Risk Aversion in Finite Markov Decision Processes Using Total Cost Criteria and Average Value at Risk

Stefano Carpin   Yin-Lam Chow   Marco Pavone

Abstract— **In this paper we present an algorithm to compute risk averse policies in Markov Decision Processes (MDP) when the total cost criterion is used together with the average value at risk (AVaR) metric. Risk averse policies are needed when large deviations from the expected behavior may have detrimental effects, and conventional MDP algorithms usually ignore this aspect. We provide conditions for the structure of the underlying MDP ensuring that approximations for the exact problem can be derived and solved efficiently. Our findings are novel inasmuch as average value at risk had never been considered in association with the total cost criterion. Our method is demonstrated in a rapid deployment scenario, whereby a robot is tasked with the objective of reaching a target location within a temporal deadline where increased speed is associated with increased probability of failure. We demonstrate that the proposed algorithm not only produces a risk averse policy reducing the probability of exceeding the expected temporal deadline, but also carefully characterizes the statistical distribution of costs, thus offering a valuable analysis and design tool.**

I. Introduction

Markov Decision Processes (MDPs) are extensively used to solve sequential stochastic decision making problems in robotics [22] and other disciplines [9]. A solution to an MDP problem instance provides a policy mapping states into actions with the property of optimizing (e.g., minimizing) in expectation a given objective function. In many practical situations a formulation based on expectation only is however not sufficient. This is specifically important when variability from the expected value detrimentally impacts the performance of the system. For example, in an autonomous navigation system, a robot attempting to minimize the expected length of the traveled path will likely travel close to obstacles, and a large deviation from the planned path may result in a collision that incurs huge loss (i.e. damage an expensive robot or fail the whole mission altogether.) Metrics that study deviations from the expected value are often referred to as *risks*. Basic solutions to the MDP problem like value iteration or policy iteration aim exclusively at optimizing the expectation without embedding computational expedients to control risk. Their policies are therefore labeled as *risk neutral*. The problem of quantifying risk associated with random variables has a rich history and is often related to the problem of managing financial assets [2]. In fact, many risk-related studies motivated by financial problems have recently found applications in domains such as robotics [19]. The term *risk aversion* refers to the preference given to stochastic realizations with limited deviation from the expected value. In risk averse optimal control one may prefer a policy with higher cost in expectation but lower deviations to one with lower cost but possibly higher deviations. Particularly in the context of robotic planning, introducing risk aversion in MDPs is crucial to guarantee mission safety. However introducing risk aversion in MDPs creates a number of additional theoretical and computational hurdles. For example, in risk averse MDPs optimal policies are no longer Markov stationary and one needs to account for history dependency in order to reach optimality.

Average Value at Risk (AVaR – also known as Conditional Value at Risk or CVaR) is a risk metric that has gained notable popularity in the area of risk averse control [2], [17]. For a given random value and a predetermined confidence level, the AVaR is the *tail average* of the distribution exceeding a given confidence level (see Section III for a formal definition). Risk averse policies considering the AVaR metric have been studied for the case of MDPs with finite horizon and discounted infinite horizon cost models. In this paper we instead consider how the AVaR metric can be applied when the total cost criterion is used in MDPs. This focus is motivated by numerous robotic applications where standard cost criteria like infinite horizon discounted cost or finite horizon cost are not an appropriate modeling option. On the contrary, we postulate that the most appropriate criterion is *total cost*, i.e., the undiscounted cost accrued during missions with random stopping times. The contribution of this paper is three-fold:

- We identify conditions for the underlying MDP ensuring that the AVaR problem is well defined when the total cost criterion is used.
- We define a surrogate MDP problem that can be efficiently solved, whose solutions approximate the optimal policy of the original problem with arbitrary precision.
- We validate our findings on a rapid robotic deployment task where the objective is to maximize the mission successful rate under a given temporal deadline [6], [8].

The rest of the paper is organized as follows. Related work is discussed in Section II, whereas the mathematical background is given in Section III. In Section IV we study the well-posedness of the problem, and formulate the risk-averse, total cost MDP problem. Our approximated solution is proposed and analyzed in Section V, and in Section VI we provide the algorithmic description. Simulations on a rapid robotic deployment platform are given in VII, whereas

Y.-L. Chow and M. Pavone are with Stanford University, Stanford, CA, USA. S. Carpin is with the School of Engineering, University of California-Merced, CA, USA.

S. Carpin is partially supported by the Army Research Lab under contract MAST-CNC-15-4-4. Y-L. Chow is partially supported by The Croucher Foundation doctoral scholarship. M. Pavone is partially supported by the Office of Naval Research, Science of Autonomy Program, under Contract N00014-15-1-2673. Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author and should not be interpreted as representing the official policies, either expressly or implied, of the funding agencies of the U.S. Government.

conclusions and future work are discussed in Section VIII.

II. Related Work

For a general introduction to MDPs the reader is referred to textbooks like [4] or more recent collections like [9]. As pointed out in the introduction, risk aversion in MDPs have been studied for over four decades, with earlier efforts focusing on exponential utility [12], mean-variance [20], and percentile risk criteria [10]. With regard to mean-variance optimization in MDPs, it was recently shown that computing an optimal policy under a variance constraint is NP-hard [15]. Recently, average value at risk was introduced in [17] in order to model the tail risk of a random outcome and to address some key limitations of the prevailing value-at-risk metric. Efficient methods to compute AVaR were later on illustrated in [18].

Leveraging the recent strides of AVaR risk modeling, there have been efforts to embed the AVaR metric in the risk-sensitive MDPs. In [3] the authors show that by augmenting the state space, there exists a history dependent optimal policy for AVaR MDP problems. However, their method is only applicable to the finite horizon cost criterion and is based on a computationally complex, dynamic programming approach. Similar techniques can be found in [5], where the authors propose another dynamic programming algorithm for finite-horizon risk-constrained MDPs where risk is measured according to AVaR. The algorithm is proven to asymptotically converge to an optimal risk-constrained policy. However, the algorithm involves computing integrals over continuous variables (Algorithm 1 in [5]) and, in general, its implementation appears particularly difficult.

A different approach is taken by [7], [16], [21] which consider a finite dimensional parameterization of control policies, and show that an AVaR MDP can be optimized to a *local* optimum using stochastic gradient descent (policy gradient). However this approach impose additional restrictions to the policy space and in general policy gradient algorithm only converges to a local optimum.

More recently, Haskell and Jain considered the broader problem of risk aversion in MDPs using a framework based on occupancy measures [13]. Their findings are applicable only when the infinite horizon discounted cost criterion is used. Nevertheless their method is based on occupancy measures and is therefore closely connected to our recent works where constrained MDPs are used to solve the multirobot rapid deployment problem [6], [8]. The solution we propose uses some of the ideas introduced in [13].

III. Preliminaries

We recap some known concepts on risk metrics and MDPs. The reader is referred to the aforementioned references for more details.

A. Risk

Consider a probability space $\mathcal{S} = (\Omega, \mathcal{F}, \mathbb{P})$, and let L^∞ be the space of all essentially bounded random variables on $\mathcal{S}$. A *risk function*[1] $\Gamma : L^\infty \to \mathbb{R}$ is a function that maps an uncertain outcome $Y \in L^\infty$ onto the real line $\mathbb{R}$. A risk function that is particularly popular in many financial applications is the *value at risk*. For $\tau \in (0, 1)$ the *value at risk* of $Y \in L^\infty$ at level τ is defined as

$$\mathrm{VaR}_\tau(Y) := \inf\{\eta \in \mathbb{R} : \Pr(Y \leq \eta) \geq \tau\}.$$

Here $\mathrm{VaR}_\tau(Y)$ represents the percentile value of outcome Y at confidence level τ. Despite its popularity, VaR_τ has a number of limitations. In particular, VaR is not a *coherent* risk measure [2] which suffers from being unstable (high fluctuation under perturbations) when Y is not normally distributed. More importantly it does not quantify the losses that might be suffered beyond its value at the τ-tail of the distribution [17]. An alternative measure that overcomes most shortcomings of VaR is *average value at risk*, defined as

$$\mathrm{AVaR}_\tau(Y) := \frac{1}{1-\tau}\int_\tau^1 \mathrm{VaR}_t(Y)dt,$$

where $\tau \in (0, 1)$ is the confidence level as before. AVaR_τ can be equivalently written as [18]

$$\mathrm{AVaR}_\tau(Y) = \min_{s\in\mathbb{R}} \left\{ s + \frac{1}{1-\tau}\mathbb{E}[(Y-s)^+] \right\}, \qquad (1)$$

where $x^+ := \max(x, 0)$. This paper relies extensively on Eq. (1) and aims at devising efficient methods to approximate the expectation in Eq. (1) when the random variable Y is the total cost of an MDP.

B. Total Cost, Transient Markov Decision Processes

For a finite set S, let $\mathbb{P}(S)$ indicate the set of mass distributions with support on S. A finite, discrete-time Markov Decision Process (MDP) is a tuple $\mathcal{M} = (X, U, \Pr, c)$ where

- X, the state space, is a finite set comprising n elements.
- U, the control space, is a collection of n finite sets $\{U(x_i)\}_{i=1}^n$. Set $U(x_i)$, $i = 1, \ldots, n$, represents the actions that can be applied when in state $x_i \in X$. The set of allowable state/action pairs is defined as

$$\mathcal{K} := \{(x, u) \in X \times U \mid u \in U(x)\}.$$

- $\Pr(y|x, u) : \mathcal{K} \to \mathbb{R}$ is the transition probability from state x to state y when action $u \in U(x)$ is applied. According to our definitions, $\Pr(\cdot|x, u) \in \mathbb{P}(X)$.
- $c : \mathcal{K} \to \mathbb{R}_{\geq 0}$ is a non-negative cost function. Specifically, $c(x, u)$ is the cost incurred when executing action $u \in U(x)$ at state x.

Let $\overline{K} := \max_{(x,u)\in\mathcal{K}} c(x, u)$ and note that the maximum is attained as $\mathcal{K}$ is a finite set. Define the set $\mathcal{H}_t$ of admissible histories up to time t by $\mathcal{H}_t := \mathcal{K} \times \mathcal{H}_{t-1}$, for $t \geq 1$, and $\mathcal{H}_0 := X$. An element of $\mathcal{H}_t$ has the form $x_0, u_0, x_1, \ldots, x_{t-1}, u_{t-1}, x_t$, and records all states traversed and actions taken up to time t. In the most general case a policy is a function $\pi : \mathcal{H}_t \to \mathbb{P}(U(x_t))$, i.e., it decides which action to take in state x_t considering the entire former history. Note that according to this definition the policy could in general be randomized. Let Π be the set of all policies, i.e., including history dependent, randomized policies. It is well known that in the standard MDP setting where an expected cost is minimized there is no loss of optimality to restrict the attention to deterministic, stationary Markovian policies,

[1] In the following the terms *risk metric* and *risk function* will be used interchangeably.

i.e., policies of the type $\pi : X \rightarrow U$. However, in the risk-averse setting one needs to consider the more general class of history-dependent policies [1]. This is achieved through a state augmentation process described later.

Following [13], we define the countable space $(\Omega, \mathbb{B}) := (\mathcal{K}^\infty, \mathbb{B}(\mathcal{K}^\infty))$. where $\mathcal{K}^\infty = \mathcal{K} \times \mathcal{K} \times \mathcal{K} \times \cdots$, is the sample space and $\mathbb{B}(\mathcal{K}^\infty)$ is the Borel field on $\mathcal{K}^\infty$. Specific trajectories in the MDP are written as $\omega \in \Omega$, and we denote by $x_t(\omega)$ and $u_t(\omega)$ the state and actions at time t along trajectory ω. In general the exact initial state x_0 is unknown. Rather it is described by an initial mass distribution β over X, i.e., $\beta \in \mathbb{P}(X)$. A policy π and initial distribution β induce a probability distribution over $(\Omega, \mathbb{B})$, that we will indicate as Pr^π_β.

In this paper we focus on transient total cost MDPs, defined as follows. Consider a partition of X into sets X^T and M, i.e., $X = X^T \cup M$ and $X^T \cap M = \emptyset$. A *transient* MDP is an MDP where *each* policy π satisfies the following two properties:

- $\sum_{t=0}^{\infty} \text{Pr}^\pi_\beta[x_t = x] < \infty$ for each $x \in X^T$, i.e., the state will eventually enter set M, and
- $P(y|x, u) = 0$ for each $x \in M$, $y \in X^T$, $u \in U(x)$, i.e., once the state enters M it cannot leave it.

A transient, *total cost* MDP is a transient MDP where

- $c(x, u) = 0$ for each $x \in M$, i.e., once the state enters M no additional cost is incurred,
- the cost associated with each trajectory ω is given by

$$c(\omega) := \sum_{t=0}^{\infty} c(x_t(\omega), u_t(\omega)).$$

Note that the cost $c(\omega)$ is a random variable depending on both the policy π and the initial distribution β. The name total cost stems from the fact that an (undiscounted) cost is incurred throughout the "lifetime" of the system (i.e., until the state hits the absorbing set).

Transient, total cost MDPs represent an alternative to the most commonly used discounted MDPs or finite horizon MDPs. As outlined in the introduction, for many robotic applications the total cost, i.e., $c(\omega)$, is the most appropriate cost function. We justify this statement by noting that most robotic tasks have finite duration but such duration is usually not known in advance. In these circumstances the finite horizon cost is inappropriate because one cannot define the length of the finite horizon upfront. Similarly, the discounted infinite horizon cost is also ill suited because the task does not continue forever and the cost shall not be exponentially discounted over time.

Without loss of generality, we assume that set M consists of a single absorbing state x_M equipped with a single action u_{x_M}, i.e., $M = \{x_M\}$ and $U(x_M) = u_{x_M}$ with $\Pr(x_M|x_M, u_{x_M}) = 1$. In the following, with a slight abuse of notation, we denote by $\mathcal{K}$ the set $\{(x, u) \in X^T \times U \mid u \in U(x)\}$, i.e., we exclude the absorbing state from the definition of $\mathcal{K}$. Moreover, we assume that for a transient, total cost MDP $\beta(x_M) = 0$, i.e., the probability of starting at the absorbing state is zero. In fact, whenerver $x_0 = x_M$ the resulting state trajectory will deterministically remain in x_M, and the corresponding cost is zero.

IV. Problem formulation

Our problem formulation relies on the following two technical assumptions necessary to establish an a-priori upper bound on the total cost incurred by any trajectory obtained under any policy, and to define an approximate problem that can be efficiently solved.

The first assumption simply requires that all costs in the transient states are positive (recall that we excluded x_M when re-defining $\mathcal{K}$.) As it will be shown later, this assumption ensures a non-zero discretization step when approximating the cumulative cost accrued by a system throughout the trajectory ω until it is absorbed in x_M.

Assumption 1 (Positivity of costs): All costs in $\mathcal{M}$ except for state x_M are positive and bounded, i.e., $\underline{K} := \min_{(x,u)\in\mathcal{K}} c(x, u) > 0$.

When considering cost criteria like finite horizon or discounted infinite horizon with a finite state space, an a-priori upper bound on the cost can be immediately established assuming that all costs are finite. However, the situation is more complex when considering the total cost case, because without introducing further hypotheses on the structure of the MDP a malicious adversarial could establish an history-dependent policy capable of invalidating any a-priori established bound on the cost[2]. The second assumption then adds a "global reachability structure" to the MDP problem. To this end, in the following it will be useful to consider the Markov Chain generated by the MDP when an input is selected for each state. For an MDP $\mathcal{M}$, select $u_1 \in U(x_1), \ldots, u_n \in U(x_n)$. The selected inputs and the transition probabilities in $\mathcal{M}$ define a finite Markov Chain that we indicate as $\mathcal{MC}_{u_1,\ldots,u_n}$. The state space of $\mathcal{MC}_{u_1,\ldots,u_n}$ is equal to X and for two states $x_i, x_j \in X$ the transition probability $\text{Pr}_{i,j}$ is defined as $\text{Pr}_{i,j} = \Pr(x_j|x_i, u_i)$ where $u_i \in U(x_i)$ is the input selected in the definition of $\mathcal{MC}_{u_1,\ldots,u_n}$ and Pr is the transition probability of the associated MDP.

Assumption 2 (Reachability of MDP): Let $\mathcal{MC}_{u_1,\ldots,u_n}$ be the Markov chain induced by the n inputs $u_i \in U(x_i)$. Then the absorbing state x_M, under Markov chain $\mathcal{MC}_{u_1,\ldots,u_n}$, is reachable from any state $x \in X^T$, for all $u_1 \in U(x_1), \ldots, u_n \in U(x_n)$.

We recall that a state j in a Markov chain is said reachable from another state i if there exists an integer $k \geq 1$ such that the probability that the chain will be in state j after k transitions is positive [11]. Note that when Assumption 2 holds, under every policy there is a path of non-zero probability connecting every state to x_M. Therefore, it is impossible to devise a policy that prevents for sure absorption for an arbitrary number of steps. This holds for all policies, including history dependent policies (see Figure 1).

Building upon the previous material, we can now define the problem we aim to solve in this paper:

> **Risk-Averse, total cost MDP** – Given a transient total cost MDP satisfying Assumptions 1-2, and an initial distribution β, determine a policy π that

[2] First note that we are seeking a uniform upper bound for all possible policies, including history dependent policies. Hence, given a tentative bound B, in the general case one could devise a history dependent policy ensuring the every trajectory generated by the policy is not absorbed in x_M in less than $B/\underline{K}$ steps, thus invalidating the bound.

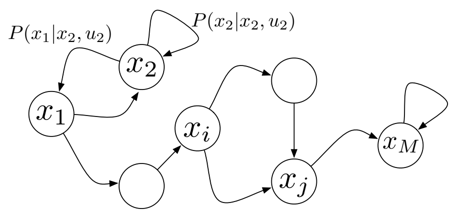

Fig. 1: Meaning of assumption 2: after one input has been chosen for every state, an associated Markov Chain $\mathcal{MC}_{u_1,\ldots,u_n}$ is defined. In this Markov Chain, the absorbing state x_M is reachable from every state, i.e., at every state there is a path of non-zero probability to x_M. The probability of a path, is given by the product of the probabilities of its edges, i.e., the probability of the path $x_i \to x_j \to x_M$ is $\Pr(x_j|x_i,u_i)\Pr(x_M|x_j,u_j)$. This requirement is imposed for every possible choice of the inputs and every policy.

minimizes $\text{AVaR}_\tau(c(\omega))$, i.e., find

$$\pi^* \in \arg\min_{\pi\in\Pi} \text{AVaR}_\tau(c(\omega)). \qquad (2)$$

Note that, under the assumption of transient total cost MDP, one can easily verify that $\mathbb{E}[c(\omega)] < \infty$. Since, by equation (1), $\text{AVaR}_\tau(c(\omega)) \leq 1/(1-\tau)\mathbb{E}[c(\omega)]$, one obtains $\text{AVaR}_\tau(c(\omega)) < \infty$ for all ω, as well. However, to derive an optimization algorithm for the computation of π^* it is necessary to formulate an a-priori upper bound for the optimal cost in (2). Assumptions 1 and 2 are introduced to ensure that such bound exists and can be computed.

V. Solution Algorithm

In this section we provide a solution algorithm for the risk averse total cost MDP in Eq. (2). Similar to the method presented in [13], we aim to solve this problem with *occupation measures*. However unlike the previously studied cases, in total cost MDP an explicit upper bound on the cumulated cost is not known in advance. This makes the characterization of feasible set impossible. Our strategy is to find a surrogate to the MDP problem in Eq. (2). By imposing an effective horizon, we construct a total cost MDP with time-out and recast this problem into bilinear programming. Furthermore we characterize the sub-optimality gap to this surrogate approximation.

A. Convergence rate to the absorbing state

Consider a selection of inputs $u_1 \in U(x_1), \ldots, u_n \in U(x_n)$ and the corresponding Markov chain $\mathcal{MC}_{u_1,\ldots,u_n}$. For each state $x \in X^T$, let $\text{MinimumPath}_{x\to x_M}(\mathcal{MC}_{u_1,\ldots,u_n})$ denote the simple (i.e., without cycles) path from x to x_M of lowest, strictly positive probability. Note that $\text{MinimumPath}_{x\to x_M}(\mathcal{MC}_{u_1,\ldots,u_n})$ exists due to Assumption 2. Let $\Pr(\text{MinimumPath}_{x\to x_M}(\mathcal{MC}_{u_1,\ldots,u_n}))$ be the probability of the path, i.e., the product of the probabilities of all the transitions along the path. Since there are n nodes and by definition the path is simple, $\text{MinimumPath}_{x\to x_M}(\mathcal{MC}_{u_1,\ldots,u_n})$ includes at most $n-1$ transitions between n nodes. Let

$$\gamma := \min_{\substack{u_k\in U(x_k),\\ k=1\ldots,n}} \min_{x\in x^T} \Pr(\text{MinimumPath}_{x\to x_M}(\mathcal{MC}_{u_1,\ldots,u_n})).$$

Note that the minimum is achieved as the minimization is over a finite set, and that γ is strictly positive due to Assumption 2. The constant γ lower bounds the probability that, under *any* policy $\pi \in \Pi$, the absorbing state is reached in no more than n steps, from *any* state $x \in X^T$. We are now in a position to state the following lemma for the convergence rate to the absorbing state.

Lemma 1 (Number of stages to reach the absorbing set): For any policy $\pi \in \Pi$ and initial distribution β,

$$\Pr{}^\pi_\beta[x_{kn} \neq x_M] \leq (1-\gamma)^k, \ \forall k \in \mathbb{N}.$$

Proof: The claim is proven by induction on k. Base case: we prove that

$$\Pr{}^\pi_\beta[x_n \neq x_M] \leq 1-\gamma.$$

Indeed, $\Pr^\pi_\beta[x_n \neq x_M] = \sum_{x\in X_T} \Pr^\pi_\beta[x_n \neq x_M|x_0 = x]\Pr^\pi_\beta[x_0 = x]$. Because of Assumption 1, for any policy π, $\Pr^\pi_\beta[x_n = x_M] \geq \gamma$, the base case follows.

For the inductive step, assume that $\Pr^\pi_\beta[x_{kn} \neq x_M] < (1-\gamma)^k$, for some $k > 1$. Then, $\Pr^\pi_\beta[x_{(k+1)n} \neq x_M] = \Pr^\pi_\beta[x_{(k+1)n} \neq x_M|x_{kn} \neq x_M]\Pr^\pi_\beta[x_{kn} \neq x_M]$. By definition of γ,

$$\Pr{}^\pi_\beta[x_{(k+1)n} \neq x_M|x_{kn} \neq x_M] \leq (1-\gamma)^{k+1},$$

and the claim follows. ■

B. Approximation bounds for surrogate problem

Our solution strategy is to solve a *surrogate* problem, whereby after a deterministic number $d \in \mathbb{N}$ of steps, the state moves to the absorbing state x_M surely. In other words, d acts as a "timeout" for the MDP problem. The surrogate problem is simpler to solve, and we will show in the following that its solution can approximate the solution of the original problem with arbitrary precision. Denote by $c^{[d]}(\omega)$ the total cost for such surrogate problem. Additionally, for the original problem, let $t^*(\omega)$ denote the *absorbing time*, i.e., the time at which the state reaches x_M. If $t^*(\omega) \leq d$, then the two processes coincide and then $c^{[d]}(\omega) = c(\omega)$. Otherwise, for each trajectory ω such that $t^*(\omega) > d$, the random process is stopped after d steps and the state goes, deterministically, to x_M at stage $d+1$. In such a case one has $c^{[d]}(\omega) \leq c(\omega)$.

We want to characterize the relation between $\text{AVaR}_\tau(c^{[d]}(\omega))$ (i.e., the risk for the surrogate problem) and $\text{AVaR}_\tau(c(\omega))$ (i.e., the risk for the original problem). To this purpose, let $c_d(\omega)$ be the total cost for the original problem up to time d, i.e.,

$$c_d(\omega) := \sum_{t=0}^{d} c(x_t(\omega), u_t(\omega)).$$

The following lemma shows the equivalence between $c^{[d]}(\omega)$ and $c_d(\omega)$.

Lemma 2 (Correspondence of costs): For any policy $\pi \in \Pi$ and any trajectory ω, $c^{[d]}(\omega) = c_d(\omega)$.

Proof: Given a policy π, for any trajectory ω, the cost cumulated up to time d is the same for both the original and the surrogate problem. After time d, both $c_d(\omega)$ and $c^{[d]}(\omega)$ do not cumulate any additional cost, then the claim follows. ■

The following theorem represents the main result of this section

Theorem 3 (Suboptimality bound): The surrogate problem approximates the original problem according to

$$\min_\pi \mathrm{AVaR}_\tau(c^{[d]}(\omega)) \leq \min_\pi \mathrm{AVaR}_\tau(c(\omega))$$
$$\leq \min_\pi \mathrm{AVaR}_\tau(c^{[d]}(\omega)) + \frac{n\overline{K}}{1-\tau}\frac{(1-\gamma)^{\lfloor (d+1)/n \rfloor}}{\gamma}.$$

Proof: The left inequality is proven by noticing that $\min_\pi \mathrm{AVaR}_\tau(c(\omega)) \geq \min_\pi \mathrm{AVaR}_\tau(c_d(\omega)) = \min_\pi \mathrm{AVaR}_\tau(c^{[d]}(\omega))$, where the equality follows from Lemma 2.

We now prove the right inequality. For any $s \in \mathbb{R}$ and policy π, one has

$$\mathbb{E}[(c(\omega)-s)^+] = \mathbb{E}[(c(\omega)-s)^+ \mid t^*(\omega) \leq d]\mathbb{P}(t^*(\omega) \leq d) + \mathbb{E}[(c(\omega)-s)^+ \mid t^*(\omega) > d]\mathbb{P}(t^*(\omega) > d). \quad (3)$$

Let $c_l(\omega) := \sum_{t=d+1}^{\infty} c(x_t(\omega), u_t(\omega))$ be the tail cumulated cost, and, as before, $c_d(\omega) := \sum_{t=0}^{d} c(x_t(\omega), u_t(\omega))$. Since the function $x \to x^+$ is sub-additive, i.e., $(x+y)^+ \leq x^+ + y^+$ and the expectation operator preserves monotonicity, one obtains the inequality

$$\begin{aligned}&\mathbb{E}[(c(\omega)-s)^+ \mid t^*(\omega) > d]\\ &\quad= \mathbb{E}[(c_d(\omega)+c_l(\omega)-s)^+ \mid t^*(\omega) > d]\\ &\quad\leq \mathbb{E}[(c_d(\omega)-s)^+ \mid t^*(\omega) > d] + \mathbb{E}[c_l(\omega) \mid t^*(\omega) > d].\end{aligned}$$

Furthermore, for each trajectory in the event set $\{\omega : t^*(\omega) \leq d\}$, one has

$$\mathbb{E}[(c(\omega)-s)^+ \mid t^*(\omega) \leq d] = \mathbb{E}\left[(c_d(\omega)-s)^+ \mid t^*(\omega) \leq d\right].$$

Collecting the results so far, one has the following inequalities:

$$\begin{aligned}&\mathbb{E}[(c(\omega)-s)^+]\\ &\leq \mathbb{E}[(c_d(\omega)-s)^+ \mid t^*(\omega) \leq d]\mathbb{P}(t^*(\omega) \leq d)\\ &\quad+ \mathbb{E}[(c_d(\omega)-s)^+ \mid t^*(\omega) > d]\mathbb{P}(t^*(\omega) > d) +\\ &\quad+ \mathbb{E}\left[c_l(\omega) \mid t^*(\omega) > d\right]\mathbb{P}(t^*(\omega) > d)\\ &= \mathbb{E}[(c_d(\omega)-s)^+] + \mathbb{E}\left[c_l(\omega) \mid t^*(\omega) > d\right]\mathbb{P}(t^*(\omega) > d)\\ &\leq \mathbb{E}[(c_d(\omega)-s)^+] + \mathbb{E}\left[c_l(\omega)\right].\end{aligned} \quad (4)$$

Equation (4) implies

$$\begin{aligned}&\min_\pi\ \mathrm{AVaR}_\tau(c(\omega))\\ &= \min_\pi \min_{s\in\mathbb{R}} \left\{ s + \frac{1}{1-\tau}\mathbb{E}[(c(\omega)-s)^+] \right\}\\ &\leq \min_\pi \min_{s\in\mathbb{R}} \left\{ s + \frac{1}{1-\tau}\left(\mathbb{E}[(c_d(\omega)-s)^+] + \mathbb{E}\left[c_l(\omega)\right]\right) \right\}\\ &\leq \min_\pi \min_{s\in\mathbb{R}} \left\{ s + \frac{1}{1-\tau}\left(\mathbb{E}[(c_d(\omega)-s)^+]\right) \right\}\\ &\quad+ \max_\pi \frac{1}{1-\tau}\mathbb{E}\left[c_l(\omega)\right]\\ &= \min_\pi \min_{s\in\mathbb{R}} \left\{ s + \frac{1}{1-\tau}\left(\mathbb{E}[(c^{[d]}(\omega)-s)^+]\right) \right\}\\ &\quad+ \max_\pi \frac{1}{1-\tau}\mathbb{E}\left[c_l(\omega)\right]\\ &= \min_\pi\ \mathrm{AVaR}_\tau(c^{[d]}(\omega)) + \max_\pi\ \frac{1}{1-\tau}\mathbb{E}\left[c_l(\omega)\right],\end{aligned}$$

where the second to last equality follows from Lemma 2. We are left with the task of upper bounding $\mathbb{E}\left[c_l(\omega)\right]$. To this purpose, one can write

$$\mathbb{E}\left[c_l(\omega)\right] \leq \overline{K} \sum_{t=d+1}^{\infty} \mathrm{Pr}_\beta^\pi(x_t \neq x_M).$$

Note that the result in Lemma 1 implies

$$\begin{aligned}\sum_{t=d+1}^{\infty} \mathrm{Pr}_\beta^\pi[x_t \neq x_M] &\leq \sum_{k=\lfloor (d+1)/n \rfloor}^{\infty} \sum_{t=kn}^{(k+1)n-1} \mathrm{Pr}_\beta^\pi[x_t \neq x_M]\\ &\leq \sum_{k=\lfloor (d+1)/n \rfloor}^{\infty} n(1-\gamma)^k\\ &= n\frac{(1-\gamma)^{\lfloor (d+1)/n \rfloor}}{\gamma}.\end{aligned}$$

The claim then follows immediately, as the above upper bound is policy-independent. ■

Note that according to Theorem 3, as $d \to \infty$, the optimal cost of the surrogate problem recovers the optimal cost of the original problem, i.e., the surrogate problem provides a consistent approximation to the original problem, with a sub optimality factor that is computable from problem data.

Lemma 2 and Theorem 3 ensure that $c^{[d]}(\omega)$ can approximate $c(\omega)$ with arbitrary precision for a sufficiently large value of d. The following theorem (see [13]) establishes that $\mathrm{AVaR}_\tau(c^{[d]}(\omega))$ can be used to approximate $\mathrm{AVaR}_\tau(c(\omega))$.

Theorem 4: AVaR is a uniformly continuous across policies in the total cost, i.e., for each $\varepsilon > 0$ there exist $\delta > 0$ such that for any policy $\pi \in \Pi$,

$$|c_1(\omega) - c_2(\omega)| < \delta \Rightarrow |\mathrm{AVaR}_\tau(c_1(\omega)) - \mathrm{AVaR}_\tau(c_2(\omega))| < \varepsilon.$$

In the next section we then show how to solve the minimization problem:

$$\min_\pi \mathrm{AVaR}_\tau(c^{[d]}(\omega)). \quad (5)$$

VI. AVaR in total cost MDPs

Leveraging the surrogate problem from the previous section, we are now in the position of adapting the results proposed in [13] to compute the AVaR for total cost MDP. An essential step to solve this optimization problem is to compute $\mathbb{E}[(c^{[d]}(\omega)-s)^+]$, which entails deriving the probability distribution for the possible costs generated by the random variable $c^{[d]}(\omega)$. This problem can be solved by suitably augmenting the state space as described in the following, and using occupancy measures. Using occupancy measures the optimal policy of an MDP is determined through the solution of a bilinear program, as explained in the following. For a given policy π and initial distribution β, define occupancy measure of $(x,u) \in \mathcal{K}$ as

$$\rho(x,u) = \sum_{t=0}^{\infty} \mathrm{Pr}\,_\beta^\pi[x_t(\omega) = x, u_t(\omega) = u)].$$

Note that $\rho(x,u)$ is non negative but is in general not a probability itself. In the following we will use occupancy measures to determine the probability distribution of the total costs $c^{[d]}(\omega)$ and then to compute the needed expectation. According to the definition, occupancy measures depend on the policy π and the initial distribution β. Given an absorbing MDP $\mathcal{M} = (X, U, \Pr, c)$, we define a new *state-augmented* absorbing MDP with additional state components tracking the accumulated total cost and current stage. Although the original MDP $\mathcal{M}$ is finite and absorbing, the set of costs $c^{[d]}(\omega)$ generated by all possible policies can be very large, and this can subsequently lead to a linear program with an unmanageable number of decision variables. To counter this problem, we introduce a discretized approximation for $c^{[d]}(\pi, \beta)$ whose error can be arbitrarily bounded. To this end, we set $\zeta = \min\left\{\underline{K}, \frac{d\bar{K}}{N'}\right\}$, where $N' \in \mathbb{N}$ is a parameter describing the desired number of discretized values to the cumulated cost. Due to Assumption 1, ζ is strictly positive. The effective number of different values is $N = \left\lceil \frac{d\bar{K}}{\zeta} \right\rceil$. This value may be higher than N' due to our definition of ζ. We then define a new MDP $\mathcal{M}'_N = (X', U', \Pr', c')$ as follows. Its state space is $X' = X \times \mathbb{N}_N \times \mathbb{N}_d$, where $\mathbb{N}_N = \{0, 1, \dots, N\}$ and $\mathbb{N}_d = \{0, 1, \dots, d\}$. Elements in the augmented states will be indicated as (x, y, z). As clarified in the following, the two additional components store the cumulated running cost (y) and stage (z). Recall from the surrogate problem we aim to approximate the original risk-averse total cost problem with a counterpart with time-out. This formulation implies that after d steps the state has entered the absorbing set, i.e., it is guaranteed that $x_d(\omega) = x_M$. Thus the value of the z component to the set is in $\mathbb{N}_d = \{0, 1, \dots, d\}$. On the other hand the input sets are defined as $U'(x, y, z) = U(x)$. X' and U' induce a new set $\mathcal{K}' = \{(x, y, z, u) \mid (x, u) \in \mathcal{K} \wedge y \in \mathbb{N}_N \wedge z \in \mathbb{N}_d\}$. The new cost function $c' : \mathcal{K}' \to \mathbb{R}_{\geq 0}$ is $c'(x, y, z, u) = c(x, u)$. The transition probability function is modified as follows.

$$\Pr'((x', y', z')|(x, y, z), u) =$$
$$\begin{cases} \Pr(x'|x, u) & \text{if } y' = y + \left\lfloor \frac{c(x,u)}{\zeta} \right\rfloor \wedge z' = z + 1 \wedge z' < d \\ 1 & \text{if } (x', y', z') = (x_M, y, d) \wedge z = d \\ 0 & \text{otherwise} \end{cases}$$

As evident from the definition of the new transition function, the new variables included in the state stores the discretized[3] running cost and the stage. Consistently with our definition of the surrogate problem, the revised transition function includes a timeout that imposes a transition to the absorbing state x_M after d steps, and from that point onwards the accrued cost does not change anymore. Note also that the additional state components y and z are deterministic functions of the previous state and control input u. Extending the formerly introduced notation, for given trajectory ω of $\mathcal{M}'_N$, we write $y_t(\omega)$ for the second component of the state at time t and $z_t(\omega)$ for the third component. Finally, for a given initial distribution β on X, we define the following new initial distribution β' on X',

$$\beta'(x, y, z) = \begin{cases} \beta(x) & \text{if } y = 0 \ \wedge z = 0 \\ 0 & \text{otherwise} \end{cases}.$$

Note that the properties of $\mathcal{M}$ carry over to $\mathcal{M}'_N$. In particular, if Assumptions 1-2 hold for $\mathcal{M}$ then they hold for $\mathcal{M}'_N$ too, and if $\mathcal{M}$ is absorbing then $\mathcal{M}'_N$ is also absorbing. Thus we indicate with X'^T its transient set of states. For a given realization ω, consider now $c_t^{[d]}(\omega) = \sum_{i=0}^{t} c(x_i, u_i)$, i.e., the true cumulative cost of the surrogate MDP problem without discretization. The following theorem establishes that even though the approximation error introduced by discretizing the running cost grows linearly with t, it is possible to bound it with arbitrary precision.

Theorem 5: For each $\varepsilon > 0$ and each $t \in \{0, \dots, d\}$, there exists a discretization step ζ such that $|\zeta y_t(\omega) - c_t^{[d]}(\omega)| < \varepsilon$.

Proof. Pick $\zeta = \varepsilon/d$. Let $e(t) = c_t^{[d]}(\omega) - \zeta y_t(\omega)$ be the approximation error at time t. Note that by definition $e(t) \geq 0$ and $e(0) = c_0^{[d]}(\omega) - \zeta y_0(\omega) = 0$. From the definition of the transition probability function P' it follows that $e(t+1) \leq e(t) + \zeta$, which implies $e(d) \leq d\zeta = \varepsilon$. Since for $t > d$ we have $e(t) = e(d)$ and the claim follows. □

A key step towards the solution of problem in Eq. (5) is therefore to derive the statistical description of the discretized total cost $y_d(\omega)$ that is used to approximate $c^{[d]}(\omega)$. This objective can be achieved by exploiting the occupancy measure for the state-augmented MDP $\mathcal{M}'$. For $(x, y, z, u) \in \mathcal{K}'$, the occupancy measure on $\mathcal{M}'$ induced by a policy π and an initial distribution β is given as:

$$\rho(x, y, z, u) = \qquad (6)$$
$$\sum_{t=0}^{\infty} \Pr{}_{\beta}^{\pi}[x_t(\omega) = x, y_t(\omega) = y, z_t(\omega) = z, u_t(\omega) = u].$$

The occupancy measure ρ is a vector in $\mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$, i.e., it is a vector with $|K'|$ non negative components. The set of legitimate occupancy vectors is constrained by the initial distribution β and defined by the policy π. It is well known [1] that these constraints can be expressed as follows:

$$\sum_{(x',y',z') \in X'^T} \sum_{u \in A(x',y',z')} \rho(x', y', z', u)[\delta_{(x,y,z)}(x', y', z') -$$
$$P'((x', y', z')|(x, y, z), u)] = \beta(x, y, z) \ \forall (x, y, z) \in X'^T$$

where $\delta_x(y) = 1$ if and only if $y = x$. For $0 \leq k \leq N$ we introduce random variables $\theta(k)$ with the property that $\theta(k) = \Pr[y_d(\omega) = k]$. This is easily achieved using occupancy measures, i.e., $\theta(k) = \sum_{(x,y,z,u) \in \mathcal{K}'} I(y = k \ \wedge z = d)\rho(x, y, z, u)$, where $I(\cdot)$ is the indicator function equal to 1 when its argument is true, and 0 otherwise. Note that by definition $\theta(k)$ is equal to $\Pr[y_d(\omega) = k]$, and by theorem 5 $y_d(\omega)$ approximates $c^{[d]}(\omega)$ with arbitrary precision. Combining the above definitions we then get to the following problem whose solution approximates Eq. (5):

$$\min_{\rho, \theta} \min_{s \in [0, \bar{K}d]} s + \frac{1}{1-\tau} \sum_{y \in \mathbb{N}_N} (y - s)^+ \theta(y) \qquad (7)$$

s.t.

$$\sum_{(x',y',z') \in X'^T} \sum_{u \in A(x',y',z')} \rho(x', y', z', u)[\delta_{(x,y,z)}(x', y', z') -$$
$$P'((x', y', z')|(x, y, z), u)] = \beta(x, y, z) \ \forall (x, y, z) \in X'^T$$
$$\theta(k) = \sum_{(x,y,z,u) \in \mathcal{K}'} I(y = k \ \wedge z = d)\rho(x, y, z, u), \ 0 \leq k \leq N.$$

[3] To be precise, the discretized running cost is scaled by ζ.

When comparing this last optimization problem with Eq. (5), the reader will note that the variable s is constrained in the interval $[0, \bar{K}d]$. Indeed, the objective function is continuous with respect to s, and it is immediate to verify[4] that the partial derivative of the objective function with respect to s is negative for $s < 0$ and positive for $s > \bar{K}d$. The objective function given in Eq. (7) is concave with respect to $\theta(y)$ and is defined over a convex feasibility set [13]. It is known that the minimum will be achieved at one vertex of the feasibility set. To the best of our knowledge, there exist no efficient methods to determine the global minimum for this class of problems. Hence, the problem can be approximately solved fixing different values of s within the range $[0, \bar{K}d]$, and then solving the corresponding linear problem over the optimization variables ρ and θ.

Comparing problem in Eq. (7) with problem in Eq. (2) one could see that the objective function in the Eq. (2) does not seem to depend on the policy π. However, the dependency on π is carried over by the occupancy measure ρ, as evident from Eq. (6). Moreover, it is well known from the theory of constrained MDPs [1] that there is a one to one correspondence between policies and occupancy measures, i.e., every policy defines a unique occupancy measure and every occupancy measure induces a policy.

VII. Simulations

To illustrate the performance of risk-averse control in the total cost MDP problem, we adopt the rapid deployment scenario considered in [6], [8]. Given a map of an environment, a graph is used to abstract and model its connectivity (see, e.g., [14]). One robot is positioned at a start vertex and is tasked to reach the goal vertex within a given temporal deadline while providing some guarantee about its the probability of successfully completing the task. When moving from vertex to vertex, the robot can choose from a set of actions, each trading off velocity with probability of success. In particular, actions with rapid transitions between two vertices have higher probability of failure; and conversely when the robot moves slowly between two vertices it has a higher probability of success. In this scenario *failure* means that the robot does not move (e.g., fails to pass through an opening), so elapsed time increases without making progress towards the goal. With given temporal deadline and success probability the robot is tasked to reach the target vertex within a certain time T with probability at least P. From a design perspective it is of interest to know if there exists a policy π achieving this objective, and to compute it. If the policy does not exist, it is of interest to know how to modify the parameters in order to make the task feasible.

In our previous work we solved this problem by modeling it using Constrained Markov Decision Processes (CMDP). In the CMDP approach one maximizes the probability of success while imposing a constraint on the temporal deadline. However, this method only returns risk-neutral policies, i.e., the resultant policies only guarantee that the temporal deadline is met in expectation, and there is no explicit control on the tail probability of the constraint.

As a radical departure from the original problem formulation, the AVaR minimization method proposed in Eq. (2) searches for a policy that is feasible with respect to the temporal deadline constraint[5] and systematically controls the worst-case variability of total travel time. Note that a policy with low success probability will have large tail probability in total travel time even if the expected temporal deadline is met. Therefore the optimal policies from AVaR minimization will have *high success probability*. This motivates the application of AVaR minimization to rapid robotic deployment.

First, note that in the devised setting the robot will eventually reach the final goal with positive probability. However due to possible failures one cannot put an a-priori bound on the random total travel time. Therefore, the total cost criterion is indeed a natural choice for this task. Moreover, Assumption 1 and 2 are easily justified because the immediate cost function (i.e., time to move) is always positive and the global reachability property follows from the graph structure.

To illustrate the performance of risk-averse deployment, two different policies are compared. Here both policies are computed using unconstrained stochastic control methodologies for which the immediate cost is the travel time between two vertices and the actions correspond to all possible node transitions on the graph. The first is the classic risk-neutral policy obtained with value iteration. The second is a risk averse policy obtained with the algorithm presented in this paper using $\tau = 0.95$. For each policy, 1000 executions are run, and the distribution of total travel time is reported. Figure 2 and 3 show the distribution for the two cases. The risk-neutral policy obtains a lower expected cost, but has a longer tail, as evidenced by the 61 instances with a cost larger or equal than 15 (assuming that $T = 15$ is the desired time to completion.) Moreover, as evidenced by the shape the histogram, costs are more spread. The risk averse policy, instead, presents less variability, as per design objectives and less than 30 instances have a cost larger or equal than 15, thus halving the weight of the tail.

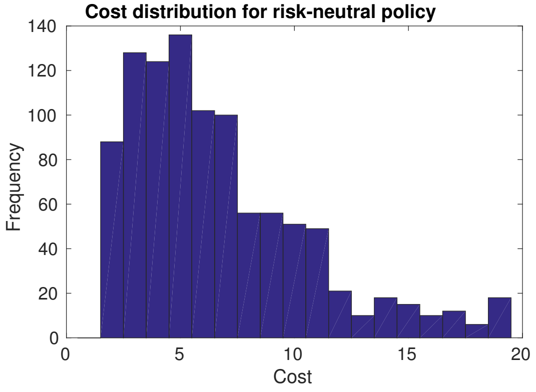

Fig. 2: Cost distribution for a risk-neutral policy

[4] In the computation one should consider the y values as constants and that $\sum_y \theta(y) = 1$.

[5] Because for any random variable Z with finite expectation, the following inequality holds: $\mathrm{AVaR}_\tau(Z) \geq \mathbb{E}[Z]$ for $\tau \in [0, 1]$. Therefore, if the solution to the AVaR minimization problem is bounded above by the temporal deadline, then the corresponding minimizer is also a feasible policy to the original problem.

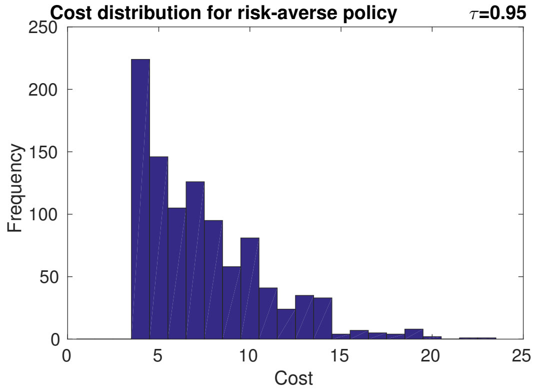

Fig. 3: Cost distribution for a risk averse policy with $\tau = 0.95$.

Importantly, when computing a risk-neutral metric using classic methods like policy iteration or value iteration, one is just provided with a policy that minimizes the expected cost (in our case time to completion), but no additional information is readily available. With our approach instead, one not only obtains a policy minimizing the AVaR criterion, but as a byproduct a statistical descriptions of the costs is obtained too. That is to say, that for each discretized completion time k, the probability $\Pr[c^{[d]} = k]$ is computed as well, thus unveiling the relationship between the time to complete the deployment task and its probability. This is shown in Figure 4 for different values of τ. Hence if the computed policy does not meet the desired performance, the designer has valuable information on how to accordingly modify the parameters T and P.

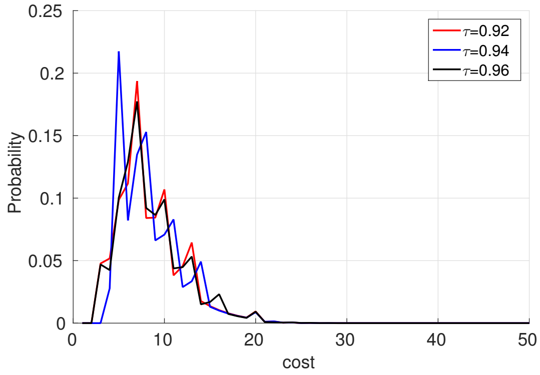

Fig. 4: Comparison of probability cost distribution for different values of τ.

VIII. Conclusions

In this paper we have considered how risk aversion in MDPs can be introduced jointly with the AVaR risk metric under the total cost criterion. Our results advance the state of the art since previously AVaR was only considered in MDPs with different cost criteria. Such extension is *crucial* becasue the total cost criterion is the most appropriate in numerous robotic scenarios, and is *non-trivial* because current algorithms from i.e., [13] and [3], only work with bound cumulated costs. Under two practical assumptions, our approximation algorithm provided sharp bounds for the sub-optimality gap. Furthermore, a rapid deployment scenario was used to demonstrate that risk-aversion gives more informative policies when compared with traditional risk-neutral metrics.

While our findings focus on risk averse MDPs with AVaR risk metric, our approach can be easily extended in multiple dimensions. In particular, by exploiting the results presented in [13], it is possible to use our approximation for a broader range of risk metrics, i.e., metrics that are uniformly continuous and law invariant. Moreover, since the algorithm we considered is based on occupancy measures, it can be easily extended also to the CMDP case. Details on these derivations will be the focus of future work.

References

[1] E. Altman. *Constrained Markov Decision Processes*. Stochastic modeling. Chapman & Hall/CRC, 1999.

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

[3] N. Bäurle and J. Ott. Markov Decision Processes with Average-Value-at-Risk Criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

[4] D. Bertsekas. *Dynamic Programming & Optimal Control*, volume 1 & 2. Athena Scientific, 2005.

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

[6] S. Carpin, M. Pavone, and B. Sadler. Rapid Multirobot Deployment with Time Constraints. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1147–1154, 2014.

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

[8] Y. Chow, M. Pavone, B. Sadler, and S. Carpin. Trading Safety Versus Performance: Rapid Deployment of Robotic Swarms with Robust Performance Constraints. *ASME Journal of Dynamical Systems, Measurements and Control*, 137(3):031005, 2015.

[9] E. Feinberg. *Handbook of Markov Decision Processes: Methods and Applications*. Springer, 2012.

[10] J. Filar, D. Krass, and K. Ross. Percentile Performance Criteria for Limiting Average Markov Decision Processes. *IEEE Transaction of Automatic Control*, 40(1):2–10, 1995.

[11] R. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.

[12] R. Howard and J. Matheson. Risk-sensitive Markov Decision Processes. *Management Science*, 18(7):356–369, 1972.

[13] W. Haskell and R. Jain. A Convex Analytic Approach to Risk-aware Markov Decision Processes. *SIAM Journal of Control and Optimization*, 2014.

[14] A. Kolling and S. Carpin. Extracting Surveillance Graphs from Robot Maps. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2323-2328, 2008.

[15] S. Mannor and J. Tsitsiklis. Algorithmic Aspects of Mean-variance Optimization in Markov Decision Processes. *European journal of Operational Research*, 231:645–653, 2013.

[16] L. Prashanth. Policy Gradients for CVaR-constrained MDPs. In *Algorithmic Learning Theory*, pages 155–169. Springer, 2014.

[17] R. Rockafeller and S. Uryasev. Optimization of Conditional Value at Risk. *Journal of Risk*, 2:21–41, 2000.

[18] R. Rockafeller and S. Uryasev. Conditional Value-at-risk for General Loss Distributions. *Journal of Banking Finance*, 26:1443–1471, 2002.

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

[20] M. Sobel. The Variance of Discounted Markov Decision Processes. *Journal of Applied Probability*, pages 794–802, 1982.

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

[22] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2006.