

# Finding good stochastic factored policies for factored Markov decision processes

Julia Radoszycki and Nathalie Peyrard and Régis Sabbadin<sup>1</sup>

**Abstract.** We propose a framework for approximate resolution of MDPs with factored state space, factored action space and additive reward, based on (i) considering stochastic factored policies (SFPs) with a given structure, (ii) using variational approximations to estimate SFP values and (iii) using local continuous optimization algorithms to compute “good” SFPs. We have implemented and tested an algorithm (CA-LBP), involving a *loopy belief propagation* algorithm and a *coordinate ascent* procedure. Experiments show that CA-LBP performs as well as a state-of-the-art algorithm dedicated to a specific sub-class of FA-FMDPs, and that CA-LBP can be applied to general FA-FMDPs with up to 100 binary state variables and 100 binary action variables.

## 1 INTRODUCTION

Markov Decision Processes (MDPs) form a suitable tool for modelling problems of sequential decision under uncertainty. However, direct application to domains like robotics, environmental management, etc. is not straightforward because problems representations are often factored. Developing methods for Markov decision processes with factored state and action spaces (FA-FMDPs, [5]) is an active topic of research. However, existing approaches often consider less general problems where only the state or the action space is factored [3] or where both spaces have the same factorization (GMDPs, [9, 2]). Some recent approaches are more general, making use of simulation [10, 1, 4] or symbolic dynamic programming [8] to solve FA-FMDPs, but they do not solve problems as large as the ones we tackle. In the framework of Dec-POMDPs, an EM algorithm was proposed that seems to scale better [7], but assumes additive value function.

We propose a framework for FA-FMDPs with additive reward, based on (i) considering stochastic factored policies (SFPs) with a given structure, (ii) using variational approximations to estimate SFP values and (iii) using local continuous optimization algorithms to compute “good” SFPs. In this framework, we propose an algorithm, CA-LBP, based on the use of Loopy Belief Propagation (LBP, [6]) for evaluation and Coordinate Ascent (CA) for optimization.

## 2 FRAMEWORK AND METHODS

The method we propose in this paper is dedicated to the resolution of FA-FMDPs whose reward function is a sum of small scope reward functions. Such an FA-FMDP is an MDP  $\langle S, \mathcal{A}, P, R \rangle$  with:

- Factored state space:  $S = \prod_{i=1}^n S_i$ , with each  $S_i$  a finite set; the state of the system at time  $t$  is noted<sup>2</sup>:  $S^t = (S_1^t, \dots, S_n^t) \in S$

- 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  $A^t = (A_1^t, \dots, A_m^t) \in \mathcal{A}$
- Factored 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  $pa_P(s_i^{t+1}) \subseteq \{s_j^t, j = 1 \dots n, s_{j'}^{t+1}, j' = 1 \dots n, j' \neq i, a_k^t, k = 1 \dots m, a_{k'}^{t+1}, k' = 1 \dots m\}$ . Synchronous arcs are allowed, but the underlying directed graph must be acyclic.
- Additive reward function:  $R(s^t, a^t) = \sum_{\alpha=1}^r R_{\alpha}(pa_R(R_{\alpha}))$ , where  $pa_R(R_{\alpha}) \subseteq \{s_i^t, i = 1 \dots n, a_j^t, j = 1 \dots m\}$ .

One important question when considering factored MDPs is policy representation. Several articles on GMDPs [9, 2], a sub-class of FA-FMDPs, consider deterministic factored policies, whose structure is defined by the graph linked to the transition. But one can easily find examples where, for a given structure, the best factored policy is stochastic. In this paper we propose to consider stochastic factored policies (SFPs) with an arbitrary structure. We will consider a given factored initial distribution on states:  $P^0(s^0) = \prod_{i=1}^n P_i^0(s_i^0)$  and look for a good SFP for this initial distribution.

An SFP represents the joint probability of choosing  $A^t = a^t$  when the system is in state  $s^t$  at time  $t$ . It is defined as:  $\delta(a^t|s^t) = \prod_{j=1}^m \delta_j(a_j^t|pa_{\delta}(a_j^t))$  where the  $\delta_j(\cdot|pa_{\delta}(a_j^t))$  are conditional probability distributions and  $pa_{\delta}(a_j^t) \subseteq \{s_i^t, i = 1 \dots n, a_k^t, k = 1 \dots m, k \neq j\}$  describes the structure of the SFP. We only consider factored policies for which the graph of dependencies corresponding to the transition and the SFP is acyclic. The purpose of this work is to find good SFPs for a given structure. For a given SFP  $\delta$ , let  $P_{\delta}^t((s, a)^{0:t})$  be the probability of a trajectory  $(s, a)^{0:t} = \langle s^0, a^0, \dots, s^t, a^t \rangle$ . The value of  $\delta$  for a given initial distribution  $P^0$ , in the finite horizon case with horizon  $T$ , is defined as:

$$\begin{aligned} V_{\delta}^{R,T}(P^0) &= \mathbb{E}_{P_{\delta}^T} \left[ \sum_{t=0}^T R(S^t, A^t) \middle| P^0, \delta \right] \\ &= \sum_{t=0}^T \sum_{\alpha=1}^r \sum_{pa_R(R_{\alpha})} b_{\alpha}^t(pa_R(R_{\alpha})) R_{\alpha}(pa_R(R_{\alpha})). \end{aligned}$$

where  $b_{\alpha}^t(pa_R(R_{\alpha}))$  is the marginal distribution over variables influencing reward  $\alpha$  at time  $t$  (it is computed by marginalizing  $P_{\delta}^T((s, a)^{0:T})$  over all other state and action variables from time 0 to time  $T$ ). We propose to use Loopy Belief Propagation (LBP, [6]) for computing approximations of these marginal distributions. In the infinite horizon case with discount  $\gamma$ , the infinite sum can be

<sup>1</sup> INRA-MIAT (UR 875), F-31326, Castanet-Tolosan, France. Julia.Radoszycki@toulouse.inra.fr

<sup>2</sup> We will note random variables with capital letters and their realizations with lower case letters. Following the same idea, notations  $pa_P(s_i^t)$  and  $pa_P(s_i^t)$  represent respectively the random variables influencing  $S_i^t$  and an instantiation of these variables.

approximated by a sum over a finite temporal horizon  $T$ , sufficiently large. When we look for stochastic factored policies that maximize this value, we are facing a continuous constrained optimization problem:

$$\begin{aligned} & \underset{\delta \in (\mathbb{R}^+)^N}{\text{maximize}} && V_\delta^{R,T}(P^0) \\ & \text{subject to} && \sum_{a_j \in \mathcal{A}_j} \delta_j(a_j | pa_\delta(a_j)) = 1 \quad \forall j, \forall pa_\delta(a_j) \end{aligned}$$

where  $N$  is the number of optimization variables (policy parameters).

**Proposition.** *The decision version of this problem is  $NP^{PP}$ -complete. See future long paper for the proof.*

Regarding the optimization algorithm, we used a cyclic coordinate ascent algorithm on the equivalent problem with inequality constraints and fewer ( $N'$ ) optimization variables. In the case of binary action variables, the constraints are parallel to the axes and, if evaluation was exact, the coordinate ascent algorithm would be guaranteed to converge to a local optimum. The combination of LBP for value approximation and Coordinate Ascent (CA) for optimization defines the CA-LBP algorithm.

### 3 EXPERIMENTS AND FUTURE WORK

We first validated CA-LBP by comparing the policy it provides to the optimal global (deterministic) policy on small GMDP and FA-FMDP problems. We also compared it with the MF-API algorithm [9] which approximately solves GMDPs. We generated 100 random GMDPs and FA-FMDPs without synchronous arcs with  $n = m = r = 6$ ,  $\forall i = 1 \dots n, |\mathcal{S}_i| = |\mathcal{A}_i| = 2$ ,  $\gamma = 0.9$ ,  $T = 20$ . Both the structure and the values of the transitions and the rewards were random (with a sparsity constraint). In the case of GMDPs, the structure of the policy must be the same as that of the transition and reward. In the case of FA-FMDPs we used a 'natural' structure for the policy, where a state variable is a parent of an action variable if (i) they are both present jointly in the scope of a local reward function or if (ii) they are both parents of a common state variable, wrt a local transition function. As shown in Table 1, for the two types of problems, CA-LBP is able to provide an SFP with good value, compared to the value of the optimal global policy. In the case of GMDP problems, MF-API leads to better results than CA-LBP (however, the order was inverted when we considered an horizon of  $T = 40$ ). Differences in implementation of MF-API and CA-LBP make it difficult to compare computational times and these are just indicative.

**Table 1:** Results on random GMDPs and FA-FMDPs. MRE=mean relative error. Scores between brackets are scores for the deterministic policies derived from the SFPs.

		CA-LBP	MF-API	global optimum
GMDP	MRE	0.0065 (0.0066)	0.0015	-
	mean time	4.45s	1.79s	115.16s
FA-FMDP	MRE	0.1228 (0.1230)	-	-
	mean time	91.8s	-	117s

We then considered a generalization of the GMDP problem of disease management in crop fields with two infection states [9], to an FA-FMDP. The difference with the experiment described in [9] is that treatment occurs before disease spread and reduces the probability of contamination of neighbouring fields. The transition table is the same as in [9] except that the probability for an uninfected field to become infected when there is no treatment is :

$P(\epsilon, p_1, n_1, p_2, n_2) = \epsilon + (1-\epsilon)(1-(1-p_1)^{n_1}(1-p_2)^{n_2})$ , where  $\epsilon$  is the probability of long distance contamination,  $p_1$  (resp.  $p_2$ ) is the probability of short distance contamination from non treated (resp. treated) fields,  $n_1$  (resp.  $n_2$ ) is the number of non treated (resp. treated) infected neighbouring fields. Table 2 shows the results for a  $5 \times 5$  grid ( $n = m = 25, N' = 512$ ) and a  $10 \times 10$  grid ( $n = m = 100, N' = 2592$ ), with a maximum of four neighbours per field. We used  $T = 20$  for LBP evaluation (using  $T = 40$  did not change the results). We compared the CA-LBP policy with the uniform policy (both actions are chosen with probability 0.5), a random stochastic policy and the greedy policy. All policies were evaluated with 4000 Monte-Carlo simulations for  $T = 40$ . CA-LBP improves the value of the greedy policy by 46% in the case of the  $5 \times 5$  grid and 32% in the case of the  $10 \times 10$  grid.

**Table 2:** Evaluation of the disease management problem policies.

		CA-LBP (time)	unif.	rand.	greedy
$5 \times 5$ $10 \times 10$	MC value	19142 (1h16)	10436	10801	13092
	MC value	69175 (43h42)	40495	40014	52191

Ongoing work consists in using a gradient ascent algorithm for optimization when the action variables are not binary, with parallelization of the computation of the gradient. We are also planning to try to extend *planning as inference* methods [7] to FA-FMDPs with large state and action spaces, and compare them with our approach. Future research could be on two extensions of our approach : partially observable problems and problems with undetermined policy structures.

### ACKNOWLEDGEMENTS

We would like to thank the referees for helpful comments on previous version and Victor Picheny for helpful advice on optimization algorithms. This work was funded by ANR-13-AGRO-0001-04.

### REFERENCES

- [1] O. Buffet and D. Aberdeen, 'The factored policy-gradient planner', *Artificial Intelligence*, **173**, 722–747, (2009).
- [2] Q. Cheng, Q. Liu, F. Chen, and A. Ihler, 'Variational planning for graph-based MDPs', in *Advances in Neural Information Processing Systems* 26, pp. 2976–2984, (2013).
- [3] J. Hoey, R. St-Aubin, A. Hu, and C. Boutilier, 'SPUDD: Stochastic planning using decision diagrams', in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, (1999).
- [4] T. Keller and P. Eyerich, 'PROST: Probabilistic Planning Based on UCT', in *Proceedings of the 22nd International Conference on Automated Planning and Scheduling*, (2012).
- [5] 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*, (2002).
- [6] 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).
- [7] A. Kumar, S. Zilberstein, and M. Toussaint, 'Scalable multiagent planning using probabilistic inference', in *Proceedings of the 22th International Joint Conference on Artificial intelligence*, (2011).
- [8] A. Raghavan, S. Joshi, A. Fern, P. Tadepalli, and R. Khordon, 'Planning in factored action spaces with symbolic dynamic programming', in *Proceedings of the 26th AAAI Conference*, (2012).
- [9] 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).
- [10] B. Sallans and G. E. Hinton, 'Reinforcement learning with factored states and actions', *Journal of Machine Learning Research*, **5**, 1063–1088, (2004).