Difference Rewards Policy Gradients

Jacopo Castellini[1], Sam Devlin[2], Frans A. Oliehoek[3], Rahul Savani[1]

[1] Dept. of Computer Science, University of Liverpool, UK [2] Microsoft Research Cambridge, UK, [3] Interactive Intelligence Group, Delft University of Technology, NL

{j.castellini, rahul.savani}@liverpool.ac.uk, sam.devlin@microsoft.com, f.a.oliehoek@tudelft.nl

Introduction

Multi-agent policy gradients (MAPG) are:

1. An established technique for **cooperative MARL** problems under the CTDE framework,
2. Not addressing **multi-agent credit assignment** [3]: an agent telling how it is affecting the overall performance.

- **Difference rewards** [5]: using a shaped reward to infer each agent contribution to the shared reward.
- COMA [2] combines MAPG with the differencing of a learned Q-function, but:
  - Learning the Q-function is a difficult problem (bootstrapping, moving targets, Q's dependence on the joint actions),
  - COMA is not exploiting knowledge about $R(s,a)$.

To overcome these potential difficulties, we propose:

- *Difference rewards REINFORCE* (Dr.Reinforce), new MARL algorithm that combines MAPG with difference rewards when $R(s,a)$ is known,
- A practical implementation, called Dr.ReinforceR, for settings where the reward function is not known upfront,
- Learning $R(s,a)$ is a simple regression problem and does not suffer from many of the above problems.

Difference Rewards Policy Gradients

The aristocrat utility [5] difference rewards method uses the shaped reward:

$$\Delta R^i(a^i|s,a^{-i}) = R(s,a) - \mathbb{E}_{b^i\sim\pi_{\theta^i}}\left[R(s,\langle a^{-i},b^i\rangle)\right]. \tag{1}$$

If the reward function $R(s,a)$ is known, we propose Dr.Reinforce: let define the *difference return* ΔG^i_t for agent i:

$$\Delta G^i_t(a^i_{t:t+T}|s_{t:t+T},a^{-i}_{t:t+T}) \triangleq \sum_{l=0}^{T}\gamma^l \Delta R^i(a^i_{t+l}|s_{t+l},a^{-i}_{t+l}), \tag{2}$$

we plug it into a modified version of the *distributed policy gradients* [4] as:

$$\theta^i \leftarrow \theta^i + \alpha\gamma^t \Delta G^i_t(a^i_{t:t+T}|s_{t:t+T},a^{-i}_{t:t+T})\nabla_{\theta^i}\log\pi_{\theta^i}(a^i_t|s_t). \tag{3}$$

However, there are cases in which $R(s,a)$ is unknown. For such settings, we propose Dr.ReinforceR:

- Learn online an additional *centralized reward network* $R_\psi(s_t,a_t)$, trained by minimizing the MSE w.r.t. the experienced reward r_t and only needed during training.
- Although having the same dimensionality of the COMA critic, learning R_ψ is a regression problem that does not involve bootstrapping or moving targets.

We can now use the learned R_ψ to compute an alternative to (1) to be used in (4) as:

$$\Delta R^i_\psi(a^i_t|s_t,a^{-i}_t) \triangleq r_t - \sum_{b^i\in A^i}\pi_{\theta^i}(b^i|s_t)R_\psi(s_t,\langle b^i,a^{-i}_t\rangle). \tag{4}$$

Convergence proof and analysis are available in [1].

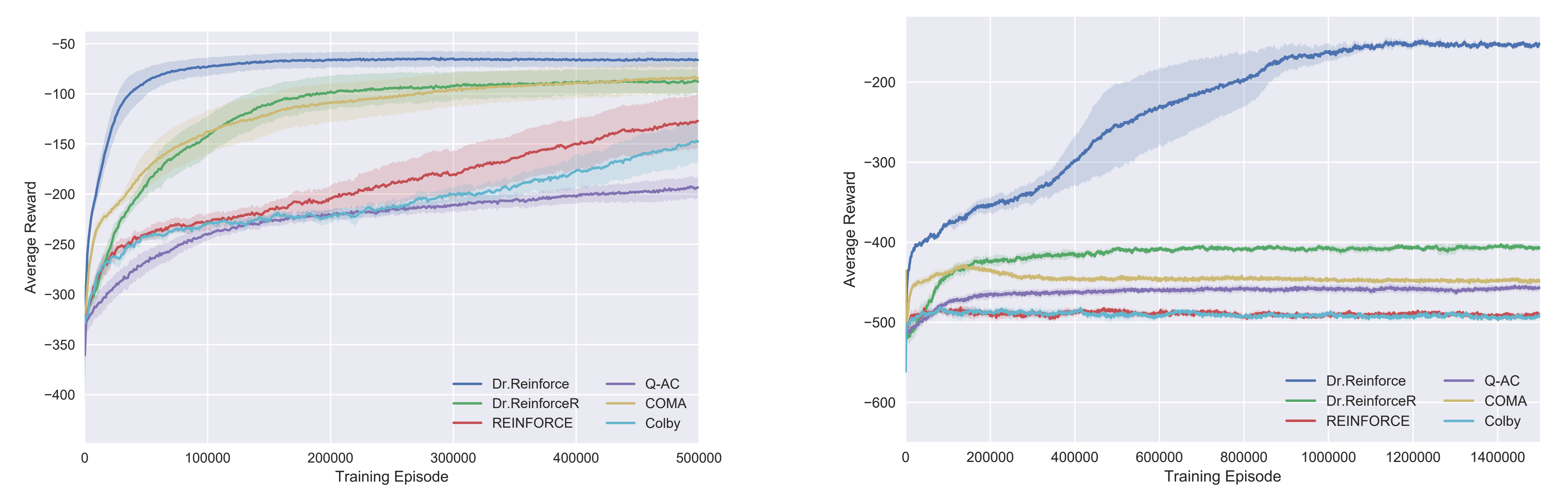
Experiments

We compare to COMA [2] and other policy gradients methods on two popular cooperative benchmark problems (full results are available in [1]):

- *Multi-Rover*: navigation over a set of landmarks,
- *Predator-Prey*: pursing of a random-moving prey.

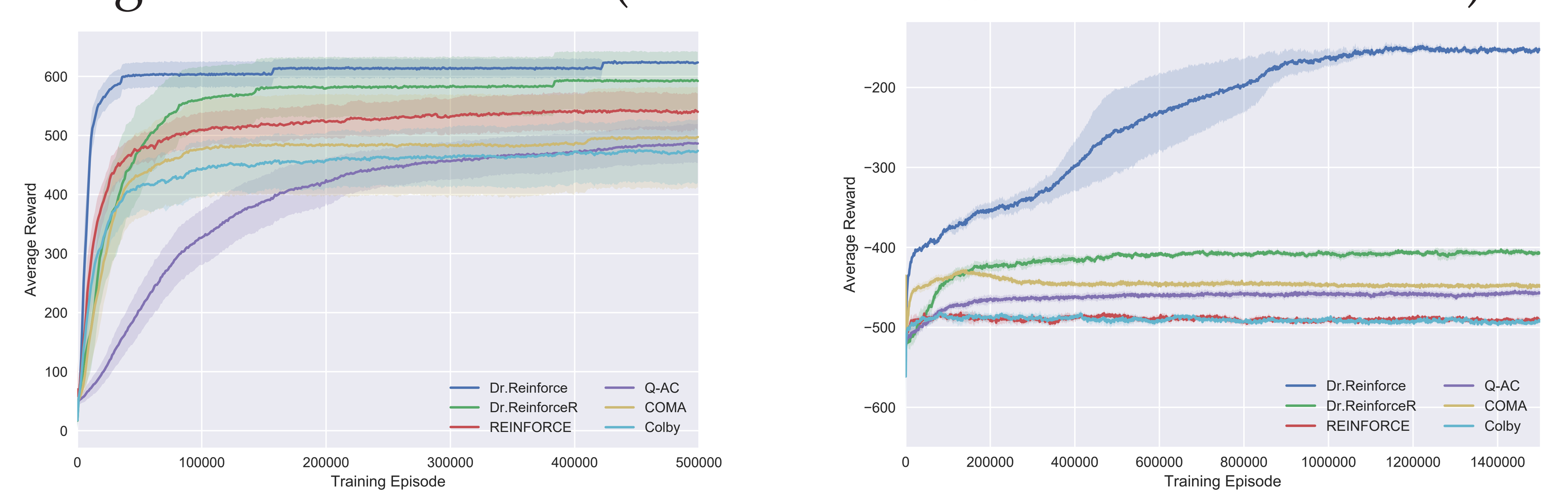
Main takeaways:

- When there are few agents, both COMA and Dr.ReinforceR are doing good, and Dr.Reinforce is outperforming all,
- With more agents instead, COMA performance is deteriorating, while Dr.ReinforceR is doing better, matching the Dr.Reinforce upper bound of Predator-Prey,
- Learning the Q-function may be problematic, as it needs to generalize well to unseen examples, and hinder the learning of optimal policies,
- There are cases in which also the reward network may not generalize properly, but it is generally easier.



$N=3$ $N=8$

Figure 2: Multi-Rover (mean and 90% confidence interval).



$N=3$ $N=8$

Figure 4: Predator-Prey (mean and 90% confidence interval).

Conclusions

1. We combined MAPG with difference rewards to tackle multi-agent credit assignment and proposed Dr.Reinforce for cases in which $R(s,a)$ is known in advance,
2. Moreover, we proposed Dr.ReinforceR for problems in which such knowledge is not available, learning a centralized reward network to predict the required reward values,
3. We analysed how learning the reward function is an easier problem than learning the Q-function as done in COMA, not presenting the difficulties related to bootstrapping or moving targets.

References

[1] J. Castellini, S. Devlin, F. A. Oliehoek, and R. Savani. Difference rewards policy gradients. *arXiv*, abs/2012.11258, 2020.

[2] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual multi-agent policy gradients. In *Proc. of the 32th AAAI Conf. on Artificial Intelligence*, AAAI'18, pages 2974–2982, 2018.

[3] D. T. Nguyen, A. Kumar, and H. C. Lau. Credit assignment for collective multiagent rl with global rewards. In *Advances in Neural Information Processing Systems 32*, NIPS'18, pages 8113–8124. 2018.

[4] L. Peshkin, K.-E. Kim, N. Meuleau, and L. P. Kaelbling. Learning to cooperate via policy search. In *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence*, UAI'00, pages 489—496, 2000.

[5] D. H. Wolpert and K. Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4:265–280, 2001.