Value-Based Planning for Teams of Agents in Stochastic Partially Observable Environments
Oliehoek, F.A.

Citation for published version (APA):
Summary

Situations in which multiple decision makers influence an environment arise in many important current and future real-world problems such as crisis management, network control, robotic teams and distributed software applications. Making decisions in such multiagent systems is of crucial interest to artificial intelligence and related fields.

This thesis is concerned with the task of computing a plan for a team of cooperative agents. Many real-world planning tasks for such teams of agents are subject to uncertainty: both the outcome of the actions and the perception of the current state of the environment may be uncertain and each of the agents may have a different partial view of this environment. Also, the agents may be uncertain with respect to each other’s actions. Such settings can be captured by the decentralized partially observable Markov decision process (Dec-POMDP), a decision-theoretic model that allows a principled treatment of the mentioned uncertainties. Unfortunately, computing an optimal plan, or joint policy, that specifies for each agent what to do in each possible situation is proven to be intractable and even finding a bounded approximation to the optimal solution is NEXP-complete. This means that for many interesting problems we have to resort to approximation methods that will not be able to guarantee a bound on the quality of the joint policy.

One option is to apply optimization methods such as genetic algorithms or cross entropy to find a joint policy. However, such methods do not exploit the structure of the problem nor do they provide any insight in how the found approximation relates to an optimal solution. Therefore, this thesis describes a value-based approach. For single-agent planning (as formalized by the Markov decision process) many algorithms exist that find an (approximate) solution by constructing an optimal value function that represents the expected cumulative reward from each state, and subsequently extracting an optimal policy from the value function. This thesis discusses how a similar procedure can be applied in decentralized settings by identifying optimal value functions for Dec-POMDPs. By using the optimal value function as the payoff function in a series of Bayesian games (BGs) the optimal policy can be found, thereby extending the solution method of Emery-Montemerlo et al. (2004), to which we refer as forward-sweep policy computation (FSPC), to include the exact setting.

It may come as no surprise that computing an optimal value function is also intractable, therefore this thesis proposes to use approximate value functions that are easier to compute. In particular, it covers $Q_{MDP}$ and $Q_{POMDP}$ and proposes
a new approximation $Q_{BG}$ and applies them in a heuristic policy search method dubbed \textit{generalized multiagent A*} (GMAA*). GMAA* unifies FSPC and multiagent A* (MAA*) (Szer et al., 2005) and works by solving BGs for different stages. In a BG for a particular stage $t$, each agent has to select an action for each of its possible histories. By setting a parameter $k$ to 1 GMAA* reduces to FSPC and gives an approximate solution, while for $k = \infty$ the behavior is identical to MAA* and the method is exact. Still, the scalability of GMAA* is limited by the fact that the BGs grow exponentially with respect to the number of agents and time (because the number of histories grows exponentially with time).

To counter the first type of growth, the thesis explores how independence between agents can be exploited: in typical problems not all agents will have to interact at the same time which leads to sparseness in interaction. We propose to exploit this sparseness by using collaborative graphical Bayesian games (CGBGs), which can be represented much more compactly than the regular BGs. For these CGBGs it is possible to efficiently find approximate solutions by converting them to a factor-graph and applying \textit{Max-Plus}, a message passing algorithm that operates on this graph.

To reduce the growth induced by the number of histories, we consider clustering histories, an idea first introduced by Emery-Montemerlo et al. (2005). However, their approach uses an ad-hoc heuristic to determine which histories to cluster and consequently finds only approximate solutions. By contrast, the work presented in this thesis identifies a criterion that \textit{guarantees} that two individual histories have the same optimal value, allowing \textit{lossless clustering} and therefore faster optimal solutions of Dec-POMDPs and solutions over longer horizons.

The thesis closes with some general conclusions and a discussion of the main directions of future work for practical Dec-POMDP solutions.