Social Influence-Based Mutual Acknowledgement Token Exchange (Student Abstract)

Shoma Mizuno[1], **Keichi Namikoshi**[1], **Yuko Sakurai**[1]

[1]Nagoya Institute of Technology
{s.mizuno.974@stn., namikoshi.keiichi@, sakurai@}nitech.ac.jp

Abstract

External incentive mechanisms have been studied as a method to promote cooperation in sequential social dilemmas involving multiple autonomous agents. Mutual Acknowledgment Token Exchange (MATE) is one such approach: by enabling agents to exchange acknowledgment tokens, it induces cooperation without additional training. However, MATE's use of fixed, manually tuned token values limits adaptability to nonstationary environments and can constrain performance. To enable a dynamically adapted token, we introduce Social Influence-based MATE (SI-MATE), which allows agents to share their individual improvement signals and to self-punishment in response to inequality. Experiments in a four-agent environment show that SI-MATE outperforms MATE across multiple metrics, including learning speed.

Introduction

Achieving cooperation in multi-agent reinforcement learning remains a significant challenge, particularly in sequential social dilemma (SSD) environments, where individual rationality conflicts with collective welfare. External incentive mechanisms to promote cooperation in SSDs have been studied (surveyed in (Mu et al. 2024)); the Mutual Acknowledgment Token Exchange (MATE) (Phan et al. 2024) is one such approach. This mechanism promotes cooperation by enabling agents to exchange tokens with neighboring agents based on their self-evaluated metrics, it's called the monotonic improvement (MI) value. However, in some tasks, MATE's performance is sensitive to a key hyperparameter—fixed, manually tuned token values. Since these cannot adapt to multi-agent nonstationarity, dynamic token values may help, but they require a principled criterion or algorithm to set them.

This study proposes Social Influence-based MATE (SI-MATE), an adaptive protocol that leverages social influence to better handle nonstationary environments. The core idea is that each agent treats the change in neighboring agents' MI as a "social influence". The agent decides whether to send a token, how many tokens to send, and how much self-punishment to impose based on this "social influence" signal. Although SI-MATE incurs communication overhead that grows linearly with the number of agents compared to MATE, we show that it improves both performance and learning speed empirically.

Preliminaries

Following MATE (Phan et al. 2024), we consider sequential social dilemma environments formulated as partially observable stochastic games (POSG). POSG is described as $M = \langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{O} \rangle$, where $\mathcal{I} = \{1, \ldots, N\}$ is the set of agents, $\mathcal{S}$ is the set of states, $\mathcal{A} = \big(\mathcal{A}_1, \ldots, \mathcal{A}_N\big) = \big(\mathcal{A}_i\big)_{i \in \mathcal{I}}$ is the set of joint actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability, $\big(\mathcal{R}_i\big)_{i \in \mathcal{I}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$ is the joint reward, $\mathcal{O}$ is the set of observations. Each agent $i \in \mathcal{I}$ has a H-step history $\tau_{t,i} \in (\mathcal{O} \times \mathcal{A}_i)^H$, individual policy $\pi_i\big(a_{t,i} \mid \tau_{t,i}\big)$ which is the probability to select an action $a_{t,i}$ given $\tau_{t,i}$ at time step t and a neighborhood $\mathcal{N}_{t,i} \subseteq \mathcal{I} \backslash i$. The agent i value of joint policy $\pi = (\pi_i)_{i \in \mathcal{I}}$ at state $s \in \mathcal{S}$ is $V_i^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_{t,i} \mid S_t = s\right]$, where $\gamma \in [0, 1)$ is the discount factor and T is the total time step of an episode. The efficiency $U = \sum_{i \in \mathcal{I}} R_i$ is the sum of individual undiscounted reward $R_i = \sum_{t=0}^{T-1} r_{t,i}$.

MATE permits token exchange between agents and bases send/respond decisions on the MI value defined in (1):

$$MI_i(\hat{r}_{t,i}) = \hat{r}_{t,i} + \gamma \hat{V}_i(\tau_{t+1,i}) - \hat{V}_i(\tau_{t,i}), \qquad (1)$$

where $\hat{r}_{t,i}$ could be either the environment reward $r_{t,i}^{\text{env}}$ or some shaped reward; and $\hat{V}_i$ is estimated value function. The MI value can be interpreted as surprise relative to the agent's own expectation (i.e., a TD-like signal). In MATE, when a single step yields a positive surprise ($MI \geq 0$), the agent sends or responds with fixed positive tokens to its neighbors as a cooperative acknowledgment; when $MI < 0$, it responds with fixed negative tokens. See Algorithm 2 in (Phan et al. 2024) for details.

Proposed Method: SI-MATE

SI-MATE introduces a social influence computed from changes in neighboring agents' MI signals. In MATE, what an agent can infer about others is essentially limited to whether tokens are received (i.e., a binary signal), making it difficult to estimate others' improvement magnitudes. In SI-MATE, an agent observes from its neighbors the change in

Algorithm 1: SI-MATE protocol for agent i at step t

Input: Current and previous MI values: $\{MI_k(r^{\text{env}}_{t',i})\}^{t}_{t'=t-1}$ for $k \in \mathcal{N}'_{t,i} = \mathcal{N}_{t,i} \cup \{i\}$; Environmental reward $r^{\text{env}}_{t,i}$; Scaling factor α

1: **Initialize:** $\hat{r}_{\text{req}} \leftarrow 0$, $\hat{r}_{\text{res}} \leftarrow 0$, $r_{\text{punish}} \leftarrow 0$
2: $U'_i(t) \leftarrow \sum_{k \in \mathcal{N}'_{t,i}} MI_k(r^{\text{env}}_{t,k})$
3: $U'_i(t-1) \leftarrow \sum_{k \in \mathcal{N}'_{t,i}} MI_k(r^{\text{env}}_{t-1,k})$
4: $\Delta U'_i(t) \leftarrow U'_i(t) - U'_i(t-1)$
5: **if** $\Delta U'_i(t) \geq 0$ **and** $r^{\text{env}}_{t,i} > 0$ **then**
6: Send request $x_i = \Delta U'_i(t)$ to all neighbors $j \in \mathcal{N}_{t,i}$
7: **else if** $\Delta U'_i(t) < 0$ **and** $r^{\text{env}}_{t,i} > 0$ **then**
8: $r_{\text{punish}} \leftarrow \alpha \Delta U'_i(t)$
9: **end if**
10: Let $\{x_j\}_{j \in \mathcal{N}_{t,i}}$ be the set of received requests from neighbors, and let $\hat{r}_{\text{req}}$ be $\max(\{x_j\} \cup \{0\})$
11: **for** each received request x_j from neighbor j **do**
12: **if** $MI_i(r^{\text{env}}_{t,i} + \hat{r}_{\text{req}}) \geq 0$ **then**
13: Send response $y_i = +x_j$ to agent j
14: **else**
15: Send response $y_i = -x_j$ to agent j
16: **end if**
17: **end for**
18: Let $\{y_j\}_{j \in \mathcal{N}_{t,i}}$ be the set of received responses for requests sent by i, and let $\hat{r}_{\text{res}}$ be $\min(\{y_j\} \cup \{0\})$
19: **return** Adjusted reward $\hat{r}_{t,i} = r^{\text{env}}_{t,i} + \hat{r}_{\text{req}} + \hat{r}_{\text{res}} + r_{\text{punish}}$

MI, $\Delta U'_i(t)$ (i.e., the MI difference from the previous step), and adjusts its own reward based on this value.

Algorithm 1 shows the SI-MATE algorithm. SI-MATE replaces Algorithm 2 of MATE with this. Relative to MATE, SI-MATE modifies the protocol in two ways:

- When the social influence improves ($\Delta U'_i(t) \geq 0$) and the agent receives a positive environment reward, it sends a positive tokens that matches $\Delta U'_i(t)$ (Line 6).
- When the social influence deteriorates ($\Delta U'_i(t) < 0$) while the agent receives a positive environment reward, it adds a self-punishment $\alpha \Delta U'_i(t)$ to its reward (Line 8,19).

The self-punishment is similar to inequity aversion (Hughes et al. 2018) and α is the scaling factor. Lines 10–18 of Algorithm 1 remain unchanged from MATE.

Experimental Results

We evaluated our method on the $Coin$[4] environment with a 5×5 gridworld (cf. (Phan et al. 2024) for details). Each agent learned a policy using an independent actor-critic method, assuming fully observable neighbors (i.e., $\mathcal{N}_{t,i} = \mathcal{I} \setminus i$ for all t). We conducted five trials each using four different methods. The parameter used in SI-MATE was $\alpha = 2$, and other hyperparameters related to learning were set according to MATE. As evaluation metrics, we used the efficiency U and "own coin rate" $P(\text{own coin}) = \frac{\text{\# collected coins with same color}}{\text{\# all collected coins}}$ based on the coins collected by each agent.

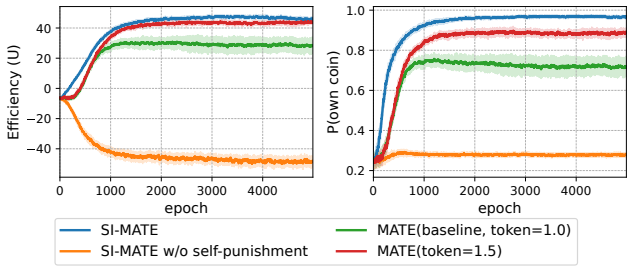

Figure 1: Comparison of learning curves. Efficiency is shown on the left, and own coin rate on the right. The blue line represents SI-MATE (our method); the orange line shows SI-MATE without self-punishment; the green line indicates MATE with token size 1.0 (baseline); and the red line corresponds to MATE with the token size tuned to 1.5 for higher efficiency.

Figure 1 shows the changes in the averaged metrics for each method over training epochs. SI-MATE outperforms existing methods in both metrics, demonstrating high coordination. Specifically, compared to the baseline of MATE's token size of 1.0, a 50% improvement in efficiency was observed. Furthermore, it achieved slightly higher efficiency than when we manually tuned MATE's token size. Furthermore, SI-MATE learns faster than existing methods and acquires cooperative policies from the early stages of training. In contrast, SI-MATE without self-punishment acquired a non-cooperative policy. This demonstrates that self-evaluation significantly contributes to our algorithm acquiring cooperative policies.

Conclusion

We propose SI-MATE, an improved MATE algorithm aimed at dynamic token size determination. This paper demonstrates that agents can acquire better token sizes during learning and achieve cooperative policies even in SSD environments like the $Coin$. Our future work is to generalize SI-MATE to enable cooperative policy acquisition in more complex environments.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 24K01112 and 25K24424 and JST CREST Grant Number JPMJCR2564.

References

Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez Guzman, E.; García Castañeda, A.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; Roff, H.; and Graepel, T. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, volume 31.

Mu, C.; Guo, H.; Chen, Y.; Shen, C.; Hu, D.; Hu, S.; and Wang, Z. 2024. Multi-agent, human–agent and beyond: A survey on cooperation in social dilemmas. *Neurocomputing*, 610: 128514.

Phan, T.; Sommer, F.; Ritz, F.; Altmann, P.; Nüßlein, J.; Kölle, M.; Belzner, L.; and Linnhoff-Popien, C. 2024. Emergent cooperation from mutual acknowledgment exchange in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 38(2): 34.