more research is concerned with receding-horizon motion planning under uncertainty with coherent risk constraints [47],

[48], with particular application in robot exploration in unstructured subterranean environments [49] (also see works on receding horizon path planning where the coherent risk measure is in the total cost [50], [51] rather than the collision avoidance constraint).

APPENDIX

A. Examples of Coherent Risk Measures

In this Appendix, we briefly review three examples of coherent risk measures that will be used in this article.

Total conditional expectation: The simplest risk measure is the total conditional expectation given by

$$\rho_t(c_{t+1}) = \mathbb{E}\left[c_{t+1} \mid \mathcal{F}_t\right]. \tag{24}$$

It is easy to see that total conditional expectation satisfies the properties of a coherent risk measure as outlined in Definition 3. Unfortunately, total conditional expectation is agnostic to realization fluctuations of the random variable c and is only concerned with the mean value of c at large number of realizations. Thus, it is a risk-neutral measure of performance.

CVaR: Let $c \in \mathcal{C}$ be a random variable. For a given confidence level $\varepsilon \in (0,1)$, value-at-risk (VaR_ε) denotes the $(1-\varepsilon)$-quantile value of the random variable $c \in \mathcal{C}$. Unfortunately, working with VaR for non-normal random variables is numerically unstable and optimizing models involving VaR is intractable in high dimensions [52].

In contrast, CVaR overcomes the shortcomings of VaR. CVaR with confidence level $\varepsilon \in (0,1)$ denoted CVaR_ε measures the expected loss in the $(1-\varepsilon)$-tail given that the particular threshold VaR_ε has been crossed, i.e., $\text{CVaR}_\varepsilon(c) = \mathbb{E}[c \mid c \geq \text{VaR}_\varepsilon(c)]$. An optimization formulation for CVaR was proposed in [52]. That is, CVaR_ε is given by

$$\begin{aligned}\rho_t(c_{t+1}) &= \text{CVaR}_\varepsilon(c_{t+1}) \\ &:= \inf_{\zeta \in \mathbb{R}} \left(\zeta + \frac{1}{\varepsilon}\mathbb{E}\left[(c_{t+1}-\zeta)_+ \mid \mathcal{F}_t\right]\right)\end{aligned} \tag{25}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$. A value of $\varepsilon \to 1$ corresponds to a risk-neutral case, i.e., $\text{CVaR}_1(c) = \mathbb{E}(c)$; whereas, a value of $\varepsilon \to 0$ is rather a risk-averse case, i.e., $\text{CVaR}_0(c) = \text{VaR}_0(c) = \text{ess inf}(c)$ [53]. Fig. 5 illustrates these notions for an example c variable with distribution $p(c)$.

EVaR: Unfortunately, CVaR ignores the losses below the VaR threshold. EVaR is the tightest upper bound in the sense of Chernoff inequality for VaR and CVaR and its dual representation is associated with the relative entropy. In fact, it was shown in [54] that EVaR_ε and CVaR_ε are equal only if there are no losses ($c \to -\infty$) below the VaR_ε threshold. In addition, EVaR is a strictly monotone risk measure; whereas, CVaR is only monotone [42]. EVaR_ε is given by

$$\rho_t(c_{t+1}) = \inf_{\zeta > 0}\left(\log\left(\frac{\mathbb{E}[e^{\zeta c_{t+1}} \mid \mathcal{F}_t]}{\varepsilon}\right)/\zeta\right). \tag{26}$$

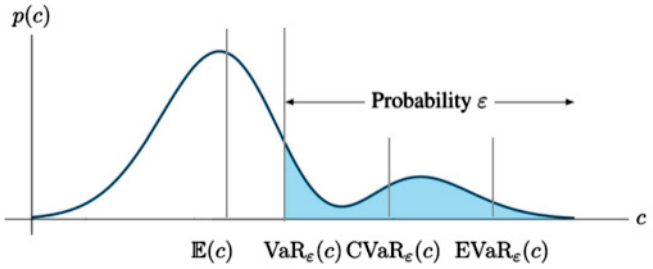


Fig. 5. Comparison of the mean, VaR, and CVaR for a given confidence $\varepsilon \in (0,1)$. The axes denote the values of the stochastic variable c and its probability density function $p(c)$. The shaded area denotes the $\%\varepsilon$ of the area under $p(c)$. The expected cost $\mathbb{E}(c)$ is much smaller than the worst case cost. VaR gives the value of c at the $(1-\varepsilon)$-tail of the distribution. But, it ignores the values of c with probability below ε. CVaR is the average of the values of VaR with probability less than ε (average of the worst-case values of c in the $(1-\varepsilon)$ tail of the distribution).

Similar to CVaR_ε, for EVaR_ε, $\varepsilon \to 1$ corresponds to a risk-neutral case; whereas, $\varepsilon \to 0$ corresponds to a risk-averse case. In fact, it was demonstrated in [41, Proposition 3.2] that $\lim_{\varepsilon \to 0} \text{EVaR}_\varepsilon(c) = \text{ess inf}(c)$.

B. Proof of Proposition 2

We present different forms of the improvement constraint in (23) for different risk measures. Note that the rest of the constraints and the cost function are linear in the decision variables ϵ and ω_T. The improvement constraint in (23) is linear in ϵ. However, its convexity or concavity in ω_T changes depending on the risk measure one considers. We recall from the previous section that in the policy evaluation step, the quantities for $V_{\gamma,\mathcal{M}}$ and $\boldsymbol{\lambda} \succeq 0$ (for conditional expectation, CVaR, and EVaR measures) and ζ for (CVaR and EVaR measures) are calculated and, therefore, fixed here.

For conditional expectation, the improvement constraint alters to

$$\begin{aligned}V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq &\sum_{\alpha \in \text{Act}} p(\alpha \mid g)\tilde{c}([s,g],\alpha) \\ &+\gamma \sum_{s' \in \mathcal{S}, g' \in \mathcal{G}} V_{\gamma,\mathcal{M}}([s,'g'])T^{\mathcal{M}}([s,'g'] \,|[s,g]) \\ &\forall s \in \mathcal{S} \quad \forall g \in G.\end{aligned} \tag{27}$$

Substituting the expression for $T^{\mathcal{M}}$ (as defined in Section V-A), i.e.,

$$T^{\mathcal{M}}([s,'g'] \,|[s,g]) = \sum_{o \in \mathcal{O}} \sum_{\alpha \in \text{Act}} O(o|s)\omega_T(g,'\alpha|g,o)T(s'|s,\alpha)$$

and $p(\alpha \mid g)$, i.e.,

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

we obtain

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o)O(o|g')\tilde{c}([s,g],\alpha)$$

$$+\gamma \sum_{s,'g,'o,\alpha} V_{\gamma,\mathcal{M}}([s,'g'])O(o|s)\omega_T(g,'\alpha|g,o)T(s'|s,\alpha)$$

$$\forall s \in \mathcal{S} \quad \forall g \in G. \tag{28}$$

The abovementioned expression is linear in ω_T as well as ϵ. Hence, I-state improvement optimization becomes a linear program for conditional expectation risk measure.

Based on a similar construction, for CVaR measure, the improvement constraint changes to

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o)O(o|g')\tilde{c}([s,g],\alpha)$$

$$+\gamma\left\{\zeta + \frac{1}{\varepsilon}\sum_{g,'s'} \left(V_{\gamma,\mathcal{M}}\left([s,'g']\right) - \zeta\right)_+ T^{\mathcal{M}}([s,'g']|[s,g])\right\}$$

$$\forall s \in \mathcal{S} \quad \forall g \in G. \tag{29}$$

After substituting the term for $T^{\mathcal{M}}$, we obtain

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o)O(o|g')\tilde{c}([s,g],\alpha)$$

$$+\gamma\left\{\zeta + \frac{1}{\varepsilon}\sum_{g,'s,'o,\alpha} \left(V_{\gamma,\mathcal{M}}\left([s,'g']\right) - \zeta\right)_+ O(o|s)\right.$$

$$\left.\times\omega_T(g,'\alpha|g,o)T(s'|s,\alpha)\right\} \quad \forall s \in \mathcal{S} \quad \forall g \in G \tag{30}$$

Furthermore, for fixed $V_{\gamma,\mathcal{M}}$, λ, and ζ, the abovementioned inequality is linear in ω_T and ϵ. Hence, (30) becomes a linear constraint rendering (23) a linear program (maximizing a linear objective subject to linear constraints), i.e., optimization problem (31) shown at the bottom of this page.

For the EVaR measure, the improvement constraint is given by

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o)O(o|g')\tilde{c}([s,g],\alpha)$$

$$+\gamma\left\{\frac{1}{\zeta}\log\left(\frac{\sum_{g,'s'} e^{\zeta V_{\gamma,\mathcal{M}}([s,'g'])}T^{\mathcal{M}}([s,'g']|[s,g])}{\varepsilon}\right)\right\}$$

$$\forall s \in \mathcal{S} \quad \forall g \in G. \tag{32}$$

Substituting the expression for $T^{\mathcal{M}}$, i.e.,

$$T^{\mathcal{M}}\left([s,'g'] \,|[s,g]\right) = \sum_{o\in\mathcal{O}}\sum_{\alpha\in\text{Act}} O(o|s)\omega_T(g,'\alpha|g,o)T(s'|s,\alpha)$$

we obtain

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon \leq \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o)O(o|g')\tilde{c}([s,g],\alpha)$$

$$+\frac{\gamma}{\zeta}\log\left(\frac{\sum_{g,'s,'o,\alpha} e^{\zeta V_{\gamma,\mathcal{M}}([s,'g'])}O(o|s)\omega_T(g,'\alpha|g,o)T(s'|s,\alpha)}{\varepsilon}\right)$$

$$\forall s \in \mathcal{S} \quad \forall g \in G \tag{33}$$

In the abovementioned inequality, the first term on the right-hand side of the is linear in ω_T and the second term on the right-hand side (logarithm term) is concave in ω_T (convex if all terms are moved to the left side, since $-\log(x)$ is convex in x). Therefore, (33) becomes a convex constraint rendering (23), a convex optimization problem (maximizing a linear objective subject to linear and convex constraints) for EVaR measures. That is, the I-state improvement optimization takes the convex optimization form of (34) shown at the top of the next page.

$$\max_{\epsilon>0,\omega_T(g,'\alpha|g,o)} \quad \langle \iota_{\text{init}}, V_{\gamma,\mathcal{M}}\rangle - \langle \lambda, \beta\rangle + \epsilon$$

subject to

Improvement Constraint:

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon - \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o)O(o|g')\tilde{c}([s,g],\alpha) \tag{31a}$$

$$-\gamma\left\{\zeta + \frac{1}{\varepsilon}\sum_{g,'s,'o,\alpha} \left(V_{\gamma,\mathcal{M}}\left([s,'g']\right) - \zeta\right)_+ O(o|s)\omega_T(g,'\alpha|g,o)T(s'|s,\alpha)\right\} \leq 0 \quad \forall s \in \mathcal{S} \quad \forall g \in G$$

Probability Constraints:

$$\sum_{(g,'\alpha)\in G\times\text{Act}} \omega_T(g,'\alpha \mid g,o) = 1 \quad \forall o \in \mathcal{O}$$

$$\omega_T(g,'\alpha \mid g,o) \geq 0 \quad \forall g' \in G, \alpha \in \text{Act}, o \in \mathcal{O}. \tag{31b}$$

$$\max_{\epsilon>0,\omega_T(g,'\alpha|g,o)} \quad \langle \iota_{\text{init}}, V_{\gamma,\mathcal{M}} \rangle - \langle \lambda, \beta \rangle + \epsilon$$

subject to

Improvement Constraint:

$$V_{\gamma,\mathcal{M}}([s,g]) + \epsilon - \sum_{\alpha,g,'o} \omega_T(g,'\alpha \mid g,o) O(o|g') \tilde{c}([s,g],\alpha) \tag{34a}$$

$$-\gamma \left\{ \frac{1}{\zeta} \log \left(\frac{\sum_{g,'s,'o,\alpha} e^{\zeta V_{\gamma,\mathcal{M}}([s,'g'])} O(o|s) \omega_T(g,'\alpha|g,o) T(s'|s,\alpha)}{\varepsilon} \right) \right\} \leq 0 \quad \forall s \in \mathcal{S} \quad \forall g \in G$$

Probability Constraints:

$$\sum_{(g,'\alpha)\in G \times \text{Act}} \omega_T(g,'\alpha \mid g,o) = 1 \quad \forall o \in \mathcal{O}$$

$$\omega_T(g,'\alpha \mid g,o) \geq 0 \quad \forall g' \in G, \alpha \in \text{Act}, o \in \mathcal{O}. \tag{34b}$$

ACKNOWLEDGMENT

Mohamadreza Ahmadi would like to thank the stimulating discussions with Dr. M. Ono at NASA Jet Propulsion Laboratory and Prof. M. Pavone at Nvidia Research-Stanford University.

REFERENCES

[1] A. Wang, A. M. Jasour, and B. Williams, "Non-Gaussian chance-constrained trajectory planning for autonomous vehicles under agent uncertainty," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6041–6048, Oct. 2020.

[2] H. Xu and S. Mannor, "Distributionally robust Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, vol. 23, pp. 2505–2513.

[3] A. Majumdar and M. Pavone, "How should a robot assess risk? Towards an axiomatic theory of risk in robotics," in *Proc. 18th Int. Symp. Robot. Res.*, 2020, pp. 75–84.

[4] M. P. Chapman, R. Bonalli, K. M. Smith, I. Yang, M. Pavone, and C. J. Tomlin, "Risk-sensitive safety analysis using conditional value-at-risk," *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6521–6536, Dec. 2022.

[5] A. Singletary, M. Ahmadi, and A. D. Ames, "Safe control for nonlinear systems with stochastic uncertainty via risk control barrier functions," *IEEE Control Syst. Lett.*, vol. 7, pp. 349–354, 2022.

[6] M. Ahmadi, X. Xiong, and A. D. Ames, "Risk-averse control via CVaR barrier functions: Application to bipedal robot locomotion," *IEEE Control Syst. Lett.*, vol. 6, pp. 878–883, 2021.

[7] P. Artzner, F. Delbaen, J. Eber, and D. Heath, "Coherent measures of risk," *Math. Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[8] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. New York, NY, USA: Wiley, 1994.

[9] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1468–1476.

[10] H. Huo and X. Guo, "Risk probability minimization problems for continuous-time Markov decision processes on finite horizon," *IEEE Trans. Autom. Control*, vol. 65, no. 7, pp. 3199–3206, Jul. 2020.

[11] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Sequential decision making with coherent risk," *IEEE Trans. Autom. Control*, vol. 62, no. 7, pp. 3323–3338, Jul. 2017.

[12] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA, USA: SIAM, 2014.

[13] W. Huang and W. B. Haskell, "Stochastic approximation for risk-aware Markov decision processes," *IEEE Trans. Autom. Control*, vol. 66, no. 3, pp. 1314–1320, Mar. 2021.

[14] L. Prashanth, "Policy gradients for CVaR-constrained MDPs," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2014, pp. 155–169.

[15] Y. Chow and M. Ghavamzadeh, "Algorithms for CVaR optimization in MDPs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3509–3517.

[16] G. C. Pflug and A. Pichler, "Time-consistent decisions and temporal decomposition of coherent risk functionals," *Math. Operations Res.*, vol. 41, no. 2, pp. 682–699, 2016.

[17] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: A CVaR optimization approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1522–1530.

[18] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes," *Math. Program.*, vol. 125, no. 2, pp. 235–261, 2010.

[19] V. Borkar and R. Jain, "Risk-constrained Markov decision processes," *IEEE Trans. Autom. Control*, vol. 59, no. 9, pp. 2574–2579, Sep. 2014.

[20] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[21] J. Fan and A. Ruszczyński, "Process-based risk measures and risk-averse control of discrete-time systems," *Math. Program.*, vol. 191, pp. 113–140, 2018.

[22] J. Fan and A. Ruszczyński, "Risk measurement and risk-averse control of partially observable discrete-time Markov systems," *Math. Methods Operations Res.*, vol. 88, no. 2, pp. 161–184, 2018.

[23] M. Ahmadi, M. Ono, M. D. Ingham, R. M. Murray, and A. D. Ames, "Risk-averse planning under uncertainty," in *Proc. IEEE Amer. Control Conf.*, 2020, pp. 3305–3312.

[24] M. Ahmadi, U. Rosolia, M. Ingham, R. Murray, and A. Ames, "Constrained risk-averse Markov decision processes," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 11718–11725.

[25] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artif. Intell.*, vol. 101, no. 1/2, pp. 99–134, 1998.

[26] O. Madani, S. Hanks, and A. Condon, "On the undecidability of probabilistic planning and related stochastic optimization problems," *Artif. Intell.*, vol. 147, no. 1, pp. 5–34, 2003.

[27] T. Osogami, "Robustness and risk-sensitivity in Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 233–241.

[28] E. Altman, *Constrained Markov Decision Processes*, vol. 7. Boca Raton, FL, USA: CRC Press, 1999.

[29] J. T. Ott, "A Markov decision model for a surveillance application and risk-sensitive Markov decision processes," Ph.D. dissertation, Karlsruhe Inst. Technol., 2010.

[30] N. Bäuerle and J. Ott, "Markov decision processes with average-value-at-risk criteria," *Math. Methods Operations Res.*, vol. 74, no. 3, pp. 361–379, 2011.

[31] D. Du and P. M. Pardalos, *Minimax and Applications*, vol. 4. Berlin, Germany: Springer, 2013.

[32] R. Horst and N. V. Thoai, "DC programming: Overview," *J. Optim. Theory Appl.*, vol. 103, no. 1, pp. 1–43, 1999.

[33] H. A. Le Thi et al., "A DC programming approach for feature selection in support vector machines learning," *Adv. Data Anal. Classification*, vol. 2, no. 3, pp. 259–278, 2008.

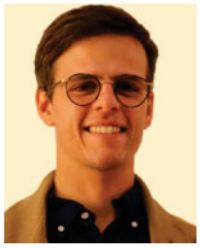
[34] J. Thai, T. Hunter, A. K. Akametalu, C. J. Tomlin, and A. M. Bayen, "Inverse covariance estimation from data with missing values using the concave-convex procedure," in *Proc. IEEE 53rd Conf. Decis. Control*, 2014, pp. 5736–5742.

[35] E. L. Lawler and D. E. Wood, "Branch-and-bound methods: A survey," *Operations Res.*, vol. 14, no. 4, pp. 699–719, 1966.
[36] D. Bertsekas, *Nonlinear Programming*, Belmont, MA, USA: Athena Scientific, 1999.
[37] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optim. Eng.*, vol. 17, no. 2, pp. 263–287, 2016.
[38] X. Shen, S. Diamond, Y. Gu, and S. Boyd, "Disciplined convex-concave programming," in *Proc. IEEE 55th Conf. Decis. Control*, 2016, pp. 1009–1014.
[39] S. Diamond and S. Boyd, "CVXPY: A python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.
[40] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*. New York, NY, USA: Academic, 1976.
[41] A. Ahmadi-Javid, "Entropic value-at-risk: A new coherent risk measure," *J. Optim. Theory Appl.*, vol. 155, no. 3, pp. 1105–1123, 2012.
[42] A. Ahmadi-Javid and M. Fallah-Tafti, "Portfolio optimization with entropic value-at-risk," *Eur. J. Oper. Res.*, vol. 279, no. 1, pp. 225–241, 2019.
[43] S. Carpin, Y. Chow, and M. Pavone, "Risk aversion in finite Markov decision processes using total cost criteria and average value at risk," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 335–342.
[44] M. Ahmadi, A. Dixit, J. W. Burdick, and A. D. Ames, "Risk-averse stochastic shortest path planning," in *Proc. IEEE 60th Conf. Decis. Control (CDC)*, Dec. 2021, pp. 5119–5204.
[45] M. Ahmadi, R. Sharan, and J. W. Burdick, "Stochastic finite state control of POMDPs with LTL specifications," 2020, *arXiv:2001.07679*.
[46] U. Rosolia, M. Ahmadi, R. M. Murray, and A. D. Ames, "Time-optimal navigation in uncertain environments with high-level specifications," in *Proc. IEEE 60th Conf. Decis. Control*, 2021, pp. 4287–4294.
[47] A. Dixit, M. Ahmadi, and J. W. Burdick, "Risk-sensitive motion planning using entropic value-at-risk," in *Proc. 20th Eur. Control Conf.*, 2021, pp. 1726–1732.
[48] A. Hakobyan, G. C. Kim, and I. Yang, "Risk-aware motion planning and control using CVaR-constrained optimization," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3924–3931, Oct. 2019.
[49] A. Dixit, D. D. Fan, K. Otsu, S. Dey, A.-A. Agha-Mohammadi, and J. W. Burdick, "STEP: Stochastic traversability evaluation and planning for risk-aware off-road navigation; results from the DARPA subterranean challenge," 2023, *arXiv:2303.01614*.
[50] P. Sopasakis, D. Herceg, A. Bemporad, and P. Patrinos, "Risk-averse model predictive control," *Automatica*, vol. 100, pp. 281–288, 2019.
[51] S. Singh, Y. Chow, A. Majumdar, and M. Pavone, "A framework for time-consistent, risk-sensitive model predictive control: Theory and algorithms," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2905–2912, Jul. 2019.
[52] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *J. Risk*, vol. 2, pp. 21–42, 2000.
[53] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *J. Bank. Finance*, vol. 26, no. 7, pp. 1443–1471, 2002.
[54] A. Ahmadi-Javid and A. Pichler, "An analytical study of norms and banach spaces induced by the entropic value-at-risk," *Math. Financial Econ.*, vol. 11, no. 4, pp. 527–550, 2017.

Mohamadreza Ahmadi received the D.Phil. (Ph.D.) degree in engineering science, control systems and aeronautics as a Clarendon Scholar from the University of Oxford, Oxford, U.K., in November, 2016.

He is currently the Founder and CEO with Caspian Autonomy LLC, Pasadena, CA, USA. He was a Postdoctoral Scholar with the Center for Autonomous Systems and Technologies (CAST), California Institute of Technology, Pasadena, CA, USA. His Ph.D. degree was followed by research positions with the University of Texas at Austin, Austin, TX, USA. His current research is on planning and control under uncertainty with application to autonomous systems.

Dr. Ahmadi is the recipient of the Sloan-Robinson Engineering Fellowship, an Edgell-Sheppee Award, and an ICES Postdoctoral Fellowship.



Ugo Rosolia received the B.S. and M.S. (*cum laude*) degrees from the Politecnico di Milano, Milan, Italy, in 2012 and 2014, respectively, and the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 2019, all in mechanical engineering.

He is currently a Postdoctoral Fellow with California Institute of Technology, Pasadena, CA, USA. He was a Visiting Scholar with Tongji University, Shanghai, China, for the Double Degree Program PoliTong from 2010 to 2011. During his M.S. degree, he was a Visiting Student for two semesters with the University of Illinois at Urbana - Champaign, Urbana, IL, USA, sponsored by a Global E3 Scholarship. In 2015, he was a Research Engineer with Siemens PLM Software, Leuven, Belgium, where he was involved in the optimal control algorithms. His current research interests include nonlinear optimal control for the centralized and decentralized system, the iterative learning control, and the predictive control.

Michel D. Ingham received the bachelor's degree in honours mechanical engineering from McGill University, Montreal, QC, Canada, in 1995, and the master's and doctoral degrees in aerospace, aeronautical and astronautical engineering from MIT's Department of Aeronautics and Astronautics, Cambridge, MA, USA, in 1998 and 2003, respectively.

He is currently the Chief Technologist of JPL's Systems Engineering Division, responsible for spearheading and coordinating research and technology efforts across the division, and integrating technology work more broadly across JPL. Prior to this role, he was the Project Software Systems Engineer for the Europa Clipper Mission, NASA's flagship mission to characterize the icy surface and underlying ocean on Jupiter's moon Europa. Since he joined JPL in 2003, he was a Software Systems Engineer and System Architect on a variety of projects, including the MoonRise robotic lunar sample return mission, the Mars Science Laboratory rover mission, and the Altair lunar lander. He has also led several NASA, JPL, and DARPA research and development activities, in the areas of model-based systems and software engineering, software architectures, and spacecraft autonomy.

Richard M. Murray (Fellow, IEEE) received the B.S. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1985 and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 1988 and 1991, respectively.

He is currently the Thomas E. and Doris Everhart Professor of Control and Dynamical Systems and Bioengineering with Caltech. His research is in the application of feedback and control to networked systems, with applications in biology and autonomy. His current projects include analysis and design of biomolecular feedback circuits, synthesis of discrete decision-making protocols for reactive systems, and design of highly resilient architectures for autonomous systems

Aaron D. Ames (Fellow, IEEE) received the B.S. degree in mechanical engineering and the B.A. degree in mathematics from the University of St. Thomas, Saint Paul, MN, USA, in 2001, and the M.A. degree in mathematics in 2006 and the Ph.D. degree in electrical engineering and computer sciences in 2006 both from the University of California at Berkeley, Berkeley, CA, USA.

He is currently the Bren Professor of Mechanical and Civil Engineering and Control and Dynamical Systems with the California Institute of Technology, Pasadena, CA, USA. His research interests span the areas of robotics, nonlinear, safety-critical control, and hybrid systems, with a special focus on applications to bipedal robotic walking both formally and through experimental validation.