Evaluation of stochastic policies for factored Markov decision processes

Julia Radoszycki[1], Nathalie Peyrard[1] and Régis Sabbadin[1]

Introduction

We are interested in the resolution of general Markov decision processes (MDP) with factored state and action spaces, called FA-FMDP [4]. We are in particular interested in reward functions which constrain the system to be in an admissible domain, defined by several conditions, the longest time as possible.

There are few algorithms able to resolve FA-FMDPs with both a reasonable complexity and a reasonable quality of approximation. Some recent algorithms described in [9], can be used with affine algebraic decision diagrams (AADDs), which are more suited to multiplicative rewards than algebraic decision diagrams (ADDs). The drawback of these algorithms is that they are designed for binary state and action variables, and do not scale to variables with more than two modalities. Most of the other existing methods for solving MDPs with large state and action spaces make the assumption of an additive reward, like approximate linear programming [3] or mean field approaches [10]. As an alternative, we propose to consider multiplicative rewards as an interesting way of modelling objectives defined as admissible domains, and because this trick, as we will see, allows to find methods of approximate policy evaluation.

Recently, several decision problems have been resolved with methods of inference in graphical models, an idea which has been recently called *planning as inference* [2]. We may cite for example [11] which proposes an EM algorithm for solving (non factored) MDPs or [6] which proposes a *belief propagation* algorithm for solving influence diagrams. Our idea is to follow this trend and propose an (approximate) policy iteration type algorithm [8] for solving FA-FMDPs with multiplicative rewards. This algorithm iterates over two steps : an evaluation step and an optimization step. In this communication, we focus on the approximate evaluation step of stochastic policies for FA-FMDPs with multiplicative rewards. The method we propose is based on the computation of normalization constants of *factor graphs* with increasing sizes by existing variational or *belief propagation* methods which are applicable on large size graphs.

1 Preliminaries

We consider a FA-FMDP [4] with multiplicative rewards as a MDP $< \mathcal{S}, \mathcal{A}, P, R, \gamma >$ where :

- $\mathcal{S}$ is the factored state space : $\mathcal{S} = \prod_{i=1}^{n} \mathcal{S}_i$, with each $\mathcal{S}_i$ a finite set ; the state of the system at time t is noted[2] : $S^t = (S_1^t, ..., S_n^t) \in \mathcal{S}$
- $\mathcal{A}$ is the factored action space : $\mathcal{A} = \prod_{j=1}^{m} \mathcal{A}_j$, with each $\mathcal{A}_j$ a finite set ; the action at time t is noted $A^t = (A_1^t, ..., A_m^t) \in \mathcal{A}$
- P is the global transition function : $P(s^{t+1}|s^t, a^t) = \prod_{i=1}^{n} P_i(s_i^{t+1}|pa_P(s_i^{t+1}))$ where $pa_P(s_i^{t+1}) \subset \{s_j^t, j = 1...n, s_k^{t+1}, k = 1...n, a_l^t, l = 1...m\}$ (we authorize synchronous arcs)
- R is the global reward function : $R(s^t, a^t) = \prod_{\alpha=1}^{k} R_\alpha\left(s_\alpha^t, a_\alpha^t\right) = \prod_{\alpha=1}^{k} R_\alpha\left(pa_R(R_\alpha^t)\right)$ where $pa_R(R_\alpha^t) \subset \{s_i^t, i = 1...n, a_j^t, j = 1...m\}$; $R(s^t, a^t)$ is supposed to be in $\mathbb{R}^+$ and bounded by R_{max}
- $\gamma \in]0; 1[$ is the discount factor.

A (stochastic) factored policy is defined by : $\delta(a^t|s^t) = \prod_{j=1}^{m} \delta_j(a_j^t|pa_\delta(a_j^t))$ where $pa_\delta(a_j^t) \subset \{s_i^t, i = 1...n, a_j^t, j = 1...m\}$. It represents the probability of choosing $A^t = a^t$ when the system is in state s^t. We limit our study to problems where the directed graph generated by the $pa_P(.)$ and $pa_\delta(.)$ functions is acyclic over state and action variables in time steps t and $t+1$. In this case, for a given factored policy δ and a given factored distribution on initial states $P^0(s^0) = \prod_{i=1}^{n} P_i(s_i^0|pa_P(s_i^0))$ (with $pa_P(s_i^0) = \varnothing \ \forall i = 1...n$), the probability distribution over the finite horizon trajectories $(s, a)_{0:t} =< s^0, a^0, ..., s^t, a^t >$ may be represented as a DBN :

$$P_\delta((s, a)_{0:t}|P^0) = \prod_{t'=0}^{t} \left(\prod_{i=1}^{n} P_i(s_i^{t'}|pa_P(s_i^{t'})) \prod_{j=1}^{m} \delta_j(a_j^{t'}|pa_\delta(a_j^{t'})) \right)$$

The objective of this paper is to provide a way to compute the value of a given factored policy δ for a given factored distribution on initial states. In a FA-FMDP this value is defined as :

$$V_\delta(P^0) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t R(s^t, a^t) \middle| P^0, \delta \right]$$

2 A method for the evaluation of stochastic policies based on the computation of normalization constants

Let us use the linearity of the expectation in order to express the value of δ for P^0 in another way :

[1] Unité de Mathématiques et Informatique Appliquées, Toulouse (UR 875) - INRA - 24 Chemin de Borde Rouge - Auzeville - CS 52627 - 31326 Castanet Tolosan cedex
{julia.radoszycki,nathalie.peyrard,regis.sabbadin}@toulouse.inra.fr

[2] We will note with capital letters the random variables and with lower case letters their realizations.

$$
\begin{aligned}
V_\delta(P^0) &= \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}[R(s^t,a^t)|P^0,\delta] \text{ since } \gamma^t \to 0 \\
&= \sum_{t=0}^{+\infty} \gamma^t \sum_{s^0\ldots s^t} \sum_{a^0\ldots a^t} P_\delta((s,a)_{0:t}|P^0) R(s^t,a^t) \\
&= \sum_{t=0}^{+\infty} \gamma^t \sum_{s^0\ldots s^t} \sum_{a^0\ldots a^t} P_\delta((s,a)_{0:t}|P^0) \prod_{\alpha=1}^{k} R_\alpha\left(pa_R(R_\alpha^t)\right)
\end{aligned}
$$

We thus have $V_\delta(P^0) = \sum_{t=0}^{+\infty} \gamma^t C(t)$ where $C(t)$ is the normalization constant of a *factor graph* [5] as the one represented in Figure 1, whose variable nodes, represented by circles, are the $s_i^0 \ldots s_i^t, i = 1 \ldots n$ and $a_j^0 \ldots a_j^t, j = 1 \ldots m$ and the factor nodes, represented by squares, are the $P_i, i = 1 \ldots n$ (repeated in each 'time slice'), $\delta_j, j = 1 \ldots m$ (repeated in each 'time slice') and $R_\alpha, \alpha = 1 \ldots k$ (only in the last 'time slice'). Any technique allowing the computation of normalization constants of *factor graphs* may then be used successively to compute $V_\delta(P^0)$. We may cite for example the *junction tree* algorithm [1] which is exact (and applicable only to problems with a reasonable treewidth), *loopy belief propagation* [5] or *generalized belief propagation* [13], *tree-reweighted belief propagation* [12] which gives an upper bound, *mean field* [7] which gives a lower bound... We propose to deal with the infinite horizon by using a cut-off time t^* managed with a parameter ϵ corresponding to the following condition : $\frac{\gamma^{t^*+1}}{1-\gamma} R_{\text{max}} \leq \epsilon$, guaranteeing an absolute error less than ϵ in the policy evaluation, in the (utopian) case where we can compute $C(t)$ exactly. The evaluation of the policy is thus obtained by the computation of t^* normalization constants.

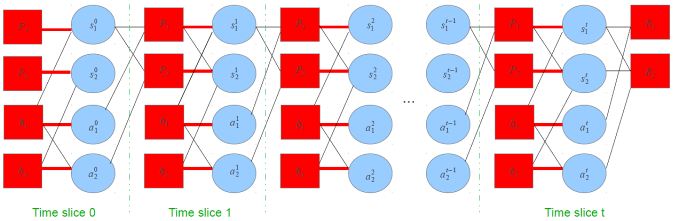

Figure 1. Example of *factor graph* representing $P_\delta((s,a)_{0:t}|P^0) \prod_{\alpha=1}^{k} R_\alpha\left(pa_R(R_\alpha^t)\right)$ for a given FA-FMDP with two state and action variables and two reward features ($n = m = k = 2$)

In the case of additive rewards : $R(s^t,a^t) = \sum_{\alpha=1}^{k} R_\alpha\left(pa_R(R_\alpha^t)\right)$, we can also rewrite the value of δ for P^0 as a sum of normalization constants :

$$
\begin{aligned}
V_\delta(P^0) &= \sum_{t=0}^{+\infty} \gamma^t \sum_{s^0\ldots s^t} \sum_{a^0\ldots a^t} P_\delta((s,a)_{0:t}|P^0) \sum_{\alpha=1}^{k} R_\alpha\left(pa_R(R_\alpha^t)\right) \\
&= \sum_{t=0}^{+\infty} \gamma^t \sum_{\alpha=1}^{k} \sum_{s^0\ldots s^t} \sum_{a^0\ldots a^t} P_\delta((s,a)_{0:t}|P^0) R_\alpha\left(pa_R(R_\alpha^t)\right)
\end{aligned}
$$

We thus have $V_\delta(P^0) = \sum_{t=0}^{+\infty} \gamma^t \sum_{\alpha=1}^{k} C_\alpha(t)$ where $C_\alpha(t)$ is the normalization constant of a *factor graph* which is slightly different from the one of figure 1, with only the reward feature α in the last time slice. With the same stopping time t^* as before, the evaluation is obtained by the computation of kt^* normalization constants.

Results of experiments on simulations will be given during the workshop, for small size problems in comparison with the MDP value (computed by matrix calculus), and for large size problems in comparison with an evaluation by Monte-Carlo simulations.

Conclusion

We have proposed an original way of evaluating a stochastic factored policy in the general case of FA-FMDPs, based on the computation of normalization constants of *factor graphs* with increasing sizes. The first results are encouraging. We are now thinking to an optimization procedure, in order to develop a policy iteration algorithm. The policy iteration algorithm iterates phases of policy evaluation and phases of greedy optimisation of the current policy δ :

$$
\forall s \in \mathcal{S}, \delta'(s) = \operatorname*{argmax}_a \left\{ R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_\delta(s') \right\}
$$

In our case where policies are stochastic, this greedy optimisation problem is a continuous optimisation problem (the parameters of the local functions $\delta_j, j = 1 \ldots m$ should be optimised). The difficulty of this optimisation problem depends on whether the function to maximise has some convexity property, which still has to be determined. In all cases, local optimisation methods (gradient-based approaches, for example), may allow to strictly improve the value of the current policy, which is generally sufficient for an approximate policy iteration algorithm to compute "satisfying" approximately optimal policies.

REFERENCES

[1] C. M. Bishop, 'Pattern Recognition and Machine Learning', chapter Graphical models, Springer, (2007).

[2] M. Botvinick and M. Toussaint, 'Planning as inference', *Trends in Cognitive Sciences*, **16**, 485–488, (2012).

[3] C. Guestrin, M. Hauskrecht, and B. Kveton, 'Solving factored MDPs with continuous and discrete variables', in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 235–242, (2004).

[4] K-E. Kim and T. Dean, 'Solving factored MDPs with large action space using algebraic decision diagrams', in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence*, PRICAI '02, pp. 80–89, (2002).

[5] F. R. Kschischang, B. J. Frey, and H-A. Loeliger, 'Factor Graphs and the Sum-Product Algorithm', *IEEE Transactions on information theory*, **47**(2), 498–519, (2001).

[6] Q. Liu and A. T. Ihler, 'Belief propagation for structured decision making', in *UAI*, pp. 523–532, (2012).

[7] N. Peyrard, *Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales*, Ph.D. dissertation, Université Joseph Fourier - Grenoble 1, 2001.

[8] M.L Puterman, *Markov decision processes*, John Wiley and Sons, 1994.

[9] A. Raghavan, S. Joshi, A. Fern, P. Tadepalli, and R. Khardon, 'Planning in factored action spaces with symbolic dynamic programming', in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, (2012).

[10] R. Sabbadin, N. Peyrard, and N. Forsell, 'A framework and a mean-field algorithm for the local control of spatial processes', *International Journal of Approximate Reasoning*, **53**(1), 66–86, (2012).

[11] M. Toussaint and A. J. Storkey, 'Probabilistic inference for solving discrete and continuous state Markov Decision Processes', in *ICML*, pp. 945–952, (2006).

[12] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, 'Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching', in *Workshop on Artificial Intelligence and Statistics*, (2003).

[13] J. S. Yedidia, W. T. Freeman, and Y. Weiss, 'Constructing free-energy approximations and generalized belief propagation algorithms', *IEEE Transactions on Information Theory*, **51**(7), 2282–2312, (2005).