ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-969

On Finding Compromise Solutions in Multiobjective Markov Decision Processes

Patrice Perny and **Paul Weng**[1]

1 INTRODUCTION

Markov Decision Processes (MDP) constitute a general model for solving planning problems under uncertainty. In its standard form, the objective is to maximize the expectation of the sum of rewards. As in practice, actions are generally valued over several dimensions (cost, time, energy consumption, reward…), MDPs have been extended to take into account multiple (generally conflicting) objectives or criteria. When several objectives must be optimized simultaneously, most of the studies on MDPs concentrate on the determination of the entire set of Pareto-optimal solutions, i.e. policies having a value vector that cannot be improved on a criterion without being downgraded on another criterion. However, the size of the Pareto set is often very large due to the combinatorial nature of the set of deterministic policies, its determination induces prohibitive response times and requires very important memory space when the number of states and/or criteria increases. Fortunately, there is generally no need to determine the entire set of Pareto-optimal policies, but only specific compromise policies achieving a good tradeoff between the possibly conflicting objectives. The quality of the compromise achieved can be measured using a scalarizing function discriminating between Pareto-optimal solutions [6].

2 BACKGROUND

An MDP [5] is described by a finite set S of states, a finite set A of actions, transition probabilities $T(s, a, s')$ of reaching state s' from state s with action a, immediate rewards $R(s, a) \in \mathbb{R}$ obtained when executing action a in state s. In this context, a *policy* π is a procedure that determines which action to choose in each state. A policy can be *deterministic*, i.e. defined as $\pi : S \to A$, or more generally, *randomized*, i.e. defined as $\pi : S \to \mathcal{P}(A)$ where $\mathcal{P}(A)$ is the set of probability distributions over A. The value of a policy π is defined by a *value function* $v^\pi : S \to \mathbb{R}$ that gives the expected discounted total reward yielded by applying π from each initial state, i.e. $\forall s \in S$, $v^\pi(s) = E(\sum_{t>0} \gamma^{t-1} R_t | \pi, s_0 = s)$ where $\gamma \in [0, 1[$ is a discount factor, R_t is a random variable giving the reward at step t and s_0 is the initial state.

In this standard framework, there exists an optimal deterministic policy that yields the best expected discounted total reward in each initial state. Solving an MDP amounts to finding one of those policies. There are three main approaches for solving MDPs [5]. Two are based on dynamic programming: value iteration and policy iteration. The last one is based on linear programming.

MDPs have been extended to take into account multiple dimensions or criteria. A multiobjective MDP (MMDP) is defined as an MDP where rewards $R(s, a)$ are now defined as vectors of $\mathbb{R}^n$ where n is the number of criteria, $R(s, a) = (R_1(s, a), \dots, R_n(s, a))$ and $R_i(s, a)$ is the immediate reward for criterion i.

Now, a policy π is valued by a value function $V^\pi : S \to \mathbb{R}^n$, which gives the expected discounted total reward vector in each state. To compare the value of policies in a given state s, the basic model adopted in most previous studies [7, 8, 1] is *Pareto-dominance* defined by: for two policies π, π', for a state s, $V^\pi(s)$ Pareto-dominates $V^{\pi'}(s)$, denoted $V^\pi(s) \succ_P V^{\pi'}(s)$ if and only if $V^\pi(s) \neq V^{\pi'}(s)$ and $\forall i = 1 \dots n$, $V_i^\pi(s) \geq V_i^{\pi'}(s)$. For a set $X \subset \mathbb{R}^n$, the set of *Pareto-optimal* vectors of X is defined by $M(X, \succ_P) = \{x \in X : \forall y \in X, \text{ not } y \succ_P x\}$.

Standard methods for MDPs can be extended to solve MMDPs although some problems remain opened as noted by [8] for methods based on dynamic programming. [7] proposed a multiobjective linear program for finding Pareto-optimal solutions in a MMDP.

Looking for all Pareto-optimal solutions can be difficult and time-consuming as $\succ_P$ being partial, there could be many Pareto-optimal solutions. In fact, there exists instances of problems where the number of Pareto-optimal deterministic policies is exponential in the number of states [4]. Besides, in practice, one is generally only interested in one particular solution among all the Pareto-optimal solutions that gives interesting tradeoffs between all the criteria. A more interesting approach for MMDP would be to directly search for that particular solution instead of finding first all the Pareto-optimal solutions.

3 SEARCH FOR COMPROMISE SOLUTIONS

We introduce the notion of *scalarizing function* that will be used to discriminate between Pareto-optimal vectors in a given state. Formally, a scalarizing function is a function $\psi : \mathbb{R}^n \to \mathbb{R}$ that defines an *overall value function* $v : S \to \mathbb{R}$ from a vector value function $V : S \to \mathbb{R}^n$ by $v(s) = \psi(V_1(s), \dots, V_n(s))$.

The most straightforward choice for ψ seems to be the weighted sum. In this case, $v(s) = \sum_{i=1}^n \lambda_i V_i(s)$ where $\lambda_i > 0, \forall i = 1 \dots n$ so as to preserve the monotonicity with respect to Pareto dominance. By linearity of the mathematical expectation and the weighted sum, optimizing v is equivalent to solving the standard MDP obtained from the MMDP where the reward function is defined as: $r(s, a) = \sum_{i=1}^n \lambda_i R_i(s, a), \forall s, a$. In that case, an optimal deterministic policy exists and standard solution methods can then be applied. However, using a weighted sum is not a good procedure for reaching balanced solutions as weighted sum is a fully compensatory operator that does not encode the idea of balanced solutions. Besides, generally good balanced solutions can only be obtained with randomized policy, which excludes the use of a weighted sum.

[1] LIP6, UPMC, France, email: firstname.surname@lip6.fr

In multiobjective optimization, the question of finding balanced solutions among the Pareto-optimal ones is a crucial issue. The standard way of generating compromise solutions in the Pareto-optimal set is to resort to the so-called reference point approach that consists in finding a feasible vector that minimizes a distance to a prescribed reference point [9, 6, 3]. This distance is given by the *Tchebycheff scalarizing function* defined for all $x \in \mathbb{R}^n$ by: $\psi(x, r, \lambda) = \max_{i=1\dots n} \lambda_i|r_i - x_i| + \varepsilon\sum_{i=1\dots n} \lambda_i|r_i - x_i|$ where $r \in \mathbb{R}^n$ is the reference point, $\lambda \in \mathbb{R}^n$ is a positive weighting vector and ε is a positive real chosen arbitrarily small. The best compromise solution $V^{T*} : S \to \mathbb{R}^n$, called *Tchebycheff-optimal*, can then be computed with $V^{T*} = \text{argmin}_V\, \psi(\sum_{s\in S} \mu(s)V(s), r, \lambda)$ where μ is a distribution probability over initial states.

To exploit this equation, we need to set properly those parameters. Generally, reference point r is taken as the *ideal point*. It can be computed with n different one-dimensional optimizations, that is we solve the MMDP as a standard MDP successively with reward function R_i for $i = 1 \dots n$. We denote $V^{i*} : S \to \mathbb{R}^n$ the optimal value function for the i-th criterion. In a state s, the ideal point then can be formally defined as follows: $v^I_i = \sum_{s\in S} \mu(s)V^{i*}_i(s)$ for all $i = 1 \dots n$. With the V^{i*}s, a second point can also be defined: $v^A_i(s) = \sum_{s\in S} \mu(s) \min_{j=1\dots n} V^{j*}_i(s)$ which is in fact an approximation of the Nadir point, i.e. a lower bound of the Pareto-optimal solutions. We use an approximation of the Nadir point as it is generally difficult to determine exactly [2]. Then the weights are defined as follows: $\lambda_i(s) = \frac{w_i}{|v^I_i(s) - v^A_i(s)|}$ where $w \in \mathbb{R}^n$ represents the weights of criteria.

As shown in [9], this way of constructing compromise solutions guarantees some nice properties. Contrary to the weighted sum, here, any Pareto-optimal solution can be reached by minimizing the Tchebycheff scalarizing function with a proper choice of w. Moreover, any Tchebycheff-optimal solution is Pareto-optimal.

4 SOLUTION METHOD AND EXPERIMENTS

As the Tchebycheff scalarizing function is not linear, solving methods based on dynamic programming can not be exploited directly. However, the multiobjective linear program proposed for MMDP [7] can be adapted to our problem. The Tchebycheff-optimal solutions can be found with the following linear program:

$$\begin{aligned}
\min \quad & z + \varepsilon \sum_{i=1\dots n} \lambda_i\big(v^I_i - \sum_{s\in S}\sum_{a\in A} R_i(s,a)x(s,a)\big) \\
\text{s.t.} \quad & z \geq \lambda_i\big(v^I_i - \sum_{s\in S}\sum_{a\in A} R_i(s,a)x(s,a)\big) \quad \forall i = 1\dots n \\
& \sum_{a\in A} x(s,a) - \gamma \sum_{s'\in S}\sum_{a\in A} x(s',a)T(s',a,s) = \mu(s) \\
& \forall s \in S \\
& x(s,a) \geq 0 \quad \forall s \in S, \forall a \in A
\end{aligned}$$

Due to the non-linearity of the Tchebycheff scalarizing function, solutions now depend on the initial state. Therefore, when the initial state s_0 is known, $\mu(s_0) = 1$ and $\mu(s) = 0$ when $s \neq s_0$; otherwise, distribution μ can be chosen as the uniform distribution over the possible initial states.

We tested our solving method on the navigation problem over a grid NxN. In this problem, the robot can choose among four actions: Left, Up, Right, Down. Figure 1 gives the transitions for action Right. The whole transition function can then be obtained by symmetry. Rewards are two-dimensional vectors whose components are randomly drawn within interval $[0, 1]$. The discount factor is set to 0.9 and the initial state is set arbitrarily to the upper left corner of the grid.

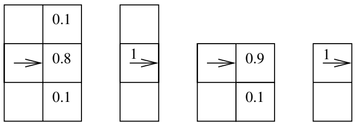

Figure 1. Transitions

As in real problems, criteria are generally conflicting, for the first set of experiments, to generate realistic random instances, we simulate conflicting criteria with the following procedure: for each state and action, one criterion (picked randomly) is drawn uniformly in $[0, 0.5]$ and the other is drawn in $[0.5, 1]$. The results over 100 experiments are represented on Figure 2. One point on that figure (a dot for weighted sum and a circle for the Tchebycheff norm) represents the optimal value function in the initial state for one instance. Naturally, for some instances, the weighted sum yields a balanced solution. But, in most cases, the weighted sum gives a bad compromise solution. Figure 2 actually shows that we do not have any control on tradeoffs obtained with a weighted sum. On the contrary, when using a Tchebycheff norm, the profile of the solutions are always balanced.

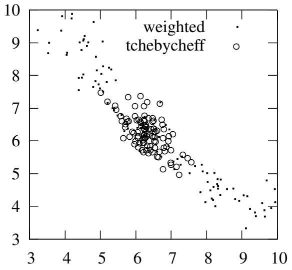

Figure 2. First experiments

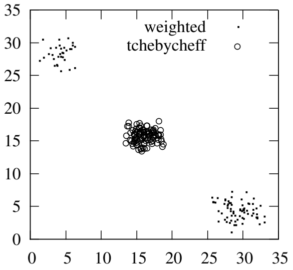

Figure 3. Second experiments

To show the effectiveness of our approach, we ran a second set of experiments on pathological instances. Rewards are drawn randomly as for the first set of experiments. Then in the initial state, for each action, we randomly pick one criterion and add a constant (here, arbitrarily set to 5). Then by construction, the value functions of all deterministic policies in the initial state are unbalanced (Figure 3). Reassuringly, Tchebycheff-optimal solutions are always well-balanced.

REFERENCES

[1] K. Chatterjee, R. Majumdar, and T.A. Henzinger, 'Markov decision processes with multiple objectives', in *STACS*, (2006).

[2] M. Ehrgott and D. Tenfelde-Podehl, 'Computation of ideal and Nadir values and implications for their use in MCDM methods', *EJOR*, **151**, 119–139, (2003).

[3] L. Galand and P. Perny, 'Search for compromise solutions in multiobjective state space graphs', in *ECAI*, pp. 93–97, (2006).

[4] P. Hansen, *Bicriterion Path Problems*, chapter Bicriterion Path Problems, 109–127, Springer, 1980.

[5] M.L. Puterman, *Markov Decision Processes - Discrete Stochastic Dynamic Programming*, John Wiley and Sons, 1994.

[6] R.E. Steuer, *Multiple criteria optimization*, John Wiley, 1986.

[7] B. Viswanathan, V.V. Aggarwal, and K.P.K. Nair, 'Multiple criteria Markov decision processes', *TIMS Studies in the management sciences*, **6**, 263–272, (1977).

[8] D.J. White, 'Multi-objective infinite-horizon discounted Markov decision processes', *Journal of mathematical analysis and applications*, **89**, 639–647, (1982).

[9] A.P. Wierzbicki, 'On the completeness and constructiveness of parametric characterizations to vector optimization problems', *OR Spektrum*, **8**, 73–87, (1986).