Towards a Unifying Model of Rationality in Multiagent Systems

Robert Loftin
Delft University of Technology
Delft, The Netherlands
R.T.Loftin@tudelft.nl

Mustafa Mert Çelikok
Aalto University
Espoo, Finland
mustafamert.celikok@aalto.fi

Frans A. Oliehoek
Delft University of Technology
Delft, The Netherlands
F.A.Oliehoek@tudelft.nl

ABSTRACT

Multiagent systems deployed in the real world need to cooperate with other agents (including humans) nearly as effectively as these agents cooperate with one another. To design such AI, and provide guarantees of its effectiveness, we need to clearly specify what types of agents our AI must be able to cooperate with. In this work we propose a generic model of *socially intelligent agents*, which are individually rational learners that are also able to cooperate with one another (in the sense that their joint behavior is Pareto efficient). We define rationality in terms of the *regret* incurred by each agent over its lifetime, and show how we can construct socially intelligent agents for different forms of regret. We then discuss the implications of this model for the development of "robust" MAS that can cooperate with a wide variety of socially intelligent agents.

KEYWORDS

Multiagent Learning, Game Theory, Human-AI Interaction

1 INTRODUCTION

Multiagent systems deployed in the real world (e.g., autonomous vehicle fleets) must be robust to the presence of other intelligent agents that are unlike themselves, such as AI's designed by other companies, or possibly human beings. Ideally, such systems should not only be robust to other heterogeneous agents, but should be able to reliably cooperate with these agents when it would be mutually beneficial. The first step in designing such robust, cooperative systems is to identify the types of agents they should be able to cooperate with. The key challenge is here is that we will often have little or no prior data regarding the behavior of other agents our system might encounter. The goal of this work is therefore to develop a formal model of *socially intelligent*[1] behavior in multiagent settings, which can serve as the basis for defining robust cooperation, and for creating agents that satisfy this definition.

In developing our social intelligence (SI) model we restrict our attention to the case of two adaptive agents playing a repeated bimatrix game. Our model imposes two requirements on these agents. The first is that each be *consistent*, in the sense that they will eventually play a best-response to any fixed partner strategy. The second is that each agent be *compatible* with some set of agents (potentially just the agent itself), such that when paired with any other member of this set, the joint payoffs will be Pareto efficient (over the set of equilibrium strategies). Our model is inspired by the *targeted learning* criteria [12, 13], with the key distinction being that we require socially intelligent agents to be consistent against all possible partner strategies. This implies that socially intelligent agents cannot simply fall back to a non-cooperative, "safe" strategy upon encountering another adaptive agent. In Section 3.1, we consider two notions of consistency against non-stationary partners. Our main theoretical contribution (Section 4) is showing how agents satisfying the SI criteria can be constructed for both definitions of consistency. We conclude with a discussion of the practical utility of this model for the design of multi-agent systems, and pose additional theoretical questions as directions for future work.

[1]While we borrow the terminology, we cannot claim that our model captures the nuances of the psychological concept of social intelligence [8].

Appears in the 14th Workshop on Optimization and Learning in Multiagent Systems (OptLearnMAS 2023), , May 29, 2023, London, England. 2023.

2 PRELIMINARIES

We consider the case where interactions between agents take the form of repeated, two-player general-sum matrix games. We consider classes of games defined by a *type space* Θ, where the current type $\theta \in \Theta$ is known to both agents. For simplicity, we assume that in all games both players have N pure strategies (henceforth "actions") available. We let $G_1(\theta)$ and $G_2(\theta)$ denote the $N \times N$ payoff matrices for players 1 and 2 in the game defined by type $\theta \in \Theta$. With a slight abuse of notation, we let $G_i(s_1, s_2; \theta) = s_1^\top G_i(\theta) s_2$ denote the expected payoffs for player $i \in 1, 2$ under the mixed strategy profile $\langle s_1, s_2 \rangle$. We assume that agents interact for a fixed number of stages $0 < T < \infty$, and let a_t^1 and a_t^2 denote the actions chosen by players 1 and 2 in stage $0 < t \leq T$. We overload a_t^1 and a_t^2 to also denote the mixed strategies that assign all probability mass to actions a_t^1 and a_t^2, such that $G_i(a_t^1, a_t^2; \theta)$ is player i's payoff at stage t given type θ. We also assume that for all $\theta \in \Theta$ and $a^1, a^2 \in [N]$, $G_i(a^1, a^2; \theta) \in [0, 1]$. We let $p(s_1, s_2; \theta) = \langle G_1(s^1, s^2; \theta), G_2(s^1, s^2; \theta) \rangle$ denote the players' payoff profile given the joint strategy $\langle s^1, s^2 \rangle$ and type θ.

Let $\mathcal{H}_t = (N \times N)^t$ be the set of histories of play of length t (with $\mathcal{H}_0 = \{\emptyset\}$), and let $\mathcal{H}_{\leq t} = \bigcup_{s=0}^{t} \mathcal{H}_s$ be the set of all histories of length at most t. An *agent* π is pair of mappings $\langle \pi_1, \pi_2 \rangle$ with $\pi_i : \Theta \times \mathcal{H}_{\leq T-1} \mapsto \Delta(N)$, where $\Delta(N)$ is the set of probability distributions over the action set $[N]$. For each type $\theta \in \Theta$, an agent defines separate *behavioral strategies* [15, Chapter 5.2.2] for controlling player 1 and 2, which implies that the agent is aware of its own player ID. Let $h_t \in \mathcal{H}_t$ be the history of play up to stage t, and let $s_t^i = \pi_i(\theta, h_{t-1})$ be player i's action distribution at t, with $a_t^i \sim s_t^i$. Each player observes its partner's actions, but not their full action distributions. We overload p to also denote the empirical average payoff profile for a history h, such that

$$p_i(h; \theta) = \frac{1}{|h|} \sum_{t=1}^{|h|} G_i(a_t^1(h), a_t^2(h); \theta), \tag{1}$$

where $|h|$ is the length of the history, and $a_t^i(h)$ is player i's action at stage t observed under h. Finally, all probabilities and expectations will be conditional on the current type θ, and the strategies π_1 and

π_2 selecting actions for players 1 and 2 respectively. These define a finite probability space over the set of final histories $h_T \in \mathcal{H}_T$.

3 CONSISTENCY AND COMPATIBILITY

Informally, we say that an agent π is *socially intelligent* (SI) if for every type $\theta \in \Theta$ it is both *consistent* with θ for every possible partner strategy, and *compatible* with itself under θ. Consistency means that π is guaranteed to perform nearly as well as some best-response to its partner's observable behavior, while compatibility means that the joint payoff profile will be nearly as good as that of some Pareto-efficient joint strategy. In this section, we will consider two formal definitions of consistency, leading to two definitions of social intelligence.

3.1 Consistent Agents

Our first criteria for social intelligence is that an agent acts as a consistent learner, and attempts to achieve a payoff nearly as large as that of the best response to its partner's strategy. This is complicated by the fact that the partner's strategy may be non-stationary (particularly if it adapts to the agent as well). We therefore consider two notions of consistency, *adversarial* and *stochastic*, that account for this non-stationary behavior in different ways. Our first definition of consistency requires that the agent be robust to such adversarial partners, and relies on the standard notion of *external regret* [6], defined as

$$R_i^{\text{ext}}(h;\theta) = \max_{a^i \in [N]} \sum_{t=1}^{|h|} \left\{ G_i(a^i, a_t^{-i}(h);\theta) - G_i(a_t^i(h), a_t^{-i}(h);\theta) \right\} \tag{2}$$

where $i \in \{1,2\}$ denotes the player ID of the agent in question, and we use $-i$ to denote the ID of its partner. Adversarial consistency simply requires that an agent have bounded external regret (i.e., that it be *Hannan consistent*) over T stages.

Definition 3.1 (Adversarial Consistency). For $\delta, \epsilon, T > 0$, an agent π is (δ, ϵ, T)-*adversarially consistent* if, for all types $\theta \in \Theta$, and all partner agents $\tilde{\pi}$, we have that $\frac{1}{T} R_i^{\text{ext}}(h_T;\theta) \leq \epsilon$ with probability at least $1-\delta$, for either player $i \in \{1,2\}$.

While external-regret is theoretically convenient, it is unlikely that most real agents (including humans) would implement a no-regret learning algorithm. More realistically, we could expect agents to build an empirical model of their partner's behavior, and act optimally with respect to this model. If the partner could be assumed to choose a fixed strategy s_{-i}, then a natural strategy for our agent would be *fictitious play* [14], under which the agent plays a best-response to its partner's empirical strategy so far. For history h, we define the fictitious play strategy $s_{\text{fp}}^i(h;\theta)$ as

$$s_{\text{fp},t}^i(h;\theta) \in \arg\max_{s \in \Delta(N)} \sum_{r=1}^{t-1} G_i(s, a_r^{-i}(h);\theta) \tag{3}$$

where we choose the $s_{\text{fp},t}^i(h;\theta)$ to be the uniform distribution over all optimal actions given h_{t-1} for player i. We then define the *stochastic regret* $R_i^{\text{sto}}(h;\theta)$ of player i under history $h \in \mathcal{H}_{\leq T}$ as

$$R_i^{\text{sto}}(h;\theta) = \sum_{t=1}^{|h|} \left\{ G_i(s_{\text{fp},t}^i(h;\theta), a_t^{-i}(h);\theta) - G_i(a_t^i(h), a_t^{-i}(h);\theta) \right\} \tag{4}$$

This is the difference between the payoff player i would have received had they followed the strategy suggested by fictitious play, rather than π_i, assuming that player $-i$'s actions would have remained unchanged. Using this notion of regret, we can now define our second notion of consistency.

Definition 3.2 (Stochastic Consistency). For $\delta, \epsilon, T > 0$, an agent π is (δ, ϵ, T)-*stochastically consistent* if, for all types $\theta \in \Theta$, and all partner agents $\tilde{\pi}$, we have that $\frac{1}{|T|} R_i^{\text{sto}}(h_T;\theta) \leq \epsilon$ with probability at least $1-\delta$, for either player $i \in \{1,2\}$.

A stochastically consistent agent does not need to be robust to adversarially chosen partner actions. If changes in the partner's strategy over time would have caused fictitious-play to perform poorly, then a stochastically consistent agent is allowed to perform poorly as well. The following lemma (proved in Appendix A) will be useful for our analysis:

Lemma 3.3. *For any history $h \in \mathcal{H}_T$, type $\theta \in \Theta$, and player $i \in \{1,2\}$, we have that $R_i^{sto}(h;\theta) \leq R_i^{ext}(h;\theta)$.*

We can also define the *expected* adversarial regret as

$$\bar{R}_i^{\text{ext}}(h;\theta) = \max_{a^i \in [N]} \sum_{t=1}^{|h|} \left\{ G_i(a^i, a_t^{-i}(h);\theta) - G_i(s_t^i(h), a_t^{-i}(h);\theta) \right\} \tag{5}$$

and the expected stochastic regret as

$$\bar{R}_i^{\text{sto}}(h_t;\theta) = \sum_{t=1}^{|h|} \left\{ G_i(s_{\text{fp},t}^i(h;\theta), a_t^{-i}(h);\theta) - G_i(s_t^i(h), a_t^{-i}(h);\theta) \right\} \tag{6}$$

Finally we have

$$R_i^{\text{ext}}(h_t;\theta) \leq \bar{R}_i^{\text{ext}}(h_t;\theta) + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}, \tag{7}$$

$$R_i^{\text{sto}}(h_t;\theta) \leq \bar{R}_i^{\text{sto}}(h_t;\theta) + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}, \tag{8}$$

w.p. at least $1-\delta$ for all $t \leq T$ simultaneously (this follows directly from [1, Lemma 4.1]. We therefore only need to bound the expected regrets to provide high-probability regret bounds.

3.2 Compatible Agents

Consistency captures the idea that an agent is a general-purpose learner. What makes such a learner socially intelligent, however, is that it is capable of cooperating with other socially intelligent agents. We therefore need a notion of cooperation that does not preclude consistency. Let $\mathcal{N}(G) \subseteq \Delta(N) \times \Delta(N)$ be the set of Nash equilibria of the stage game G. For any $s \in \mathcal{N}(G(\theta))$, if both players act according to their respective components of s at each stage, then neither will incur any external (or stochastic) regret in expectation. For a fully cooperative game G (with $G_1 = G_2$), $\mathcal{N}(G)$ will contain all globally optimal strategy profiles. It may, however, also contain strategies that are highly suboptimal, but where neither player can

improve the joint payoff by changing their individual strategy, as in the fully cooperative 2×2 game:

	a^2	b^2
a^1	2	0
b^1	0	1

As in [12], we therefore define compatibility in terms of the *Pareto optimal Nash equilibra* (PONE) [11] of G, which we denote as the set $\mathcal{P}(G) \subseteq \mathcal{N}(G)$. We say that $s \in \mathcal{P}(G)$ if and only if $s \in \mathcal{N}(G)$, and there does not exist $s^{'} \in \mathcal{N}(G)$ such that $G_1(s^{'}) > G_1(s)$ *and* $G_2(s^{'}) > G_2(s)$. This means that s is a PONE if it is a Nash equilibrium of G, and it is not *strongly* Pareto-dominated by any other Nash equilibrium of G. This gives us the following definition of compatibility:

Definition 3.4 (Compatibility). For $\delta, \epsilon, T > 0$, two agents π and $\pi^{'}$ are (δ, ϵ, T)-*compatible* if, when π_i and $\pi^{'}_{-i}$ play together, for any type $\theta \in \Theta$ we have that w.p. at least $1 - \delta$, $\exists s \in \mathcal{P}(G(\theta))$ s.t.

$$\text{s.t. } p_i(s; \theta) - p_i(h_T; \theta) \leq \epsilon, \tag{9}$$

for either $i = 1$ or $i = 2$.

A pair of agents is compatible if, when paired together, with high-probability over their path of play h_T there will exist some PONE that does not ϵ-dominate their observed payoff profile $p(h_T; \theta)$. Note that this definition of compatibility is very similar to that provided in [12], but is now approximate, and defined over a finite time horizon.

4 SOCIALLY INTELLIGENT AGENTS

It is natural to model an existing population of agents as a set of compatible, but otherwise heterogeneous agents. We therefore introduce the more general idea of a socially intelligent *class* of agents that are compatible with any other member of their class:

Definition 4.1. A set C of agents forms am (adversarially or stochastically) *socially intelligent class* of agents w.r.t. Θ if, for some $\delta, \epsilon, T > 0$, each agent $\pi \in C$ is (adversarially or stochastically) (δ, ϵ, T)-consistent for all $\theta \in \Theta$, and any two agents $\pi, \pi' \in C$ are (δ, ϵ, T)-compatible over Θ. An individual agent π is called *socially intelligent* if it forms a socially intelligent class $\{\pi\}$ with itself.

For this notion of social intelligence to be meaningful, it must be possible to construct agents that satisfy the SI criteria. For both definitions of consistency, we will show that agents using a specific *fallback strategy* satisfy these criteria. For type space Θ, we first define a function $s(\theta) \in \mathcal{P}(G(\theta))$ that maps from each type $\theta \in \Theta$ to a PONE strategy profile under that type. We can think of $s(\theta)$ as a "convention" the agent or agents have settled upon for the game $G(\theta)$. Given a consistent agent $\bar{\pi}$, the fallback strategy plays $s_i(\theta)$ at every stage so long as its partner plays the corresponding strategy $s_{-i}(\theta)$. If its partner eventually fails to play $s_{-i}(\theta)$, the fallback strategy switches to $\bar{\pi}_i$ for all subsequent stages.

If $s_{-i}(\theta)$ is a mixed strategy, directly testing for deviation from $s_{-i}(\theta)$ is not straightforward. Instead, the fallback strategy examines the regret the agent has incurred so far, and switches if this exceeds a time-dependent threshold. As $s(\theta)$ is a Nash equilibrium of $G(\theta)$, we would expect each agent to have small regret when both play according to $s(\theta)$. Specifically, we have:

Lemma 4.2. *For any $\delta, T > 0$, if both players follow strategy $s(\theta)$ at each stage, then then with probability at least $1 - \delta$ we have*

$$\bar{R}_i^{ext}(h_t; \theta) \leq \sqrt{2T \ln \frac{2}{\delta}} \tag{10}$$

for all $t \leq T$ and $i \in \{1, 2\}$, and w.p. at lest $1 - \delta$ we have

$$R_i^{ext}(h_t; \theta) \leq 2\sqrt{2T \ln \frac{4}{\delta}} \tag{11}$$

for all $t \leq T$ $i \in \{1, 2\}$.

This follows from an application of the Azuma-Hoeffding inequality (shown in Appendix B). Combined with Lemma 3.3 this also provides a bound on the stochastic regret as well. For both definitions of consistency, we will use Lemma 4.2 to show that the fallback strategy defined by $s(\theta)$ is compatible with itself.

4.1 Stochastic Social Intelligence

We first derive a fallback strategy for the case of stochastic regret, which will serve as a template for the adversarial case. By definition, fictitious play has stochastic regret of zero. Therefore, the strategy π^{fp} which implements fictitious play (with uniform tie-breaking) for each player is (δ, ϵ, T)-stochastically consistent for any $\delta, \epsilon, T > 0$. We define the stochastic fallback strategy $\pi_i^{T,\epsilon}$ for player i as follows:

(1) While $R_i^{\text{sto}}(h_{t-1}; \theta) < \epsilon T - 1$, play $s_i(\theta)$.
(2) If $R_i^{\text{sto}}(h_{t-1}; \theta) \geq \epsilon T - 1$, follow π_i^{fp} for all subsequent stages.

Theorem 4.3. *For any $\delta, T > 0$, let $\epsilon_0 \geq 2\sqrt{\frac{2}{T} \ln \frac{4}{\delta}}$, and let $\epsilon = \epsilon_0 + \frac{1}{T}$. Then the stochastic fallback agent $\pi^{T,\epsilon}$ is stochastically (δ, ϵ, T)-socially intelligent.*

Note that $\pi^{T,\epsilon}$ will only deviate if $R_i^{\text{sto}}(h_{t-1}; \theta) \geq \epsilon_0 T$ for some $t \leq T$. By Lemmas 3.3 and 4.2, we have that the probability of this happening for either player is at most δ, and so $\pi^{T,\epsilon}$ is (δ, ϵ_0, T)-compatible. As π_i^{fp} will incur no stochastic regret, and since the maximum regret incurred in a single stage is 1, the maximum possible stochastic regret incurred by $\pi^{T,\epsilon}$ will be ϵT surely. Therefore $\pi^{T,\epsilon}$ is (δ, ϵ, T)-stochastically consistent. Since $\epsilon > \epsilon_0$, $\pi^{T,\epsilon}$ is also stochastically (δ, ϵ, T)-SI.

4.2 Adversarial Social Intelligence

The case of adversarial regret is somewhat more complex. Here we base our fallback strategy on the *multiplicative weights* [5] update rule, defined as:

$$s^i_{\text{mw},k}(h_t; \theta) = s^i_{\text{mw},k}(h_{t-1}; \theta) \exp\left(-\eta G_i(k, a^{-i}_{t-1}(h))\right) \tag{12}$$

for $k \in N$, where $s^i_{\text{mw}}(h_0; \theta)$ is the uniform strategy. Define $\pi^{\text{mw},T}$ as the agent that plays $s^i_{\text{mw}}(h_t; \theta)$ with learning rate $\eta = \sqrt{8 \ln(N/T)}$. The expected external regret of $\pi^{\text{mw},T}$ is bounded as

$$\bar{R}_i^{\text{ext}}(h_T; \theta) \leq \sqrt{\frac{T}{2} \ln N} \tag{13}$$

surely, by [1, Theorem 2.2]. Similar to the stochastic case, we then define the adversarial fallback strategy $\pi^{T,\epsilon}$ as follows:

(1) While $\bar{R}_i^{\text{ext}}(h_t; \theta) \leq \epsilon T - \sqrt{\frac{T}{2} \ln N} - 1$, play $s_i(\theta)$.

(2) Otherwise, switch to $\pi^{\text{mw},T}$ for all subsequent stages.

Theorem 4.4. *For any $\delta, T > 0$, let $\epsilon_0 \geq \sqrt{\frac{2}{T}\ln\frac{2}{\delta}}$, and let $\epsilon_1 = \epsilon_0 + \sqrt{\frac{1}{2T}\ln N} + \frac{1}{T}$. Then for $\epsilon = \epsilon_1 + \sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$, the adversarial fallback agent $\pi'^{,T,\epsilon_1}$ is stochastically (δ, ϵ, T)-socially intelligent.*

The proof of is similar to that for Theorem 4.3. By the definition of ϵ_1, $\pi'^{,T,\epsilon_1}$ will only deviate when playing with itself if at some point $t \leq T$ one player incurs an expected external regret of at least ϵ_0, and by Lemma 4.2 that will occur with probability at most δ. Therefore, $\pi'^{,T,\epsilon_1}$ is (δ, ϵ_0, T)-compatible. We also have that the total expected external regret of the MW agent $\pi^{\text{mw},T}$ is at most $\sqrt{(T/2)\ln N}$. This means that if $\pi'^{,T,\epsilon_1}$ switches at stage t, then the maximum possible expected external regret incurred by $\pi'^{,T,\epsilon_1}$ will be less than $\bar{R}_i^{\text{ext}}(h_t;\theta) + \sqrt{\frac{T}{2}\ln N}$. Since $\pi^{\text{mw},T}$ will always switch just before this point is reached, its total expected regret will be less than ϵ_1 surely, and will be less than ϵ w.p. $1-\delta$. As $\epsilon \geq \epsilon_0$, we have that the adversarial fallback strategy $\pi'^{,T,\epsilon_1}$ is adversarially (δ, ϵ, T)-socially intelligent.

5 DISCUSSION

While we have described a specific approach to designing socially intelligent agents, there are likely many other ways these criteria could be satisfied. Even restricting ourselves to the fallback strategies considered here, different socially intelligent classes described by different mappings $s(\theta)$ would yield very different behaviors. A critical theoretical and practical question then is whether we could design a single agent capable of learning to cooperate with any socially intelligent agent. Under our current definition of social intelligence, this reduces to the problem of learning to cooperate with any consistent agent [2]. To see this, note that for any socially intelligent class C, and any arbitrary joint action sequence $\sigma \in \{N \times N\}^k$, we could construct another class C' that initially play the "secret code" sequence σ_i, and immediately fall back to some arbitrary consistent strategy if the other player fails to do so. This then raises the related question of how robust a socially intelligent class of agents can be to stochastic or adversarial perturbations of actions taken within an interaction. It may be possible to establish lower bounds on the probability that cooperation between consistent agents will fail for a given noise distribution.

6 RELATED WORK

Our model is closely related to the previous targeted learning model [12, 13], which defines similar compatibility and consistency criteria. The main difference is that targeted learning only requires consistency against a specific target class of partners, which generally would not include the agent itself, or other adaptive agents. We also require that cooperation and consistent learning occur over a fixed time horizon T, rather than asymptotically. These differences mean that a hypothetical "universally cooperative" agent might be able to leverage the consistency of an SI agent to achieve cooperation without a prearranged convention.

This work is partly motivated by the practical challenge of using reinforcement learning to train agents that are able to cooperate with previously unseen partners, a problem sometimes described as *ad hoc teamwork* or *zero-shot coordination*. A key challenge in using RL for such scenarios is the need to construct populations of training partners (generally trained with RL themselves) that capture the range of cooperative behaviors in the target task [3, 16]. At present, construction of these populations is guided by heuristics that encourage diversity in the strategy space [2, 4, 10], but do not capture the ability of other agents to adapt to the behavior of others. By enforcing such a consistency requirement as our SI model does, we would hope to create more realistic training partners for cooperative multiagent RL.

[2]This is known to be impossible in general (see [9]).

7 CONCLUSIONS

This work has presented a novel framework for understanding the behavior of rational agents in multiagent scenarios. We have shown that it is possible to construct classes of consistent learning agents that are also able to reliably cooperate with one another. Our social intelligence model raises several important theoretical questions that could be explored in future work. These include determining whether we can design a single agent that can learn to cooperate with any socially intelligent partner, and providing lower bounds on how robust cooperation can be to noisy interactions. Future work could also consider practical realizations of the social intelligence model for multiagent reinforcement learning, training teams of adaptive agents that (approximately) satisfy the SI criteria.

ACKNOWLEDGMENTS

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022. This work was also supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI; grants 319264, 313195, 305780, 292334, 328400, 28400) and the Finnish Science Foundation for Technology and Economics (KAUTE).

REFERENCES

[1] Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge university press.
[2] Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. 2023. Generating Diverse Cooperative Agents by Learning Incompatible Policies. In *The Eleventh International Conference on Learning Representations*.
[3] Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob Foerster. 2021. K-level Reasoning for Zero-Shot Coordination in Hanabi. *Advances in Neural Information Processing Systems* 34 (2021), 8215–8228.
[4] Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J Wu, and Jakob Nicolaus Foerster. 2023. Adversarial Diversity in Hanabi. In *The Eleventh International Conference on Learning Representations*.
[5] Yoav Freund and Robert E Schapire. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* 29, 1-2 (1999), 79–103.
[6] James Hannan. 1957. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games* 3, 2 (1957), 97–140.
[7] Wassily Hoeffding. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* 58, 301 (1963), 13–30.
[8] John F. Kihlstrom and Nancy Cantor. 2000. *Social Intelligence*. Cambridge University Press, 359–379. https://doi.org/10.1017/CBO9780511807947.017
[9] Robert Loftin and Frans A Oliehoek. 2022. On the Impossibility of Learning to Cooperate with Adaptive Partner Strategies in Repeated Games. In *International Conference on Machine Learning*. PMLR, 14197–14209.
[10] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. 2021. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*. PMLR, 7204–7213.

[11] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.
[12] Rob Powers and Yoav Shoham. 2004. New Criteria and a New Algorithm for Learning in Multi-Agent Systems. *Advances in Neural Information Processing Systems* 17 (2004).
[13] Rob Powers and Yoav Shoham. 2005. Learning against opponents with bounded memory.. In *The Nineteenth International Joint Conference on Artificial Intelligence*. 817–822.
[14] Julia Robinson. 1951. An iterative method of solving a game. *Annals of mathematics* 54, 2 (1951), 296–301.
[15] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.
[16] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in Neural Information Processing Systems* 34 (2021).

APPENDICES

A PROOF OF LEMMA 3.3

We define $P_i^a(h;\theta)$ as

$$P_i^a(h;\theta) = \sum_{t=1}^{|h|} G_i(a, a_t^{-i}(h);\theta), \tag{14}$$

and

$$P_i^{\text{ext}}(h;\theta) = \max_{a\in N} P_i^a(h;\theta). \tag{15}$$

We also define $P_i^{\text{sto}}(h;\theta)$ as

$$P_i^{\text{sto}}(h;\theta) = \sum_{t=1}^{|h|} G_i(s^i_{\text{fp},t}(h;\theta), a_t^{-i}(h);\theta) \tag{16}$$

We prove by induction on T that $P_i^{\text{ext}}(h_T;\theta) \geq P_i^{\text{sto}}(h;\theta)$. This is trivially true for $T = 1$, as $s^i_{\text{fp},1}(h_T;\theta)$ is simply the uniform distribution over all N actions. Define the regret of fictitious play w.r.t. action $a \in N$ as

$$R_i^a(h;\theta) = P_i^a(h;\theta) - P_i^{\text{sto}}(h;\theta) \tag{17}$$

and the external regret of fictitious play $R_i(h;\theta)$ as

$$R_i(h;\theta) = P_i^{\text{ext}}(h;\theta) - P_i^{\text{sto}}(h;\theta). \tag{18}$$

We also define the set $A_i(h,\theta)$ as

$$A_i(h;\theta) = \arg\max_{a\in N} R_i^a(h;\theta). \tag{19}$$

We observe that

$$R_i(h_{T+1};\theta) = \max_{a\in N} \left\{ G_i(a, a_{T+1}^{-i}(h_{T+1});\theta) \right. \tag{20}$$

$$\left. -G_i(s^i_{\text{fp},T+1}(h_{T+1};\theta), a_T^{-i}(h);\theta) + R_i^a(h_T;\theta) \right\} \tag{21}$$

By the definition of fictitious play, we have that for any action $a \in A_i(h_T;\theta)$, the probability of a under $(s^i_{\text{fp},T+1}(h_{T+1};\theta)$ is > 0. Assuming that $R_i(h_T;\theta) \geq 0$, this means that for all $a \in A_i(h;\theta)$, $R_i^a(h_T;\theta) \geq 0$, and there exists $a' A_i(h_T;\theta)$ such that

$$G_i(a', a_{T+1}^{-i}(h_{T+1});\theta) - G_i(s^i_{\text{fp},T+1}(h_{T+1};\theta), a_T^{-i}(h);\theta) \geq 0. \tag{22}$$

This in implies that $R_i^{a'}(h_{T+1};\theta) \geq 0$, which in turn implies that $R_i(h_{T+1};\theta) \geq 0$. This means that the payoff of the best action in hindsight is always at least as large as the accumulated payoff of fictitious play, and proves Lemma 3.3. □

B PROOF OF LEMMA 4.2

Here the type θ will be implicit. For $i \in \{1,2\}$, we define V_t^i as

$$V_t^i = G_i(s_t^i, s_t^{-i}) - G_i(s_t^i, a_t^{-i}) \tag{23}$$

We can see that $\mathrm{E}[V_t^i | h_{t-1}] = 0$. We can then have that

$$\bar{R}_t^{\text{ext}} = \max_{a\in N} \sum_{r=1}^{t} \left\{ G_i(a, s_r^{-i}) - G_i(s_r^i, a_r^{-i}) \right\} \tag{24}$$

$$= \max_{a\in N} \sum_{r=1}^{t} \left\{ G_i(a, s_r^{-i}) - G_i(s_r^i, s_r^{-i}) + G_i(s_r^i, s_r^{-i}) - G_i(s_r^i, a_r^{-i}) \right\} \tag{25}$$

$$= \sum_{r=1}^{t} \left\{ G_i(s_r^i, s_r^{-i}) - G_i(s_r^i, a_r^{-i}) \right\} = \sum_{r=1}^{t} V_r^i \tag{26}$$

$$\leq \sqrt{\frac{2}{T} \ln \frac{1}{\delta}} \tag{27}$$

with probability $1 - \delta$ for all $t \leq T$ simultaneously.

This follows from the fact that $|V_t^i| \in [0,1]$ and the "maximal" Azuma-Hoeffding inequality [7]. The second equality follows from the fact that $\langle s_t^i, s_t^{-i} \rangle = s(\theta)$ is a Nash equilibrium. The first bound of Lemma 4.2 follows from a union bound over the probability for both players, while the second bound combines this with Equation 7. □