Dominance-constrained Markov decision processes

William B. Haskell and Rahul Jain
(ISE, EE & ISE) Departments
University of Southern California
(haskell,rahul.jain)@usc.edu

Abstract—We are interested in risk constraints for discrete time Markov decision processes (MDPs). Starting with the average reward case, we argue that stochastic dominance constraints are natural risk constraints for MDPs. Specifically, we constrain the empirical distribution of reward to dominate a benchmark distribution in the increasing concave stochastic order. We argue that the optimal policy for the dominance-constrained MDP is a stationary randomized policy. Further, the optimal policy can be computed from a linear program in the space of occupation measures, where the dominance constraint is represented by linear inequalities. The dual of this linear program is computed and is shown to be close to the usual linear programming form of the average dynamic program.

The set of feasible state-action pairs is written as
$$\Psi = \{(s, a) \in S \times A : a \in A_s\}.$$ 

The state and action at time $t$ are written $s_t$ and $a_t$, respectively. The stationary transition kernel is $P(j|s, a)$ which gives the probability of visiting state $j \in S$ given the state-action pair $(s, a)$.

Let $H_t$ be the set of admissible histories at time $t$, $H_0 = S$ and $H_t = H_{t-1} \times A \times S$ for all $t \geq 1$. A particular history is written $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t) \in H_t$. Let $\Pi^{HR}$ be the set of history-dependent randomized policies, mappings $\pi : H_t \to \mathcal{P}(A)$, where $\mathcal{P}(A)$ is the set of probability measures on $A$. The class of stationary randomized (Markov) policies is written $\Pi^{SR}$. $\pi \in \Pi^{SR}$ only depends on history through the current state, $\pi : S \to \mathcal{P}(A)$.

Let $r : \Psi \to \mathbb{R}$ be a reward function. The classic average reward maximization problem is
$$\max_{\pi \in \Pi^{HR}} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} r(s_t, a_t) \right]. \quad (1)$$

It is known that a policy in $\Pi^{SR}$ is optimal for problem (1) under suitable conditions. [14] develops the theory of Markov decision processes including problem (1) in great detail.

Convex analytic methods are well studied for problem (1) with many types of constraints. [4], [5] discusses a rigorous theory of convex optimization for constrained MDPs with general state and action spaces. [1] focuses on constrained MDP. [11] present a monograph on linear programming methods for MDPs in Borel spaces. The connection to linear programming in infinite-dimensional spaces, and infinite-dimensional linear programming duality, is emphasized. MDPs with expected constraints and pathwise constraints, also called hard constraints, are considered in [13] using convex analytic methods. An inventory system is detailed to motivate the theoretical results.

I. INTRODUCTION

Markov decision processes (MDP) are a natural and powerful framework for stochastic control. In the present paper, we focus on formulating and solving MDPs with risk constraints. Convex analytic methods have been applied to MDPs to create optimization problems in a space of occupation measures, and these methods have been successful at handling many types of constraints (see [4], [5], [11], [13]). Our specific goal is to find risk constraints for MDPs that are amenable to convex analytic methods. It turns out that stochastic dominance constraints for MDPs are readily handled by the existing theory. In particular, stochastic dominance constraints are equivalent to a continuum of linear constraints on occupation measures, as we will see.

We begin with a general setting for a discrete time MDP. Let $S$ be a finite state space and let $A$ be a finite set of actions. We focus on finite state and action spaces to emphasize risk management instead of the technical intricacies of general state and action spaces. Let $A_s \subset A$ be the set of admissible actions for each state $s \in S$. The state and action at time $t$ are written $s_t$ and $a_t$, respectively. The stationary transition kernel is $P(j|s, a)$ which gives the probability of visiting state $j \in S$ given the state-action pair $(s, a)$.
Risk management for MDPs has been considered from many perspectives in the literature. [8] studies variance penalties on the rewards in MDPs. The optimal policy is obtained by solving a nonlinear program in the occupation measure. In [16], the mean-variance trade-off in MDPs is further explored in a Pareto optimality sense. In [9], the risk of a policy in an MDP is evaluated in terms of its probability of visiting certain undesirable states called “error states”. Specifically, risk is determined by the probability of entering the error states. A learning algorithm is developed to obtain good stationary policies.

Infinite horizon discounted MDP with expected utility constraints are examined in [12]. A constrained optimal policy is shown to exist using a Lagrangian saddle-point approach. The CVaR of the total cost in a finite horizon MDP is constrained in [3]. It is argued that convex analytic methods do not apply to this problem type and an offline iterative algorithm is employed to solve for the optimal policy. [15] develops Markov risk measures for finite horizon and infinite horizon discounted MDP. Dynamic programming equations are derived for both models that reflect the risk aversion, and policy iteration is shown to solve the infinite horizon problem.

Our notion of risk-constrained MDPs differs from this literature survey. In the present paper, we are interested in the empirical distribution of reward, rather than in its expectation, variance, or other summary statistics. We make two main contributions. First, we show that stochastic dominance constraints readily mesh with the constrained MDP theory. More immediately, we show that dominance-constrained MDPs can be solved with linear programming problems over occupation measures. Second, we compute the duals of the preceding linear programs to gain more insight. These duals are close to the linear programming form of the dynamic programming optimality equations. New pricing decision variables appear in the dual in an intuitive way, corresponding to the dominance constraint.

This paper is organized as follows. In section 2, we consider stochastic dominance constraints for average reward infinite horizon MDP. We formalize our model, and then show that the optimal policy can be computed from a linear program. In section 3, we carry out a parallel development for discounted reward infinite horizon MDP with stochastic dominance constraints. Section 4 brings multivariate stochastic dominance into play using the same toolset from the preceding sections. The paper concludes with a discussion of numerous future research directions in section 5.

II. AVERAGE REWARD

In this section, we will consider stochastic dominance constraints in problem (1). To begin, let \( q : \Psi \rightarrow \mathbb{R} \) be an additional reward function that is possibly different from the function \( r \) in the objective of problem (1). We are interested in the empirical distribution of reward \( q \),

\[
\limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} q(s_t, a_t).
\]

Notice that an expectation does not appear in the preceding expression, it is a random variable.

We would like to put a risk constraint on the empirical distribution \( \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} q(s_t, a_t) \). For this purpose, recall the definition of the increasing concave stochastic order. For random variables \( X, Y \in \mathbb{R} \), \( X \) dominates \( Y \) in the increasing concave (icv) stochastic order if \( \mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)] \) for all increasing concave functions \( u : \mathbb{R} \rightarrow \mathbb{R} \) such that both expectations exist. Let \( Y \) be a fixed reference random variable to act as a benchmark for the empirical distribution of rewards. The resulting dominance constrained MDP is

\[
\max_{\pi \in \mathcal{H}_R} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[ \sum_{t=0}^{T-1} r(s_t, a_t) \right] \quad \text{s.t.} \quad \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} q(s_t, a_t) \geq_{icv} \mathbb{E}[Y].
\]

We will approach problem (2) - (3) by showing that a stationary randomized policy is optimal, and that such a policy can be obtained from a linear program.

A. Optimal policy

In terms of occupation measures, stochastic dominance is the most immediate type of risk constraint. Let \( x \) denote a long-run average occupation measure on \( \Psi \), and \( \mathbb{E}_x \{ \cdot \} \) denote expectation with respect to \( x \), i.e.:

\[
\mathbb{E}_x[f(\bar{s}, \bar{a})] = \sum_{(s,a) \in \Psi} x(s,a) f(s,a).
\]

It is clear that the expectation of any function with respect to \( x \) is a linear function in \( x \). In our case, we will show that constraint (3) can be written in terms of the occupation measure \( x \) as

\[
\mathbb{E}_x[(q(s,a) - \eta)_+] \geq \mathbb{E}[(Y - \eta)_+], \quad \forall \eta \in \mathbb{R},
\]

where \((x)_- = \min \{0, x\}\); where each function

\[
\mathbb{E}_x[(q(s,a) - \eta)_+] = \sum_{(s,a) \in \Psi} (q(s,a) - \eta)_+ x(s,a)
\]
is linear in \( x \). We will assume that the benchmark \( Y \) has finite support \( \mathcal{Y} \subset \mathbb{R} \) throughout so that (3) is equivalent to finitely many linear inequalities
\[
\mathbb{E}_x \left[ (q(s,a) - \eta) \right] \geq \mathbb{E} \left[ (Y - \eta) \right], \quad \forall \eta \in \mathcal{Y}, \quad (5)
\]
by [6, Proposition 3.2].
Problem (2) - (3) can be written in terms of occupation measures as
\[
\max_{x \geq 0} \quad \sum_{(s,a) \in \Psi} r(s,a) x(s,a) \quad (6)
\]
s.t. \[
\sum_{a \in A_s} x(j,a) = \sum_{s \in S} \sum_{a \in A_s} P(j \mid s, a) x(s,a), \quad \forall j \in S, \quad (7)
\]
\[
\sum_{(s,a) \in \Psi} x(s,a) = 1, \quad (8)
\]
\[
\mathbb{E}_x \left[ (q(s,a) - \eta) \right] \geq \mathbb{E} \left[ (Y - \eta) \right], \quad \forall \eta \in \mathcal{Y}, \quad (9)
\]
For emphasis, problem (6) - (9) is a linear program (LP).
We have the following result which justifies searching for policies in \( \Pi^{SR} \), and thus justifies problem (6) - (9).

**Theorem II.1.** Suppose the underlying Markov chain is unichain.
(a) If \( x^* \) is an optimal solution to problem (6) - (9), then there exists an optimal policy \( \pi^* \in \Pi^{SR} \) for problem (2) - (3) with
\[
\pi^*(a \mid s) = \frac{x^*(s,a)}{\sum_{k \in A_s} x^*(s,k)}
\]
for \( \sum_{k \in A_s} x^*(s,k) > 0 \) and \( \pi^*(a \mid s) \) is arbitrary otherwise.
(b) If \( \pi^* \) is optimal for problem (2) - (3), then there exists a \( \hat{\pi} \in \Pi^{SR} \) which is also optimal and induces an optimal solution to problem (6) - (9).

**Proof:** Follows from the proof of [14, Theorem 8.9.6], since the constraints (9) are just linear constraints on \( x \).
(a) Let \( \mathcal{X} \) be the set of all \( x \) which satisfy (6) - (9).
By [14, Corollary 8.9.5], since \( x^* \in \mathcal{X} \) there is a policy \( \pi^* \in \Pi^{SR} = \Pi^{HR} \) that has limiting distribution \( x^* \). Since \( x^* \) satisfies (9), it follows that \( \pi^* \) is optimal for problem (2) - (3). Then, [14, Theorem 8.8.6] guarantees that \( \pi^* \) has the form given above.
(b) Since \( \pi^* \in \Pi^{HR} \), [14, Corollary 8.9.5] gives the existence of a \( \pi \in \Pi^{SR} \) with the same limiting distribution. It follows that the corresponding steady state distribution \( x \) is optimal for problem (6) - (9).

Problem (6) - (9) differs from the rest of the literature on optimization with stochastic dominance constraints. Measures are the decision variables in problem (6) - (9), while random variable-valued mappings are the decision variables in the rest of the literature. This distinction is not academic, as the next subsection will show.

**B. Duality**
Duality for problem (6) - (9) is immediate from LP duality. As discussed in [14, Chapter 8], the dual of the linear programming problem (6) - (8) without the dominance constraint is
\[
\min g \quad (10)
\]
s.t. \[
g + h(s) - \sum_{j \in S} P(j \mid s, a) h(j) \geq r(s,a), \quad (11)
\]
\[
\forall (s,a) \in \Psi,
\]
where \( g \in \mathbb{R} \) and \( h \in \mathbb{R}^{[S]} \). We can adapt this result to compute the dual of (6) - (9). There is no difficulty with the Slater condition for (5) because these constraints are linear and LP duality does not depend on a constraint qualification. In contrast, in [6], [7] there is significant difficulty with the Slater condition for stochastic dominance constraints because the decision variable is a random variable-valued mapping.

In the following theorem, let \( U \) be the nonnegative cone of functions generated by \( (\xi - \eta) \) for all \( \eta \in \mathcal{Y} \).

**Theorem II.2.** The dual to problem (6) - (9) is
\[
\min_{u \in U} g - \mathbb{E} [u(Y)] \quad (12)
\]
s.t. \[
r(s,a) + u(q(s,a)) \quad (13)
\]
\[
\leq g + h(s) - \sum_{j \in S} P(j \mid s, a) h(j), \quad \forall (s,a) \in \Psi,
\]
and strong duality holds between problem (6) - (9) and problem (12) - (13).

For emphasis, problem (12) - (13) is an LP since the function \( u(\xi) = \sum_{\eta \in \mathcal{Y}} \lambda(\eta) (\xi - \eta) \) exists in a finite-dimensional space. Problem (12) - (13) is equivalent to
the explicit LP

\[
\begin{align*}
\min_{\lambda \geq 0} & \quad g - \mathbb{E} \left[ \sum_{\eta \in \mathcal{Y}} \lambda (\eta) (Y - \eta) \right] \\
\text{s.t.} & \quad r(s, a) + \sum_{\eta \in \mathcal{Y}} \lambda (\eta) (q(s, a) - \eta) \\
& \quad \leq g + h(s) - \sum_{j \in S} P(j \mid s, a) h(j), \\
& \quad \forall (s, a) \in \Psi,
\end{align*}
\]

where the variables \( \lambda \in \mathbb{R}^{|\mathcal{S}|} \) replace \( u \). The role of the utility function \( u \) in problem (12) - (13) is intuitive. The function \( u \) serves as an additional pricing variable for the performance function \( q(s, a) \), and the total reward is treated as if it were \( r(s, a) + u(q(s, a)) \). In this sense, problem (12) - (13) leads to a new version of the dynamic programming optimality equations for average reward.

### III. Discounted Reward

In this section, we carry out a parallel development for infinite horizon discounted reward with stochastic dominance constraints. For a discount factor \( \gamma \in (0, 1) \), we are now interested in

\[
\begin{align*}
\max_{\pi \in \Pi^{HR}} & \quad \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \\
\text{s.t.} & \quad \sum_{t=0}^{\infty} \gamma^t q(s_t, a_t) \geq \text{icv}, Y.
\end{align*}
\]

Let \( \alpha \in \mathbb{R}^{|\mathcal{S}|} \) represent an initial distribution on \( S \). From [14, Section 6.9] we can write problem (14) - (15) in terms of discounted occupation measures as:

\[
\begin{align*}
\max_{x \geq 0} & \quad \sum_{(s, a) \in \Psi} r(s, a) x(s, a) \\
\text{s.t.} & \quad \sum_{a \in A_s} x(j, a) - \alpha(j) \\
& \quad = \sum_{s \in S} \sum_{a \in A_s} \gamma P(j \mid s, a) x(s, a), \\
& \quad \forall j \in S, \\
& \quad \mathbb{E}_x [(q(s, a) - \eta)_-] \leq \mathbb{E} [(Y - \eta)_-], \\
& \quad \forall \eta \in \mathcal{Y}.
\end{align*}
\]

Problem (16) - (18) is similar to problem (6) - (9) because both problems are linear programs in occupation measures, though the occupation measures are interpreted differently. The occupation measures in problem (6) - (9) represent the long-run proportion of time spent in each state-action pair. The occupation measures in problem (16) - (18) represent the expected discounted number of visits to each state-action pair. The following result confirms a stationary policy is optimal in the discounted constrained case and justifies problem (16) - (18).

**Theorem III.1.** (a) If \( x^* \) is an optimal solution to problem (16) - (18), then there exists an optimal policy \( \pi^* \in \Pi^{SR} \) for problem (14) - (15) with

\[
\pi^*(a \mid s) = \frac{x^*(s, a)}{\sum_{k \in A_s} x^*(s, k)}
\]

for \( \sum_{k \in A_s} x^*(s, k) > 0 \) and \( \pi^*(a \mid s) \) is arbitrary otherwise.

(b) If \( \pi^* \) is optimal for problem (14) - (15), then there exists a \( \tilde{\pi} \in \Pi^{SR} \) which is also optimal and induces an optimal solution to problem (16) - (18).

**Proof:** Follows from the proof of [1, Theorem 3.3], since problem (16) - (18) has finitely many constraints.

It is well known that the dual of the linear programming problem (16) - (18) without the dominance constraints is

\[
\begin{align*}
\min_{a \in U} & \quad \sum_{j \in S} \alpha(j) v(j) \\
\text{s.t.} & \quad v(s) - \sum_{j \in S} \gamma P(j \mid s, a) v(j) \geq r(s, a), \\
& \quad \forall (s, a) \in \Psi,
\end{align*}
\]

in the decision variables \( v \in \mathbb{R}^{|\mathcal{S}|} \). In the next theorem, we compute the dual of problem (16) - (18). In analogy with the previous section, the dominance constraint (18) induces a new pricing variable in the dual.

**Theorem III.2.** The dual to problem (16) - (18) is

\[
\begin{align*}
\min_{a \in U} & \quad \sum_{j \in S} \alpha(j) v(j) - \mathbb{E}[u(Y)] \\
\text{s.t.} & \quad v(s) - \sum_{j \in S} \sum_{a \in A_s} \gamma P(j \mid s, a) v(j) \geq r(s, a), \\
& \quad \forall (s, a) \in \Psi,
\end{align*}
\]

and strong duality holds between problem (16) - (18) and problem (22) - (23).

Problem (22) - (23) leads to a new version of the optimality equations for discounted reward. The classic optimality equations for the primal version of problem (14) without the dominance constraint (15) are

\[
v(s) = \max_{a \in A_s} \left\{ r(s, a) + \sum_{j \in S} \gamma P(j \mid s, a) v(j) \right\},
\]
for all $s \in S$. The modified optimality equations implied by problem (16) - (18) are

$$v(s) = \max_{a \in A_s} \left\{ \frac{r(s, a) + u(q(s, a))}{\sum_{j \in S} P(j | s, a) v(j)} \right\},$$

for all $s \in S$, where the decision variable $u$ is common to all $s \in S$.

### IV. Multivariate Dominance

We have so far treated $q : \Psi \to \mathbb{R}$ as a single performance measure. However, many MDPs have multiple performance measures suggesting a vector-valued function $q : \Psi \to \mathbb{R}^n$. For example, in a queueing network we want to manage the vector of backlogs at every server, rather than a summary statistic or performance at a single server. We are interested in the empirical distribution of reward $q$,

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} q(s_t, a_t),$$

which is now a random vector. We would like to put a multivariate stochastic dominance constraint on this random vector $X, Y \in \mathbb{R}^n$. $X$ dominates $Y$ in the multivariate increasing concave order if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for all increasing concave functions $u : \mathbb{R}^n \to \mathbb{R}$ such that both expectations exist.

The problem is that multivariate icv does not have a convenient characterization in terms of a simple family of functions as was the case for univariate icv. To obtain a manageable semi-infinite programming problem, we consider a relaxation of multivariate icv. Let $\{u(\cdot; \xi)\}_{\xi \in \Xi}$ be a family of increasing concave functions $u(\cdot; \xi) : \mathbb{R}^n \to \mathbb{R}$ parametrized by $\xi \in \Xi \subseteq \mathbb{R}^m$ where $\Xi$ is compact. For an example of $\{u(\cdot; \xi)\}_{\xi \in \Xi}$ consider the increasing polyhedral concave functions with a fixed number of faces.

The benchmark $Y$ is now treated as a random vector but has the same interpretation as a target distribution, and we continue to assume $Y$ has finite support. Our new control problem is

$$\max_{\pi \in \Pi_{SR}} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

subject to:

$$\mathbb{E} \left[ u \left( \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} q(s_t, a_t); \xi \right) \right] \geq \mathbb{E} [u(Y; \xi)],$$

for all $\xi \in \Xi$.

Problem (24) - (25) can be written in terms of occupation measures as:

$$\max_{x \geq 0} \sum_{(s, a) \in \Psi} r(s, a) x(s, a)$$

subject to:

$$\sum_{a \in A_s} x(j, a) = \sum_{s \in S} \sum_{a \in A_s} P(j | s, a) x(s, a),$$

for all $j \in S$,

$$\sum_{(s, a) \in \Psi} x(s, a) = 1,$$

$$\mathbb{E}[u(q(s, a); \xi)] \geq \mathbb{E}[u(Y; \xi)],$$

for all $\xi \in \Xi$.

Problem (26) - (29) is a linear semi-infinite programming problem (LSIP). The next theorem follows from our earlier reasoning, it does not matter that there are now infinitely many constraints.

**Theorem IV.1.** Suppose the underlying Markov chain is unichain.

(a) If $x^*$ is an optimal solution to problem (26) - (29), then there exists an optimal policy $\pi^* \in \Pi_{SR}$ for problem (24) - (25) with

$$\pi^*(a | s) = \frac{x^*(s, a)}{\sum_{k \in A_s} x^*(s, k)}$$

for $\sum_{k \in A_s} x^*(s, k) > 0$ and $\pi^*(a | s)$ is arbitrary otherwise.

(b) If $\pi^*$ is optimal for problem (24) - (25), then there exists a $\bar{\pi} \in \Pi_{SR}$ which is also optimal and induces an optimal solution to problem (26) - (29).

Duality for problem (26) - (29) follows from LSIP duality. We assume the Slater condition holds for (29) for technical convenience, i.e. there is a feasible occupation measure $\bar{x}$ with $\mathbb{E}_{\bar{x}} [u(q(s, a); \xi)] > \mathbb{E}[u(Y; \xi)]$ for all $\xi \in \Xi$. For the following result, let $\text{cl cone} \{u(\cdot; \xi)\}_{\xi \in \Xi}$ denote the closure of the cone of the functions $\{u(\cdot; \xi)\}_{\xi \in \Xi}$.

**Theorem IV.2.** The dual to problem (26) - (29) is

$$\min_{g} g - \mathbb{E}[u(Y)]$$

subject to:

$$r(s, a) + u(q(s, a)) \leq g + h(s) - \sum_{j \in S} P(j | s, a) h(j),$$

$$\forall (s, a) \in \Psi,$$

$$u \in \text{cl cone} \{u(\cdot; \xi)\}_{\xi \in \Xi}.$$

Strong duality holds between problem (26) - (29) and problem (30) - (32).
The decision variables $u$ in problem (30) - (32) are in general infinite-dimensional. We continue to interpret $u$ as a pricing decision, only now it prices vectors instead of scalars.

V. Conclusion

In this paper, we manage risk in MDPs by using stochastic dominance. This modeling of risk is natural in many contexts, and also attractive because stochastic dominance-constrained MDPs lead to linear programs. In fact, this result does not depend on the choice of the increasing concave stochastic order. Many integral stochastic orders will give dominance-constrained MDPs that can be solved with linear programs, such as the increasing stochastic order. When the reward $q$ is vector-valued, then we could use the convex or supermodular orders. The corresponding dual will have new pricing variables that depend on the specific integral stochastic order, the way that the pricing variables in problems (12) - (13) and (22) - (23) depend on the increasing concave order.

There are three main directions for our future research. First, we would like to consider some practical aspects of solving the problems presented in this paper. We acknowledge that the linear programming approach is intractable for MDPs with large state and action spaces. To address this issue, we will adapt approximate linear programming for the large scale linear programs in this paper. Second, we would like to extend this development to general state and action spaces. Many MDPs are naturally set on continuous state and action spaces, for example. Finally, our ultimate goal is to handle online data-driven optimization for the dominance-constrained MDP in this paper. In practice, the transition probabilities of underlying MDPs will not be known and must be learned online. We want an effective learning policy with a guarantee on the speed at which the dominance constraint is satisfied.

References