Approximate Difference Rewards For Scalable Multiagent Reinforcement Learning

Arambam James Singh, Akshat Kumar and Hoong Chuin Lau, School of Computing & Information Systems, Singapore Management University
{arambamjs.2016, akshatkumar, hclau}@smu.edu.sg

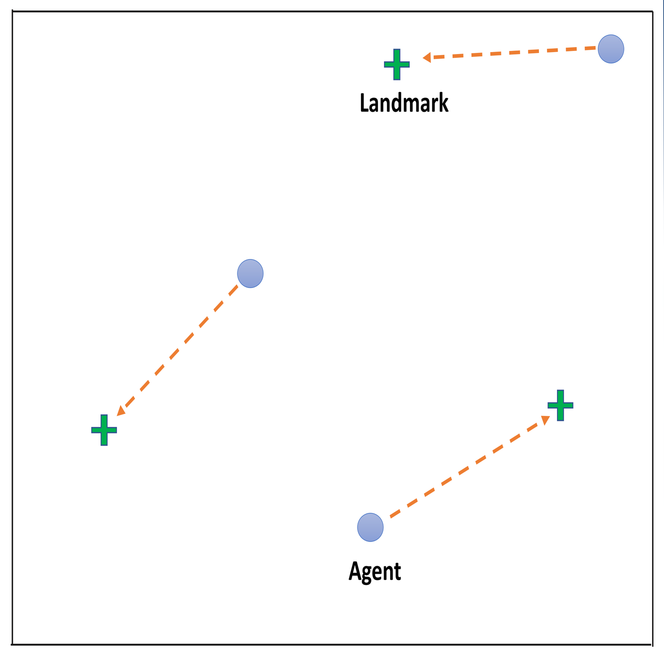
Introduction

We address the problem of multiagent credit assignment in large scale multiagent system. Our main contributions are:

- An approach to learn a differentiable reward model by exploiting the collective nature of interactions among agents.
- A principled method to analytically compute shaped rewards from the reward model.
- A model-based RL approach that uses learned shaped rewards addressing credit assignment problem.

❖Motivating Domain :

Air Traffic Control



Cooperative Navigation

❖Challenges :

- Empirical reward signal is not effective in addressing multiagent credit assignment problem.
- The credit assignment problem becomes more challenging with large number of agents.
- Current proposed approaches either do not scale well for large agent settings or their credit assignment mechanism is not effective.

Our work address these challenges.

Count Variables

❖ State count variable :

$$\mathrm{n}_t(s) = \sum_{m=1}^{M} \mathbb{I}[s_t^m = s; \boldsymbol{s}_t], \forall s \in S$$

❖ State-action count variable :

$$\mathrm{n}_t(s,a) = \sum_{m=1}^{M} \mathbb{I}[s_t^m = s, a_t^m = a; \boldsymbol{s}_t, \boldsymbol{a}_t], \forall s \in S$$

System Reward Approximator

❖System Reward :

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

❖Loss Function for Reward Approximator :

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

Approximate Difference Reward

❖Difference rewards (DRs) :

$$D^m\left(s_t^m, a_t^m\right) = r\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) - r\left(\boldsymbol{s}_t^{-m}\cup d_s, \boldsymbol{a}_t^{-m}\cup d_a\right)$$

❖Difference rewards with count variable:

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

❖Difference rewards for state-action :

$$D_t(s,a) = r\left(\mathbf{n}_t^{SA}\right) - r\left(\mathbf{n}_t^{SA} - \mathcal{I}^{sa} + \mathcal{I}^{d_s d_a}\right)$$

❖ Approximate difference rewards :

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

Policy Gradient with DRs

❖ Return with difference rewards :

$$R_t^{dr} = \sum_{i=0}^{\infty}\gamma^i\left(\sum_{s\in S}\sum_{a\in A}\mathrm{n}_{t+i}(s,a)\cdot D_{t+i}(s,a)\right)$$

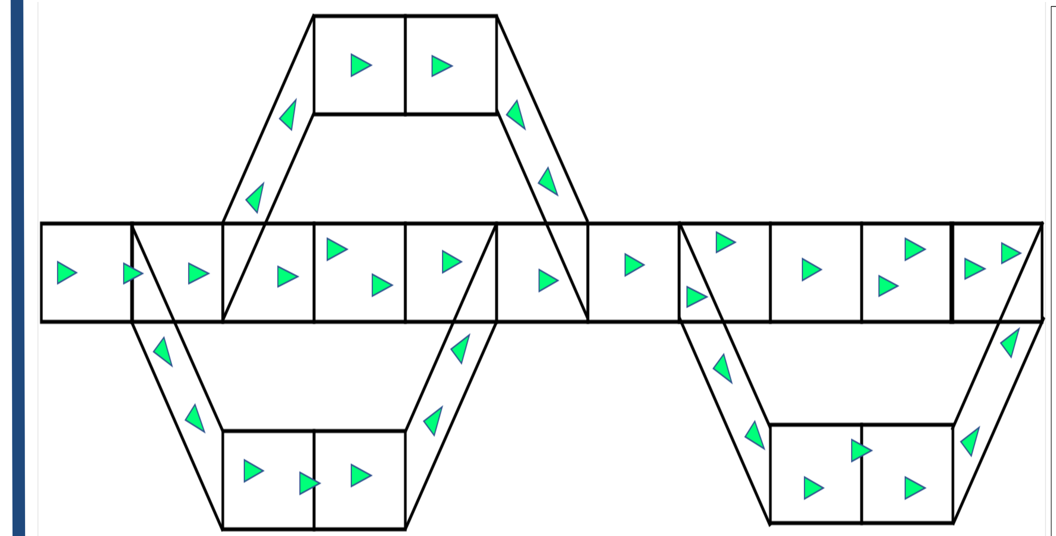
❖ Policy gradient :

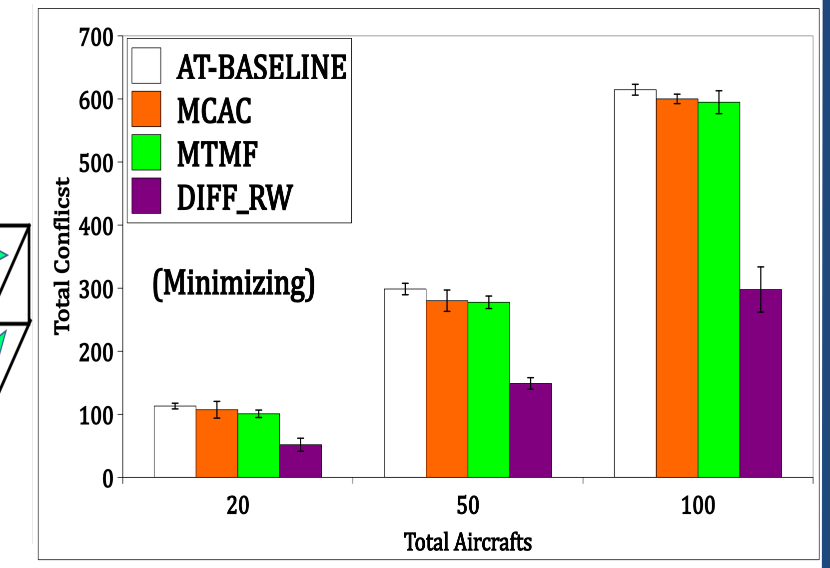
$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\boldsymbol{s}_{0:\infty},\boldsymbol{a}_{0:\infty}}\left[\sum_{t=0}^{\infty}\sum_{s\in S}\sum_{a\in A}\mathrm{n}_t(s,a)\cdot\nabla_\theta\log\pi_\theta(a\mid s_t)\cdot R_t^{dr}\right]$$

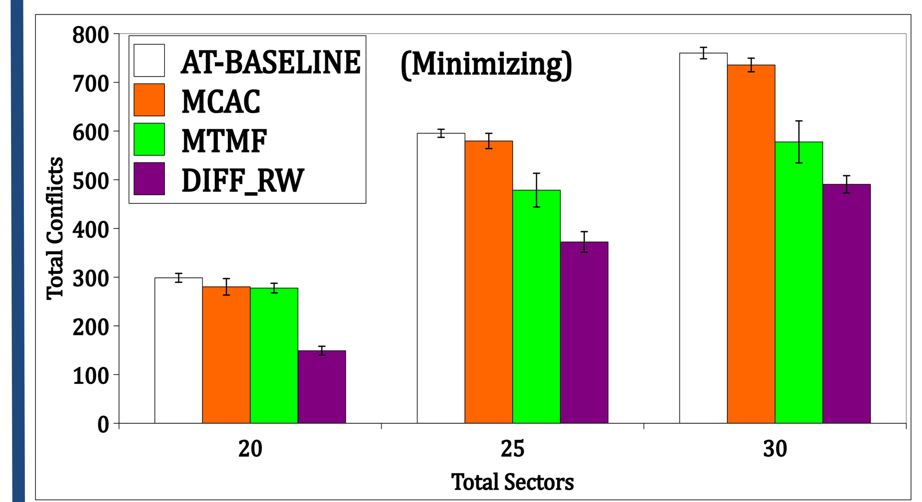
Experiments

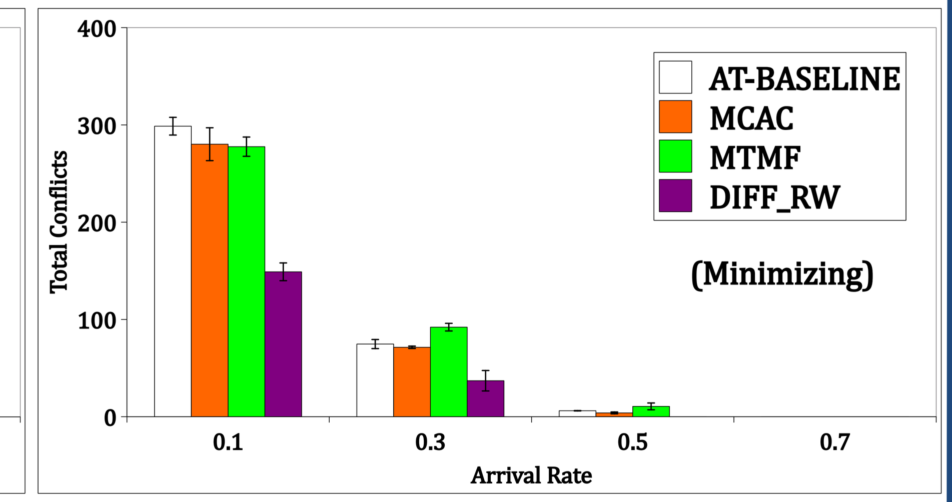
❖Air Traffic Control :

Synthetic Data:

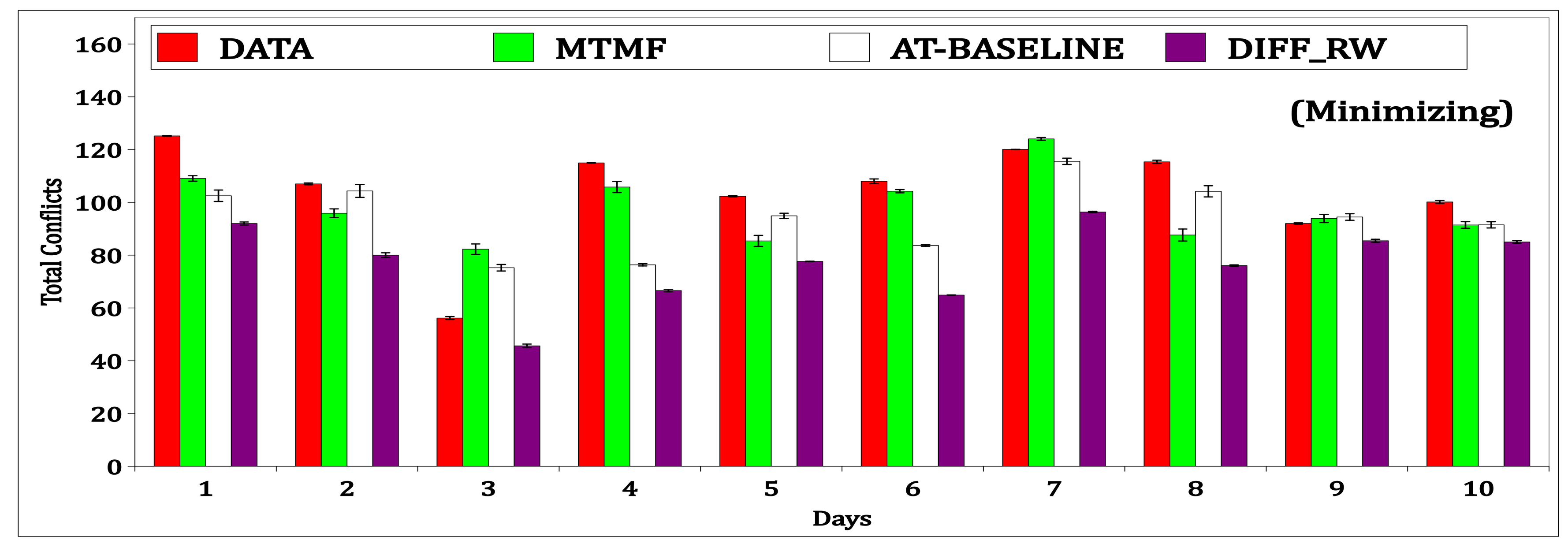




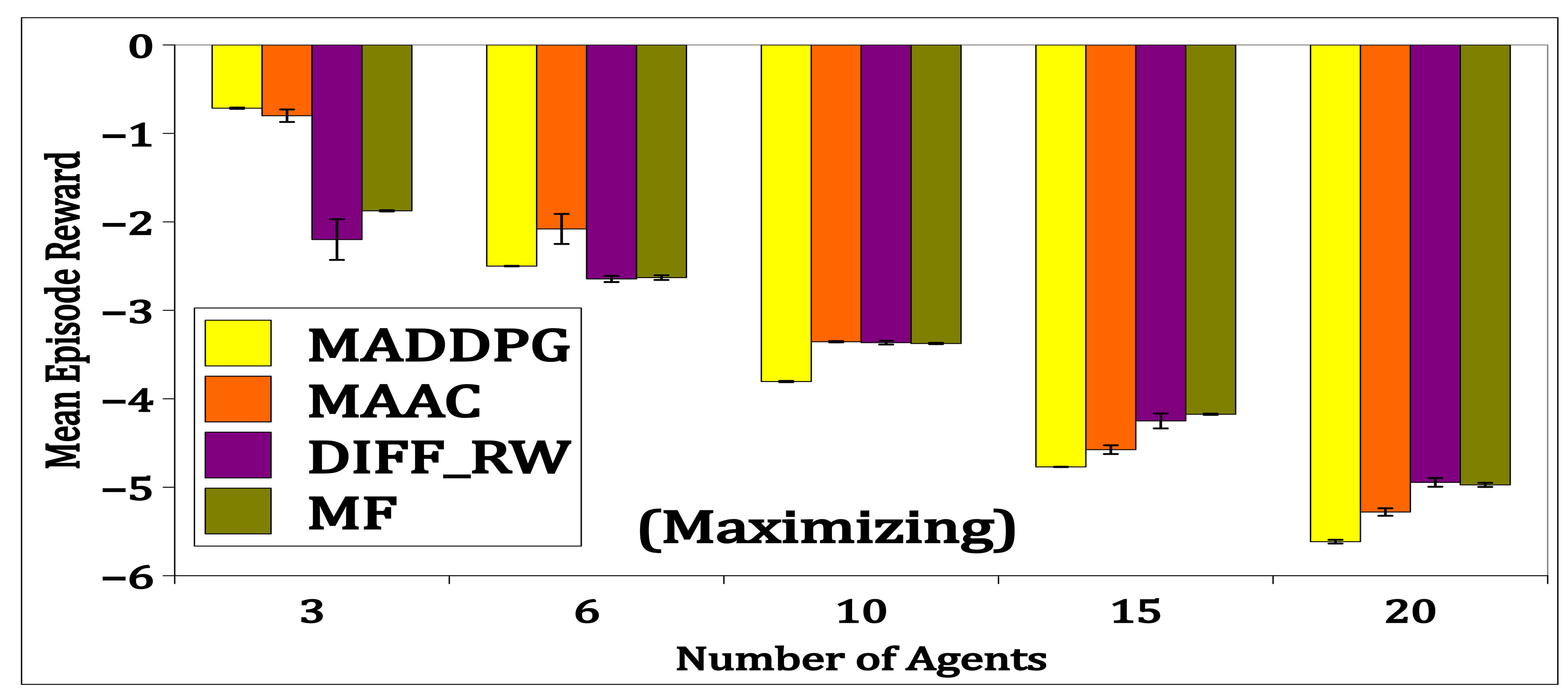




Real world dataset (1 month data):



❖Cooperative Navigation :



❖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.