Deep Spite

Yann Billeter
Alexander Born

Abstract

Existing multi-agent reinforcement learning research commonly assumes rationality of the agents. While the design of the stochastic games includes cooperative, competitive and mixed settings, policy design does not consider irrational behaviour harming the agent's own score. We propose a novel Nash-suboptimal learning policy, allowing the agents to be myopic regarding their objective goals. We introduce the element of spite and analyse its propagation through the evolution of the environment. Motivated by the conceptual similarity of anti-social and adversarial behaviour, we investigate the effect of spiteful training on robustness to adversarial attacks.

1. Introduction

Multi-agent Reinforcement Learning (MARL) allows researchers and practitioners to solve complex tasks in robotics and control, communication, social policies and finance, and are even applied to climate change negotiations (Zhang et al.).

While, the use of policies converging towards the Nash equilibrium, such as in Nash Q learning, seems to be a silent agreement regarding agents' behaviour, there are a few potential arguments for why one should be motivated to study irrational behaviour in the context of (multi-agent) reinforcement learning. One apparent reason is that human decision-making has repeatedly and consistently been shown to diverge strongly from the game-theoretic solution to various problems (Reed et al., 2013; Pesendorfer, 2006). Hence, rationality might constitute a limitation for reinforcement learning (RL) systems: In settings with human-AI interaction for instance, one ought to be motivated to train RL-agents with policies that perform well in interactions with irrational or adversarial agents. It seems likely to us that MARL training which does not contain examples of irrationality would, upon deployment, be subject to distributional shift in the other agent's behaviour. This is notable because distributional shift in the agent environment is a well-documented failure-mode of current deep RL systems (Amodei et al., 2016).

The contribution of our research is twofold. First, we address the aforementioned arguments through novel application of spiteful behaviour. We define spite as suboptimal actions of the spiteful agent, as well as sacrifice of the agent's own score to gain an opportunity to harm another agent. As the environment is only partially known to the agents, we propose to resort to randomness in selection of a victim of a spiteful act. Second, we combine our proposition of random spiteful acts with deliberate adversarial acts of rogue agents, motivated by the intuition that a rational attack of an adversary might be countered by the random irrationality of an agent or, in the worst case, being amplified by the same irrationality.

2. Models and Methods

To our best knowledge, topics of altruism, spite and punishment are widely discussed in economics and social sciences, but are not a research focus in the field of Reinforcement Learning (RL), aside from (Franzmeyer et al., 2022), where the authors study altruism in RL. Specifically, the authors propose a duality of agents with a leader agent, executing its own policies, and an altruistic agent, who aims to enable the leader by maximizing the leader's choice. Similar to our intuition, the authors assume a partially observable environment, thus emphasizing on larger space of choices. Contrary to our ideas, the supporting agent does not get own rewards, but the reward of the leader.

Similarly, the authors of (Hughes et al.) and (Peysakhovich & Lerer, 2017) have investigated pro-social behaviour. The former authors conduct research on multi-agent games, where each agent has a choice of optimal short-term strategies, which can lead to suboptimal long-term outcome for the group by introducing separate policies for cooperation and defection. The latter authors implement consequential reward policies to encourage cooperation in multi-episodes games, following the intuition of the tit-for-tat strategy (Axelrod & Hamilton, 1981), well-studied in the Prisoner's Dilemma setting.

The dynamics of spite are the subject of focus in (Fulker et al.), where the authors define social and spiteful strategies, but without further references to RL. The inclusion of irrational actions is deemed an important research direction

in (Wong et al.).

Multi-agent deep deterministic policy gradient (MADDPG)

In our research, we estimate the optimal policies for Markov Decision Processes. Following (Lowe et al.) we expand the objective J for rewards R and the state-action value function Q

$$J(\theta) = \mathbb{E}_s[R(s,a)]$$
$$\nabla_\theta J(\theta) = \mathbb{E}_s[\nabla_\theta \mu_\theta(a|s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}]$$

for a deterministic policy μ, actions a and state s to an environment with N agents and optimize this objective via gradient descent.

Spite

We study the effects of spite by splitting spiteful behaviour in three stages. In each stage, we hinge on the idea of the advantage function for the agents as an indicator for spite. In Stage 1 we model the effects of *stubborn spite*. Our motivation is that an agent might become spiteful, if its rewards and state-action value unfavourably mismatch. Thus, we compute spite factor $\mathbf{sf}$ and argue that a spiteful agent might wish to deflect from optimal actions, with higher spite factors inducing higher deflection rates. For a batch $\mathcal{B}$ of actions, we compute $\mathbf{da}$, representing the amount of deflecting actions. We create an index set $\mathcal{I}$ by randomly choosing samples from $\mathcal{B}$ and update actions $\tilde{a}$ by random permutations within the selected samples. In the last step, we recompute Q^μ with updated actions, using $(\tilde{a}_i, a_j)$ notation to account for the fact, that unchanged actions still contribute to Q^μ.

$$\mathbf{sf}_i = \max\left(\tanh(\frac{r_i - Q_i^\mu}{Q_i^\mu}), 0\right) \tag{1}$$
$$\mathbf{da}_i = \lfloor \mathbf{sf}_i * |\mathcal{B}| \rfloor \tag{2}$$
$$\mathcal{I} = (\mathcal{U}[1, |\mathcal{B}|])_{1,\dots,\mathbf{da}_i} \tag{3}$$
$$\tilde{a}_i = \mathcal{P}(a_i), a_i \in \mathcal{B}_\mathcal{I} \tag{4}$$
$$\tilde{Q}_i^\mu = Q_i^\mu(\tilde{a}_i, a_j) \tag{5}$$

In Stage 2 we model the effects of *harmful spite*. Once again, we start with mismatched expectations as the trigger, but decide not to alter the actions and instead directly sacrifice a part of agent's Q_i^μ in exchange for directly impacting a random agent. Thus, equations (1), (2) and (3) are unchanged, while we compute penalty fee $\mathbf{d}(s)$ for the agent's spiteful actions, the inflicted harm $\mathbf{h}(s)$ on a randomly selected actor and update the respective Q values. For brevity, we omit the dependence of penalty and harm on the state s and write shorthand $\mathbf{d}$ and $\mathbf{h}$.

$$j = \mathcal{U}[N \setminus i] \tag{6}$$
$$\mathbf{d}_i = \alpha * Q_i^\mu \tag{7}$$
$$\mathbf{h}_j = \beta * \mathbf{d}_i \tag{8}$$
$$\tilde{Q}_i^\mu = Q_i^\mu - \mathbf{d}_i \tag{9}$$
$$\tilde{Q}_j^\mu = Q_j^\mu - \mathbf{h}_j \tag{10}$$

In Stage 3 - *utility spite* - we argue that contrary to the idea of random spiteful behaviour, spiteful actions are not blind and erratic, but actually satisfy a subjective, short-term utility function $U(s,r)$ depending on the state s and the expected reward r. We propose an asymmetric utility function dependent on the expected reward and possible negative impact on other agents. The asymmetry arises from the fact, that the harm $\mathbf{h}$ inflicted to the neighbours of the spiteful agent should be greater than the fee $\mathbf{d}$ deducted from the agent's current score. Combining both desiderata, we propose the following adaptation to the state-action value function:

$$\tilde{Q}(s,a,r) = Q(s,a) + U(s,r),$$
$$U(s,r) = \max\{\alpha * (Q_i^\mu - r), \beta * (\mathbf{h} - \mathbf{d})\}$$

where $Q_i^\mu(\tilde{a}_i, a_j)$ denotes once again Q function updated by random actions. As we do not compute (1), we compute (2) as $\mathbf{da}_i = \lfloor \mathcal{U}[0,1] * |\mathcal{B}| \rfloor$. As agents have only partial observations of their environment, we suggest that the best guess an agent can make about its spiteful actions is the difference in the rewards or the state-action values compared to a suboptimal $\tilde{Q}$. Once again, we use deflecting actions as updates and argue that the agent can directly assess its own Q values, but estimates others' values via its rewards. Thus, we imply that the spiteful agent might falsely assume similarity of its rewards and other agents' state values.

$$\mathbf{h} = Q_i^\mu - Q_i^\mu(\tilde{a}_i, a_j)$$
$$\mathbf{d} = r_i - Q_i^\mu(\tilde{a}_i, a_j)$$

Adversarial attacks

To investigate the impact of spite on adversarial attacks, we adopt the framework described (Liu et al.). We assume that an adversary can take control of a pretrained agent (the victim) and modify its actions. Three strategies for the selection of adversarial actions are implemented.

We implement the *randomly timed attacks* and *strategically timed attacks*, which aim to decrease the performance of the multi-agent system by deteriorating the performance of the compromised agent, thus referred to as *self-destructive attacks*. In *randomly timed attacks*, the adversary replaces the actions of the victim at random time steps with off-distribution actions. The *strategically timed attack* aims to maximize the negative impact by attacking at strategically

selected time steps. For this, we first compute the c-function defined in (Lin et al., 2017):

$$c(s_t) = \max_{a_t}(\pi_m(s_t, a_t)) - \min_{a_t}(\pi_m(s_t, a_t))$$

The adversary attacks if $c > \tau$, where τ is some pre-defined attack threshold. π_m denotes the policy of agent m.

The third attack is the *counterfactual reasoning-based attack* from (Liu et al.).In this attack, the adversary attempts to predict how the compromised agent's actions will affect the other agents' actions. It then aims to maximize the destructive impact of the attack. To this end, an additional RL agent is trained with the following reward function,

$$r_{\text{att}} = \sum_{t}^{l} \gamma^t D_{\text{KL}}(p(a_t^{-m}|a_t^m, s_t)\|p(a_t^{-m}|a_t^{*m}, s_t))$$

m is the index of the agent in the system and γ a discount factor. The agent learns to replace the original actions $a_t^m, \ldots, a_{t+l}^m$ with the best set of adversarial actions $a_t^{*m}, \ldots, a_{t+l}^{*m}$. Its reward function encourages actions that maximize the divergence between the original policy and the adversarial one. The additional adversarial agent is trained using deep deterministic policy gradient (DDPG) (Lillicrap et al., 2015).

Games, environments and benchmarking

We provide a brief overview of the environments, games and the benchmarking process used in our study.

LB-Foraging: Level based foraging (Christianos et al., 2020) is a 2D-grid based game, where try to collect items near them. An item can only be collected when the sum of levels of agents involved is higher than the item's level.

MPE: Multi-agent particle environments (Lowe et al.) are simple particle worlds with different game modes. In the cooperative setting, N agents navigate across the world, trying to cover all landmarks and avoiding collisions. In mixed setting, N good agents compete against one adversary in a navigation game. The competitive setting is a simple predator-prey model, where N predators hunt M preys.

Benchmarking: We run each game 10 times for each stage of spite, and benchmark it against the standard implementation in EPyMARL (Papoudakis et al., 2021). We initially considered running experiments for SMAC environment (Samvelyan et al.), but refrained from doing so due to computational resource constraints.

3. Results

We answer several research questions through thorough analysis of our numerical experiments.

1) Does the introduction of spite lower the average rewards, while benefiting few individuals? We observe spiteful agents outperforming the non-spiteful baseline nearly in each game. Notably, *utility spite* shows either best performance or is the runner-up in most of the games. See Fig. 1.

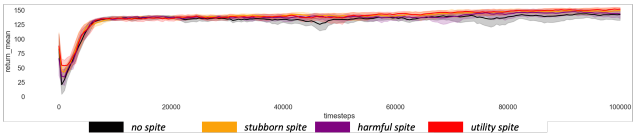

Figure 1. Mean returns in SIMPLEADVERSARY with 16 agents. *Utility spite* and *stubborn spite* clearly outperform non-spiteful benchmark

We analyse $Q-$values of each agent to understand the detailed dynamics of the individuals, c.f. Fig. 2. First, we observe that also on the individual level, spiteful agents seem to perform better. Second, we notice increased larger variance, suggesting, that in fact, agents experience runs, where they are worse off, while being spiteful. We report *utility spite* to perform stable across games, while *harmful spite* had significant problems in cooperative LB-Foraging game with 6 agents. Overall, we find *harmful spite* being the least robust method to model spiteful behaviour.

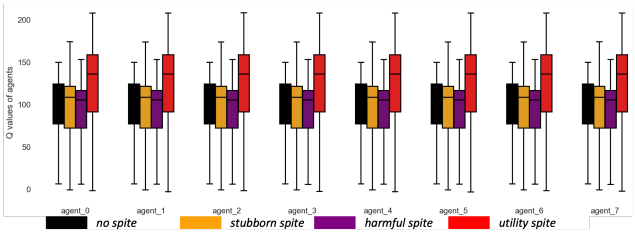

Figure 2. Q-values for each agent in SIMPLESPREAD with 8 agents

2) Does spite propagate in the environment through harming interactions with grieving neighbours? We again observe heterogeneous propagation of spite through different environments. *Stubborn spite* and *harmful spite* tend to propagate extremely fast until each agent is spiteful, while *utility spite* tend to stay low and decay over a few episodes. Results are displayed in Fig. 3.

3) Does spite increase robustness against adversarial attacks? We observe mixed results for effects of spite against adversarial attacks, with averaged returns over an epoch being sometimes close for all approaches and sometimes in favour for spiteful agents, see Fig. 4 for a comparison.

Notably, *stubborn spite* seems to increase robustness against attacks in cooperation-based games.

Spiteful agents can perform significantly worse than the baseline. We explain this as follows: The KL-attack tends

Figure 3. Propagation of spite in 8x8 LBFORAGING with 2 players and 2 food locations. Cooperative scenario on top, default below.

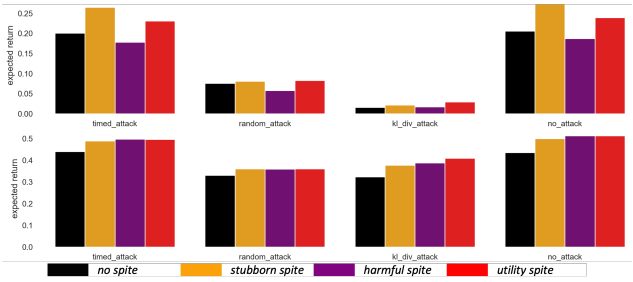

Figure 4. Expected returns in 8x8 LBFORAGING with 2 players and 2 food locations under different attacks. Cooperative scenario on top, default below.

to display much larger negative impact than the other attacks. This can be explained by the fact that this attack aims to maximize the divergence between the compromised agent's actions and the learned distribution of actions. Hence, spiteful policies are unlikely to provide robustness beyond the improvements described in 3.1.

The self-destructive attacks aim to lower the reward of the compromised agent, which may trigger spite in the attacked agent. As all three stages of spite include a decrease in the agent's reward, further decreasing the reward beyond the impact of the attack.

The comparatively worse performance of stage 2 spite in the cooperative setting might be attributed to propagation of spite. When attacked, an agent implementing stage 2 spite will aim to harm other agents, in turn triggering spiteful behaviour in the other agents.

4) Are the findings consistent across cooperative, competitive and mixed-settings tasks? We present our finding in the table below. For mean returns, we present the best learner printed in bold as well as the runner-up. While being spiteful is, in general, beneficial, simply harming a random agent does not tend to be a winning tactic, over either random disagreement or utility-based approach.

4. Discussion

Surprisingly, our results do not support the initial intuition, that spiteful actions would overall lower the average rewards. Contrary, we see positive effects of spite, which might be actually an indication for spite as regularizer in cooperative

Game	Mode	# Agents	Average Returns
Foraging	cooperative	2	**stubborn**, utility
Foraging	competitive	2	**harmful**, utility
Foraging	cooperative	6	**utility**, non-spiteful
Foraging	competitive	6	**utility**, harmful
MPE	competitive	6	**utility**, non-spiteful
MPE	cooperative	8	**stubborn**, utility
MPE	mixed	16	**utility**, stubborn

Table 1. Overview of best average returns over games

behaviour. The results indicate, that additional randomness in actions, as well as utility-driven spite are well-suited methods to model spite, while harm inflicted on a random victim rarely is the leading approach. Our intuition regarding individual returns is indeed correct, and we can report higher $Q-$values for individual agents.

The propagation of *utility spite* deviates significantly from the dynamics of *stubborn spite* and *harmful spite*, with only one game where *utility spite* affects all agents. We attribute it to the fact that our asymmetric utility function leaves the choice not to spite as long as the training updates are sufficiently good, and to spite when the agent encounters a plateau of its loss function. While more evidence is required, this finding might actually improve the overall learning strategies in MARL by scheduling spiteful actions in the later stages of training.

The results regarding the effects of spite on adversarial robustness are inconclusive. While a slight benefit is present, we deem further research necessary. An interesting avenue for further development could be to base future formulations of spite on adversarial attacks, such as the counterfactual reasoning-based attack.

Since publication of the original MADDPG paper (Lowe et al.), more advanced methods for MARL have been published. The extension of the methodology developed in this paper presents a possible avenue for future research into spiteful MARL.

Our research is limited to partially observable environments with limited number of agents, which have no knowledge over the score of other individuals. Intuitively, in a denser populated environment where scores are either known or could be imagined, the dynamics of spite as proxy of envy are of great interest, and will be in the focus of further research by the authors.

5. Summary

Our novel approach of spiteful agents shows promising results, outperforming plain implementation of MADDPG in several games and different environments. We attribute these findings to increased robustness through action permu-

tation as well as learnable utility function to activate spiteful behaviour, considering both ideas desirable in complex MARL games. Further findings indicate a potential gain in the training process which can be achieved by scheduling spiteful behaviour in the later stages of the training process. The aforementioned scheduling as well as analysis of more densely populated environments are subject to future work in the exciting field of MARL.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. Concrete problems in AI safety. *Computing Research Repository (CoRR)*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565.

Axelrod, R. and Hamilton, W. D. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981. doi: 10.1126/science.7466396. URL https://www.science.org/doi/abs/10.1126/science.7466396.

Christianos, F., Schäfer, L., and Albrecht, S. V. Shared experience actor-critic for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Franzmeyer, T., Malinowski, M., and Henriques, J. F. Learning altruistic behaviours in reinforcement learning without external rewards. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KxbhdyiPHE.

Fulker, Z., Forber, P., Smead, R., and Riedl, C. Spite is contagious in dynamic networks. 12 (1):260. ISSN 2041-1723. doi: 10.1038/s41467-020-20436-1. URL https://www.nature.com/articles/s41467-020-20436-1.

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. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2018/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning, 2015. URL https://arxiv.org/abs/1509.02971.

Lin, Y., Hong, Z., Liao, Y., Shih, M., Liu, M., and Sun, M. Tactics of adversarial attack on deep reinforcement learning agents. *CoRR*, abs/1703.06748, 2017.

Liu, T., McCalmon, J., Rahman, M. A., Lischke, C., Halabi, T., and Alqahtani, S. Weaponizing actions in multi-agent reinforcement learning: Theoretical and empirical study on security and robustness. In Aydoğan, R., Criado, N., Lang, J., Sanchez-Anguix, V., and Serramia, M. (eds.), *PRIMA 2022: Principles and Practice of Multi-Agent Systems*, Lecture Notes in Computer Science, pp. 347–363.

Springer International Publishing. ISBN 9783031212031. doi: 10.1007/978-3-031-21203-1_21.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6382–6393. Curran Associates Inc. ISBN 9781510860964.

Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S. V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

Pesendorfer, W. Behavioral economics comes of age: A review essay on "advances in behavioral economics". *Journal of Economic Literature*, 44(3):712–721, 2006. ISSN 00220515. URL http://www.jstor.org/stable/30032350.

Peysakhovich, A. and Lerer, A. Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975*, 2017.

Reed, D. D., Niileksela, C. R., and Kaplan, B. A. Behavioral economics: a tutorial for behavior analysts in practice. *Behav Anal Pract*, 6(1):34–54, 2013.

Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., and Whiteson, S. The StarCraft multi-agent challenge. URL http://arxiv.org/abs/1902.04043. type: article.

Wong, A., Bäck, T., Kononova, A. V., and Plaat, A. Deep multiagent reinforcement learning: challenges and directions. ISSN 1573-7462. doi: 10.1007/s10462-022-10299-x. URL https://doi.org/10.1007/s10462-022-10299-x.

Zhang, T., Williams, A., Phade, S., Srinivasa, S., Zhang, Y., Gupta, P., Bengio, Y., and Zheng, S. AI for global climate cooperation: Modeling global climate negotiations, agreements, and long-term cooperation in RICE-n. URL https://papers.ssrn.com/abstract=4189735.