Constrained Markov Decision Processes

Nelson Gonçalves

April 16, 2007

Topics

Within the Markov Decision Processes framework, agents attempt to find policies maximizing a given reward. But what if the multiple objectives, possibly conflicting, are considered? This is a common situation in communication networks, project management and multi-robot team coordination

In this presentation, the framework of Constrained Markov Decision Processes (CMDP's) is introduced to deal with such dynamical, multi-objective, decision problems. A small example, using mobile robots, is presented to illustrate the application of CMDP's.

Examples of mutiple objective, dynamic, decision problems:

- Project managment
  - Fullfill project objectives
  - Constrained budget
  - Constrained material resources
  - Constrained human resources
- Communication networks
  - Maximize data throughoutput, for example
  - Constrain message delays
  - Constrain different types of data
  - Constrain power compsumption
- Multirobot teams coordination
  - Execute assigned task
  - Limited number of robots
  - Robots have limited capabilities
  - The task can impose aditional constraints (time, for example)

Markov Decision Process:

- Discrete and finite space state, X
- Finite set of actions, A
- $A(x)$ are the actions available at state $x \in X$, $A(x) \subset A$
- Set of state-action pairs, $\mathcal{K} = \{(x, a) : x \in X, a \in A(x)\}$
- Transition probabilities, $\mathcal{P}_{xay}$
- Immediate cost, $c : \mathcal{K} \mapsto \mathbb{R}$

The system history at time instant t is given by:

$$h_t = (y_1, a_1, y_2, a_2, \ldots, y_{t-1}, a_{t-1}, y_t) \tag{1}$$

A policy u is a sequence $u = (u_1, u_2, \ldots)$ where the element u_t is the probability of selection action a given the system history, $u_t(a|h_t)$. There are different classes of policies:

- U_M : Markov policies, in which u_t is a function of the current state y_t
- U_S : Stationary policies, in which u_t is independent of the time
- U_D : Stationary and deterministic policies, wich define the map $g : X \mapsto A$

These classes verify the relation:

$$U_D \subset U_S \subset U_M \tag{2}$$

Let β be the distribuition of the initial state, x_0. Define for any initial distribuition β and policy u:

$$f_\alpha(\beta, u; x, a) := (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1} P_\beta^u(X_t = x, A(x) = a) \tag{3}$$

Then $f_\alpha(\beta, u)$ can be viewed as the probability (*occupation*) measure, assigning the probability $f_\alpha(\beta, u; x, a)$ to each pair (x, a). Thus the **discounted cost** can be expressed as:

$$C_\alpha(\beta, u) = \sum_{x \in X} \sum_{a \in A(x)} f_\alpha(\beta, u; x, a) c(x, a) = \langle f(\beta, u), c \rangle \tag{4}$$

where f and c are vectors with dimension $|\mathcal{K}|$. To simplify notation, we assume α and β are fixed.

A Constrained Markov Decision Process is similar to a Markov Decision Process, with the difference that the policies are now those that verify additional cost constraints. That is, determine the policy u that:

$$\begin{aligned} &min\, C(u) \\ &s.t. \quad D(u) \leq V \end{aligned} \tag{5}$$

where $D(u)$ is a vector of cost functions and V is a vector , with dimension N_c, of constant values.

Using the discounted cost, an CMDP can be shown to be equivalent to the linear program:

$$\begin{aligned} & min \, \langle \rho, C \rangle \\ & s.t. \quad \langle \rho, D_n \rangle \leq V_n, \quad n = 1, \ldots, N_c \\ & \quad \rho \in \mathcal{Q} \end{aligned} \tag{6}$$

where C and D_n are imediate cost vectors, both with dimension $|\mathcal{K}|$.

The vector ρ is defined as:

$$\begin{cases} \sum_{y \in X} \sum_{a \in A(y)} \rho(y,a)(\delta_x(y) - \alpha \mathcal{P}_{yax}) = (1-\alpha)\beta(x), \ \forall x \in X \\ \\ \rho(y,a) \geq 0, \forall y, a \end{cases} \tag{7}$$

If the first line is summed over x, we obtain $\sum_{a,x} \rho(x,a) = 1$ for all $\rho \in \mathcal{Q}$. Thus, each element $\rho(x,a)$ can be considered a probability of selecting action a in state x.

From the elements $\rho(x, a) \in \mathcal{Q}$, the stationary optimal policy u can be determined:

$$u(a|x) = \frac{\rho(x, a)}{\sum_{a \in A(x)} \rho(x, a)} \tag{8}$$

if $\sum_{a \in A(x)} \rho(x, a) \neq 0$. Otherwise select an arbitrary value for $u(a|x)$, but ensuring that $\sum_{a \in A(x)} u(a|x) = 1$.

Dominating Policy:
Supose that for any $u \in U$, any of the previous cost criterions and for an initial distribuiton β, there exists another policy $\overline{u} \in \overline{U}$ such that:

$$C(\beta, \overline{u}) \leq C(\beta, u) \quad \textit{and} \quad D(\beta, \overline{u}) \leq D(\beta, u) \tag{9}$$

Then $\overline{U}$ is said to be dominating over U. An important result is that: *Markov policies are dominating for any cost criterion wich is a function of the marginal distribuition of states and actions*.

Completeness of Stationary Policies:
Given a set U of policies, define $L_U = \{f(u) : u \in U\}$. It can be shown that for the discounted cost:

$$L_M = L_S = \overline{co} L_D \tag{10}$$

where L_M are the set of Markov policies, L_S the set of stationary policies and L_D the set of deterministic policies.
This is an usefull result because it allows to reduce the classes of policies under consideration. For example, if using the discounted cost we gain nothing by using more general, non-stationary, policies.

Uniformaly Optimal Policy:
Because now the policy is determined by additional constraints, for some values of β, or even initial states $x_0 \in X$, the optimal policy might not exist (it is not feasible). This is not a problem with traditional MDP's. A policy that is optimal for all initial states is said to be an uniformly optimal policy.

Consider the following example:
Two robots must pickup a box and move it to a destiny position. When moving the box, they must move together to avoid dropping the box.

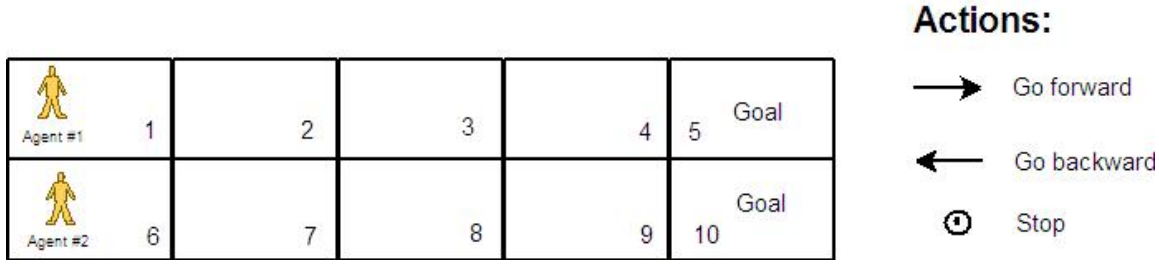

Figure: Box carrying example

The state x, is the tuple(y_1, y_2) where y_1 and y_2 are the agents positions:

$$x \in X = \{(1,6),(1,7),\ldots,(5,10)\} = \{1,\ldots,25\} \tag{11}$$

Identically, the action at each state a, is the tuple (a'_1, a'_2) :

$$\begin{array}{rcl} a \in A & = & \{s_1 s_2, s_1 f_2, s_1 b_2, f_1 s_2, b_1 s_2, f_1 f_2, b_1 b_2, f_1 b_2, b_1 f_2\} \\ & = & \{1,2,3,4,5,6,7,8,9\} \end{array} \tag{12}$$

All actions are equiprobable. At some states not all of the actions are not available.

The minimization imediate cost (objective) is:

$$c(x,a) = \begin{cases} 0 & \textit{if } x = 25 \textit{ and } a = s_1 s_2 \\ 10 & \textit{otherwise} \end{cases} \tag{13}$$

The constraint imediate cost is:

$$d(x,a) = \begin{cases} 0 & \textit{if agents get parallel} \\ 1 & \textit{if agents not parallel but get closer} \\ 5 & \textit{otherwise} \end{cases} \tag{14}$$

The CMDP is then:

$$
\begin{aligned}
& inf \ \langle \rho, c \rangle \\
& s.t. \\
& D\rho \leq 1 \\
& A_{eq}\rho = b_{eq} \\
& \rho \geq 0
\end{aligned}
\tag{15}
$$

where the last two constraints represent $\rho \in \mathcal{Q}$.

Solving the linear program, we obtain the policy:

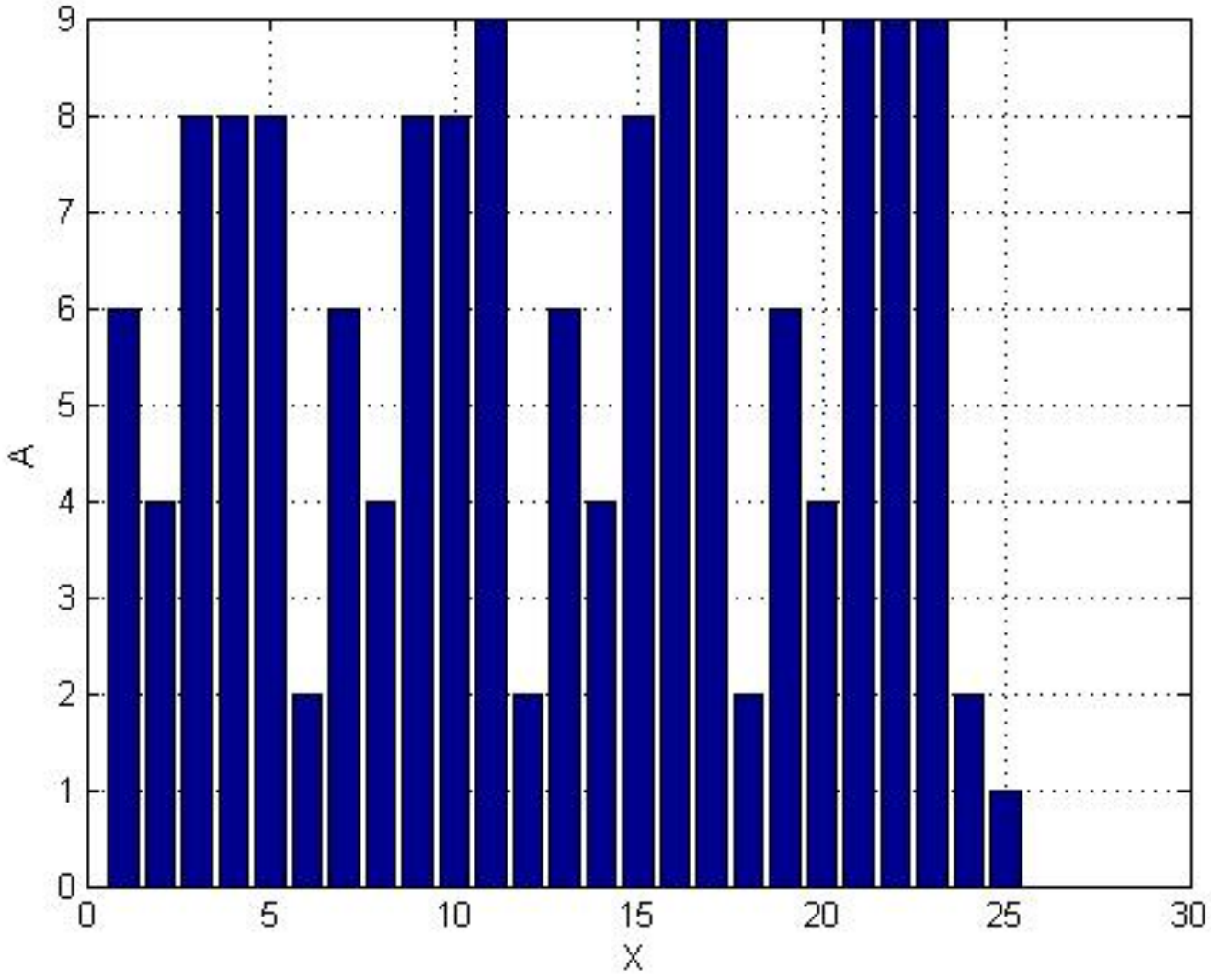

Figure: Optimal policy

Could this problem be solved using an MDP ?

Constrained Markov Decision Processes
Eitan Altman
Chapman & Hall/RC, 1999

Robustness of Policies in Constrained Markov Decision Processess
Alexander Zadorojniy and Adam Shwartz
IEEE Transactions on Automatic Control, Vol. 51, No. 4, April 2006