Approximate Difference Rewards for Scalable Multiagent Reinforcement Learning

Extended Abstract

Arambam James Singh
Singapore Management University
arambamjs.2016@smu.edu.sg

Akshat Kumar
Singapore Management University
akshatkumar@smu.edu.sg

Hoong Chuin Lau
Singapore Management University
hclau@smu.edu.sg

ABSTRACT

We address the problem of *multiagent credit assignment* in a large scale multiagent system. Difference rewards (DRs) are an effective tool to tackle this problem, but their exact computation is known to be challenging even for small number of agents. We propose a scalable method to compute difference rewards based on aggregate information in a multiagent system with large number of agents by exploiting the symmetry present in several practical applications. Empirical evaluation on two multiagent domains—air-traffic control and cooperative navigation, shows better solution quality than previous approaches.

KEYWORDS

Reinforcement learning; multiagent systems

ACM Reference Format:
Arambam James Singh, Akshat Kumar, and Hoong Chuin Lau. 2021. Approximate Difference Rewards for Scalable Multiagent 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 many real world applications, large group of agents interact with each other to achieve a common goal. For example, aircraft coordination in busy air space is required to ensure certain minimum separation among aircrafts [4], in maritime traffic management [11, 12], the common goal is to reduce congestion for safety of navigation problems. Such problems can be modeled as a cooperative multiagent reinforcement learning (MARL) problem. Specific to cooperative multiagent system with shared rewards is the critical problem of *multiagent credit assignment* [1, 3, 5]. In a cooperative setting, all agents receive the same global reward based on their joint action, as a result, the individual contribution of each agent (and its specific actions) to the global reward is unclear. In a cooperative multiagent systems, *difference rewards* (DRs) [2, 6, 14] are an effective tool to tackle the multiagent credit assignment problem. It quantifies the contribution of an agent to the system reward, computed as the difference between the system reward and a *counterfactual* value of the system reward when the particular agent's impact is removed from the system. The counterfactual value can be computed by replacing the agent's state-action with a default state-action. Although difference rewards effectively address the credit assignment problem, the computation of counterfactual term is very challenging because it requires access to reward function or performing additional simulations, which in model-free RL is not available to agents or computationally expensive.

In many large scale multiagent systems, an agent's behavior is mainly influenced by the *aggregate* information about neighboring agents rather than identity of agents. Such features can be observed in several applications. For example, in air traffic control and maritime traffic control, most of the agents can be considered as homogeneous (or belonging to a small number of types) [11]. Our proposed approach precisely exploits such symmetries present in large multiagent systems.

We address the problem of multiagent credit assignment in a scalable multiagent system. We exploit the property that for homogeneous agents, agent dynamics is primarily based on the aggregate information of the agent population (such as count of agents that are in the same state and take same actions), and develop new loss functions to train the reward function approximator in such settings. Using the learned reward function approximator, we develop a principled method that can efficiently approximate *different rewards* without requiring any extra simulation, or domain expertise.

2 APPROXIMATE DIFFERENCE REWARDS

We first show how to learn a function approximator for the system reward based only on reward signals from the simulator and by exploiting the aggregate nature of interaction among agents. We then develop techniques to approximate difference rewards from such a reward approximator.

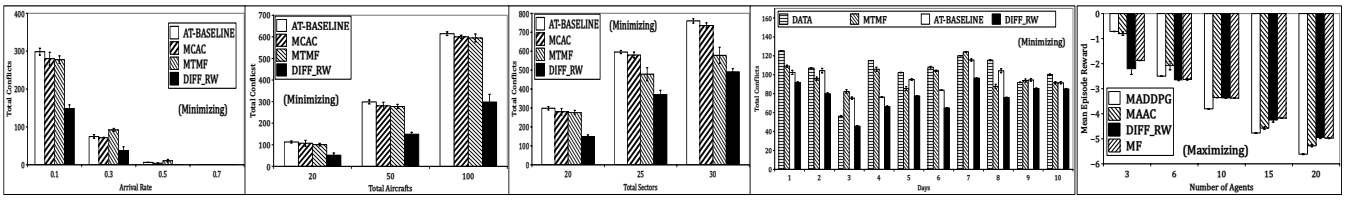
Learning System Reward Approximator: In many application domains, the global reward is decomposable. For example, in the air traffic domain, if two aircrafts are closer than a threshold distance, then penalty is given to each aircraft [4]. For such settings, we assume the global reward is decomposed into local reward received by each agent which depends on agent's local state and action, as well as the aggregate statistic of the agent population. Since agents are homogeneous, we can aggregate the reward over all the agents in a particular state-action (s, a) as:

$$r\left(\mathbf{n}_t^{SA}\right) = \sum_{s \in S} \sum_{a \in A} \mathrm{n}_t(s, a) \cdot \tilde{r}(s, a, \mathbf{n}_t^S) \tag{1}$$

Where $\mathrm{n}_t(s, a)$ is number of agents in state s and taking action a at time t. $\mathbf{n}_t^S = \langle \mathrm{n}_t(s) \rangle_{\forall s}$ and $\mathbf{n}_t^{SA} = \langle \mathrm{n}_t(s, a) \rangle_{\forall s, \forall a}$ denote the state and state-action count tables respectively. We can learn a reward function approximator $r_{\mathbf{w}}$ using samples $\xi \in \mathcal{B}$ collected during simulation, and trained with the loss function:

$$\tilde{\mathcal{L}}(\mathbf{w}) = M \sum_{\xi \in \mathcal{B}} \sum_{s \in S} \sum_{a \in A} \mathrm{n}_\xi(s, a) \cdot \left(\tilde{r}(s, a, \mathbf{n}_\xi^S) - r_{\mathrm{w}}(s, a, \mathbf{n}_\xi^S) \right)^2 \tag{2}$$

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.



(a) Varying arrival rate (b) Varying agent population (c) Varying number of sectors (d) Real world instances (e) Cooperative Navigation

Figure 1

Computing Difference Rewards: For our homogeneous agent setting, using the state-action count table, $\mathbf{n}_t^{SA}$, derived from the joint state-action pair $(\boldsymbol{s}_t, \boldsymbol{a}_t)$, we can write difference rewards as:

$$D^m(s_t^m, a_t^m) = r_{\mathrm{w}}\left(\mathbf{n}_t^{SA}\right) - r_{\mathrm{w}}\left(\mathbf{n}_t^{SA-(s_t^m, a_t^m)+(d_s, d_a)}\right) \tag{3}$$

where, $\mathbf{n}_t^{SA-(s_t^m, a_t^m)+(d_s, d_a)}$ is like a counterfactual state-action count table obtained by replacing the current state and action (s_t^m, a_t^m) of agent m with a default state and action (d_s, d_a). There can be many agents in the same state taking same action, and thus sharing the same DR value. Direct evaluation of above expression is computationally expensive. Therefore, we propose a gradient based method to compute the DR.

$$D_t(s, a) \approx \frac{1}{M} \cdot \left(\frac{\partial r_{\mathrm{w}}\left(\tilde{\mathbf{n}}_t^{SA}\right)}{\partial \tilde{\mathbf{n}}_t^{SA}(s, a)} - \frac{\partial r_{\mathrm{w}}\left(\tilde{\mathbf{n}}_t^{SA}\right)}{\partial \tilde{\mathbf{n}}_t^{SA}(d_s, d_a)} \right) \tag{4}$$

Where M is total agent population size. Once difference rewards are computed, it can be easily integrated in any policy gradient based method. The resulting approach computes returns R_t using the learned difference rewards (rather than empirical rewards from the simulator). This results in a model-based RL where difference rewards provide a counterfactual value highlighting an agent's contribution to the team's reward.

3 EXPERIMENTS

We evaluate our proposed approach on both synthetic and real world instances of air-traffic control problem, and in a continuous state cooperative navigation problem.

Air traffic control (ATC): We follow similar settings used in [4] where only aircraft speed is controlled, not the altitude. We use following baselines — a domain specific approaches AT-BASELINE [4], MCAC [10] and MTMF (multitype mean field RL) [13], which are designed specifically for homogenous agents population.

Synthetic instance experiments: Figure (1a) shows result for experiments with 50 aircraft agents, 20 sector map, and with varying arrival rate. Arrival rate essentially denote traffic intensity, lower arrival rate has highest traffic intensity. We observe the expected trend of total conflicts decreasing with increasing arrival rate. We also observe the performance gap of DIFF-RW with other approaches decrease with increasing arrival rate. At arrival rate = 0.1, due to high frequency of aircraft arrivals, most of the baseline approaches suffer from high conflicts. This setting require tighter coordination among aircrafts, which is better achieved by DIFF-RW

Figure (1b) shows result for setting with 20 sector map, fixed arrival rate = 0.1, and with varying aircraft population. In this setting, we observe the empirical evidence of DIFF-RW performing better with higher agent population. At lower population setting, almost all approach perform equally well with marginal differences. But for large population setting, other baselines suffer due to lack of efficient credit assignment.

We also tested with increasing number of sectors. Figure (1c) shows results with fixed arrival rate = 0.1 and aircraft population as 50. We observe that AT-BASELINE suffers most among other approaches. This is because in difficult instances, parameter sharing based method lacks coordination among agent without explicit credit assignment, even though it is scalable. Similar to previous results, DIFF-RW performed best for different map sizes.

Real data experiments: We evaluate our approach on air space surrounding one of the busiest airport Heathrow, London. The data for 30 days is obtained from *Flightradar*24 [1]. We use 20 days data for training and 10 days for testing. We use bluesky air-traffic simulator [7] to simulate, and learn the policy from training data. We then evaluate the learned policy on 10 separate testing days. Figure (1d) shows result of our approach compared against baselines (MCAC was slightly worse than MTMF , to avoid clutter, its bars are omitted). We observe in most of the days all approaches perform equal or better than DATA (which is the replay of the historical dataset). DIFF-RW is able to achieve much better solution quality than other baselines.

Cooperative navigation domain: We also evaluate our approach on cooperative multiagent navigation [9]. The state space in this environment is continuous, therefore we use tile coding based technique for discretization. For this domain we use following baselines—MADDPG [9], MAAC [8], and mean field multiagent RL (MF) [15]. Figure (1e) shows results with varying agent population; y-axis denotes mean episode reward. For small agent population $n = 3$, MADDPG and MAAC perform better than MF , DIFF-RW. This is because our approximation of DR for small number of agents may not be accurate. However, with increasing agent population, solution quality of MADDPG and MAAC drops, and DIFF-RW and MF improves. This trend is an empirical evidence of the accuracy of our DR method with increasing agent population.

ACKNOWLEDGMENTS

This research is supported by the Agency for Science, Technology and Research (A*STAR), Fujitsu Limited and the National Research Foundation Singapore as part of the A*STAR-Fujitsu- SMU Urban Computing and Engineering Centre of Excellence.

[1] www.flightradar24.com

REFERENCES

[1] Adrian K. Agogino and Kagan Tumer. 2004. Unifying temporal and structural credit assignment problems. In *International Joint Conference on Autonomous Agents and Multiagent Systems*. 980–987.

[2] Adrian K. Agogino and Kagan Tumer. 2008. Analyzing and visualizing multiagent rewards in dynamic and stochastic domains. *Journal of Autonomous Agents and Multi-Agent Systems* 17, 2 (2008), 320–338.

[3] Drew Bagnell and Andrew Y Ng. 2006. On local rewards and scaling distributed reinforcement learning. In *Advances in Neural Information Processing Systems*. 91–98.

[4] Marc Brittain and Peng Wei. 2019. Autonomous Separation Assurance in An High-Density En Route Sector: A Deep Multi-Agent Reinforcement Learning Approach. In *IEEE Intelligent Transportation Systems Conference*. 3256–3262.

[5] Yu-Han Chang, Tracey Ho, and Leslie P Kaelbling. 2004. All learning is local: Multi-agent learning in global reward games. In *Advances in neural information processing systems*. 807–814.

[6] Mitchell K. Colby, Theodore Duchow-Pressley, Jen Jen Chung, and Kagan Tumer. 2016. Local Approximation of Difference Evaluation Functions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. ACM, 521–529.

[7] Jacco M Hoekstra and Joost Ellerbroek. 2016. Bluesky atc simulator project: an open data and open source approach. In *Proceedings of the 7th International Conference on Research in Air Transportation*, Vol. 131. FAA/Eurocontrol USA/Europe, 132.

[8] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, PMLR, 2961–2970.

[9] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 6379–6390.

[10] Duc Thien Nguyen, Akshat Kumar, and Hoong Chuin Lau. 2018. Credit assignment for collective multiagent RL with global rewards. In *Advances in Neural Information Processing Systems*. 8102–8113.

[11] Arambam James Singh, Akshat Kumar, and Hoong Chuin Lau. 2020. Hierarchical Multiagent Reinforcement Learning for Maritime Traffic Management. In *19th International Conference on Autonomous Agents and MultiAgent Systems*. IFAAMAS, 1278–1286.

[12] Arambam James Singh, Duc Thien Nguyen, Akshat Kumar, and Hoong Chuin Lau. 2019. Multiagent Decision Making For Maritime Traffic Management. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 6171–6178.

[13] Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E. Taylor, and Nidhi Hegde. 2020. Multi Type Mean Field Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS*. IFAAMAS, 411–419.

[14] David H. Wolpert and Kagan Tumer. 2001. Optimal Payoff Functions for Members of Collectives. *Advances in Complex Systems* 4, 2-3 (2001), 265–280.

[15] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, Vol. 80. PMLR, 5567–5576.