ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1724

Decoupling a Resource Constraint Through Fictitious Play in Multi-Agent Sequential Decision Making

Frits de Nijs[1] and **Matthijs T. J. Spaan**[1] and **Mathijs M. de Weerdt**[1]

Abstract. When multiple independent agents use a limited shared resource, they need to coordinate and thereby their planning problems become coupled. We present a resource assignment strategy that decouples agents using marginal utility cost, allowing them to plan individually. We show that agents converge to an expected cost curve by keeping a history of plans, inspired by fictitious play. This performs slightly better than a state-of-the-art best-response approach and is significantly more scalable than a preallocation Mixed-Integer Linear Programming formulation, providing a good trade-off between performance and quality.

1 INTRODUCTION

When multiple agents must coordinate under a shared resource constraint, individually tractable problems become tightly coupled through the dependency on the resource consumption of all other agents. In problems where agents have the ability to compute and execute their own plans, these agents may be used to decouple the problem into efficiently solvable sub-problems [4].

Resource-constrained agents can be decoupled by preallocating resources a priori. Wu and Durfee [5] present a Mixed-Integer Linear Programming (MILP) formulation to optimally preallocate resources. Unfortunately, preallocating resources still has an exponential complexity, which prevents application to real-world scale problems. To overcome these restrictions, we proposed an on-line conflict resolution approach by planning a best-response policy to the likelihood of successfully executing constrained actions [3]. While this results in efficiently computable policies, the assignment of such a state-independent success probability may be overly pessimistic.

In this paper we propose to look at the marginal utility gained as a consequence of being assigned a resource. By comparing this utility to that of other agents, they can make an informed decision on the distribution of resources. We use this idea to decouple agents by computing a marginal utility cost for the resource. The key insight is that a cost allows agents to compute an expected resource assignment also based on their state. Convergence of the cost function is obtained by keeping a history of expected states, similar to fictitious play.

2 PROBLEM DESCRIPTION

We define Resource Constrained Multi-agent Markov Decision Processes (RC-MMDPs) as an extension of finite horizon MMDPs [2].

Each individual agent i is modeled as a Markov Decision Process specified by tuple $M_i = \langle S_i, A_i, P_i, R_i \rangle$. The current state of an agent is an element $s_{i,j}$ of set S_i containing a finite number of possible states. In any state the agent can choose one of the finite number of actions $a_{i,j}$ contained in set A_i. The transition function $P_i(s_{i,l} \mid s_{i,j}, a_{i,k})$ defines the probability that agent i ends up in state $s_{i,l}$ from state $s_{i,j}$ by choosing action $a_{i,k}$. Agents are rewarded for their choice through reward function $R_i(s_{i,j}, a_{i,k})$ which returns a real-valued utility.

The independent agent problems are coupled through a resource constraint, turning it into an RC-MMDP problem. RC-MMDP problems are specified by tuple $\langle \mathcal{M}, c, L, h \rangle$. Set $\mathcal{M}$ contains the n individual agent problems, $\mathcal{M} = \langle M_1, M_2, \ldots, M_n \rangle$. The binary cost function $c(a_{i,j})$ is set to 1 if action $a_{i,j}$ uses the resource. We require that all agents have an action with $c(a_\emptyset) = 0$ to ensure feasibility of the model. The non-negative resource consumption limit L_t specifies the maximum consumption at any time t in finite horizon h.

Because the agents are cooperative, the goal of the agents is to maximize the sum of individual agent utilities over the entire horizon. A policy $\pi(\mathbf{s}, t)$ specifies for joint state $\mathbf{s} = \langle s_1, s_2, \ldots, s_n \rangle$ at time t which (feasible) joint action $\mathbf{a} = \langle a_1, a_2, \ldots, a_n \rangle$ the agents should take. Action $\mathbf{a}$ is feasible at time t if $c(\mathbf{a}) \leq L_t$, $c(\mathbf{a}) = \sum_{i=1}^{n} c(a_i)$.

The goal of RC-MMDP planning is to compute an optimal policy π^*, which returns the feasible joint action with the highest expected value for every possible joint state and time. We define the expected value of state $\mathbf{s}$ by following policy π as $V_\pi[\mathbf{s}, t]$, with $V_\pi[\mathbf{s}, h] = 0$. Given this, we define the expected value of taking action $\mathbf{a}$ in state $\mathbf{s}$ as

$$Q_\pi[\mathbf{s}, \mathbf{a}, t] = \mathcal{R}(\mathbf{s}, \mathbf{a}) + \sum_{\mathbf{s}' \in \mathcal{S}} \left(\mathcal{P}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) \cdot V_\pi[\mathbf{s}', t+1] \right). \tag{1}$$

3 MARGINAL UTILITY COST PLANNING

To improve on the preallocation algorithm, we propose to have the agents agree on the marginal utility cost u of the resource. The agents include this cost in their action selection. When for two actions it holds that (for decoupled policy π_i of agent i)

$$\begin{aligned} Q_{\pi_i}[s, a_1, t] > Q_{\pi_i}[s, a_2, t], \text{ and} \\ Q_{\pi_i}[s, a_1, t] - c(a_1) \cdot u < Q_{\pi_i}[s, a_2, t] - c(a_2) \cdot u, \end{aligned} \tag{2}$$

the agent will choose action a_2, even though it prefers a_1 in the unconstrained case. In general we are looking for the marginal utility cost $u_{\mathbf{s},t}$ which makes the sum of resource consumption induced over the preferred actions for joint state $\mathbf{s}$ at time t fit in the resource limit:

$$\begin{aligned} \max_{u_{\mathbf{s},t}} \; & \sum_{i=1}^{n} Q_{\pi_i}[\mathbf{s}_i, a_{i,j}, t] \\ \text{s.t.} \; & \sum_{i=1}^{n} c\left(\arg\max_{a_{i,j} \in A_i} \left(Q_{\pi_i}[\mathbf{s}_i, a_{i,j}, t] - c(a_{i,j}) \cdot u_{\mathbf{s},t} \right) \right) \leq L_t \\ & u_{\mathbf{s},t} \geq 0. \end{aligned} \tag{3}$$

[1] Delft University of Technology, email: f.denijs@tudelft.nl

Because we know there exists an action $c(a_\emptyset) = 0$, we are guaranteed that a feasible cost exists. The cost $u_{\mathbf{s},t}$ can be computed by sorting the expected future marginal utility values of the agents, and assigning the preferred action to each agent until the constraint is reached. The marginal utility of the agent that consumes the last remaining resource is equal to the cost $u_{\mathbf{s},t}$ that prevents overconsumption in state $\mathbf{s}$.

Of course, changing the executed actions of some agents can make their state trajectories deviate substantially from their plans. Therefore, our key idea is that agents coordinate on the expected resource cost $E[u_t]$ at plan time. Because the expected cost depends on the expected joint states that the agents visit, which in turn depends on their policy, we first let agents plan for the unconstrained case where $E[u_t] = 0, \forall t$. The resulting policies are then evaluated to obtain an informed prior over the reachable states. Let a prior over the starting states $p_{i,1}$ be given for each agent. Since the number of reachable states is (typically) exponential in the number of agents, we propose to perform Monte Carlo sampling to obtain an approximation of the probability distribution $p_t(\mathbf{s})$. Given this prior, the expected resource cost subject to the joint policy $\pi = \langle \pi_1, \pi_2, \ldots, \pi_n \rangle$ is

$$E_\pi[u_t] = \sum_{\mathbf{s}} p_t(\mathbf{s}) \cdot u_{\mathbf{s},t}, \tag{4}$$

where $u_{\mathbf{s},t}$ is determined by solving Equation 3. The agents can then re-plan their policies taking into account this resource cost by applying the modified Bellman equation

$$V_{\pi_i}[\mathbf{s}_i, t] = \max_{a_{i,j} \in A_i} \left(Q_{\pi_i}[\mathbf{s}_i, a_{i,j}, t] - c(a_{i,j}) \cdot E_\pi[u_t] \right). \tag{5}$$

The joint policy is derived by planning all agents individually using Value Iteration with this modified Bellman equation. Since each iteration modifies the expected value at time t, the expected cost $E_\pi[u_t]$ also needs to be updated to reflect future values. Therefore, cost $E_\pi[u_t]$ is computed on the basis of the newest $V_\pi[\mathbf{s}, t+1]$, before the Bellman equation is applied for time t.

This process changes where resource constraints restrict agents' actions. Thus, these steps should be repeated until convergence of the expected cost function. It is easy to imagine that the cost function may oscillate between extremes if we only consider the previous prior. Therefore, to ensure convergence, we keep the history of all past samples, inspired by fictitious play [1]. Each prior can be seen as the adversary 'nature' performing her actions as a consequence of our choices. By remembering all past plays, eventually the full strategy of nature is obtained. Thus, let p^k be the probability distribution over states in iteration (or play) k, then we maintain the set $\mathbb{P} = \langle p^1, p^2, \ldots, p^k \rangle$, and compute the expected cost as

$$E_\pi[u_t] = \sum_{j=1}^{k} \sum_{\mathbf{s} \in p_t^j} \frac{p_t^j(\mathbf{s})}{k} \cdot u_{\mathbf{s},t}. \tag{6}$$

4 EMPIRICAL EVALUATION

To evaluate the performance of this algorithm we compare it against an optimal preallocation MILP [5] and our Best-response planner [3] on an energy-consumption planning problem. In this setting a population of electric heaters must be controlled to keep the aggregate consumption below a power constraint, while satisfying consumers' heat demands. The power constraint may arise due to fluctuating supply of renewable sources like wind or solar. In the experiments we measure the time to compute a policy for 4 agents, and its quality. The Fictitious Play and Best-response algorithms are set to perform at most 10 iterations, computing 1000 priors each iteration.

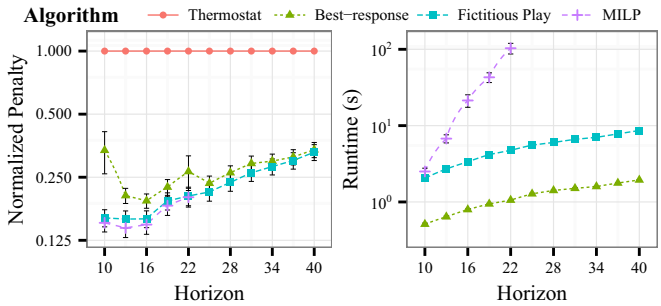

Figure 1. Algorithm performance for increasing horizon: policy quality normalized to the thermostat policy (left), and wall-clock computation time (right). Both plots on a log scale, lower values are better.

Figure 1 presents the mean and standard error of both runtime and policy quality. The policy quality metric penalizes the total amount of deviation of the current temperature from the set-point temperature. The quality is normalized to the myopic strategy of using thermostat controllers with an on-line prioritized load-shedding system to keep the resource demand below the limit.

Based on the MILP formulation, we expect that a linear increase in the length of the horizon results in an exponential growth of the runtime. We observe this exponential scaling in the right plot; several instances of $h = 22$ could not be solved within 30 minutes. The other algorithms have polynomial complexity, and are able to solve each instance within at most 10 seconds. Nevertheless, the policies found by Fictitious Play are almost as good as the MILP policies, and significantly better than Best-response for short horizon instances.

5 CONCLUSIONS AND FUTURE WORK

This paper introduces a decoupling algorithm for multi-agent planning problems under hard resource constraints based on fictitious play. The algorithm computes a time-dependent cost for resources which is used to decouple individual policies so that they can be computed in polynomial time. We compared against two state-of-the-art approaches, and found that the fictitious play algorithm produces policies which are not significantly worse than an optimal preallocation decoupling while requiring exponentially less runtime.

For future work we intend to adapt the fictitious play algorithm to handle stochastic resource levels and multiple resources.

ACKNOWLEDGEMENTS

Support of this research by network company Alliander is gratefully acknowledged.

REFERENCES

[1] U. Berger, 'Brown's original fictitious play', *Journal of Economic Theory*, **135**(1), 572–578, (2007).
[2] C. Boutilier, 'Planning, Learning and Coordination in Multiagent Decision Processes', in *TARK*, pp. 195–210, (1996).
[3] F. de Nijs, M. T. J. Spaan, and M. M. de Weerdt, 'Best-Response Planning of Thermostatically Controlled Loads under Power Constraints', in *AAAI*, pp. 615–621, (2015).
[4] F. A. Oliehoek, S. J. Witwicki, and L. P. Kaelbling, 'Influence-Based Abstraction for Multiagent Systems', in *AAAI*, pp. 1422–1428, (2012).
[5] J. Wu and E. H. Durfee, 'Resource-Driven Mission-Phasing Techniques for Constrained Agents in Stochastic Environments', *JAIR*, **38**, 415–473, (2010).