Conditional Return Policy Search for TI-MMDPs with Sparse Interactions[1]

Joris Scharpff[a] Diederik M. Roijers[b] Frans A. Oliehoek[c]
Matthijs T. J. Spaan[a] Mathijs M. de Weerdt[a]

[a] *Delft University of Technology*
[b] *University of Oxford*
[c] *University of Amsterdam, University of Liverpool*

When cooperative teams of agents are planning in uncertain domains, they must coordinate to maximise their (joint) team value. In several problem domains, such as maintenance planning [6], the full state of the environment is assumed to be known to each agent. Such *centralised* planning problems can be formalised as multi-agent Markov decision processes (MMDPs) [1], in which the availability of complete and perfect information leads to highly-coordinated policies. However, these models suffer from exponential joint action spaces as well as a state that is typically exponential in the number of agents. This is especially an issue when *optimal* policies are required. In this paper, we identify a significant MMDP sub-class whose structure we compactly represent and exploit via locally-computed upper and lower bounds on the optimal policy value. We exploit both the compact representation, and the upper and lower bounds to formulate a new branch-and-bound policy search algorithm we call *conditional return policy search (CoRe)*. CoRe typically requires less runtime than the available alternatives and finds solutions to previously unsolvable problems [5].

We consider *transition independent* MMDPs (TI-MMDPs). In TI-MMDPs, agent rewards depend on joint states and actions, but transition probabilities are *individual*. Our key insight is that we can exploit the reward structure of TI-MMDPs by decomposing the *returns* of all execution histories – i.e., all possible state/action sequences from the initial time step to the planning horizon – into components that depend on local states and actions. To do so, we build on three key observations. 1) Contrary to the optimal value function, returns *can* be decomposed without loss of optimality, as they depend only on local states and actions of execution sequences. This allows a compact representation of rewards and efficiently computable bounds on the optimal policy value via a data structure we call the *conditional return graph* (CRG). 2) In TI-MMDPs agent interactions are often sparse and/or local, typically resulting in very compact CRGs. 3) In many problems the state space is transient, i.e., states can only be visited once, leading to a directed, acyclic transition graph. With our first two key observations this often gives rise to *conditional reward independence* – the absence of further reward interactions – and enables agent decoupling during policy search.

In order to represent the returns compactly with local components, we first partition the reward function into additive components $\mathcal{R}_i$ and assign them to agents. The *local* reward for an agent $i \in N$ is given by $\mathcal{R}_i = \{R^i\} \cup \mathcal{R}_i^e$, where R^i is the reward function that only depends on agent i and $\mathcal{R}_i^e$ is the set of interaction reward functions assigned to i (restricted to a subset of those R^e where $i \in e$, i.e., those functions that depend on i have at least one other agent in its scope, e). The sets $\mathcal{R}_i$ are disjoint sub-sets of the reward functions, $\mathcal{R}$. Given a disjoint partitioning $\bigcup_{i \in N} \mathcal{R}_i$ of rewards, the *Conditional Return Graph (CRG)* ϕ_i is a *directed acyclic graph* with for every stage t of the decision process a node for every reachable local state s_i, and for every local transition $(s^i, a^i, \hat{s}^i)$, a tree compactly representing all transitions of the agents in scope in $\mathcal{R}_i$. The tree consists of two parts: an action tree that specifies all dependent local joint actions, and an influence tree, that contains the relevant local state transitions included in the respective joint action.

[1] Full version published in the proceedings of AAAI 2016 [5].

An example CRG for one time step is given in Fig. 1. The graph represents all possible transitions that can effect the rewards in $\mathcal{R}_i$, given the local transitions of the state of agent i (in this case only from s_0^1 to s_1^1). The labels on the path to a leaf node of an influence tree, via a leaf node of the action tree, sufficiently specify the joint transitions of the agents e in scope of the functions $R^e \in \mathcal{R}^i$, such that we can compute the reward $\sum_{R^e \in \mathcal{R}_i} R^e(s^e, \vec{a}^e, \hat{s}^e)$. The wildcard, $*^2$, represents any action of agent 2 for which there is no interaction reward, i.e., all reward functions depending on both agent 1 and agent 2 yield 0. In the paper, we prove that CRGs are indeed a compact representation of histories, and even more so when interactions are sparse.

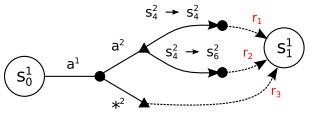

Figure 1: *Example of a CRG for the transition for one agent of a two-agent problem, where R_1 only depends on a^2.*

In addition to storing rewards compactly, we use CRGs to bound the optimal expected value. Specifically, the maximal (resp. minimal) attainable return from a joint state s_t onwards, is an upper (resp. lower) bound on the value. Moreover, the sum of bounds on local returns bounds the global return and thus the optimal value. We define them as $U(s^i) = \max_{(s^e, \vec{a}_t^e, \hat{s}^e) \in \phi_i(s^i)} \left(\mathcal{R}_i(s^e, \vec{a}_t^e, \hat{s}^e) + U(\hat{s}^i)\right)$, such that $\phi_i(s^i)$ denotes the set of transitions available from state $s^i \in s^e$ (ending in $\hat{s}^i \in \hat{s}^e$), in the corresponding CRG. The bound on the optimal value for a joint transition $(s, \vec{a}, \hat{s})$ of all agents is $U(s, \vec{a}_t, \hat{s}) = \sum_{i \in N} \left(\mathcal{R}_i(s^e, \vec{a}_t^e, \hat{s}^e) + U(\hat{s}^i)\right)$, and lower bound L is defined similarly over minimal returns. Note that a bound on the joint returns automatically implies a bound on the value.

We combine the above, together with conditional reward independence, in our *Conditional Return Policy Search (CoRe)* algorithm. CoRe performs a branch-and-bound search over the joint policy space, represented as a DAG with nodes s_t and edges $\langle \vec{a}_t, \hat{s}_{t+1} \rangle$, such that finding a joint policy corresponds to selecting a subset of action arcs from the CRGs (corresponding to $\vec{a}_t$ and $\hat{s}_{t+1}$). First, however, the CRGs ϕ_i are constructed for the local rewards $\mathcal{R}_i$ of each agent $i \in N$, assigned heuristically to obtain balanced CRGs. The generation of the CRGs follows a recursive procedure, during which we store upper and lower bounds on the local returns. During the subsequent policy search, CoRe detects when subsets of agents become conditionally reward independent, and recurses on these subsets separately.

When we compare CoRe to previously available methods, we observe that CoRe can both solve instances that could not previously be solved [5], and that CoRe can solve instances that could be solved by existing methods a lot faster. For example, in the sample of our results presented in Figure 2 we compare the runtime of CoRe to that of SPUDD [2] using a problem-tailored encoding [6] for instances of the maintenance planning problem.

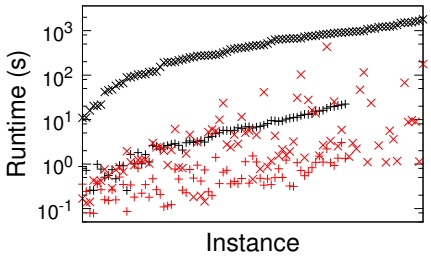

Figure 2: *Experimental results: the runtime of CoRe (red) versus that of SPUDD (black), on the same 2-(+) and 3-agent (×) instances.*

Finally, inspired by the success of CoRe for single-objective TI-MMDPs, we have shown [3] that we can extend our earlier work on multi-objective (TI-)MMDPs [4], using CoRe as a subroutine, to solve significantly larger multi-objective problem instances as well. We thus conclude that CoRe is vital to keeping both single- and multi-objective TI-MMDPs tractable.

References

[1] Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *TARK*, pages 195–210, 1996.

[2] Jesse Hoey, Robert St-Aubin, Alan Hu, and Craig Boutilier. SPUDD: Stochastic planning using decision diagrams. In *UAI*, pages 279–288, 1999.

[3] Diederik M. Roijers. *Multi-Objective Decision-Theoretic Planning*. PhD thesis, Univ. of Amsterdam, 2016.

[4] Diederik M. Roijers, Joris Scharpff, Matthijs T. J. Spaan, Frans A. Oliehoek, Mathijs de Weerdt, and Shimon Whiteson. Bounded approximations for linear multi-objective planning under uncertainty. In *ICAPS*, pages 262–270, 2014.

[5] Joris Scharpff, Diederik M. Roijers, Frans A. Oliehoek, Matthijs T. J. Spaan, and Mathijs M. de Weerdt. Solving transition-independent multi-agent MDPs with sparse interactions. In *AAAI*, pages 3174–3180, 2016.

[6] Joris Scharpff, Matthijs T. J. Spaan, Mathijs M. de Weerdt, and Leentje Volker. Planning under uncertainty for coordinating infrastructural maintenance. In *ICAPS*, pages 425–433, 2013.