Attention Actor-Critic algorithm for Multi-Agent Constrained Co-operative Reinforcement Learning*

Extended Abstract

P. Parnika*
Mindtree Ltd., India
parnika3103@gmail.com

Raghuram Bharadwaj Diddigi*
Indian Institute of Science, India
raghub@iisc.ac.in

Sai Koti Reddy Danda*
IBM Research, India
d.saikotireddy@in.ibm.com

Shalabh Bhatnagar
Indian Institute of Science, India
shalabh@iisc.ac.in

ABSTRACT

In this work, we consider the problem of computing optimal actions for Reinforcement Learning (RL) agents in a co-operative setting, where the objective is to optimize a common goal. However, in many real-life applications, the agents are also required to satisfy certain constraints specified on their actions. Under this setting, the objective of the agents is to not only learn the actions that optimize the common objective but also meet the specified constraints. In recent times, the Actor-Critic algorithm with an attention mechanism has been successfully applied to obtain optimal actions for RL agents in multi-agent environments. In this work, we extend this algorithm to the constrained multi-agent RL setting.

KEYWORDS

Multi-agent Reinforcement Learning; Attention Mechanism

ACM Reference Format:
P. Parnika*, Raghuram Bharadwaj Diddigi*, Sai Koti Reddy Danda*, and Shalabh Bhatnagar. 2021. Attention Actor-Critic algorithm for Multi-Agent Constrained Co-operative Reinforcement Learning: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

In a multi-agent constrained co-operative RL setting, the agents have to learn actions that not only minimize the expected total discounted cost but also respect the constraints specified. One approach to satisfy the constraints is to construct a modified cost as a linear combination of the original cost and the constraint costs. However, the weights to be associated with the costs are not known upfront and need to be learned in a trial-and-error fashion. We alleviate this problem by considering the Lagrangian formulation of the problem and training dual Lagrange parameters that act as weights for the constraint costs. In this work, we propose an Actor-Critic algorithm for computing the optimal actions for agents that makes use of the attention mechanism. A brief overview of the comparison of our work with other works in the literature is provided in Table 1.

References	Features
[7, 11, 16, 18, 19]	Deep RL algorithms for multi-agent setting. Attention mechanism not considered.
[13, 14, 17]	Deep RL algorithms with Attention for multi-agent setting. Constrained setting not considered.
[3–5]	RL algorithms for single-agent constrained setting. Multi-agent constrained setting not considered.
[1, 15, 23]	Deep RL algorithms for single-agent constrained setting. Multi-agent constrained setting not considered.
[2, 6, 9, 10, 12, 21, 22, 24]	RL algorithms for multi-agent constrained setting. Attention mechanism not considered.
Our Work	Deep RL algorithm with Attention mechanism for multi-agent Constrained setting.

Table 1: Comparison with other works in the Literature

2 MODEL

We now briefly discuss the constrained co-operative multi-agent setting [8] described by the tuple $< n, S, A, T, k, c_1, \ldots, c_m, \gamma >$. Here, n denotes the number of agents in the environment, $S = S_1 \times S_2 \times \cdots S_n$ is the joint state space, $A = A_1 \times \cdots \times A_n$ denotes the joint action space and T is the probability transition matrix. Single-stage cost function (k) where $k(s, a)$ denotes the cost incurred when joint action a is taken in state s. Moreover, $c_1, \ldots, c_m$ denote the single-stage cost functions for the constraints. Note that both the main cost function (k) and constraint costs ($c_1, \ldots, c_m$) depend on the joint action of the agents. Finally, γ denotes the discount factor. Let $\pi_i : S_i \rightarrow \Delta(A_i)$ denote the policy of agent i, where for a given state of agent i, $\pi_i(s_i)$ is a probability distribution over its actions. We now define the total discounted cost (J) for a joint policy $\pi = (\pi_1, \ldots, \pi_n)$ as follows:

$$J(\pi) = E\Big[\sum_{t=0}^{\tau} \gamma^t k(s_t, \pi(s_t))\Big], \tag{1}$$

where $E[.]$ is the expectation over the entire trajectory of states with initial state $s_0 \sim d_0$, where d_0 is a given probability distribution

*Equal contribution by the first three authors. Full paper is available at [20].

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

over states, τ is a finite stopping time and s_t is the joint state at time t. The m constraints on the system are defined as follows:

$$E\Big[\sum_{t=0}^{\tau}\gamma^t c_j(s_t,\pi(s_t))\Big] \leq \alpha_j,\ \forall j \in 1,\ldots,m, \tag{2}$$

where $\alpha_1,\ldots,\alpha_m$ are pre-specified thresholds.

The objective of the agents in the multi-agent constrained cooperative RL setting is to compute a joint policy $\pi^* = (\pi_1^*,\ldots,\pi_n^*)$ that is the solution to the optimization problem

$$\min_{\pi\in\Pi} J(\pi) = E\Big[\sum_{t=0}^{\tau}\gamma^t k(s_t,\pi(s_t))\Big] \tag{3}$$

$$\text{s.t}\quad E\Big[\sum_{t=0}^{\tau}\gamma^t c_j(s_t,\pi(s_t))\Big] \leq \alpha_j,\ \forall j \in 1,\ldots,m,$$

where Π is the set of all joint policies. We make use of Lagrangian approach to solve this constrained problem. The pseudo-code of our proposed algorithm is described in Algorithm 1.

Algorithm 1 Multi-Agent Constrained Attention Actor-Critic (MACAAC)

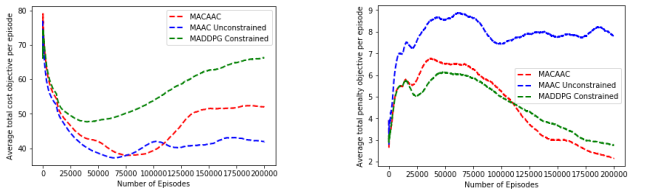
$E \leftarrow$ Maximum number of episodes.
$L \leftarrow$ Length of an episode.
$U \leftarrow$ Steps per update.
$\theta_i \leftarrow$ policy parameters of the agent i, $i = 1,\ldots,n$.
UpdateCritic: Subroutine to update the critic parameters.
UpdateActors: Subroutine to update the policy parameters of all the agents.
$Q_{\eta_j} \leftarrow$ Q-value of constrained cost associated with constraint j, $j = 1,\ldots,m$.
$\beta_t \leftarrow$ Slower timescale step-size at time step t.
Initialize Lagrange parameters $\lambda_1,\ldots,\lambda_m$.
Create μ parallel environments.
Initialize replay buffer, D.
$u \leftarrow 0$
for $ep = 1,2,\ldots,E$ **do**
  Obtain initial observations o_i^e for all agents i in each environment e
  for $t = 1,2,\ldots,L$ **do**
    Obtain actions $a_i^e \sim \pi_{\theta_i}(.|o_i^e)$, $\forall i = 1,\ldots,n$, $\forall e = 1,\ldots,\mu$
    Execute actions and get $(o_i^{*,e}, k^e, c_1^e, c_2^e,\ldots,c_m^e)\quad \forall i,e$
    Let $r^e = k^e + \sum_{j=1}^{m}\lambda_j c_j^e,\quad \forall e$
    Store $(o_i^e, a_i^e, r^e, c_1^e, c_2^e,\ldots,c_m^e, o_i^{*,e})$, $\forall i,e$ in D
    $o_i^e = o_i^{*,e}, \forall i,e$
    $u+ = \mu$
    if (u% U) < μ **then**
      Sample minibatch (B) from D
      Get next actions $a_1^{'},\ldots,a_n^{'}$
      UpdateCritic(B, $a_1^{'},\ldots,a_n^{'}$)
      UpdateActors(B)
      for $j = 1,\ldots,m$ **do**
        $\lambda_j \leftarrow \max(0, \lambda_j + \beta_t(Q_{\eta_j} - \alpha_j))$

3 EXPERIMENTS AND RESULTS

In the constrained version of Cooperative Navigation [16] that we consider, there are 5 agents and 5 targets that are randomly generated in a continuous environment at the beginning of each episode. The objective of the agents is to navigate towards the targets in a co-operative manner such that all targets are covered. The length of each episode is 25 time steps and the single-stage cost at each time step is the sum of the distance to the nearest agent, over all the targets. Therefore, the agents have to learn to navigate towards the targets in such a way that all target positions are covered. However, we include a single-stage penalty of 1 when there is a collision between the agents (and 0 otherwise). The penalty threshold (α) is set to 3 in our experiments. This means that the expected total penalty over all the episodes must be less than or equal to 3. The discount factor is set to 0.99. We refer to the main cost that the agents are minimizing as 'cost' and the constrained cost as the 'penalty'. For comparison purposes, we also implement the constrained version of MADDPG [16] algorithm, which we refer to as 'MADDPG-C'. Moreover, to better analyze the results, we also report the results on an un-constrained version of Multi-agent Attention Actor-Critic [13] where there is no penalty incurred for collisions among the agents, which we simply refer to as 'Unconstrained'.

3.1 Discussion

In Figure 1a, we observe that the total cost approaches convergence for all the three algorithms. The 'Unconstrained' algorithm achieves the smallest average cost as there is no penalty for collisions in this case. Therefore, the agents can move freely in the continuous space and navigate quickly towards the targets. This can also be observed in Figure 1b, where we see that the average penalty of the 'Unconstrained' algorithm is the highest. In Figure 1b, we see that the average penalty comes down as the training progresses for the constrained algorithms (MADDPG-C and our proposed MACAAC), while for the 'unconstrained' algorithm it almost remains constant. This is the effect of Lagrange parameters that are learnt in the constrained setting.



(a) Expected total cost **(b) Expected total penalty**

Figure 1: Performance of Algorithms on Constrained Cooperative Navigation during the training.

4 ACKNOWLEDGEMENTS

Raghuram Bharadwaj Diddigi was supported by a fellowship grant from the Centre for Networked Intelligence (a Cisco CSR initiative) of the Indian Institute of Science, Bangalore. Shalabh Bhatnagar was supported by the J.C.Bose Fellowship, the RBCCPS, IISc and a research project from DST, India.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization. *arXiv preprint arXiv:1705.10528* (2017).

[2] Pritee Agrawal, Pradeep Varakantham, and William Yeoh. 2016. Scalable greedy algorithms for task/resource constrained multi-agent stochastic planning. (2016).

[3] Shalabh Bhatnagar. 2010. An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes. *Systems & Control Letters* 59, 12 (2010), 760–766.

[4] Shalabh Bhatnagar and K Lakshmanan. 2012. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications* 153, 3 (2012), 688–708.

[5] Vivek S Borkar. 2005. An actor-critic algorithm for constrained Markov decision processes. *Systems & control letters* 54, 3 (2005), 207–213.

[6] Craig Boutilier and Tyler Lu. 2016. Budget allocation using weakly coupled, constrained Markov decision processes. (2016).

[7] Gang Chen. 2019. A New Framework for Multi-Agent Reinforcement Learning–Centralized Training and Exploration with Decentralized Execution via Policy Distillation. *arXiv preprint arXiv:1910.09152* (2019).

[8] Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, Prabuchandran K.J., and Shalabh Bhatnagar. 2019. Actor-Critic Algorithms for Constrained Multi-agent Reinforcement Learning. *arXiv preprint arXiv:1905.02907* (2019).

[9] Raghuram Bharadwaj Diddigi, D Reddy, Prabuchandran KJ, and Shalabh Bhatnagar. 2019. Actor-Critic Algorithms for Constrained Multi-agent Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1931–1933.

[10] Dmitri A Dolgov and Edmund H Durfee. 2011. Resource Allocation Among Agents with MDP-Induced Preferences. *arXiv preprint arXiv:1110.2767* (2011).

[11] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2017. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926* (2017).

[12] Michael Fowler, Pratap Tokekar, T Charles Clancy, and Ryan K Williams. 2018. Constrained-Action POMDPs for Multi-Agent Intelligent Knowledge Distribution. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–8.

[13] Shariq Iqbal and Fei Sha. 2019. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, Long Beach, California, USA, 2961–2970. http://proceedings.mlr.press/v97/iqbal19a.html

[14] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*. 7254–7264.

[15] Qingkai Liang, Fanyu Que, and Eytan Modiano. 2018. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480* (2018).

[16] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems*. 6379–6390.

[17] Hangyu Mao, Zhengchao Zhang, Zhen Xiao, and Zhibo Gong. 2019. Modelling the dynamic joint policy of teammates with attention multi-agent ddpg. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1108–1116.

[18] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. 2020. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics* (2020).

[19] Afshin OroojlooyJadid and Davood Hajinezhad. 2019. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963* (2019).

[20] P. Parnika, Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, and Shalabh Bhatnagar. 2021. Attention Actor-Critic algorithm for Multi-Agent Constrained Co-operative Reinforcement Learning. *arXiv e-print arXiv:2101.02349* (2021).

[21] D Sai Koti Reddy, Amrita Saha, Srikanth G Tamilselvam, Priyanka Agrawal, and Pankaj Dayama. 2019. Risk averse reinforcement learning for mixed multi-agent environments. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2171–2173.

[22] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295* (2016).

[23] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074* (2018).

[24] Ruohan Zhang, Yue Yu, Mahmoud El Chamie, Behçet Açikmese, and Dana H Ballard. 2016. Decision-Making Policies for Heterogeneous Autonomous Multi-Agent Systems with Safety Constraints.. In *IJCAI*. 546–553.