We address the problem of multiagent credit assignment in large scale multiagent systems. 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 addresses these challenges.

### System Reward Approximator

**System Reward**: 

\[ r(\mathbf{n}_t^{SA}) = \sum_{s \in S} \sum_{a \in A} n_t(s, a) \cdot \tilde{r}(s, a, \mathbf{n}_t^{S}) \]

**Loss Function for Reward Approximator**: 

\[ \hat{L}(w) = M \sum_{\xi \in B} \sum_{s \in S} n_{\xi}(s, a) \cdot (\tilde{r}(s, a, \mathbf{n}_t^{S}) - r_w(s, a, \mathbf{n}_t^{S}))^2 \]

### Approximate Difference Reward

**Difference rewards (DRs)**: 

\[ D^m(s_t^m, a_t^m) = r(s_t, a_t) - r(s_t^m \cup d_s, a_t^m \cup d_a) \]

**Difference rewards with count variable**: 

\[ D^m(s_t^m, a_t^m) = r_w(\mathbf{n}_t^{SA}) - r_w(\mathbf{n}_t^{SA-(s_t^m,a_t^m)+(d_s,d_a)}) \]

**Difference rewards for state-action**: 

\[ D_t(s, a) = r(\mathbf{n}_t^{SA}) - r(\mathbf{n}_t^{SA - T^a + T^d}) \]

**Approximate difference rewards**:

\[ D_t(s, a) \approx \frac{1}{M} \left( \frac{\partial r_w(\mathbf{n}_t^{SA})}{\partial n_t^{SA}(s, a)} - \frac{\partial r_w(\mathbf{n}_t^{SA})}{\partial n_t^{SA}(d_s, d_a)} \right) \]

### Policy Gradient with DRs

**Return with difference rewards**:

\[ R_t^{dr} = \sum_{i=0}^{\infty} \gamma^i \left( \sum_{s \in S} \sum_{a \in A} n_{t+i}(s, a) \cdot D_{t+i}(s, a) \right) \]

**Policy gradient**:

\[ \nabla_{\theta} J(\pi_{\theta}) = \mathbb{E}_{s_0, \ldots, a_0} \left[ \sum_{i=0}^{\infty} \sum_{s \in S} \sum_{a \in A} n_t(s, a) \cdot \nabla_{\theta} \log \pi_{\theta}(a | s_t) \cdot R_t^{dr} \right] \]

### 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.