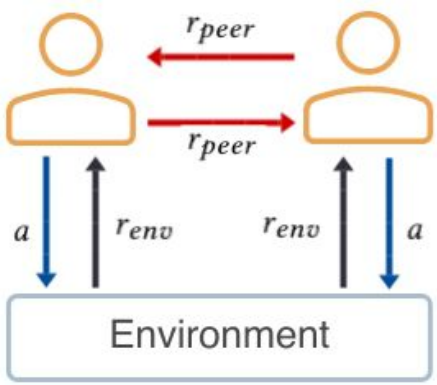
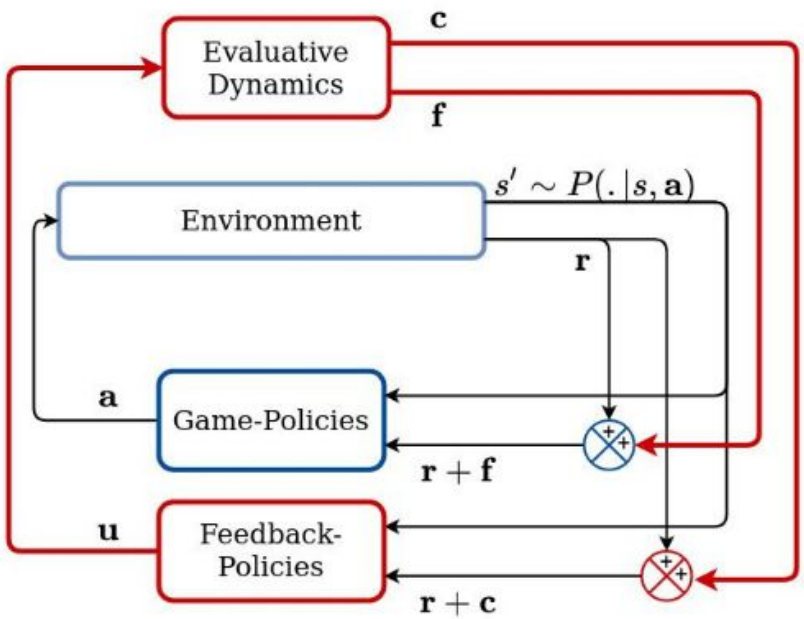
LIEF: Learning to Influence through Evaluative Feedback

Ramona Merhej and Mohamed Chetouani

Inter-Agent Rewards





How can agents *learn* an effective *rewarding policy* to increase cooperation in multi-agent reinforcement learning?

Related Work

Influencing an opponent

- **Opponent Modelling:**

→ gradient-based opponent:

- Zhang, C. et al., 2010
- Foerster, J. et al., 2018
- Letcher, A. et al., 2018

→ non gradient-based opponent:

- Xie, A. et al., 2020

Previous works:
Influencing opponent through regular *actions*

Our work:
Influencing opponent through *rewards*

Optimal reward functions

- **Inverse Reinforcement Learning (IRL):**

→ *given optimal policy or trajectories, can we recover R?* (Ng, A. Y., & Russell, S. J., 2000)

- **AutoRL:**

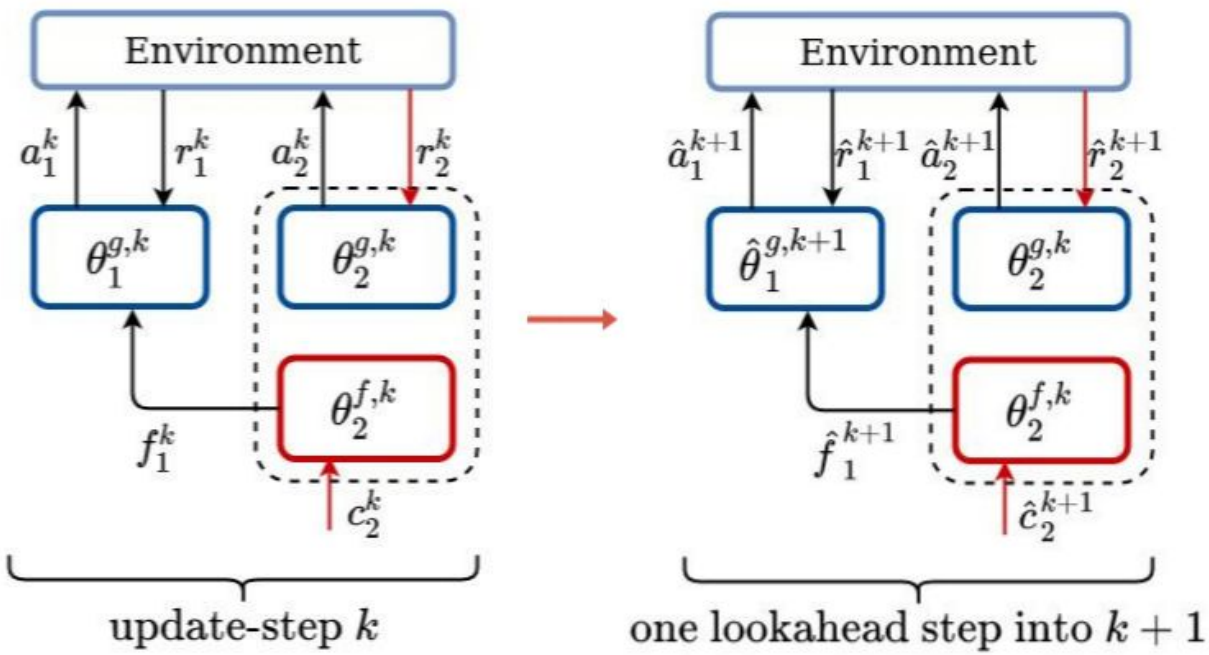
→ *given a goal or a task, can we recover R?*
(Chiang, H. T. L. et al., 2019)

- **Adversarial RL:**

→ *given optimal adversarial policy or trajectories and R_env, can we recover R_adv?* (Rakhsha, A. et al., 2020; Zhang, X. et al., 2020)

Our work:
→ *given a gradient-based opponent and without prior access to a desired opponent policy, can we recover R_peer?*

The Model



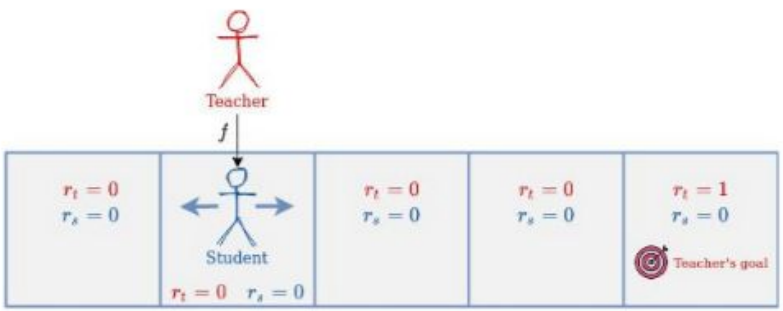
Game policy:

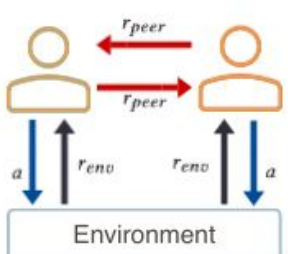
$$\sum_{t=0}^{T} \gamma_g^t (r_t^k + f_t^k)$$

Feedback policy:

$$\sum_{t=0}^{T} \gamma_f^t (r_t^{k+1} + c_t^{k+1}) - \sum_{t=0}^{T} \gamma_f^t (r_t^k + c_t^k)$$

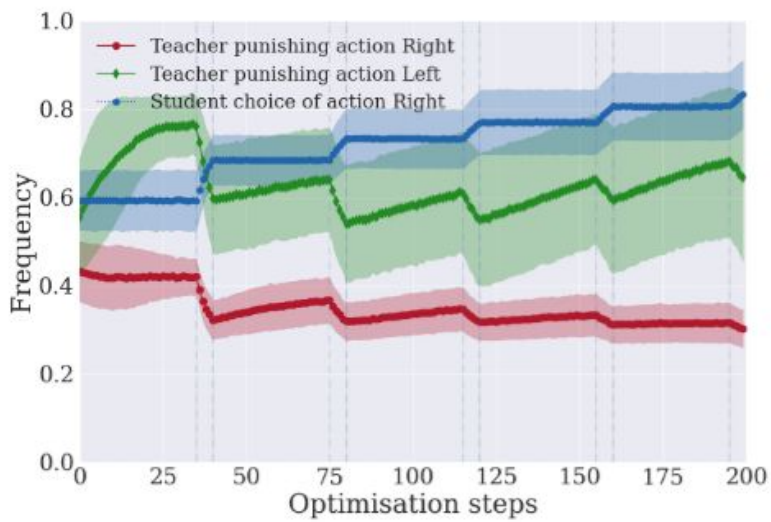
Experiments 1/2: Teacher-Student





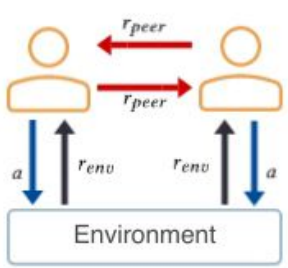
$$f = 0 \rightarrow c = 0$$

$$f = -1 \rightarrow c = -1$$

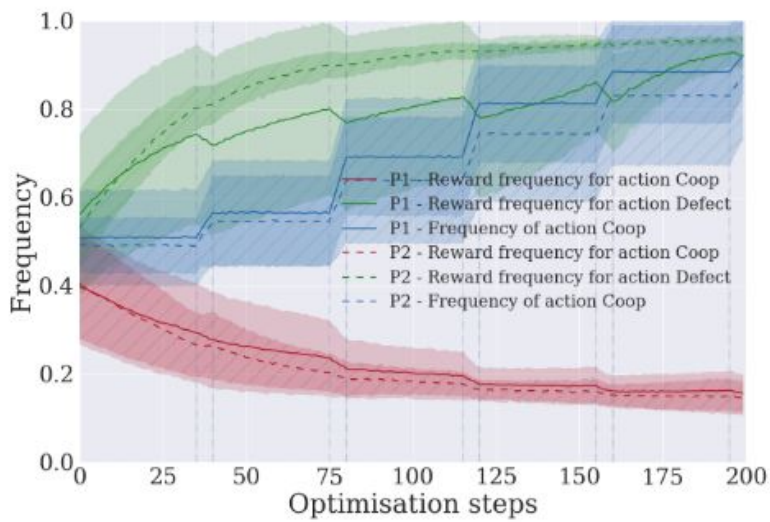


Experiments 2/2: Iterated Prisoner's Dilemma

Actions	A2 - C	A2 - D
A1 - C	$(-1, -1)$	$(-3, 0)$
A1 - D	$(0, -3)$	$(-2, -2)$



$$f = 0 \quad \rightarrow c = 0$$
$$f = -3 \rightarrow c = -1$$



Conclusion

Performing one look-ahead step allows agents to correctly reward each other in simple environments without a prior knowledge of the optimal opponent policy
→ *How does this method scale for larger state spaces?*
→ *How does it scale with n players?*

Only negative peer rewarding was tested
→ *How would agents learn to reward using positive feedback instead?*

The feedback dynamics, i.e., the ratio between the value of the given feedback and its cost are crucial for effective learning
→ *At what point do the immediate costs of sending feedback become larger than their long term benefits?*

References

- Chiang, H. T. L., Faust, A., Fiser, M., & Francis, A. (2019). Learning navigation behaviors end-to-end with autorl. IEEE Robotics and Automation Letters, 4(2), 2007-2014.

- Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., & Mordatch, I. (2017). Learning with opponent-learning awareness. arXiv preprint arXiv:1709.04326.

- Letcher, A., Foerster, J., Balduzzi, D., Rocktäschel, T., & Whiteson, S. (2018). Stable opponent shaping in differentiable games. arXiv preprint arXiv:1811.08469.

- Ng, A. Y., & Russell, S. J. (2000, June). Algorithms for inverse reinforcement learning. In Icml (Vol. 1, p. 2).

- Rakhsha, A., Radanovic, G., Devidze, R., Zhu, X., & Singla, A. (2020, November). Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In International Conference on Machine Learning (pp. 7974-7984). PMLR.

- Xie, A., Losey, D. P., Tolsma, R., Finn, C., & Sadigh, D. (2020). Learning Latent Representations to Influence Multi-Agent Interaction. arXiv preprint arXiv:2011.06619.

- Zhang, C., & Lesser, V. (2010, July). Multi-agent learning with policy prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 24, No. 1).

- Zhang, X., Ma, Y., Singla, A., & Zhu, X. (2020, November). Adaptive reward-poisoning attacks against reinforcement learning. In International Conference on Machine Learning (pp. 11225-11234). PMLR.

Thank you for your attention !