Constrained Markov Decision Processes with Total Expected Cost Criteria

Eitan Altman, Said Boularouk, Didier Josselin

Constrained Markov Decision Processes with Total Expected Cost Criteria

Eitan Altman
INRIA, Univ Cote d'Azur; LIA, Univ of Avignon, and LINCS, France
Eitan.Altman@inria.fr

Said Boularouk
INRIA and LIA, Univ of Avignon, France
S.B@gmail.com

Didier Josselin
CNRS and LIA, Univ of Avignon, France
Didier.josselin@univ-avignon.fr

ABSTRACT

We study in this paper a multiobjective dynamic programming where all the criteria are in the form of total expected sum of costs till absorption in some set of states $\mathcal{M}$. We assume that instantaneous costs are strictly positive and make no assumption on the ergodic structure of the Markov Decision Process. Our main result is to extend the linear program solution approach that was previously derived for transient CMDPs (Constrained Markov Decision Processes) to general ergodic structure. Several (additive) cost metrics are defined and (possibly randomized) routing policies are sought which minimize one of the costs subject to constraints over the other objectives.

1. INTRODUCTION

When one has to select a path between a source S and a destination R, one often has several criteria. In road traffic problems it may be the minimization of the delay as well as the tolls. In communication networks it may be the minimization of delays, of loss probabilities of packets, of blocking probabilities of calls or of starvation probabilities in streeming video traffic. This motivates us to study the more general framework of CMDPs (Constrained Markov Decision Processes) which can be solved by transforming it to a linear program. The existing theory for solving such problems requires strong assumptions on the ergodic structure of the problem. In particular, for the shortest path multiobjective problem these conditions translate to a restrictive condition on the topology of the graph which fails to hold if there are cycles in the network. Our contribution is to extend existing solution methods to our multiobjective CMDP.

2. MODEL

Constrained Markov decision process (CMDP) A CMDP is described by the following objects. There is a set $\mathbf{X}$ of states which we asssume to be finite, a finite set $\mathbf{A}(x)$ of actions available at state x, a set of transition probabilities $\{P_{x,a,y}\}$, where $x, y \in \mathbf{X}, a \in \mathbf{A}(x)$. $P_{x,a,y}$ denotes the probability to move from state x to state y if action a is chosen at state x.

We consider $K+1$ cost criteria where one cost criterion, C, will be minimized subject to constraints on the other costs $D_1, ..., D_K$ of the form $D_k(u) \leq W_k$ where W_k are given constants.

Instantaneous costs In order to define the costs we first introduce the instantaneous costs c and $d^k, k = 1, ..., K$ where c and d^k are each strictly positive functions of the state-and action.

Histories and Policies. Define $h_n = (x_0, a_0, x_1, a_1, ..., x_{n-1}, a_{n-1}, x_n)$ to be a history of length $n+1$ where $a_k \in \mathbf{A}(x_k)$. A policy u is a sequence $(u_0, u_1, ...)$ where $u_m(\cdot|h_m)$ is a probability measure over the action set $\mathbf{A}(x_m)$ conditioned on the observed history h_m. Define the class of stationary policies: w is a stationary policy if the dependence of w_k on the history h_k is only through the current state x_k. Under a stationary policy w, the state process is a Markov chain with a transition probability matrix $P(w)$ satisfying

$$P_{xy}(w) = \sum_a w(a|x) P_{xay}$$

Each distribution β over the initial states x_0 and policy u define a probability measure P^u_β over the set of histories H. We denote by E^u_β the corresponding expectation operator. We shall use the notation X_n, A_n, H_n to denote the stochastic state process, the stochastic action process and the history stochastic process.

Cost criteria We consier in this paper the total expected cost till a set $\mathcal{M}$ of states is reached for the first time. More precisely, we assume that the set $\mathbf{X}$ of all states is the disjoint union of the two sets $\mathbf{X}'$ and $\mathcal{M}$. Let $T_\mathcal{M}$ the time till some state within $\mathcal{M}$ is reached for the first time. The total expected costs till absorption in $\mathcal{M}$ are defined as

$$C(\beta, u) = E^u_\beta \sum_{t=1}^{T_\mathcal{M}} [c(X_t, A_t)]$$

$$D^k(\beta, u) = E^u_\beta \sum_{t=1}^{T_\mathcal{M}} [d^k(X_t, A_t)]$$

We shall assume without loss of generality that the set $\mathcal{M}$ is absorbing, i.e. $P_{xay} = 0$ for all $x \in \mathcal{M}, a \in \mathbf{A}(x)$ and $y \notin \mathcal{M}$. We further assume that $c(x, a) = d^k(x, a) = 0$ for all $x \in \mathcal{M}$, $a \in \mathbf{A}(x)$, $k = 1, ..., K$. We then have

$$C(\beta, u) = E^u_\beta \sum_{t=1}^{\infty} [c(X_t, A_t)]\ ,\ D^k(\beta, u) = E^u_\beta \sum_{t=1}^{\infty} [d^k(X_t, A_t)]$$

For a stochastic matrix P and the set of states $\mathcal{M}$, we de-

VALUETOOLS 2019 Palma de Mallorca, Spain

fine by ${}_{\mathcal{M}}P$ the Taboo matrix which is obtained by replacing in P the entries of each column $x \in \mathcal{M}$ with zeros.

3. OCCUPATION MEASURE

For any initial distribution β, any policy u, state x and set of actions $\mathcal{A} \in \mathbf{A}(x)$ define

$$p_\beta^u(t,x,\mathcal{A}) := P_\beta^u(X_t = x, A_t \in \mathcal{A}, T_{\mathcal{M}} > t)$$

$$f(\beta,u,x,\mathcal{A}) = \sum_{t=1}^{\infty} p_\beta^u(t;x,\mathcal{A})$$

and $s(\beta,u,x) = f(\beta,u;x,\mathbf{A}(x))$. Define the following polyhedron $\mathbf{Q}(\beta)$ to be the set of non-negative measures over the set of state action pairs, that satisfy for all $x \in \mathbf{X}$,

$$\sum_{a\in\mathbf{A}(x)} \rho(x,a) = \sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}(x)} \rho(y,a)(P_{y,a,x}1\{x \in \mathbf{X}'\}) + \beta(x). \tag{1}$$

4. MAIN RESULTS

Below we identify a linear program that allows to compute the optimal value and an optimal stationary policy for CMDP. A similar result was already available in [1] but required the strong assumption that $s(\beta,u)$ is finite for any u. This excludes the shortest path problem in which policies that include cycles may have infinite cost. In order to handle such situations we note that one may assume throughout that c and d are uniformly bounded below by some positive constant $\underline{c}$. The following Lemma shows that although we do not assume that all policies have finite occupation measures, those having infinite occupation measures are not optimal. It further shows that one may restrict the search of solutions to CMDP to stationary policies.

LEMMA 1. *(i) Fix some initial distribution β and a policy u. Then either $C(\beta,u)$ is infinite or $f(\beta,u)$ satisfies (1) and both $f(\beta,u)$ and $s(\beta,u)$ are finite measures.*
(ii) Fix some initial distribution β and a stationary policy w. Then $s(\beta,w;x)$ is the minimal solution to

$$r = \beta + r_{\mathcal{M}}P(w), \qquad r \geq 0 \tag{2}$$

(in matrix notation, where r and β are row vectors).

Proof. (i) We have

$$C(\beta,u) = \sum_{x,a} c(x,a) f(\beta,u;x,a) \geq \underline{c} \sum_{x} s(\beta,u;x) = \underline{c} E_\beta^u[T_{\mathcal{M}}]$$

where $\underline{c} := min_{x,a} c(x,a)$. Thus if $C(\beta,u)$ is finite then $s(\beta,u)$ and $f(\beta,u)$ are indeed finite measures. That $f(\beta,u)$ satisfies (1) follows by noting that

$$p_\beta^u(t;x) = \sum_{y,a} p_\beta^u(t-1;y,a) P_{y,a,x} 1\{x \in X'\}$$

and taking the sum over t.
(ii) See Lemma 7.1 (i) in p. 76 in [1]. Although the statement in that reference is for for transient MDPs, the proof does carry on to our framework. ∎

THEOREM 2. *Choose any initial distribution β and policy u. Then either $C(\beta,u) = \infty$ or there exists a stationary policy w such that $C(\beta,w) \leq C(\beta,u)$,*

Proof. This is an extension of thm 8.1 in p.100 of [1]. Choose some policy u. Assume $C(\beta,u)$ is finite. Consider a stationary policy w satisfying $f(\beta,u;x,a) = s(\beta,u;x)w(a|x)$. Note that f and s are finite measures otherwize $C(\beta,u)$ would be infinite. It then follows that

$$s(\beta,u;x) = \beta(x) + \sum_{y\in\mathbf{X}} s(\beta,u;y)_{\mathcal{M}}P_{yx}(w).$$

see the derivation of eq 8.6 page 102 in [1]. Note that in [1], $s(\beta,u;x)$ is finite for all β,u and x but eq. 8.6 in [1] holds also in our framework. Hence by Lemma 1, for all x, $s(\beta,w;x) \leq s(\beta,u;x)$. Thus $f(\beta,w;x,a) \leq f(\beta,u;x,a)$ for all state action pairs x,a. As the costs c and d are strictly positive, we conclude that

$$C(\beta,u) = \sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}(y)} f(\beta,u;y,a)c(y,a)$$

$$\geq \sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}(y)} f(\beta,w;y,a)c(y,a) = C(\beta,w)$$

$$\textit{and similarly,} \quad D_k(\beta,u) \geq D_k(w), \quad k = 1,...,K$$

∎

Consider the following LP:
$LP(\beta)$: Find the infimum $\mathcal{C}^*$ of $C(\rho) :=< \rho, c >$ over ρ subject to

$$\mathcal{D}^k(\rho) :=< \rho, d^k > \leq W_k, \quad k = 1,...,K, \quad \rho \in \mathbf{Q}(\beta)$$

where $\mathbf{Q}(\beta)$ is defined in (1).

THEOREM 3. *(i) To each policy $u \in U$ in CMDP there corresponds a point $\rho(u) := f(\beta,u)$ in $\mathbf{Q}(\beta)$ whose corresponding costs are the same:*

$$C(\beta,u) = \mathcal{C}(\rho), \qquad D(\beta,u) = \mathcal{D}(\rho). \tag{3}$$

(ii) Conversely, for each $\rho \in \mathbf{Q}(\beta)$, there corresponds a stationary policy $w(\rho)$ whose performance is at least as good as the one of ρ. More precisely, select a stationary policy $w(\rho)$ so that

$$w^{\rho^*}(a|y) = \frac{\rho(y,a)}{\sum_{a'\in\mathbf{A}(y)} \rho(y,a')} \tag{4}$$

for all y for which the denominator is finite and strictly positive. Then

$$C(\beta,w(\rho)) \leq \mathcal{C}(\rho), \qquad D(\beta,w(\rho)) \leq \mathcal{D}(\rho).$$

(iii) The optimal value $C(\beta)$ of CMDP is equal to the optimal value C^* of LP(β). Let ρ^* be an optimal solution of LP(β). Then the stationary policy w^{ρ^*} given in eq. (4) is optimal for CMDP.

Proof. (i) From Lemma 1 (i) it follows that either $C(\beta,u)$ is infinite, or $f(\beta,u)$ satisfies (1) and hence is in $\mathbf{Q}(\beta)$. in both cases (3) holds. (ii) follpws from Lemma 1 (ii). (iii) We may assume that there exists a policy u such that $D_k(u) \leq W_k$ and $C(u)$ is finite, otherwise the statement is trivial. The statement follows then from (i) and (ii) of the Theorem. ∎

5. REFERENCES

[1] E. Altman, Constrained Markov Decision Processes, Chapman and Hall/CRC, 1999.