Stochastic Optimal Control With Dynamic, Time-Consistent Risk Constraints

Yin-Lam Chow, Marco Pavone

Abstract— In this paper we present a dynamic programing approach to stochastic optimal control problems with dynamic, time-consistent risk constraints. Constrained stochastic optimal control problems, which naturally arise when one has to consider multiple objectives, have been extensively investigated in the past 20 years; however, in most formulations, the constraints are formulated as either risk-neutral (i.e., by considering an expected cost), or by applying static, single-period risk metrics with limited attention to "time-consistency" (i.e., to whether such metrics ensure rational consistency of risk preferences across multiple periods). Recently, significant strides have been made in the development of a rigorous theory of dynamic, *time-consistent* risk metrics for multi-period (risk-sensitive) decision processes; however, their integration within constrained stochastic optimal control problems has received little attention. The goal of this paper is to bridge this gap. First, we formulate the stochastic optimal control problem with dynamic, time-consistent risk constraints and we characterize the tail subproblems (which requires the addition of a Markovian structure to the risk metrics). Second, we develop a dynamic programming approach for its solution, which allows to compute the optimal costs by value iteration. Finally, we discuss both theoretical and practical features of our approach, such as generalizations, construction of optimal control policies, and computational aspects. A simple, two-state example is given to illustrate the problem setup and the solution approach.

I. Introduction

Constrained stochastic optimal control problems naturally arise in several domains, including engineering, finance, and logistics. For example, in a telecommunication setting, one is often interested in the maximization of the throughput of some traffic subject to constraints on delays [1], [2], or seeks to minimize the average delays of some traffic types, while keeping the delays of other traffic types within a given bound [3]. Arguably, the most common setup is the optimization of a *risk-neutral expectation* criterion subject to a *risk-neutral* constraint [4], [5], [6]. This model, however, is not suitable in scenarios where risk-aversion is a key feature of the problem setup. For example, financial institutions are interested in trading assets while keeping the *riskiness* of their portfolios below a threshold; or, in the optimization of rover planetary missions, one seeks to find a sequence of divert and driving maneuvers so that the rover drive is minimized and the *risk* of a mission failure (e.g., due to a failed landing) is below a user-specified bound [7].

A common strategy to include risk-aversion in constrained problems is to have constraints where a static, single-period

Y.-L. Chow and M. Pavone are with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305, USA. Email: {ychow, pavone}@stanford.edu.

risk metric is applied to the future stream of costs; typical examples include variance-constrained stochastic optimal control problems (see, e.g., [5], [8], [9]), or problems with probability constraints [4], [5]. However, using static, single-period risk metrics in multi-period decision processes can lead to an over or under-estimation of the true dynamic risk, as well as to a potentially "inconsistent" behavior (whereby risk preferences change in a seemingly irrational fashion between consecutive assessment periods), see [10] and references therein. In [11], the authors provide an example of a portfolio selection problem where the application of a static risk metric in a multi-period context leads a risk-averse decision maker to (erroneously) show risk neutral preferences at intermediate stages.

Indeed, in the recent past, the topic of *time-consistent* risk assessment in multi-period decision processes has been heavily investigated [12], [13], [14], [15], [16], [17], [18]. The key idea behind time consistency is that if a certain outcome is considered less risky in all states of the world at stage k, then it should also be considered less risky at stage k [10]. Remarkably, in [15], it is proven that any risk measure that is time consistent can be represented as a composition of one-step conditional risk mappings, in other words, in multi-period settings, risk (as expected) should be compounded over time.

Despite the widespread usage of constrained stochastic optimal control and the significant strides in the theory of dynamic, time-consistent risk metrics, their integration within constrained stochastic optimal control problems has received little attention. The purpose of this paper is to bridge this gap. Specifically, the contribution of this paper is threefold. First, we formulate the stochastic optimal control problem with dynamic, time-consistent risk constraints and we characterize the tail subproblems (which requires the addition of a Markovian structure to the risk metrics). Second, we develop a dynamic programming approach for the solution, which allows to compute the optimal costs by value iteration. There are two main reasons behind our choice of a dynamic programing approach: (a) the dynamic programming approach can be used as an analytical tool in special cases and as the basis for the development of either exact or approximate solution algorithms; and (b) in the risk-neutral setting (i.e., both objective and constraints given as expectations of the sum of stage-wise costs) the dynamic programming approach appears numerical convenient with respect to other approaches (e.g., with respect to the convex analytic approach [1]) and allows to build all (Markov) optimal control strategies [5]. Finally, we discuss both theo-

retical and practical features of our approach, generalizations, construction of optimal control policies, and computational aspects. A simple, two-state example is given to illustrate the problem setup and the solution approach.

The rest of the paper is structured as follows. In Section II we present background material for this paper, in particular about dynamic, time-consistent risk measures. In Section III we formally state the problem we wish to solve, while in Section IV we present a dynamic programming approach for the solution. In Section V we discuss several aspects of our approach and provide a simple example. Finally, in Section VI, we draw our conclusions and offer directions for future work.

II. PRELIMINARIES

In this section we provide some known concepts from the theory of Markov decision processes and of dynamic risk measures, on which we will rely extensively later in the paper.

A. Markov Decision Processes

A finite Markov Decision Process (MDP) is a four-tuple $(S, U, P, U(\cdot))$, where S, the state space, is a finite set; U, the control space, is a finite set; for every $x \in S$, $U(x) \subseteq U$ is a nonempty set which represents the set of admissible controls when the system state is x; and, finally, $Q(\cdot|x, u)$ (the transition probability) is a conditional probability on S given the set of admissible state-control pairs, i.e., the sets of pairs (x, u) where $x \in S$ and $u \in U(x)$.

Define the space H_k of admissible histories up to time k by $H_k = H_{k-1} \times S \times U$, for $k \geq 1$, and $H_0 = S$. A generic element $h_{0,k} \in H_k$ is of the form $h_{0,k} = (x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k)$. Let Π be the set of all deterministic policies with the property that at each time k the control is a function of $h_{0,k}$. In other words, $\Pi := \Big\{ \{\pi_0 : H_0 \to U, \pi_1 : H_1 \to U, \ldots\} | \pi_k(h_{0,k}) \in U(x_k) \text{ for all } h_{0,k} \in H_k, k \geq 0 \Big\}$.

B. Time-Consistent Dynamic Risk Measures

This subsection follows closely the discussion in [15]. Consider a probability space $(\Omega, \mathcal{F}, P)$, a filtration $\mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_N \subset \mathcal{F}$, and an adapted sequence of random variables Z_k, $k \in \{0, \cdots, N\}$. We assume that $\mathcal{F}_0 = \{\Omega, \emptyset\}$, i.e., Z_0 is deterministic. In this paper we interpret the variables Z_k as stage-wise costs. For each $k \in \{1, \cdots, N\}$, define the spaces of random variables with finite pth order moment as $\mathcal{Z}_k := L_p(\Omega, \mathcal{F}_k, P)$, $p \in [1, \infty]$; also, let $\mathcal{Z}_{k,N} := \mathcal{Z}_k \times \cdots \times \mathcal{Z}_N$.

The fundamental question in the theory of dynamic risk measures is the following: how do we evaluate the risk of the subsequence $Z_k, \ldots, Z_N$ from the perspective of stage k? Accordingly, the following definition introduces the concept of dynamic risk measure (here and in the remainder of the paper equalities and inequalities are in the almost sure sense).

Definition II.1 (Dynamic Risk Measure). *A dynamic risk measure is a sequence of mappings $\rho_{k,N} : \mathcal{Z}_{k,N} \to \mathcal{Z}_k$, $k \in$*

$\{0, \ldots, N\}$*, obeying the following monotonicity property:*

$$\rho_{k,N}(Z) \leq \rho_{k,N}(W) \text{ for all } Z, W \in \mathcal{Z}_{k,N} \text{ such that } Z \leq W.$$

The above monotonicity property is arguably a natural requirement for any meaningful dynamic risk measure. Yet, it does not imply the following notion of *time consistency*:

Definition II.2 (Time Consistency). *A dynamic risk measure $\{\rho_{k,N}\}_{k=0}^N$ is called time-consistent if, for all $0 \leq l < k \leq N$ and all sequences $Z, W \in \mathcal{Z}_{l,N}$, the conditions*

$$
\begin{aligned}
& Z_i = W_i, \ i = l, \cdots, k-1, \ and \\
& \rho_{k,N}(Z_k, \cdots, Z_N) \leq \rho_{k,N}(W_k, \cdots, W_N),
\end{aligned}
\tag{1}
$$

imply that

$$\rho_{l,N}(Z_l, \cdots, Z_N) \leq \rho_{l,N}(W_l, \cdots, W_N).$$

In other words, if the Z cost sequence is deemed less risky than the W cost sequence from the perspective of a future time k, and they yield identical costs from the current time l to the future time k, then the Z sequence should be deemed as less risky at the current time l, as well. The pitfalls of time-inconsistent dynamic risk measures have already been mentioned in the introduction and are discussed in detail in [19], [20], [10].

The issue then is what additional "structural" properties are required for a dynamic risk measure to be time consistent. To answer this question we need one more definition:

Definition II.3 (Coherent one-step conditional risk measures). *A coherent one-step conditional risk measures is a mapping $\rho_k : \mathcal{Z}_{k+1} \to \mathcal{Z}_k$, $k \in \{0, \ldots, N\}$, with the following four properties:*

- *Convexity: $\rho_k(\lambda Z + (1 - \lambda)W) \leq \lambda \rho_k(Z) + (1 - \lambda)\rho_k(W)$, $\forall \lambda \in [0, 1]$ and $Z, W \in \mathcal{Z}_{k+1}$;*
- *Monotonicity: if $Z \leq W$ then $\rho_k(Z) \leq \rho_k(W)$, $\forall Z, W \in \mathcal{Z}_{k+1}$;*
- *Translation invariance: $\rho_k(Z+W) = Z+\rho_k(W)$, $\forall Z \in \mathcal{Z}_k$ and $W \in \mathcal{Z}_{k+1}$;*
- *Positive homogeneity: $\rho_k(\lambda Z) = \lambda \rho_k(Z)$, $\forall Z \in \mathcal{Z}_{k+1}$ and $\lambda \geq 0$.*

We are now in a position to state the main result of this section.

Theorem II.4 (Dynamic, time-consistent risk measures). *Consider, for each $k \in \{0, \cdots, N\}$, the mappings $\rho_{k,N} : \mathcal{Z}_{k,N} \to \mathcal{Z}_k$ defined as*

$$
\begin{aligned}
\rho_{k,N} = Z_k + \rho_k(Z_{k+1} + \rho_{k+1}(Z_{k+2} + \ldots + \\
\rho_{N-2}(Z_{N-1} + \rho_{N-1}(Z_N))\ldots)),
\end{aligned}
\tag{2}
$$

where the ρ_k's are coherent one-step risk measures. Then, the ensemble of such mappings is a time-consistent dynamic risk measure.

Proof. See [15]. □

Remarkably, Theorem 1 in [15] shows (under weak assumptions) that the "multi-stage composition" in equation (2) is indeed *necessary for time consistency*. Accordingly, in

the remainder of this paper, we will focus on the *dynamic, time-consistent risk measures* characterized in Theorem II.4.

With dynamic, time-consistent risk measures, since at stage k the value of ρ_k is $\mathcal{F}_k$-measurable, the evaluation of risk can depend on the whole past (even though in a time-consistent way). On the one hand, this generality appears to be of little value in most practical cases, on the other hand, it leads to optimization problems that are intractable from a computational standpoint (and, in particular, do not allow for a dynamic programing solution). For these reasons, in this paper we consider a (slight) refinement of the concept of dynamic, time-consistent risk measure, which involves the addition of a Markovian structure [15].

Definition II.5 (Markov dynamic risk measures). *Let $\mathcal{V} := L_p(S, \mathcal{B}, P)$ be the space of random variables on S with finite pth moment. Given a controlled Markov process $\{x_k\}$, a Markov dynamic risk measure is a dynamic, time-consistent risk measure if each coherent one-step risk measure $\rho_k : \mathcal{Z}_{k+1} \to \mathcal{Z}_k$ in equation (2) can be written as:*

$$\rho_k(V(x_{k+1})) = \sigma_k(V(x_{k+1}), x_k, Q(x_{k+1}|x_k, u_k)), \quad (3)$$

for all $V(x_{k+1}) \in \mathcal{V}$ and $u \in U(x_k)$, where σ_k is a coherent one-step risk measure on $\mathcal{V}$ (with the additional technical property that for every $V(x_{k+1}) \in \mathcal{V}$ and $u \in U(x_k)$ the function $x_k \mapsto \sigma_k(V(x_{k+1}), x_k, Q(x_{k+1}|x_k, u_k))$ is an element of $\mathcal{V}$).

In other words, in a Markov dynamic risk measures, the evaluation of risk is not allowed to depend on the whole past.

Example II.6. *An important example of coherent one-step risk measure satisfying the requirements for Markov dynamic risk measures (Definition II.5) is the mean-semideviation risk function:*

$$\rho_k(V) = \mathbb{E}[V] + \lambda \left(\mathbb{E}\left[[V - \mathbb{E}[V]]_+^p \right] \right)^{1/p}, \quad (4)$$

where $p \in [1, \infty)$, $[z]_+^p := (\max(z, 0))^p$, and $\lambda \in [0, 1]$.

Other important examples include the conditional average value at risk and, of course, the risk-neutral expectation [15].

Accordingly, in the remainder of this paper we will restrict our analysis to Markov dynamic risk measures.

III. PROBLEM STATEMENT

In this section we formally state the problem we wish to solve. Consider an MDP and let $c : S \times U \to \mathbb{R}$ and $d : S \times U \to \mathbb{R}$ be functions which denote costs associated with state-action pairs. Given a policy $\pi \in \Pi$, an initial state $x_0 \in S$, and an horizon $N \geq 1$, the cost function is defined as

$$J_N^\pi(x_0) := \mathbb{E}\left[\sum_{k=0}^{N-1} c(x_k, u_k) \right],$$

and the risk constraint is defined as

$$R_N^\pi(x_0) := \rho_{0,N}\Big(d(x_0, u_0), \ldots, d(x_{N-1}, u_{N-1}), 0 \Big),$$

where $\rho_{k,N}(\cdot)$, $k \in \{0, \ldots, N-1\}$, is a Markov dynamic risk measure (for simplicity, we do not consider terminal costs,

even though their inclusion is straightforward). The problem we wish to solve is then as follows:

> **Optimization problem** $\mathcal{OPT}$ — Given an initial state $x_0 \in S$, a time horizon $N \geq 1$, and a risk threshold $r_0 \in \mathbb{R}$, solve
>
> $$\min_{\pi \in \Pi} \quad J_N^\pi(x_0)$$
> $$\text{subject to} \quad R_N^\pi(x_0) \leq r_0.$$

If problem $\mathcal{OPT}$ is not feasible, we say that its value is $\overline{C}$, where $\overline{C}$ is a "large" constant (namely, an upper bound over the N-stage cost). Note that, when the problem is feasible, an optimal policy always exists since the state and control spaces are finite. When $\rho_{0,N}$ is replaced by an expectation, we recover the usual risk-neutral constrained stochastic optimal control problem studied, e.g., in [4], [5]. In the next section we present a dynamic programing approach to solve problem $\mathcal{OPT}$.

IV. A DYNAMIC PROGRAMMING ALGORITHM FOR RISK-CONSTRAINED MULTI-STAGE DECISION-MAKING

In this section we discuss a dynamic programming approach to solve problem $\mathcal{OPT}$. We first characterize the relevant value functions, and then we present the Bellman's equation that such value functions have to satisfy.

A. Value functions

Before defining the value functions we need to define the tail subproblems. For a given $k \in \{0, \ldots, N-1\}$ and a given state $x_k \in S$, we define the *sub-histories* as $h_{k,j} := (x_k, u_k, \ldots, x_j)$ for $j \in \{k, \ldots, N\}$; also, we define the *space of truncated policies* as $\Pi_k := \Big\{ \{\pi_k, \pi_{k+1}, \ldots\} | \pi_j(h_{k,j}) \in U(x_j) \text{ for } j \geq k \Big\}$. For a given stage k and state x_k, the cost of the tail process associated with a policy $\pi \in \Pi_k$ is simply $J_N^\pi(x_k) := \mathbb{E}\left[\sum_{j=k}^{N-1} c(x_j, u_j) \right]$. The risk associated with the tail process is:

$$R_N^\pi(x_k) := \rho_{k,N}\Big(d(x_k, u_k), \ldots, d(x_{N-1}, u_{N-1}), 0 \Big),$$

which is *only* a function of the current state x_k and does *not* depend on the history $h_{0,k}$ that led to x_k. This crucial fact stems from the assumption that $\{\rho_{k,N}\}_{k=0}^{N-1}$ is a Markov dynamic risk measure, and hence the evaluation of risk only depends on the *future* process and on the present state x_k (formally, this can be easily proven by repeatedly applying equation (3)). Hence, the tail subproblems are *completely* specified by the knowledge of x_k and are defined as

$$\min_{\pi \in \Pi_k} \quad J_N^\pi(x_k) \quad (5)$$
$$\text{subject to} \quad R_N^\pi(x_k) \leq r_k(x_k), \quad (6)$$

for a given (undetermined) threshold value $r_k(x_k) \in \mathbb{R}$ (i.e., the tail subproblems are specified up to a threshold value). We are interested in characterizing a "minimal" set of *feasible* thresholds at each step k, i.e., a "minimal" interval of thresholds for which the subproblems are feasible.

The minimum risk-to-go for each state $x_k \in S$ and $k \in \{0, \ldots, N-1\}$ is given by:

$$\underline{R}_N(x_k) := \min_{\pi \in \Pi_k} R_N^\pi(x_k).$$

Since $\{\rho_{k,N}\}_{k=0}^{N-1}$ is a Markov dynamic risk measure, $\underline{R}_N(x)$ can be computed by using a dynamic programming recursion (see Theorem 2 in [15]). The function $\underline{R}_N(x_k)$ is clearly the lowest value for a feasible constraint threshold. To characterize the upper bound, let:

$$\rho_{\max} := \max_{k \in \{0, \ldots, N-1\}} \max_{(x,u) \in S \times U} \rho_k(d(x,u)).$$

By the monotonicity and translation invariance of Markov dynamic risk measures, one can easily show that

$$\max_{\pi \in \Pi_k} R_N^\pi(x_k) \le (N-k)\rho_{\max} := \overline{R}_N.$$

Accordingly, for each $k \in \{0, \ldots, N-1\}$ and $x_k \in S$, we define the set of feasible constraint thresholds as

$$\Phi_k(x_k) := [\underline{R}_N(x_k), \overline{R}_N], \quad \Phi_N(x_N) := \{0\}.$$

(Indeed, thresholds larger than $\overline{R}_N$ would still be feasible, but would be redundant and would increase the complexity of the optimization problem.)

The value functions are then defined as follows:

- If $k < N$ and $r_k \in \Phi_k(x_k)$:

$$V_k(x_k, r_k) = \min_{\pi \in \Pi_k} J_N^\pi(x_k)$$
$$\text{subject to} \quad R_N^\pi(x_k) \le r_k(x_k);$$

  the minimum is well-defined since the state and control spaces are finite.
- iI $k \le N$ and $r_k \notin \Phi_k(x_k)$:

$$V_k(x_k, r_k) = \overline{C};$$

- when $k = N$ and $r_N = 0$:

$$V_N(x_N, r_N) = 0.$$

Clearly, for $k = 0$, we have the definition of problem $\mathcal{OPT}$.

B. Dynamic programming recursion

In this section we prove that the value function can be computed by dynamic programming. Let $B(S)$ denote the space of real-valued bounded functions on S, and $B(S \times \mathbb{R})$ denote the space of real-valued bounded functions on $S \times \mathbb{R}$. For $k \in \{0, \ldots, N-1\}$, we define the dynamic programming operator $T_k[V_k] : B(S \times \mathbb{R}) \mapsto B(S \times \mathbb{R})$ according to the equation:

$$T_k[V_{k+1}](x_k, r_k) := \inf_{(u,r') \in F_k(x_k, r_k)} \left\{ c(x_k, u) + \sum_{x_{k+1} \in S} Q(x_{k+1}|x_k, u) V_{k+1}(x_{k+1}, r'(x_{k+1})) \right\},$$
(7)

where $F_k \subset \mathbb{R} \times B(S)$ is the set of control/threshold *functions*:

$$F_k(x_k, r_k) := \left\{ (u, r') \Big| u \in U(x_k), r'(x') \in \Phi_{k+1}(x') \text{ for } \right.$$
$$\left. \text{all } x' \in S, \text{ and } d(x_k, u) + \rho_k(r'(x_{k+1})) \le r_k \right\}.$$

If $F_k(x_k, r_k) = \emptyset$, then $T_k[V_{k+1}](x_k, r_k) = \overline{C}$.

Note that, for a given state and threshold constraint, set F_k characterizes the set of feasible pairs of actions and subsequent constraint thresholds. Feasible subsequent constraint thresholds are thresholds which if satisfied at the next stage ensure that the current state satisfies the given threshold constraint (see [6] for a similar statement in the risk-neutral case). Also, note that equation (7) involves a functional minimization over the Banach space $B(S)$. Indeed, since S is finite, $B(S)$ is isomorphic with $\mathbb{R}^{|S|}$, hence the minimization in equation (7) can be re-casted as a regular (although possibly large) optimization problem in the Euclidean space. Computational aspects are further discussed in the next section.

We are now in a position to prove the main result of this paper.

Theorem IV.1 (Bellman's equation with risk constraints). *Assume that the infimum in equation* (7) *is attained. Then, for all $k \in \{0, \ldots, N-1\}$, the optimal cost functions satisfy the Bellman's equation:*

$$V_k(x_k, r_k) = T_k[V_{k+1}](x_k, r_k).$$

Proof. The proof style is similar to that of Theorem 3.1 in [4]. The proof consists of two steps. First, we show that $V_k(x_k, r_k) \ge T_k[V_{k+1}](x_k, r_k)$ for all pairs $(x_k, r_k) \in S \times \mathbb{R}$. Second, we show $V_k(x_k, r_k) \le T_k[V_{k+1}](x_k, r_k)$ for all pairs $(x_k, r_k) \in S \times \mathbb{R}$. These two results will prove the claim.

Step (1). If $r_k \notin \Phi_k(x_k)$, then, by definition, $V_k(x_k, r_k) = \overline{C}$. Also, $r_k \notin \Phi_k(x_k)$ implies that $F_k(x_k, r_k)$ is empty (this can be easily proven by contradiction). Hence, $T_k[V_{k+1}](x_k, r_k) = \overline{C}$. Therefore, if $r_k \notin \Phi_k(x_k)$,

$$V_k(x_k, r_k) = \overline{C} = T_k[V_{k+1}](x_k, r_k), \qquad (8)$$

i.e., $V_k(x_k, r_k) \ge T_k[V_{k+1}](x_k, r_k)$.

Assume, now, $r_k \in \Phi_k(x_k)$. Let $\pi^* \in \Pi_k$ be the optimal policy that yields the optimal cost $V_k(x_k, r_k)$. Construct the "truncated" policy $\bar{\pi} \in \Pi_{k+1}$ according to:

$$\bar{\pi}_j(h_{k+1,j}) := \pi_j^*(x_k, \pi_k^*(x_k), h_{k+1,j}), \quad \text{for } j \ge k+1.$$

In other words, $\bar{\pi}$ is a policy in Π_{k+1} that acts as prescribed by π^*. By applying the law of total expectation, we can write:

$$V_k(x_k, r_k) = \mathbb{E}\left[\sum_{j=k}^{N-1} c(x_j, \pi_j^*(h_{k,j})) \right] =$$
$$c(x_k, \pi_k^*(x_k)) + \mathbb{E}\left[\sum_{j=k+1}^{N-1} c(x_j, \pi_j^*(h_{k,j})) \right] =$$
$$c(x_k, \pi_k^*(x_k)) + \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{N-1} c(x_j, \pi_j^*(h_{k,j})) \Big| h_{k,k+1} \right] \right].$$

Note that $\mathbb{E}\left[\sum_{j=k+1}^{N-1} c(x_j, \pi_j^*(h_{k,j})) \,\Big|\, h_{k,k+1}\right] = J_N^{\bar{\pi}}(x_{k+1})$. Clearly, the truncated policy $\bar{\pi}$ is a feasible policy for the tail subproblem

$$\min_{\pi \in \Pi_{k+1}} \quad J_N^{\pi}(x_{k+1})$$
$$\text{subject to} \quad R_N^{\pi}(x_{k+1}) \leq R_N^{\bar{\pi}}(x_{k+1}).$$

Collecting the above results, we can write

$$\begin{aligned} V_k(x_k, r_k) &= c(x_k, \pi_k^*(x_k)) + \mathbb{E}[J_N^{\bar{\pi}}(x_{k+1})] \\ &\geq c(x_k, \pi_k^*(x_k)) + V_{k+1}(x_{k+1}, R_N^{\bar{\pi}}(x_{k+1})) \\ &\geq T_k[V_{k+1}](x_k, r_k), \end{aligned}$$

where the last inequality follows from the fact that $R_N^{\bar{\pi}}(\cdot)$ can be viewed as a valid threshold function in the minimization in equation (7).

Step (2). If $r_k \notin \Phi_k(x_k)$, equation (8) holds and, therefore, $V_k(x_k, r_k) \leq T_k[V_{k+1}](x_k, r_k)$.

Assume $r_k \notin \Phi_k(x_k)$. For a given pair (x_k, r_k), where $r_k \in \Phi_k(x_k)$, let u^* and $r'^{,*}$ the minimizers in equation (7) (here we are exploiting the assumption that the minimization problem in equation (7) admits a minimizer). By definition, $r'^{,*}(x_{k+1}) \in \Phi_{k+1}(x_{k+1})$ for all $x_{k+1} \in S$. Also, let $\pi^* \in \Pi_{k+1}$ the optimal policy for the tail subproblem:

$$\min_{\pi \in \Pi_{k+1}} \quad J_N^{\pi}(x_{k+1})$$
$$\text{subject to} \quad R_N^{\pi}(x_{k+1}) \leq r'^{,*}(x_{k+1}).$$

Construct the "extended" policy $\bar{\pi} \in \Pi_k$ as follows:

$$\bar{\pi}_k(x_k) = u^*, \text{ and } \bar{\pi}_j(h_{k,j}) = \pi_j^*(h_{k+1,j}) \text{ for } j \geq k+1.$$

Since π^* is an optimal, and a fortiori feasible, policy for the tail subproblem (from stage $k+1$) with threshold function $r'^{,*}$, the policy $\bar{\pi} \in \Pi_k$ is a feasible policy for the tail subproblem (from stage k):

$$\min_{\pi \in \Pi_k} \quad J_N^{\pi}(x_k)$$
$$\text{subject to} \quad R_N^{\pi}(x_k) \leq r_k.$$

Hence, we can write

$$V_k(x_k, r_k) \leq J_N^{\bar{\pi}}(x_k) =$$
$$c(x_k, \bar{\pi}_k(x_k)) + \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{N-1} c(x_j, \bar{\pi}_j(h_{k,j})) \,\Big|\, h_{k,k+1}\right]\right].$$

Note that $\mathbb{E}\left[\sum_{j=k+1}^{N-1} c(x_j, \bar{\pi}_j(h_{k,j})) \,\Big|\, h_{k,k+1}\right] = J_N^{\pi^*}(x_{k+1})$. Hence, from the definition of π^*, one easily obtains:

$$V_k(x_k, r_k) \leq c(x_k, \bar{\pi}_k(x_k)) + \mathbb{E}\left[J_N^{\pi^*}(x_{k+1})\right] = c(x_k, u^*) +$$
$$\sum_{x_{k+1} \in S} Q(x_{k+1}|x_k, u^*) V_{k+1}(x_{k+1}, r'^{,*}(x_{k+1})) =$$
$$T_k[V_{k+1}](x_k, r_k).$$

Collecting the above results, the claim follows. $\qquad\square$

Remark IV.2 (On the assumption in Theorem IV.1). *In Theorem IV.1 we assume that the infimum in equation (7) is attained. This is indeed true under very weak conditions (namely that $U(x_k)$ is a compact set, $\sigma_k(\nu(x_{k+1}), x_{k+1}, Q)$*

is a lower semi-continuous function in Q, $Q(x_k, u_k)$ is continuous in u_k and the stage-wise cost c and d are lower semi-continuous in u_k). The proof of this statement is omitted in the interest of brevity and is left for a forthcoming publication.

V. DISCUSSION

In this section we present possible generalizations, we show how to construct optimal policies, we discuss computational aspects, and we present a simple two-state example for machine repairing.

A. Generalizations

1) Time-varying setup: One can readily extends the results of this paper to the case where the state space, the control space, the transition probalities and the stage costs are a function of time, i.e., S_k, U_k, $Q_k(\cdot|x_k, u_k)$, c_k, and d_k. Besides keeping track of the stage index, the definition in Section IV and the proof of Theorem IV.1 would be virtually identical.

2) Risk-sensitive objectives and multiple risk constraints: In some scenarios, one might be interested in minimizing a risk-sensitive objective, i.e., in applying a dynamic, time-consistent risk measure the cost objective. The optimization problem would then read as follows:

> **Optimization problem $\mathcal{OPT}^{\text{rs}}$** — Given an initial state $x_0 \in S$, a time horizon $N \geq 1$, and a risk threshold $r_0 \in \mathbb{R}$, solve
> $$\min_{\pi \in \Pi} \quad \rho_{0,N}^{\text{Obj}}\Big(c(x_0, u_0), \ldots, c(x_{N-1}, u_{N-1}), 0\Big)$$
> subject to $\quad \rho_{0,N}\Big(d(x_0, u_0), \ldots, d(x_{N-1}, u_{N-1}), 0\Big) \leq r_0,$

where $\rho_{0,N}^{\text{Obj}}$ is a Markov dynamic risk measure for the objective function. Note that without the risk constraint the problem reduces to the risk-sensitive stochastic optimal control problem studied in [15]. One can state for $\mathcal{OPT}^{\text{rs}}$ a theorem virtually identical to Theorem IV.1; the proof of Theorem IV.1 carries over almost unaltered to this case thanks to the interchangeability Lemma of coherent risk measures (refer to [21] Proposition 6.47).

One can further generalize problem $\mathcal{OPT}^{\text{rs, multi}}$ by including multiple risk constraints. The problem would then read as follows:

> **Optimization problem $\mathcal{OPT}^{\text{rs,multi}}$** — Given an initial state $x_0 \in S$, a time horizon $N \geq 1$, risk thresholds $r_0^j \in \mathbb{R}$, $j \in \{1, \ldots, M\}$, solve
> $$\min_{\pi \in \Pi} \quad \rho_{0,N}^{\text{Obj}}\Big(c(x_0, u_0), \ldots, c(x_{N-1}, u_{N-1}), 0\Big)$$
> subject to $\quad \rho_{a_j, b_j}^j\Big(d(x_{a_j}, u_{a_j}), \ldots, d(x_{b_j}, u_{b_j}), 0\Big) \leq r_0^j,$

where $j \in \{1, \ldots, M\}$, $a_j, b_j \in \{0, \ldots, N-1\}$, $a_j < b_j$, and the risk metrics ρ_{a_j, b_j}^j are Markov dynamic risk measures. Problem $\mathcal{OPT}^{\text{rs, multi}}$ can be again solved via a Bellman equation, however the expression for the dynamic programing operator becomes quite cumbersome. Details are left for a future publication.

B. Construction of optimal policies

Under the assumption of Theorem IV.1, optimal control policies can be constructed as follows. For any given $x_k \in S$ and $r_k \in \Phi_k(x_k)$, let $u^*(x_k, r_k)$ and $r'(x_k, r_k)(\cdot)$ be the minimizers in equation (7) (recall that r' is a function).

Theorem V.1 (Optimal policies). *Assume that the infimum in equation (7) is attained. Let $\pi \in \Pi$ be a policy recursively defined as follows:*

$$\pi_k(h_{0,k}) = u^*(x_k, r_k) \text{ with } r_k = r'(x_{k-1}, r_{k-1})(x_k),$$

when $k \in \{1, \ldots, N-1\}$, and

$$\pi(x_0) = u^*(x_0, r_0),$$

for a given threshold $r_0 \in \Phi_0(x_0)$. Then, π is an optimal policy for problem $\mathcal{OPT}$ with initial condition x_0 and constraint threshold r_0.

Proof. As usual for dynamic programming problems, the proof uses induction arguments (see, in particular, [22] and [6, Theorem 4] for a similar proof in the risk-neutral case).

Consider a tail subproblem starting at stage k, for $k = 0, \ldots, N-1$; for a given initial state $x_k \in S$ and constraint threshold $r_k \in \Phi_k(x_k)$, let $\pi^{k,r_k} \in \Pi_k$ be a policy recursively defined as follows:

$$\pi_j^{k,r_k}(h_{k,j}) = u^*(x_j, r_j) \text{ with } r_j = r'(x_{j-1}, r_{j-1})(x_j),$$

when $j \in \{k+1, \ldots, N-1\}$, and

$$\pi_k^{k,r_k}(x_k) = u^*(x_k, r_k).$$

We prove by induction that π^{k,r_k} is optimal. Clearly, for $k = 0$, such result implies the claim of the theorem.

Let $k = N-1$ (base case). In this case the tail subproblem is:

$$\min_{\pi \in \Pi_{N-1}} \quad c(x_{N-1}, \pi(x_{N-1}))$$

$$\text{subject to} \quad d(x_{N-1}, \pi(x_{N-1})) \le r_{N-1}.$$

Since, by definition, $r'(x_N)$ and $V_N(x_N, r_N)$ are identically equal to zero, and due to the positive homogeneity of one-step conditional risk measures, the above tail subproblem is identical to the optimization problem in the Bellman's recursion (7), hence $\pi^{N-1, r_{N-1}}$ is optimal.

Assume as induction step that $\pi^{k+1, r_{k+1}}$ is optimal for the tail subproblems starting at stage $k+1$ with $x_{k+1} \in S$ and $r_{k+1} \in \Phi_{k+1}(x_{k+1})$. We want to prove that π^{k,r_k} is optimal for the tail subproblems starting at stage k with initial state $x_k \in S$ and constraint threshold $r_k \in \Phi_k(x_k)$. First, we prove that π^{k,r_k} is a feasible control policy. Note that, from the recursive definitions of π^{k,r_k} and $\pi^{k+1, r_{k+1}}$, one has

$$R_N^{\pi^{k,r_k}}(x_{k+1}) = R_N^{\pi^{k+1, r'(x_k, r_k)(x_{k+1})}}(x_{k+1}).$$

Hence, one can write:

$$R_N^{\pi^{k,r_k}}(x_k) = d(x_k, u^*(x_k, r_k)) + \rho_k(R_N^{\pi^{k,r_k}}(x_{k+1})) =$$
$$d(x_k, u^*(x_k, r_k)) + \rho_k\left(R_N^{\pi^{k+1, r'(x_k, r_k)(x_{k+1})}}(x_{k+1})\right) \le$$
$$d(x_k, u^*(x_k, r_k)) + \rho_k\left(r'(x_k, r_k)(x_{k+1})\right) \le r_k, \tag{9}$$

where the first inequality follows from the inductive step and the monotonicity of coherent one-step conditional risk measures, and the last step follows from the definition of u^* and r'. Hence, π^{k,r_k} is a feasible control policy (assuming initial state $x_k \in S$ and constraint threshold $r_k \in \Phi_k(x_k)$). As for its cost, one has, similarly as before,

$$J_N^{\pi^{k,r_k}}(x_{k+1}) = J_N^{\pi^{k+1, r'(x_k, r_k)(x_{k+1})}}(x_{k+1}).$$

Then, one can write:

$$J_N^{\pi^{k,r_k}}(x_k) = c(x_k, u^*(x_k, r_k)) + \mathbb{E}\left[J_N^{\pi^{k,r_k}}(x_{k+1})\right] =$$
$$c(x_k, u^*(x_k, r_k)) + \mathbb{E}\left[J_N^{\pi^{k+1, r'(x_k, r_k)(x_{k+1})}}(x_{k+1})\right] =$$
$$c(x_k, u^*(x_k, r_k)) + \mathbb{E}\left[V_{k+1}(x_{k+1}), r'(x_k, r_k)(x_{k+1})\right] =$$
$$T_k[V_{k+1}](x_k, r_k) = V_k(x_k, r_k), \tag{10}$$

where the third equality follows from the inductive step, the fourth equality follows form the definition of the dynamic programming operator in equation (7), and the last equality follows from Theorem IV.1. Since policy π^{k,r_k} is feasible and achieves the optimal cost, it is optimal. This concludes the proof. $\square$

Note that the optimal policy in the statement of Theorem V.1 can be written in "compact" form without the aid of the extra variable r_k. Indeed, for $k = 1$, by defining the threshold transition function $\mathcal{R}_1(h_{0,1}) := r'(x_0, r_0)(x_1)$, one can write $r_1 = \mathcal{R}_1(h_{0,1})$. Then, by induction arguments, one can write, for any $k \in \{1, \ldots, N\}$, $r_k = \mathcal{R}_k(h_{0,k})$, where $\mathcal{R}_k$ is the threshold transition function at stage k. Therefore, the optimal policy in the statement of Theorem V.1 can be written as $\pi(h_{0,k}) = u^*(x_k, \mathcal{R}_k(h_{0,k}))$, which makes explicit the dependency of π over the process history.

Interestingly, if one views the constraint thresholds as state variables, the optimal policies of problem $\mathcal{OPT}$ have a Markovian structure with respect to the augmented control problem.

C. Computational issues

In our approach, the solution of problem $\mathcal{OPT}$ entails the solution of two dynamic programing problems, the first one to find the lower bound for the set of feasible constraint thresholds (i.e., the function $\underline{R}(x)$, see Section IV), and the second one to compute the value functions $V_k(x_k, r_k)$. The latter problem is the most challenging one since it involves a functional minimization. However, as already noted, since S is finite, $B(S)$ is isomorphic with $\mathbb{R}^{|S|}$, and the functional minimization in the Bellman operator (7) can be re-casted as an optimization problem in the Euclidean space. This problem, however, can be large and, in general, is not convex. However, assume that function $d(x, u)$ is convex. Since in a Markov dynamic risk measure the risk mappings ρ_k are convex functions, one can show that the constraint set $F_k(x_k, r_k)$ is a convex set. Furthermore, by assuming that the state transition for the system trajectories $\{x_k\}$ is affine in the control u, the value functions V_k are convex functions and the minimization in the Bellman's equation (7) is a

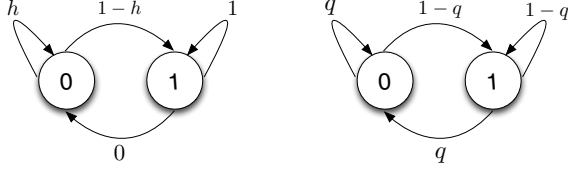convex optimization problem. The details will be given in a forthcoming paper.

Fig. 1. Left figure: transition probabilities for control $u = 0$. Right figure: transition probabilities for control $u = 1$. Circles represent states. The transition probabilities satisfy $1 \geq q > h \geq 0$.

D. System maintenance example

Finally, we illustrate the above concepts with a simple two-stage (i.e., $N = 2$) example that represents the problem of scheduling maintenance operations for a given system. The state space is given by $S = \{0, 1\}$, where $\{0\}$ represents a normal state and $\{1\}$ represents a failure state; the control space is given by $U = \{0, 1\}$, where $\{0\}$ means "do nothing" and $\{1\}$ means "perform maintenance". The transition probabilities are given in Figure 1 for some $1 \geq q > h \geq 0$. Also, the cost functions and the constraint cost functions are as follows:

$$c(0,0) = c(1,0) = 0, \quad c(0,1) = c(1,1) = c_2,$$
$$d(0,1) = d(0,0) = 0, \quad d(1,0) = d(1,1) = c_1 \in (0,1).$$

The terminal costs are zero. The one-step conditional risk measures is the mean semi-deviation (see equation (4)) with fixed $\lambda \in [0,1]$ and $p \in [1, \infty)$. We wish to solve problem $\mathcal{OPT}$ for this example.

Note that, for any λ and p, function

$$f(x) := \lambda x (1-x)^{1/p} + (1-x)$$

is a non-increasing function in $x \in [0,1]$. Therefore, $f(q) \leq f(p) \leq f(0)$. At stage $k = 2$, $V_2(1, r_2) = V_2(0, r_2) = 0$, and $\Phi_2(1) = \Phi_2(0) = \{0\}$. At stage $k = 1$,

$$V_1(0, r_1) = \begin{cases} 0 & \text{if } r_1 \geq 0, \\ \bar{C} & \text{else.} \end{cases}$$
$$V_1(1, r_1) = \begin{cases} 0 & \text{if } r_1 \geq c_1, \\ \bar{C} & \text{else.} \end{cases}$$

Also, $\Phi_1(0) = [0, \infty)$ and $\Phi_1(1) = [c_1, \infty)$. At stage $k = 0$, define $K^{(x)} := f(x)c_1$ (hence $K^{(0)} = c_1$) and

$$E_x(r'(0), r'(1)) := r'(0)x + r'(1)(1-x)$$
$$M_x(r'(0), r'(1)) := \left(\begin{array}{c} (1-x)[r'(1) - E_x(r'(0), r'(1))]_+^p \\ +x[r'(0) - E_x(r'(0), r'(1))]_+^p \end{array} \right)^{1/p};$$

hence, $E_0(r'(0), r'(1)) = r'(1)$ and $M_0(r'(0), r'(1)) = 0$.

Then, we can write

$$F_0(0, r_0) = \emptyset \quad \text{if } r_0 < K^{(q)}$$
$$F_0(0, r_0) = \{(1, r') : r'(0) \in [0, \infty), r'(1) \in [c_1, \infty),$$
$$E_q(r'(0), r'(1)) + \lambda M_q(r'(0), r'(1)) \leq r_0\}$$
$$\text{if } K^{(q)} \leq r_0 < K^{(h)}$$
$$F_0(0, r_0) = \{(1, r') : r'(0) \in [0, \infty), r'(1) \in [c_1, \infty),$$
$$E_q(r'(0), r'(1)) + \lambda M_q(r'(0), r'(1)) \leq r_0\}$$
$$\bigcup \{(0, r') : r'(0) \in [0, \infty), r'(1) \in [c_1, \infty),$$
$$E_h(r'(0), r'(1)) + \lambda M_h(r'(0), r'(1)) \leq r_0\}$$
$$\text{if } r_0 \geq K^{(h)}$$

$$F_0(1, r_0) = \emptyset \quad \text{if } r_0 < c_1 + K^{(q)}$$
$$F_0(1, r_0) = \{(1, r') : r'(0) \in [0, \infty), r'(1) \in [c_1, \infty),$$
$$c_1 + E_q(r'(0), r'(1)) + \lambda M_q(r'(0), r'(1)) \leq r_0\}$$
$$\text{if } c_1 + K^{(q)} \leq r_0 < c_1 + K^{(0)}$$
$$F_0(1, r_0) = \{(1, r') : r'(0) \in [0, \infty), r'(1) \in [c_1, \infty),$$
$$c_1 + E_q(r'(0), r'(1)) + \lambda M_q(r'(0), r'(1)) \leq r_0\}$$
$$\bigcup \{(0, r') : r'(0) \in [0, \infty), r'(1) \in [c_1, \infty),$$
$$c_1 + r'(1) \leq r_0\}$$
$$\text{if } r_0 \geq c_1 + K^{(0)}$$

As a consequence,

$$V_0(1, r_0) = \begin{cases} \bar{C} & \text{if } r_0 < c_1 + K^{(q)} \\ c_2 & \text{if } c_1 + K^{(q)} \leq r_0 < c_1 + K^{(0)} \\ 0 & \text{if } r_0 \geq K^{(0)} \end{cases}$$

$$V_0(0, r_0) = \begin{cases} \bar{C} & \text{if } r_0 < K^{(q)} \\ c_2 & \text{if } K^{(q)} \leq r_0 < K^{(h)} \\ 0 & \text{if } r_0 \geq K^{(h)} \end{cases}$$

Therefore, for $V_0(1, c_1 + K^{(q)})$, the infimum of the Bellman's equation is attained with $u = 1$, $r'(0) = 0$, $r'(1) = c_1$. For $V_0(0, K^{(h)})$, the infimum of the Bellman's equation is attained with $u = 0$, $r'(0) = 0$, $r'(1) = c_1$. Note that, as expected, the value function is a decreasing function with respect to the risk threshold.

VI. CONCLUSIONS

In this paper we have presented a dynamic programing approach to stochastic optimal control problems with dynamic, time-consistent (in particular Markov) risk constraints. We have shown that the optimal cost functions can be computed by value iteration and that the optimal control policies can be constructed recursively. This paper leaves numerous important extensions open for further research. First, it is of interest to study how to carry out the Bellman's equation efficiently; a possible strategy involving convex programming has been briefly discussed. Second, to address problems with large state spaces, we plan to develop approximate dynamic programing algorithms for problem $\mathcal{OPT}$. Third, it is of both theoretical and practical interest to study the relation between

stochastic optimal control problems with time-consistent and time-inconsistent constraints, e.g., in terms of the optimal costs. Fourth, we plan to extend our approach to the case with partial observations and an infinite horizon. Finally, we plan to apply our approach to real settings, e.g., to the architectural analysis of planetary missions or to the risk-averse optimization of multi-period investment strategies.

REFERENCES

[1] E. Altman. *Constrained Markov Decision Processes*. Boca Raton, FL: Chapman & Hall/CRC, 1999.

[2] Y. A. Korilis and A. A. Lazar. On the existence of equilibria in noncooperative optimal flow control. *J. ACM*, 42(3):584–613, May 1995.

[3] P. Nain and K. Ross. Optimal priority assignment with hard constraint. *Automatic Control, IEEE Transactions on*, 31(10):883 – 888, oct 1986.

[4] R. Chen and G. Blankenship. Dynamic programming equations for discounted constrained stochastic control. *IEEE Transaciton of Automatic Control*, 49(5):699–709, 2004.

[5] A. Piunovskiy. Dynamic programming in constrained markov decision process. *Control and Cybernetics*, 35(3):646–660, 2006.

[6] R. Chen and E. Feinberg. Non-randomized policies for constrained markov decision process. *Mathematical Methods in Operations Research*, 66:165–179, 2007.

[7] M. Pavone Y. Kuwata and J. Balaram. A risk-constrained multi-stage decision making approach to the architectural analysis of mars missions. In *IEEE Conference on Decision and Control*, 2012.

[8] M. Sniedovich. A Variance-Constrained Reservoir Control Problem. *Water Resources Research*, 16:271–274, 1980.

[9] S. Mannor and J. N. Tsitsiklis. Mean-Variance Optimization in Markov Decision Processes. In *International Conference on Machine Learning*, 2011.

[10] P. Huang, D. A. Iancu, M. Petrik, and D. Subramanian. The Price of Dynamic Inconsistency for Distortion Risk Measures. *ArXiv e-prints*, June 2011.

[11] B. Rudloff, A. Street, and D. Valladao. Time consistency and risk averse dynamic decision models: Interpretation and practical consequences, 2011.

[12] A. Ruszczynski and A. Shapiro. Optimization of risk measures. Risk and Insurance 0407002, EconWPA, 2004.

[13] A. Ruszczynski and A. Shapiro. Conditional risk mappings. *Mathematics of operations research*, 21(3):544–561, 2006.

[14] A. Ruszczynski and A. Shapiro. Optimization of convex risk functinos. *Mathematics of operations research*, 31(3):433–452, 2006.

[15] A. Ruszczynski. Risk averse dynamic programming for markov decision process. *Journal of Mathematical Programming*, 125(2):235–261, 2010.

[16] A. Shapiro. Minimax and risk averse multistage stochastic programming. *European Journal of Operational Research*, 219(3):719–726, 2012.

[17] P. Cheridito and M. Kupper. Composition of time consistent dynamic monetary risk measures in discrete time. *International Journal of Theoretical and Applied Finance*, 14(1):137–162, 2011.

[18] H. Föllmer and I. Penner. Convex risk measures and the dynamics of their penalty functions. *Statistics & Decisions*, 24(1):61–96, 2012/09/15 2006.

[19] P. Cheridito and M. Stadjie. Time inconsistency of var and time-consistent alternatives. *Finance Research Letters.*, 6(1):40–46, 2009.

[20] A. Shapiro. On a time consistency concept in risk averse multi-stage stochastic programming. *Operations Research Letters*, 37(3):143–147, 2009.

[21] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[22] D. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific, 2005.