Option-Critic in Cooperative Multi-agent Systems

Extended Abstract

Jhelum Chakravorty[1,3], Patrick Nadeem Ward[1,3], Julien Roy[2,3], Maxime Chavelier-Boisvert[3], Sumana Basu[1,3], Andrei Lupu[1,3], Doina Precup[1,3,4]
[1] McGill University, [2] University of Montreal, [3] Mila, [4] DeepMind

ABSTRACT

We investigate planning and learning temporal abstractions in cooperative multi-agent systems using common information approach and report the competitive performance of our proposed algorithm with baselines in grid-world environment.

ACM Reference Format:
Jhelum Chakravorty[1,3], Patrick Nadeem Ward[1,3], Julien Roy[2,3], Maxime Chavelier-Boisvert[3], Sumana Basu[1,3], Andrei Lupu[1,3], Doina Precup[1,3,4]. 2020. Option-Critic in Cooperative Multi-agent Systems. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), Auckland, New Zealand, May 2020,* IFAAMAS, 3 pages.

INTRODUCTION

We leverage *common information approach* [3] to address temporal abstraction in cooperative multi-agent systems. In particular, we address the planning problem in options framework [5] for the Decentralized Partially Observable Markov Decision Process (Dec-POMDP) and propose a model-free learning of temporally abstracted policies. The common information approach circumvents the combinatorial nature of the decentralized system by converting it into an equivalent centralized POMDP. We provide a dynamic programming formulation and argue the existence of an optimal option-policy. We analyze the convergence of our proposed algorithm (DOC) and validate the results with empirical experiments using cooperative multi-agent grid-world environments.

Denote by $\mathcal{E}(\boldsymbol{\omega}_t \mu_t, \mathbf{s}_t)$ the event that joint-option $\boldsymbol{\omega}_t$ is executed at time instant t at joint-state $\mathbf{s}_t$ until its termination, after which a new joint-option is chosen according to option-policy μ_t at the resultant joint-state. The *dynamic team problem* that we are interested to solve is to choose policies that maximize the the infinite-horizon discounted reward: $\mathcal{R}^{\mu_t}$ as given by

$$\mathcal{R}^{\mu_t} = \sup_{\mu_t \in \mathcal{M}} \sum_{\boldsymbol{\omega}_t \in \Omega} \mu_t(\boldsymbol{\omega}_t | \mathbf{s}_t) \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \,|\, \mathcal{E}(\boldsymbol{\omega}_0 \mu_0, \mathbf{s}_0) \right], \quad (1)$$

DEC-POMDP PLANNING WITH TEMPORAL ABSTRACTION

The Common Information Approach [3] is an effective way to solve a Dec-POMDP in which the agents share a common pool of information, updated, for example via broadcasting, in addition to *private* information available only to each individual agent. A *fictitious coordinator* observes the common information and suggests a *prescription* (in our case the Markov joint-option policy μ_t). The joint-option $\boldsymbol{\omega}_t$ is chosen from μ_t and is communicated to all agents j, who in turn generate their own action a_t^j according to their local (private) information, and their own observation $o_t^j : a_t^j \sim \pi_t^j(a_t^j | o_t^j)$. A *locally fully observable* agent chooses its action a_t^j based on its own state s_t^j or embedding e_t^j according to $a_t^j \sim \pi_t^j(a_t^j | s_t^j)$ The notion of a centralized fictitious coordinator transforms the Dec-POMDP into an equivalent centralized POMDP, so one can exploit mathematical tools from stochastic optimization such as dynamic programming to find an optimal solution.

The common information-based belief on the joint-state $\mathbf{s}_t \in \mathcal{S}$ is defined as $b_t^c(\mathbf{s}) := \mathbb{P}(\mathbf{s}_t = \mathbf{s} \,|\, \mathcal{I}_t^c)$, where $\mathcal{I}_t^c$ is the common information at time t, given by $\mathcal{I}_t^c = \{\tilde{\mathbf{o}}_{1:t-1}, \boldsymbol{\omega}_{1:t-1}\}$, where $\tilde{o}_t^j$ is the *broadcast symbol* of agent j. Consequently, $\mathcal{I}_{t-1}^c \subseteq \mathcal{I}_t^c$. b_t^c evolves in a Bayesian manner. Using the argument of [3, Lemma 1], we can show that the coordinated system is a POMDP with prescriptions μ_t and observations

$$\tilde{\mathbf{o}}_t = \tilde{h}_t(\mathbf{s}_t, \mu_t), \quad (2)$$

where $\tilde{h}_t$ is a *Bayesian filter.*

The optimal policy of the coordinated centralized system is the solution of a suitable dynamic program which has a fixed-point. In order to formulate this program, we need to show that b_t^c is an *information state*, i.e. a sufficient statistic to form, with the current joint-option μ_t, a future belief b_{t+1}^c.

Common-belief based option-value

The *option-value upon arrival*, U^μ, and the *option-value*, Q^μ, are defined below, where $\beta_{\text{none}}^{\boldsymbol{\omega}_t}(\mathbf{s}_t)$ is the probability that no agent terminates in $\mathbf{s}_t$.

$$U^{\mu_t}(b_t^c, \boldsymbol{\omega}_t) := \sum_{\mathbf{s}_t \in \mathcal{S}} U^{\mu_t}(\mathbf{s}_t, \boldsymbol{\omega}_t) b_t^c(\mathbf{s}_t) = \sum_{\mathbf{s}_t \in \mathcal{S}} \Big[\beta_{\text{none}}^{\boldsymbol{\omega}_t}(\mathbf{s}_t) Q^{\mu_t}(\mathbf{s}_t, \boldsymbol{\omega}_t) b_t^c(\mathbf{s}_t)$$

$$+ (1 - \beta_{\text{none}}^{\boldsymbol{\omega}_t}(\mathbf{s}_t)) \max_{\mathcal{T} \in \text{Pow}(\mathcal{J})} \max_{\boldsymbol{\omega}_t' \in \Omega(\mathcal{T})} Q^{\mu}(\mathbf{s}_t, \boldsymbol{\omega}_t') b_t^c(\mathbf{s}_t) \Big]. \quad (3)$$

Define operators $\mathcal{B}^{\mu_t}$ as follows:

$$[\mathcal{B}^{\mu_t} Q^{\mu_t}](b_t^c, \boldsymbol{\omega}_t)$$

$$:= \gamma \sum_{\mathbf{s}_t \in \mathcal{S}} \sum_{\mathbf{o}_t \in O} \Bigg(\sum_{\mathbf{br}_t \in \{0,1\}^J} \sum_{\mathbf{a}_t \in \mathcal{A}} \pi_t^{b, \boldsymbol{\omega}_t}(\mathbf{br}_t | \mathbf{o}_t) \pi_t^{\boldsymbol{\omega}_t}(\mathbf{a}_t | \mathbf{o}_t)$$

$$f_t(\mathbf{o}_t, \mathbf{s}_t, \boldsymbol{\omega}_{t-1}) \sum_{\mathbf{s}_{t+1} \in \mathcal{S}} b_{t+1}^c(\mathbf{s}_{t+1}) \big(p_t^{\mathbf{a}_t}(\mathbf{s}_t, \mathbf{s}_{t+1}) U^{\mu_t}(\mathbf{s}_{t+1}, \boldsymbol{\omega}_t) \big) \Bigg) b_t^c(\mathbf{s}_t).$$

$$r^{\boldsymbol{\omega}_t}(b_t^c) := \sum_{\mathbf{s}_t \in \mathcal{S}} \sum_{\mathbf{o}_t \in O} \sum_{\mathbf{br}_t \in \{0,1\}^J} \sum_{\mathbf{a}_t \in \mathcal{A}} \pi_t^{b, \boldsymbol{\omega}_t}(\mathbf{br}_t | \mathbf{o}_t) \pi_t^{\boldsymbol{\omega}_t}(\mathbf{a}_t | \mathbf{o}_t)$$

$$r^{\mathbf{a}_t, \mathbf{br}_t}(\mathbf{s}_t) f_t(\mathbf{o}_t, \mathbf{s}_t, \boldsymbol{\omega}_{t-1}) b_t^c(\mathbf{s}_t). \quad (4)$$

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand.

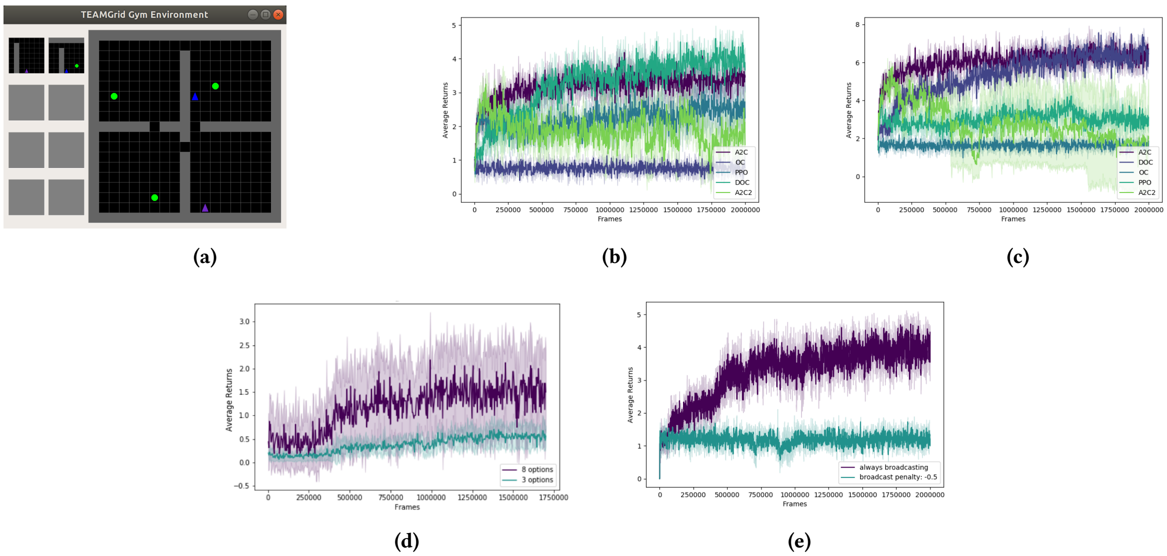

Figure 1: (a) *TEAMGrid FourRooms*, (b) average returns with 2 agents and 3 goals, (c) average returns with 3 agents and 5 goals (d) DOC: increasing number of options improved average returns, (e) DOC average returns with always broadcasting (broadcast penalty 0.0) and intermittent broadcasting (broadcast penalty = -0.5).

Q^{μ_t} in (3) is the solution of the following Bellman update:

$$Q^{\mu_t}(b_t^c, \omega_t) = r^{\omega_t}(b_t^c) + [\mathcal{B}^{\mu_t} Q^{\mu_t}](b_t^c, \omega_t), \quad (5)$$

where $f_t(\mathbf{o}_t, \mathbf{s}_t, \omega_{t-1})$ can be expressed recursively $f_t(\mathbf{o}_t, \mathbf{s}_t, \omega_{t-1}) := \sum_{a_{t-1} \in \mathcal{A}} \eta(\mathbf{o}_t | , \mathbf{s}_t, \mathbf{a}_{t-1}) \pi_{t-1}^{\omega_{t-1}}(\mathbf{a}_{t-1} | \mathbf{o}_{t-1}) f_{t-1}(\mathbf{o}_{t-1}, \mathbf{s}_{t-1}, \omega_{t-2})$ and $r^{\mathbf{a}_t, \mathbf{br}_t}(\mathbf{s}_t)$ is the immediate reward of choosing action $\mathbf{a}_t$ and broadcast symbol $\mathbf{br}_t$ in state $\mathbf{s}_t$. The optimal values corresponding to U^μ and Q^μ are defined as usual.

One can show using Cauchy-Schwartz inequality that $\mathcal{B}^{\mu_t}$ is a contraction, which is instrumental in showing the following theorem.

Theorem 0.1. *For a cooperative Dec-POMDP with options*

(1) *The optimal state-value is the fixed point solution of the following dynamic program.*

$$V^*(b_t^c) := \max_{\mu_t \in \mathcal{M}^+} \sum_{\omega_t \in \Omega} \mu_t(\omega_t | b_t^c)$$
$$\left[r^{\omega_t}(b_t^c) + \gamma \sum_{\tilde{\mathbf{o}}_t \in \mathcal{O} \cup \{\varnothing\}} \mathbb{P}(\tilde{\mathbf{o}}_t | b_t^c, \omega_t) V^*(b_{t+1}^c) \right], \quad (6)$$

where $\mathcal{M}^+$ is the space of joint option-policies and the notations have usual meaning.

(2) *There exists a* time-homogeneous *Markov joint-option policy μ^*, based on common information b_t^c, which is optimal.*

LEARNING IN DEC-POMDPS WITH OPTIONS

Our proposed algorithm for learning options, called *Distributed Option Critic* (DOC), builds on the *option-critic* architecture [2] and leverages the assumption of factored actions of agents in the distributed intra-option policy and termination function updates. The centralized option evaluation is presented from the coordinator's point of view. The agents learn to complete a cooperative task by learning in a model-free manner. In the *centralized option evaluation* step, the centralized critic (coordinator) evaluates in *temporal difference* (TD) manner [1] the performance of all agents via a shared reward (plus a broadcast penalty in case of costly communication) using the common information. Each agent updates its parameterized intra-option policy, broadcast policy and termination function through *distributed option improvement* using their private information.

Following [4, Theorem 1], one can show *Distributed gradient descent in a cooperative Dec-POMDP with options and with factored agents leads to local optima.* DOC uses one-step off policy temporal difference in centralized option evaluation and the *convergence of DOC* relies on showing that the expected value of TD-error $\delta := r^{\omega_k}(\mathbf{s}) + \gamma U(\mathbf{s}_{k+1}, \omega_k) - Q(\mathbf{s}_k, \omega_k)$ equals $r^{\omega_t}(b_k^c) + \gamma \mathbb{E}[U(b_{k+1}^c, \omega_t) \mid b_k^c] - Q(b_k^c, \omega_k)$.

Next, note that the by definition of intra-option Q-learning with full observability (e.g. see [5, Theorem 3]), we have that for any $\varepsilon \in \mathbb{R}_{>0}$, $\max_{\mathbf{s}'', \omega''} |Q(\mathbf{s}'', \omega'') - Q^*(\mathbf{s}'', \omega'')| < \varepsilon$. The rest of the proof follows by showing that the expected value of $r^{\omega_k}(\mathbf{s}) + \gamma U(\mathbf{s}'_{k+1}, \omega_k)$ converges to Q^*.

EXPERIMENTS

We evaluate empirically the merits of DOC in cooperative multi-agent tasks, and compare it to its single-agent counterpart, option-critic (OC), advantage actor-critic (A2C), A2C with central critic (A2C2) and proximal policy optimization (PPO). We created *TEAMGrid FourRooms* where the agents need to uncover multiple unknown targets and collect reward when all targets are uncovered. Fig. 1 shows that DOC performs competitively in this environment.

REFERENCES

[1] Klaas Apostol. 2012. *Temporal Difference Learning.* SaluPress.
[2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The Option-Critic Architecture. In *AAAI*.
[3] A. Nayyar, A. Mahajan, and D. Teneketzis. 2013. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Trans. Automat. Control* 58, 7 (jul 2013), 1644–1658.
[4] Leonid Peshkin, Kee-Eung Kim, Nicolas Meuleau, and Leslie Pack Kaelbling. 2000. Learning to Cooperate via Policy Search. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI'00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 489–496.
[5] Richard Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence* 112 (1999), 181–211.