On the Impossibility of Learning to Cooperate with Adaptive Partner Strategies in Repeated Games

Robert Loftin [1]   **Frans A. Oliehoek** [1]

Abstract

Learning to cooperate with other agents is challenging when those agents also possess the ability to adapt to our own behavior. Practical and theoretical approaches to learning in cooperative settings typically assume that other agents' behaviors are stationary, or else make very specific assumptions about other agents' learning processes. The goal of this work is to understand whether we can reliably learn to cooperate with other agents without such restrictive assumptions, which are unlikely to hold in real-world applications. Our main contribution is a set of impossibility results, which show that no learning algorithm can reliably learn to cooperate with all possible adaptive partners in a repeated matrix game, even if that partner is guaranteed to cooperate with *some* stationary strategy. Motivated by these results, we then discuss potential alternative assumptions which capture the idea that an adaptive partner will only adapt *rationally* to our behavior.

1. Introduction

Artificial agents deployed in real-world settings must be able to reliably interact and cooperate with humans (as well as other artificial agents). While machine learning can allow such agents to adapt to others' behavior, learning may fail when other agents are equally capable of adapting to our agent's own behavior. In this work, we consider whether an agent can reliably learn to cooperate with other adaptive agents when the strategies of those agents are unknown, and the minimal assumptions about others' strategies needed to ensure such cooperation. We show that, even when another agent's strategy is known be "cooperative" in a reasonable sense, there is no algorithm which is guaranteed to learn

to cooperate without additional assumptions. At the same time, we argue that the assumptions under which existing approaches can reliably learn to cooperate (de Farias & Megiddo, 2003; Even-Dar et al., 2005) are too restrictive for many realistic settings, and suggest alternative constraints that may be more easily satisfied in practice.

We focus on the case where two agents, who we will refer to as *Alice* and *Bob*, must work together to achieve some common goal. Our challenge is to choose a "good" strategy for Alice, without knowledge of Bob's strategy. A common way to approach such problems is to assume that Bob will play his half of some optimal joint strategy, and so have Alice play her half of this strategy as well. This *objective* approach may fail, however, when the task admits multiple, incompatible solutions, and Alice has no way to know in advance which solution Bob has chosen. For this reason, some works (Mescheder et al., 2011; Poland & Hutter, 2006) have advocated framing the problem as one of single-agent learning, where Alice's strategy adapts to Bob's observed behavior. This *subjective* approach becomes problematic, however, when Bob's strategy can also adapt to Alice's behavior. The key theoretical result of this work in Theorem 4.7, which shows that if Bob's strategy is allowed to be adaptive, then there is *no strategy* Alice can choose that will reliably cooperate with every strategy Bob could follow.

More specifically, we show that for some infinitely repeated matrix games, there is no adaptive strategy Alice can follow that will always achieve the same average payoff as the best strategy she could have followed given knowledge of Bob's strategy. The key idea is that as the game progresses, Bob may continuously increase the time over which Alice must commit to a single strategy before his behavior converges, such that Alice is never able to gain information about the long-term payoffs of potential strategies. This stands in contrast to existing results, which guarantee near-optimal payoff under the assumption that there exists some bounded time horizon over which Alice can evaluate possible strategies against Bob's behavior. We will see that this assumption is violated even by seemingly reasonable learning strategies, such as fictitious play (Robinson, 1951) (see Section 4.2), that Bob could be expected to follow.

In Section 4.2, we introduce the concept of an *open-ended*

---
[1]Department of Intelligent Systems, Delft University of Technology, Delft, South-Holland, The Netherlands. Correspondence to: Robert Loftin <R.T.Loftin@tudelft.nl>.

strategy, which weakens the assumption of a bounded evaluation horizon, and formalize the problem of no-regret learning against an adaptive, open-ended partner. In Sections 5.1 and 5.2, we prove results for two restricted classes of adaptive strategies Alice could follow, while in Section 5.3 we prove our main impossibility result. Finally, in Section 6 we discuss potential alternatives to bounded-horizon assumptions, based on the idea that adaptive elements of Bob's strategy must be *rational* with respect to the shared goal.

2. Preliminaries

We restrict ourselves to the case where Alice must learn to cooperate with an unknown partner (Bob) in an infinitely repeated matrix game. Let A and B be the *finite* action sets for Alice and Bob respectively. With a slight abuse of notation, we will denote by G both the $|A| \times |B|$ payoff matrix of the stage game, and the infinitely repeated game defined by G. Let $G(a, b) \in [0, 1]$ be the payoff Alice receives for taking action a when Bob takes action b. Let a_n and b_n be the actions taken by Alice and Bob in stage n, such that $G(a_n, b_n)$ is the payoff received by Alice in stage n. We assume that Alice knows the values of the payoff matrix G, and only needs to learn about Bob's strategy. Our main results assume that Bob's strategy can be chosen arbitrarily, and so we will not explicitly consider Bob's payoffs.

Let $\mathcal{H}_n = (A \times B)^n$ be the set of histories of play of length n, and let $\mathcal{H} = \bigcup_{n=0}^{\infty} \mathcal{H}_n$ be the set of all *finite* histories. Let $|h|$ be the length of h, and let $a(h)$ be Alice's most recent action in h. We overload h to also denote the *event* that history $h \in \mathcal{H}$ occurs during play. We represent the strategies followed by Alice and Bob as *behavioral strategies* (Shoham & Leyton-Brown, 2008), that is, as functions $\pi : \mathcal{H} \mapsto \Delta(A)$ and $\phi : \mathcal{H} \mapsto \Delta(B)$, where $\Delta(X)$ is the set of probability distributions over X. When Alice implements a learning algorithm $\mathcal{A}$, we let $\mathcal{A}_G$ denote the strategy she would follow when playing a specific game G. We let $\mathcal{A}_{G,E}$ denote Alice's strategy when $\mathcal{A}$ additionally depends on a set E of *expert* strategies. Finally, we assume Alice and Bob's respective strategies are defined for all histories $h \in \mathcal{H}$, including histories that occur with probability zero.

We evaluate strategies based on the infinite-time limit of Alice's expected per-stage payoff. When Alice follows strategy π, while Bob follows strategy ϕ, for history $h \in \mathcal{H}$ we denote Alice's conditional expected average payoff as

$$V_G(\pi, \phi | h) = \liminf_{s \to \infty} \mathrm{E}_{\pi, \phi} \left[\frac{1}{s} \sum_{n=|h|+1}^{|h|+s} G(a_n, b_n) : h \right]. \quad (1)$$

For convenience, we let $V_G(\pi, \phi) = V_G(\pi, \phi | \emptyset)$, where $\emptyset$ is the initial (empty) history. The limit inferior is used because the limit of the average payoff may not always exist.

We define all probabilities and expectations with respect to

a common probability space $(\Omega, \mathcal{F})$, where $\Omega = (A \times B)^*$, the set of all *infinite* paths of play. The history h_n up to stage n is a random variable w.r.t. $(\Omega, \mathcal{F})$, that is, $h_n : \Omega \mapsto \mathcal{H}_n$. For all $n \in \mathbb{N}$, we let $\mathcal{F}_n \subset \mathcal{F}$ be the smallest σ-algebra containing all sets $\{\omega \in \Omega : h_n(\omega) = h\}$ for all $h \in \mathcal{H}_n$. $\mathcal{F}_0, \mathcal{F}_1, \dots$ is then the natural filtration (Kallenberg, 1997) of the stochastic process $h_0, h_1, \dots$. We let $\mathrm{Pr}_{\pi,\phi}\{\xi \mid h\}$, for $\xi \in \mathcal{F}$, denote the probability measure induced by π and ϕ on $(\Omega, \mathcal{F})$, conditioned on $h \in \mathcal{H}$. We will omit one or both strategies when they are clear from the context.

3. Related Work

Existing frameworks for learning to cooperate with other agents make a wide variety of prior assumptions about those agents' strategies. Unsurprisingly, these frameworks are forced to make trade-offs between computational and sample efficiency, and robustness to different partners. At one end of the spectrum lie recent, practical approaches based on deep reinforcement learning, which assume that other agents will act (near-)optimally for the current task. Under this assumption, the most straightforward approach is to choose a strategy that approximates one component of some optimal joint solution (Foerster et al., 2019; Hu & Foerster, 2020) to this task. As this may fail when there are multiple, incompatible solutions, some approaches select strategies that perform well against synthetic distributions over near-optimal strategies (Canaan et al., 2019; Strouse et al., 2021; Cui et al., 2021), or against learned models of real agents (Carroll et al., 2019). Other techniques attempt to address specific sources of miscoordination, such as task symmetries (Hu et al., 2020), asymmetric roles (Genter et al., 2011), and arbitrary conventions (Hu et al., 2021).

More general (though less scalable) is the Bayesian approach, modelling the interaction from a single agent's perspective as a POMDP, where the strategies of other agents are components of the latent state (Macindoe et al., 2012; Panella & Gmytrasiewicz, 2016). In the rational learning framework (Kalai & Lehrer, 1993), each agent follows a Bayes-optimal strategy given some prior over the other agent's strategy. The related *targeted learning* criteria (Powers & Shoham, 2004; 2005) simply require that an agent eventually learn to play optimally against any member of some target class of strategies. A major limitation of these approaches is the difficulty in specifying priors (or target classes) for all agents that admit other agents' own learning strategies. For example, in rational learning, Alice's prior must assign positive probability to Bob's true belief over Alice's strategy, which in turn depends on Alice's prior itself, leading to a circular dependence that is difficult to satisfy.

At the other end of the spectrum are approaches which treat other agents as simply being part of the (non-stationary) environment itself, and apply single-agent reinforcement

learning within the interaction. These include methods that model the problem as one of learning in a non-episodic POMDP (Mescheder et al., 2011). Learning in POMDPs without resets is possible by evaluating potential strategies over an ever-expanding time horizon (Even-Dar et al., 2005). Similarly, the ϵ-greedy *strategic experts algorithm* (de Farias & Megiddo, 2003) and the exponentially weighted *follow or explore* algorithm (Poland & Hutter, 2005) evaluate expert strategies over expanding horizons. As we discuss in Section 4.1, all these approaches assume that even when other agents are adaptive, there is still some fixed, or slowly growing time horizon over which strategies can be evaluated. As with targeted and rational learning, the challenge is that these assumptions are generally not satisfied when other agent's implement equally capable learning algorithms.

4. Adaptive Regret

A key conceptual challenge in multi-agent learning is determining what constitutes a "good" strategy for one agent to follow given our assumptions about the strategies of other agents. In this work, we implicitly assume that Bob's strategy will be cooperative, and that there is some corresponding strategy Alice could follow which would yield satisfactory, if not optimal, average payoff. Therefore, a natural choice for Alice is to follow a *no-regret* adaptive strategy that is guaranteed to eventually do as well in expectation as the best fixed strategy Alice could have chosen. As we are primarily concerned with the case where Bob's strategy is also adaptive, we need a notion of regret that captures Bob's dependence on Alice's previous actions.

Definition 4.1. Let π and ϕ be the strategies followed by Alice and Bob respectively in a repeated game G, and let E be a set of alternative *expert* strategies for Alice. Alice's *adaptive regret*[1] with respect to the set E is then:

$$\mathcal{R}_{G,E}^{\text{adapt}}(\pi, \phi) = \max_{e \in E} V_G(e, \phi) - V_G(\pi, \phi), \quad (2)$$

The adaptive regret compares the expected average payoff Alice receives when following π against the expected average payoff she could have received had she followed the best expert strategy $e \in E$ from the beginning of the game repeated game. This is in contrast to the more common *external regret* (Hannan, 1957), defined as

$$\mathcal{R}_{G,E}^{\text{ext},s} = \max_{e \in E} \mathbb{E}_{\pi,\phi} \left[\frac{1}{s} \sum_{n=1}^{s} G(e, b_n) - G(a_n, b_n) \right], \quad (3)$$

which assumes that Bob's actions would remain unchanged had Alice always followed strategy e rather than π. Note that, when each $e \in E$ simply takes some fixed action at every stage, we overload e to denote this action as well.

As Bob's strategy is initially unknown, our goal is to choose a strategy π for Alice that minimizes her worst-case adaptive regret over all strategies ϕ that Bob could follow. For some sufficiently constrained (e.g., finite) set of experts E, we might hope to find an *adaptive* strategy π^* which guarantees that Alice suffers no adaptive regret, that is:

$$\sup_{\phi} \mathcal{R}_{G,E}^{\text{adapt}}(\pi^*, \phi) \leq 0. \quad (4)$$

The following example, however, shows that Bob's ability to simultaneously adapt to Alice's behavior as well means that such a strategy π^* will not always exist.

Example 1. Consider the game G defined by the matrix

	b^1	b^2	b^3
a^1	2	0	1
a^2	0	2	1

and Bob's *grim trigger* strategies ϕ_1 and ϕ_2. ϕ_1 plays b^1 so long as the Alice plays a^1, and switches to b^3 for all future stages if she ever plays a^2. ϕ_2 does the opposite, playing b^2 so long as Alice plays a^2, and otherwise switching to b^3. Let $E = \{a^1, a^2\}$, where a^1 and a^2 are overloaded to denote the fixed-action strategies that play actions a^1 and a^2. Observe that $V_G(a^1, \phi_1) = 2$ and $V_G(a^2, \phi_2) = 2$, because in both cases Bob never switches to action b^3. There is, however, no strategy π which simultaneously achieves $V_G(\pi, \phi_1) = 2$ and $V_G(\pi, \phi_2) = 2$. If Alice plays a^1 at any point, then $V_G(\pi, \phi_2) = 1$, but if she never plays a^1, then $V_G(\pi, \phi_1) = 1$. The best Alice can do is to select between strategies a^1 and a^2 with equal probability, such that $V_G(\pi, \phi_1) = V_G(\pi, \phi_2) = \frac{3}{2}$, and Alice always suffers an adaptive regret of $2 - \frac{3}{2} \geq 0$. [2]

While such grim-trigger strategies may seem unrealistic in a cooperative setting, it *is* reasonable to imagine that an agent might switch to a sub-optimal "safety" strategy if their partner initially fails to behave as they had expected. To address this issue from a theoretical perspective, we can either: 1) explicitly rule out such strategies, or 2) use a performance metric that does not penalize Alice for early mistakes, even when these mistakes have permanent consequences. In this section we focus on the former approach, leaving discussion of the later formulation to Section 5.4.

4.1. Flexible Strategies

As discussed in Section 3, there are several learning algorithms which Alice could implement to try to minimize her adaptive regret. The regret guarantees provided for these algorithms share a common assumption that there is some bounded time horizon over which each expert can

---

[1](Dekel et al., 2012) provide a related notion of *policy regret* over finite time horizons for deterministic ϕ.

[2]Similar grim-trigger examples were provided in (Poland & Hutter, 2005) and (Dekel et al., 2012).

be reliably evaluated. Concretely, the *strategic experts algorithm* (de Farias & Megiddo, 2003) can guarantee no adaptive regret under the assumption that Bob's strategy is *flexible* with respect to E in the game G.

Definition 4.2. Bob's strategy ϕ is *flexible* w.r.t. E in game G if there exist constants $\mu_e, \forall e \in E$, $r > \frac{1}{4}$ and c such that, for all histories $h \in \mathcal{H}$

$$E_{a,b \sim e, \phi} \left[\left| \frac{1}{s} \sum_{n=|h|+1}^{|h|+s} G(a_n, b_n) - \mu_e \right| : h \right] \leq cs^{-r} \quad (5)$$

for each expert $e \in E$.

Flexibility implies that no matter what actions Alice has taken in the past, there is some τ such that if Alice follows $e \in E$ for τ stages, e's average payoff over this period will be a good estimate of the limit of e's average payoff. Furthermore, flexibility means that this τ is fixed, and independent of the history h. We can see that the grim-trigger strategies from Example 1 are not flexible, as the value of μ_e would depend on whether the history h included the triggering condition. Similar assumptions of a bounded *evaluation horizon* have been used to provide no-regret guarantees for non-episodic POMDPs (Even-Dar et al., 2005) and against strategies that depend on only Alice's last k actions (Powers & Shoham, 2005; Dekel et al., 2012). The *follow or explore* algorithm (Poland & Hutter, 2005) does allow us to weaken this assumption slightly, guaranteeing no-regret when the evaluation horizon grows slower than $O(|h|^{\frac{1}{4}})$.

4.2. Open-Ended Strategies

The problem with flexibility (Definition 4.2) and related assumptions is that many learning algorithms fail to satisfy them, because their behavior depends on the full history of play. As an example, consider fictitious play (Robinson, 1951), which plays a best response to the empirical distribution over its partner's previous actions.

Proposition 4.3. *Fictitious play is not flexible in all fully cooperative repeated games.*

Proof. Let G be the repeated game from Example 1. If Bob's strategy ϕ^{fp} implements fictitious play, then from any history h, $V_G(a^2, \phi^{\text{fp}}|h) = 2$. For any positive c, r, let Alice first play action a^1 for $n = 2(2c)^{\frac{1}{r}}$ stages. If Alice plays a^2 for the next $\frac{n}{2}$ stages, she will receive an average payoff of 0 over this period, since Bob will not switch to b^2 until Alice plays a^2 for $n+1$ steps. The error in Alice's estimate of a^2's average long-term payoff will be 2. On the other hand, the error bound $c(\frac{n}{2})^{-r} = \frac{1}{2}$. Therefore, no constants c, r can satisfy Equation 5 for all h, and so ϕ^{fp} is not flexible. $\square$

Assuming that Bob's strategy is flexible therefore rules out the possibility that Bob is using fictitious play (at least with

unbounded memory), and indeed rules out many natural learning algorithms that Bob might implement. We therefore consider whether it is possible to relax the assumption of a fixed evaluation horizon, and still guarantee that Alice achieves optimal average payoff in the limit. Rather than assuming that Bob's strategy is flexible with respect a set of experts E, we make the weaker assumption that Bob's strategy is *open-ended* with respect to E.

Definition 4.4. Bob's strategy ϕ is *open-ended* w.r.t. experts E in game G if, $\forall e \in E$, there exists μ_e such that

$$V_G(e, \phi|h) = \mu_e \quad (6)$$

for all histories $h \in \mathcal{H}$.

The following propositions (proven in Appendices A and B) show that being open-ended is a strictly weaker condition than flexibility, and that it allows for learning strategies which are not always flexible.

Proposition 4.5. *In any repeated game G, if Bob's strategy is flexible with respect to a set of experts E, then it is also open-ended w.r.t. E.*

Proposition 4.6. *For all fully cooperative, finite-action repeated games G, fictitious play is open-ended w.r.t. the set of all fixed-action strategies.*

When Bob's strategy is open-ended, if at any point Alice commits to following a single expert $e \in E$ for all future stages, then she is guaranteed to receive an average payoff of μ_e in expectation. Our main result, however, shows that this fact alone does not allow for an algorithm which is guaranteed to do as well in expectation as the best expert in E against all possible open-ended strategies.

Theorem 4.7. *For any $N \geq 1$, there exists an $N \times N$ repeated game G, with a finite set of experts E, such that, for any learning algorithm $\mathcal{A}$, and any $\delta > 0$, we can construct a strategy ϕ such that*

$$\mathcal{R}_{G,E}^{\text{adapt}}(\mathcal{A}_{G,E}, \phi) \geq \frac{1}{2} \left[\frac{N-2}{N} - \delta \right]^2 \quad (7)$$

and such that ϕ is open-ended w.r.t. E in G.

Theorem 4.7, which we will prove in Section 5.3, means that in some games, no matter what strategy Alice follows, there will always be an open-ended strategy Bob can choose such that Alice's expected average payoff is less than that of the best expert. Furthermore, for a large enough game and small enough value of δ, Alice's regret can be made arbitrarily close to $\frac{1}{2}$ (recall that payoffs are bounded in $[0, 1]$). Here δ is a parameter of the strategy ϕ that can be chosen arbitrarily, and is an artifact of our proof technique.

5. Impossibility Results

To prove Theorem 4.7, it will be sufficient to show that no-regret learning against an open-ended strategy is impossible in an $N \times N$ coordination game G, where $G(a,b) = 1$ if $a = b$ and $G(a,b) = 0$ otherwise. The set E will consist of N fixed-action strategies, where again with a slight abuse of notation we denote by $e \in E$ the expert that in each stage takes action $e \in \{1, \ldots, N\}$. We will demonstrate that for any algorithm $\mathcal{A}$ that Alice could choose, it is always possible for Bob to choose a strategy that is open-ended, but nonetheless prevents Alice's strategy $\mathcal{A}_{G,E}$ from doing as well as the best expert $\bar{e}$ in E. Before proving Theorem 4.7, we will first prove that that no-regret learning is impossible for two special classes of algorithms. These are algorithms that are either *passive*, and eventually commit to following a fixed expert, or *active*, and never stop exploring.

5.1. Passive Algorithms

By definition, when Bob's strategy is open-ended, if Alice commits to a fixed expert e for all future stages, then the limit of her expected average payoff will be μ_e. Therefore, a potential approach to reliable learning against an open-ended partner is to <u>eventually commit to following the best expert so far for all future stages.</u> We will show, however, that such *passive* strategies can always suffer large regret against certain open-ended strategies.

Definition 5.1. We say that a learning algorithm $\mathcal{A}$ is *passive almost surely* if, for any game G and set of experts E, when Alice plays $\mathcal{A}_{G,E}$ against any strategy ϕ, we have that

$$\Pr\{\exists s, \forall t > s, a_t = a_s\} = 1 \tag{8}$$

If Alice's strategy $\mathcal{A}_{G,E}$ is passive a.s., then with probability 1 there will be some stage s after which Alice will follow the same expert in every stage. In this section, we will show that for any passive algorithm $\mathcal{A}$ Alice could implement, we can construct a strategy ϕ for the coordination game such that Alice will suffer non-zero regret for playing $\mathcal{A}_{G,E}$ against ϕ. Our main tool for dealing with passive algorithms will be Lemma 5.2, which captures the intuition that if $\mathcal{A}$ is passive a.s., then there must be some finite number of stages within which Alice will commit to always following a single expert with high probability.

Lemma 5.2. *Assume that, when Alice plays strategy $\mathcal{A}_{G,E}$ against Bob's strategy ϕ, we have*

$$\Pr\{\exists s, \forall t > s, a_t = a_s\} = 1 - \gamma \tag{9}$$

for some $\gamma \in [0, 1)$. Then for any $\delta > 0$, there exists τ s.t.

$$\Pr\{\exists s > \tau, a_s \neq a_\tau\} \leq \gamma + \delta \tag{10}$$

The proof of Lemma 5.2 is provided in Appendix C. Note that Lemma 5.2 is more general than we need for the passive case, and in Section 5.3 we will use it to prove bounds that apply to any algorithm. For passive algorithms, we can use Lemma 5.2 with $\gamma = 0$ to prove the following:

Theorem 5.3. *Assume that algorithm $\mathcal{A}$ is passive a.s., then for any $N \geq 1$, there exists an $N \times N$ game G and set of experts E such that, for any $\delta > 0$, we can construct a strategy ϕ such that*

$$\mathcal{R}_{G,E}^{\mathrm{adapt}}(\mathcal{A}_{G,E}, \phi) \geq \frac{N-2}{N} - \delta, \tag{11}$$

with ϕ open-ended w.r.t. E in G.

Proof. Let G be the $N \times N$ coordination game, and let E be the set of fixed-action experts. Let ϕ^{uni} be the uniform random strategy. We fix an arbitrary value of $\delta > 0$. Let C_s denote the event that Alice's actions have converged by stage s. Since $\mathcal{A}$ is passive a.s., by Lemma 5.2 there exists τ such that when Alice plays $\mathcal{A}_{G,E}$, we have $\Pr_{\phi^{\mathrm{uni}}}\{\neg C_\tau\} \leq \gamma + \delta$ (with $\gamma = 0$ since $\mathcal{A}$ is passive a.s.). Henceforth, let τ denote the value satisfying Equation 10 for our choice of δ.

We next define the ==switching== strategies. For action $e \in [1, N]$, the switching strategy ϕ_e^τ is defined as

$$\phi_e^\tau(h) = \begin{cases} e & |h| \geq \tau \wedge a(h) = e \\ a \sim [1, N] & \text{otherwise} \end{cases} \tag{12}$$

The strategy ϕ_e^τ follows the uniform strategy for the first τ stages, and then continues to select actions randomly at each stage unless Alice took action e in the previous stage, in which case ϕ_e^τ takes action e. Note that for any e, ϕ_e^τ is open ended w.r.t. E, with $\mu_e = 1$, and $\mu_{e'} = \frac{1}{N}$. The latter is due to the fact that when Alice plays any action other than e, ϕ_e^τ will sample its next action uniformly at random.

We next define the *convergence probabilities* p_e as

$$p_e = \Pr_{\phi^{\mathrm{uni}}}\{a_\tau = e | C_\tau\}, \tag{13}$$

that is, p_e is the probability that the Alice converges to action e (when playing against ϕ^{uni}), given that her actions converge by stage τ. We then note that

$$\Pr_{\phi_e^\tau}\{C_\tau \wedge a_\tau \neq e\} = \Pr_{\phi^{\mathrm{uni}}}\{C_\tau \wedge a_\tau \neq e\}. \tag{14}$$

If Alice converges to $e' \neq e$ by stage τ, then for each subsequent stage $s > \tau$, Bob's action distributions will be the same under both strategies, that is, $\phi^{\mathrm{uni}}(h_s) = \phi_e^\tau(h_s)$. So all finite histories h will have equal probabilities under both measures. For the same reason

$$\lim_{s \to \infty} \mathrm{E}_{\phi_e^\tau}\left[\frac{1}{s}\sum_{n=1}^{s} G(a_n, b_n) : C_\tau \wedge a_\tau \neq e\right] = \frac{1}{N}. \tag{15}$$

We can then see that

$$\Pr_{\phi_{\tilde{e}}^{\tau}}\{\neg C_{\tau} \vee a_{\tau} = e\} = 1 - \Pr_{\phi_{\tilde{e}}^{\tau}}\{C_{\tau} \wedge a_{\tau} \neq e\} \qquad (16)$$

$$= 1 - \Pr_{\phi^{\mathrm{uni}}}\{C_{\tau} \wedge a_{\tau} \neq e\} \qquad (17)$$

$$= 1 - \Pr_{\phi^{\mathrm{uni}}}\{a_{\tau} \neq e | C_{\tau}\} \Pr_{\phi^{\mathrm{uni}}}\{C_{\tau}\} \qquad (18)$$

$$= 1 - (1 - p_e)[1 - \Pr_{\phi^{\mathrm{uni}}}\{\neg C_{\tau}\}] \qquad (19)$$

$$\leq 1 - (1 - p_e)[1 - \gamma - \delta] \qquad (20)$$

$$\leq \gamma + \delta + p_e \qquad (21)$$

Define $\tilde{e} \in \arg\min_{e \in E} p_e$. Since $\sum_{e=1}^{N} p_e = 1$, we have that $p_{\tilde{e}} \leq \frac{1}{N}$. As the maximum average payoff is 1, we have

$$V_G(\mathcal{A}_{G,E}, \phi_{\tilde{e}}^{\tau}) \leq \gamma + \delta + p_{\tilde{e}} + \frac{1}{N}\Pr_{\phi_{\tilde{e}}^{\tau}}\{C_{\tau} \wedge a_{\tau} \neq \tilde{e}\} \quad (22)$$

$$\leq \gamma + \delta + p_{\tilde{e}} + \frac{1}{N} \qquad (23)$$

$$\leq \gamma + \delta + \frac{2}{N} \qquad (24)$$

Here, we have decomposed the expected average payoff into the case where Alice converges to a fixed action $a \neq \tilde{e}$, for which we showed that the payoff is at most $\frac{1}{N}$, and the case where she fails to do so (where we can only use the trivial bound of 1). Finally, we have that

$$\mathcal{R}_{G,E}^{\mathrm{adapt}}(\mathcal{A}_{G,E}, \phi_{\tilde{e}}^{\tau}) = \mu_{\tilde{e}} - V_G(\mathcal{A}_{G,E}, \phi_{\tilde{e}}^{\tau}) \qquad (25)$$

$$\geq 1 - \left[\gamma + \delta + \frac{2}{N}\right] \qquad (26)$$

$$= \frac{N-2}{N} - \gamma - \delta \qquad (27)$$

where setting $\gamma = 0$ gives the desired result. $\square$

5.2. Active Algorithms

As we saw in the previous section, a learning algorithm that eventually commits to a single expert risks missing opportunities for a larger payoff. In this section we show that the alternative, following an active learning strategy that never stops exploring, can fail in a different way.

Definition 5.4. We say that a learning algorithm $\mathcal{A}$ is *active almost surely* if, for any game G and set of experts E, when Alice plays $\mathcal{A}_{G,E}$ against any strategy ϕ, we have that

$$\Pr\{\forall s, \exists t > s, a_t \neq a_s\} = 1 \qquad (28)$$

If Alice's strategy $\mathcal{A}_{G,E}$ is active a.s., then for any stage s where the she follows expert e_s, there will be a stage $t > s$ where she follows a different expert $e_t \neq e_s$. An active strategy will never commit to a fixed expert indefinitely, and must eventually deviate from its current expert.

The issue is that if Bob is aware of Alice's strategy, he can predict at which stage Alice is likely to switch from her

current expert. Bob can then require that Alice continue to follow her current expert past this stage in order to learn anything about that expert's long-term payoffs. Again, let ϕ^{uni} be the uniformly random strategy. The following lemma shows that we can always predict, with high confidence, when Alice will deviate from her current expert.

Lemma 5.5. *Assume that, when Alice plays strategy $\mathcal{A}_{G,E}$ against Bob's strategy ϕ, we have*

$$\Pr\{\forall s, \exists t > s, a_t \neq a_s\} = \gamma, \qquad (29)$$

for some $\gamma > 0$. Let C denote the event that Alice's behavior eventually converges to some fixed action. For any history h with positive probability under $\mathcal{A}_{G,E}$ and ϕ, and any $\delta > 0$, there exists σ such that

$$\Pr\{\forall s \in [1, \sigma], a_{|h|} = a_{|h|+s} \mid h \wedge \neg C\} \leq \delta \qquad (30)$$

The proof of Lemma 5.5 is analogous to that of Lemma 5.2, and is provided in Appendix D. Using Lemma 5.5, we can now prove the following theorem:

Theorem 5.6. *Assume that algorithm $\mathcal{A}$ is active a.s., then for any $N \geq 1$, there exists an $N \times N$ game G and set of experts E such that, for any $\delta > 0$, we can construct a strategy ϕ such that*

$$\mathcal{R}_{G,E}^{\mathrm{adapt}}(\mathcal{A}_{G,E}, \phi) \geq \frac{N-1}{N} - \delta \qquad (31)$$

with ϕ open-ended w.r.t. E in G.

Before getting into the proof of Theorem 5.6, we briefly note that the theorem shows how the assumption of a finite evaluation horizon can be exploited. Imagine Alice is using the *strategic experts algorithm* (de Farias & Megiddo, 2003). Each time Alice selects an expert to evaluate, she also selects the time-horizon over which she will evaluate that expert. Because Bob has complete knowledge of Alice's strategy, no matter how large a horizon she selects, Bob can always ensure that Alice learns nothing about the final performance of the current expert over this horizon. Theorem 5.6 alone is sufficient to show that the algorithms presented in (de Farias & Megiddo, 2003), (Even-Dar et al., 2005), and (Poland & Hutter, 2005) can suffer positive adaptive regret against an open-ended partner, as they are each active almost surely.

Proof. From the Bob's perspective, we can divide the learning process into *intervals* i, such that a new interval starts every time Alice switches to following a different expert ($i = 0$ starts at the beginning of the game). Let h^i be the history up to the stage s_i at which the ith interval started. Assuming that the learner's behavior does not converge (the event $\neg C$), Lemma 5.2 means for each interval i, there must exist σ_i such that, conditioned on h^i, the probability that the

Alice will switch experts (and interval i will end) by stage $s_i + \sigma_i$ is greater that $1 - \delta$ (for any $\delta > 0$).

To show that learning is impossible, we define a strategy $\phi_{\mathcal{A}}$[3] that, in each interval i, follows the uniform strategy ϕ^{uni} for the first σ_i steps, and then follows expert e_i that the learner is following in i until the interval ends. Given a sequence $\{\delta_i\}_{i \in \mathbb{N}}$, we define each σ_i such that

$$\Pr\left\{\forall s \in [1, \sigma_i], a_{|h|} = a_{|h|+s}|h^i \wedge \neg C\right\} \leq \delta_i \quad (32)$$

Let $\iota(s)$ be the interval in which stage s occurs (note that both $\iota(s)$ and σ_i are random variables on $(\Omega, \mathcal{F})$). Clearly $\phi_{\mathcal{A}}$ is open ended, because if we commit to following any expert e from stage s, then after at most $\sigma_{\iota(s)}$ stages, $\phi_{\mathcal{A}}$ will switch to e as well, and so for all $e \in E$, $\mu_e = 1$.

Let R_i^s denote the total payoff received in interval i, truncated to stage s if s occurs during or before interval i (so that $R_i^s = 0$ if $s_i > s$). Let D_s be the event that, for all intervals i such that $s_i \leq s$, either $s_{i+1} \leq s_i + \sigma_i$, or $s < s_i + \sigma_i$. More intuitively, D_s is the event that, up to stage s, Alice has never followed a single expert for longer than Bob predicted she would. We then see that

$$V_G(\mathcal{A}_{G,E}, \phi_{\mathcal{A}}) = \liminf_{n \to \infty} \mathrm{E}\left[\frac{1}{n}\sum_{s=1}^{n} G(a_s, b_s)\right] \quad (33)$$

$$\leq (1 - \gamma) + \gamma \liminf_{n \to \infty} \mathrm{E}\left[\frac{1}{n}\sum_{s=1}^{n} G(a_s, b_s) : \neg C\right] \quad (34)$$

$$= (1 - \gamma) + \gamma \liminf_{n \to \infty} \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{\iota(n)} R_i^n : \neg C\right] \quad (35)$$

$$\leq (1 - \gamma) + \gamma \liminf_{n \to \infty} \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{\iota(n)} R_i^n : D_n \wedge \neg C\right] \quad (36)$$

$$+ \Pr\{\neg D_n \mid \neg C\}\mathrm{E}\left[\sum_{i=1}^{\iota(n)} R_i^n : \neg D_n \wedge \neg C\right] \quad (37)$$

$$\leq (1 - \gamma) + \gamma\frac{1}{N} + \gamma \liminf_{n \to \infty} \Pr\{\neg D_n \mid \neg C\} \quad (38)$$

In Equation 34 we are decomposing the expected payoff into the case where Alice's behavior does and does not converge. In Equation 37, we are decomposing the expectation into two cases: 1) the case D_n where there is no interval up to stage n in which Alice followed e_i for more that σ_i stages, and 2) the case $\neg D_n$ where there is such an interval. In the former case, Bob only ever plays the uniform base strategy ϕ^{uni}, yielding an expected average payoff of $\frac{1}{N}$. In the later case, Alice could observe and adapt to e^*, and so the maximum average payoff is 1. To bound

---

[3](Chang & Kaelbling, 2001) define a similar predictive exploiter strategy for hill-climbing learners.

$\liminf_{n \to \infty} \Pr\{\neg D_n \mid \neg C\}$, note that

$$\Pr\{\neg D_n | \neg C\} \leq \Pr\left\{\bigcup_{i=1}^{n}\{s_{i+1} > s_i + \sigma_i\}|\neg C\right\} \quad (39)$$

$$\leq \sum_{i=1}^{n} \Pr\{s_{i+1} > s_i + \sigma_i \mid \neg C\} \quad (40)$$

$$\leq \sum_{i=1}^{n} \delta_i \quad (41)$$

where the first inequality uses the fact that there can be at most n intervals observed in the first n stages, and so if we have not observed a violation of the predicted deviation horizon within the first n intervals, we will not have observed it within the first n stages. Setting $\delta_i = \frac{\delta}{2^{i+1}}$ gives us $\liminf_{n \to \infty} \Pr\{\neg D_n \mid \neg C\} \leq \delta$. The use of an exponentially decaying parameter δ_i means that Alice faces a receding evaluation horizon as learning progresses, and so with high probability $\mathcal{A}_{G,E}$ will only ever observe the uniform random strategy ϕ^{uni}. Finally, we have that

$$\mathcal{R}_{G,E}^{\text{adapt}}(\mathcal{A}_{G,E}, \phi_{\mathcal{A}}) = 1 - V_G(\mathcal{A}_{G,E}, \phi_{\mathcal{A}}) \quad (42)$$

$$\geq 1 - (1 - \gamma) - \gamma\left[\frac{1}{N} + \delta\right] \quad (43)$$

$$= \gamma\left[\frac{N-1}{N} - \delta\right] \quad (44)$$

where setting $\gamma = 1$ gives us the desired result. $\square$

5.3. Proof of Theorem 4.7

Most learning algorithms described in the literature fall into one of the two special classes described above. It is entirely possible, however, to define a learning algorithm that can be both active and passive, and so we must address this general case. Given a passive algorithm $\mathcal{A}$, and an active algorithm $\mathcal{A}'$, we can define a new algorithm which starts by randomly choosing whether to use $\mathcal{A}$ or $\mathcal{A}'$ for the rest of the interaction. Such an algorithm could either converge or explore forever, each with probability $\frac{1}{2}$.

To lower-bound the regret of such *mixed* learning strategies, we simply combine the partner strategies used for the active and passive cases. Note that Equations 27 and 44 both depend on the probability γ that Alice's behavior fails to converge when played against the uniform random strategy ϕ^{uni}. Therefore, for any algorithm $\mathcal{A}$, we can construct a strategy ϕ for the $N \times N$ coordination game, such that

$$\mathcal{R}_{G,E}^{\text{adapt}}(\mathcal{A}_{G,E}, \phi) \geq \max\left\{\frac{N-2}{N} - \gamma - \delta, \gamma\left[\frac{N-1}{N} - \delta\right]\right\} \quad (45)$$

We can simplify the result by replacing the second term with $\gamma\left[\frac{N-2}{N} - \delta\right]$. As the first term decreases linearly with γ,

and the second increases linearly with γ, we can minimize Equation 45 over possible values of γ by setting

$$\gamma = \left[\frac{N-2}{N} - \delta\right]\left(1 + \frac{N-2}{N} - \delta\right)^{-1} \quad (46)$$

the value at which both bounds are equal. Finally, plugging this value into Equation 45 we get

$$\mathcal{R}^{\text{adapt}}_{G,E}(\mathcal{A}_{G,E}, \phi) \geq \left[\frac{N-2}{N} - \delta\right]^2\left(1 + \frac{N-2}{N} - \delta\right)^{-1} \quad (47)$$

$$\geq \frac{1}{2}\left[\frac{N-2}{N} - \delta\right]^2 \quad (48)$$

which completes the proof. □

5.4. Open-Ended Regret

Rather than explicitly requiring that Bob will forgive any initial mistakes Alice might make, a more general solution is to redefine the regret itself so that it does not penalize Alice for making such mistakes. We can do this by defining Alice's regret for playing strategy π instead of strategy $e \in E$ in terms of the expected average payoff that e can guarantee starting from any history h.

Definition 5.7. In a repeated game G, if Bob follows strategy ϕ, Alice's *open-ended regret*[4] for playing strategy π w.r.t. a set of experts strategies E is:

$$\mathcal{R}^{\text{oe}}_{G,E}(\pi, \phi) = \max_{e \in E} \inf_{h \in \mathcal{H}} V_G(e, \phi|h) - V_G(\pi, \phi) \quad (49)$$

The first term is a lower bound on the expected payoff expert e can receive if an adversary were to select Alice's actions for a finite number of initial stages. If Bob's strategy ϕ can enter an unrecoverable state, the adversary will force him into this state before control is handed over to e. This value is useful because it tells us how well Alice could eventually do once she has identified a good response to ϕ, no matter how her initial exploratory behavior has affected Bob's behavior. Note that $\mathcal{R}^{\text{oe}}_{G,\phi}(\pi, \phi)$ can be negative even when $V_G(e, \phi) \geq V_G(\pi, \phi)$, as the open-ended regret accounts for worst-case histories that may never be reached. We now have the following corollary to Theorem 4.7:

Corollary 5.8. *For any $N \geq 1$, there exists an $N \times N$ game G and experts E such that, for any learning algorithm $\mathcal{A}$, and any $\delta > 0$, we can construct a strategy ϕ such that*

$$\mathcal{R}^{\text{oe}}_{G,E}(\mathcal{A}_{G,E}, \phi) \geq \frac{1}{2}\left[\frac{N-2}{N} - \delta\right]^2 \quad (50)$$

Proof. Consider the game G, Bob's strategy ϕ, and experts E which satisfy Equation 7 for $\mathcal{A}_{G,E}$ and some $\delta > 0$.

---

[4](Powers & Shoham, 2005) suggest a similar notion of regret as an alternative to finite history-dependence constraints.

ϕ is open-ended w.r.t. E, and so for all experts $e \in E$, $V_G(e, \phi|h) = \mu_e$ for all $h \in \mathcal{H}$. Therefore, for all $e \in E$

$$\inf_{h \in \mathcal{H}} V_G(e, \phi|h) = \mu_e \quad (51)$$

We then have that

$$\mathcal{R}^{\text{oe}}_{G,E}(\mathcal{A}_{G,E}, \phi) = \max_{e \in E} \mu_e - V_G(\mathcal{A}_G, \phi) \quad (52)$$

$$= \mathcal{R}^{\text{adapt}}_{G,E}(\mathcal{A}_{G,E}, \phi) \quad (53)$$

$$\geq \frac{1 - \delta}{2}\left[\frac{N-2}{N} - \delta\right]^2 \quad (54)$$

for any $N \geq 1$, which is the desired result. □

5.5. Computability

A potential limitation of this these results is that when $\mathcal{A}$ is neither passive nor active almost surely, the predictive strategy $\phi_{\mathcal{A}}$ may not be Turing computable, even when $\mathcal{A}_{G,E}$ itself can be implemented as a probabilistic Turing machine (Arora & Barak, 2009). If $\mathcal{A}$ is active a.s., we could implement $\phi_{\mathcal{A}}$ by iteratively computing the distribution over paths of play over the next $n, n+1, n+2, \ldots$ stages until the probability that $\mathcal{A}$ has switched experts is greater than $1 - \delta_{\iota(s)}$. When $\mathcal{A}$ can decide at random whether it is ever going to commit to an expert, from some histories the probability of deviation may never reach $1 - \delta_{\iota(s)}$, and the simulation process may never terminate. Therefore, Theorem 4.7 only applies if we admit non-computable strategies. It remains an open question whether there exists a computable learning algorithm $\mathcal{A}$ that can guarantee no adaptive regret against all computable, open-ended partner strategies.

6. Rational Adaptation

Our results highlight the fundamental limitation of approaching multi-agent learning as a non-stationary, single-agent learning problem. Even when we can assume that Bob's computational constraints mean that his strategy must admit a bounded evaluation horizon, Alice's own computational constraints may prevent her from implementing a no-regret algorithm over a long-enough timescale to exploit this horizon. We suggest instead constraining Bob's ability to adapt to Alice's behavior by requiring that any adaptive strategy be "rational" with respect to Bob's beliefs about Alice's strategy. Here we briefly discuss two potential models of such *rational adaptation*, and their implications for learning.

Computational Rationality As strategies which condition on Alice's behavior will generally be more difficult to implement than those that do not, we can imagine that Bob will only choose an adaptive strategy if he expects it to outperform any non-adaptive strategy. We can formalize this idea by assuming that Alice and Bob's strategies

are implemented by finite state machines (or Turing machines) $\mathcal{M}_\pi$ and $\mathcal{M}_\phi$, and replacing the repeated game with the *machine game* (Shoham & Leyton-Brown, 2008) G^M, where $G^M(\mathcal{M}_\pi, \mathcal{M}_\phi) = V_G(\pi, \phi)$. Let ρ be Bob's *belief* over Alice's possible strategies. We can say that ϕ is *computationally rational* w.r.t. ρ if for any $\phi' \neq \phi$, either $G^M(\rho, \mathcal{M}_\phi) \geq G^M(\rho, \mathcal{M}_{\phi'})$, or $G^M(\rho, \mathcal{M}_\phi) = G^M(\rho, \mathcal{M}_{\phi'})$ and $|\mathcal{M}_\phi| \leq |\mathcal{M}_{\phi'}|$, where

$$G^M(\rho, \mathcal{M}_\phi) = \mathrm{E}_{\mathcal{M}_\pi \sim \rho}[G(\mathcal{M}_\pi, \mathcal{M}_\phi)] \qquad (55)$$

and $|\mathcal{M}_\phi|$ is the size of the machine that implements ϕ. ϕ is not only a best-response to ρ, but is the least complex strategy out of the best-responses to ρ. Applied to Example 1, if $\rho(a^1) = 1$, then the grim-trigger strategy ϕ_1 is not computationally rational, as it is more complex than the strategy that always plays b^1, with greater payoff. If $\rho(a^1) = \rho(a^2) = \frac{1}{2}$ however, the strategy ϕ which adapts to Alice's is rational, even if it is more complex than any stationary strategy.

Cognitive Hierarchies   In place of, or in addition to computational constraints on Bob's strategy, we might also place constraints on Bob's beliefs about Alice's strategy. Under the *cognitive hierarchies* model (Camerer et al., 2004), a level-k rational strategy is a best-response to some belief over level-$0, \ldots, k$ strategies. Applied to repeated games, level-0 strategies could consist of all stationary strategies, with level-1 strategies adapting their partner's stationary strategy, and higher-level strategies reasoning about their partner's adaptive behavior. Rather than bounding the memory or evaluation horizon of Bob's strategy, we might instead bound the maximum level k of his strategy in the cognitive hierarchy. This assumption of *hierarchical rationality* admits complex, adaptive strategies, while potentially ruling out pathological strategies which are not a best-response to any set of strategies Alice could plausibly follow. We leave as a goal for future work the development of learning algorithms with regret guarantees under the assumption that Bob's strategy is hierarchically or computationally rational.

7. Conclusion

In this work we have shown that there is no algorithm which can be guaranteed to learn to cooperate with an arbitrary partner in a repeated game, even when this partner is always willing to forgive mistakes made by the learner. By providing a more complete understanding of the limits of the subjective approach to multi-agent learning, this work serves as a starting point for the exploration of new theoretical frameworks that specifically account for the intentions behind other agents' adaptive behavior. Ultimately, we hope this line of research will lead to novel algorithms that can reliably learn cooperate with humans and other artificial agents. Most importantly, we hope to motivate the development of learning algorithms that are robust to both adaptive and non-adaptive partners, without relying on unrealistic assumptions about these agents.

References

Arora, S. and Barak, B. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

Bartle, R. G. *The elements of integration and Lebesgue measure*. John Wiley & Sons, 2014.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Canaan, R., Togelius, J., Nealen, A., and Menzel, S. Diverse agents for ad-hoc cooperation in Hanabi. In *IEEE Conference on Games*, pp. 1–8, 2019.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chang, Y.-H. and Kaelbling, L. P. Playing is believing: The role of beliefs in multi-agent learning. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Cui, B., Hu, H., Pineda, L., and Foerster, J. K-level reasoning for zero-shot coordination in Hanabi. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

de Farias, D. P. and Megiddo, N. How to combine expert (or novice) advice when actions impact the environment. In *Advances in Neural Information Processing Systems*, volume 17, 2003.

Dekel, O., Tewari, A., and Arora, R. Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Conference on Machine Learning*, 2012.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Reinforcement learning in POMDPs without resets. In *The Nineteenth International Joint Conference on Artificial Intelligence*, pp. 690–695, 2005.

Foerster, J., Song, F., Hughes, E., Burch, N., Dunning, I., Whiteson, S., Botvinick, M., and Bowling, M. Bayesian action decoder for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 1942–1951, 2019.

Genter, K., Agmon, N., and Stone, P. Role-based ad hoc teamwork. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

Hannan, J. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–140, 1957.

Hu, H. and Foerster, J. N. Simplified action decoder for deep multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2020.

Hu, H., Lerer, A., Peysakhovich, A., and Foerster, J. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pp. 4399–4410, 2020.

Hu, H., Lerer, A., Cui, B., Pineda, L., Brown, N., and Foerster, J. Off-belief learning. In *International Conference on Machine Learning*, pp. 4369–4379, 2021.

Kalai, E. and Lehrer, E. Rational learning leads to Nash equilibrium. *Econometrica: Journal of the Econometric Society*, pp. 1019–1045, 1993.

Kallenberg, O. *Foundations of modern probability*. Springer, 1997.

Macindoe, O., Kaelbling, L. P., and Lozano-Pérez, T. POM-CoP: Belief space planning for sidekicks in cooperative games. In *Artificial Intelligence and Interactive Digital Entertainment Conference*, 2012.

Mescheder, D., Tuyls, K., and Kaisers, M. Opponent modeling with POMDPs. In *The 23rd Belgium–Netherlands Conference on Artificial Intelligence*, pp. 152–159, 2011.

Panella, A. and Gmytrasiewicz, P. Bayesian learning of other agents' finite controllers for interactive POMDPs. In *The Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Poland, J. and Hutter, M. Defensive universal learning with experts. In *International Conference on Algorithmic Learning Theory*, pp. 356–370, 2005.

Poland, J. and Hutter, M. Universal learning of repeated matrix games. In *The 15th Annual Machine Learning Conference of Belgium and The Netherlands*, 2006.

Powers, R. and Shoham, Y. New criteria and a new algorithm for learning in multi-agent systems. *Advances in Neural Information Processing Systems*, 17, 2004.

Powers, R. and Shoham, Y. Learning against opponents with bounded memory. In *The Nineteenth International Joint Conference on Artificial Intelligence*, pp. 817–822, 2005.

Robinson, J. An iterative method of solving a game. *Annals of mathematics*, 54(2):296–301, 1951.

Shoham, Y. and Leyton-Brown, K. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14502–14515, 2021.

A. Proof of Proposition 4.5

Let $r > \frac{1}{4}$, c and μ_e (for all $e \in E$) be values satisfying Equation 5 for Bob's strategy ϕ in the repeated game G. For any expert strategy $e \in E$, history $h \in \mathcal{H}$, and integer s, we have that

$$\mathrm{E}_{a,b\sim e,\phi}\left[\frac{1}{s}\sum_{n=|h|+1}^{|h|+s}G(a_n,b_n)\ :\ h\right] - \mu_e = \mathrm{E}_{a,b\sim e,\phi}\left[\frac{1}{s}\sum_{n=|h|+1}^{|h|+s}G(a_n,b_n) - \mu_e\ :\ h\right] \tag{56}$$

$$\leq \mathrm{E}_{a,b\sim e,\phi}\left[\left|\frac{1}{s}\sum_{n=|h|+1}^{|h|+s}G(a_n,b_n) - \mu_e\right|\ :\ h\right] \tag{57}$$

$$\leq cs^{-r} \tag{58}$$

Reversing this argument, we also have that

$$\mathrm{E}_{a,b\sim e,\phi}\left[\frac{1}{s}\sum_{n=|h|+1}^{|h|+s}G(a_n,b_n)\ :\ h\right] - \mu_e \geq -\mathrm{E}_{a,b\sim e,\phi}\left[\left|\frac{1}{s}\sum_{n=|h|+1}^{|h|+s}G(a_n,b_n) - \mu_e\right|\ :\ h\right] \tag{59}$$

$$\geq -cs^{-r} \tag{60}$$

Finally, using Equation 58 we get

$$V_G(e,\phi|h) - \mu_e = \liminf_{s\to\infty}\mathrm{E}_{a,b\sim e,\phi}\left[\frac{1}{s}\sum_{n=|h|+1}^{|h|+s}G(a_n,b_n)\ :\ h\right] - \mu_e \tag{61}$$

$$\leq \liminf_{s\to\infty}cs^{-r} = 0 \tag{62}$$

while using Equation 60 gives us

$$V_G(e,\phi|h) - \mu_e \geq \liminf_{s\to\infty}-cs^{-r} = 0 \tag{63}$$

Therefore, $V_G(e,\phi|h) - \mu_e = 0$ for any $e \in E$, and so ϕ is also open-ended with respect to E. $\square$

B. Proof of Proposition 4.6

Let E be the set of $|A|$ fixed-action experts in G. Let α_s be the empirical strategy up stage s, defined as $\alpha_s(a) = \frac{1}{s}\sum_{n=1}^{s}1_{\{a_n=a\}}$. For each expert $e \in E$, let $\mu_e = \max_{b\in B}G(e,b)$. Let ϕ_{fp} be the strategy that implements fictitious play. We need to show that, for each $e \in E$, and for each history $h \in \mathcal{H}$, we have $\mu_e - V_G(e,\phi^{\mathrm{fp}}|h) = 0$. The main issue is showing that, for any initial history h, when Alice commits to expert e, Bob's fictitious play strategy will eventually play a best-response to e. We define the $B(e) = \arg\max_b G(e,b)$, and then define the minimum regret $\epsilon(e)$ as

$$\epsilon(e) = \min_{b\notin B(e)}[\mu_e - G(e,b)] \tag{64}$$

that is, $\epsilon(e)$ is the minimum regret over all of Bob's sub-optimal pure-strategy responses to e. Assume that from stage $n = |h|$ Alice follows expert e for all subsequent stages. Then for any s we have

$$\alpha_{|h|+s}(e) = \frac{1}{|h|+s}\left[s + \sum_{n=1}^{|h|}1_{\{a_n=e\}}\right] \tag{65}$$

$$\geq \frac{s}{|h|+s} \tag{66}$$

In stage $s > 1$, fictitious play plays Bob's best response to α_{s-1}. Therefore, for Bob to select a sub-optimal action $b \notin B(e)$ in stage $s+1$, it will need to be the case that

$$G(\alpha_s,b') - G(\alpha_s,b) \leq 0 \tag{67}$$

for all $b' \in B(e)$, where $G(\alpha_s, b)$ is shorthand for $\sum_a \alpha_s(a)G(a, b)$. Starting from an initial history h, let Alice follow expert e for $s = \left\lceil \frac{|H|}{\epsilon(e)} \right\rceil + 1$ stages. We then have that

$$G(\alpha_{|h|+s}, b') - G(\alpha_{|h|+s}, b) = \alpha_{|h|+s}(e)\left[G(e, b') - G(e, b)\right] + \sum_{a \neq e} \alpha_{|h|+s}(a)\left[G(a, b') - G(a, b)\right] \tag{68}$$

$$\geq \alpha_{|h|+s}(e)\epsilon(e) - \left[1 - \alpha_{|h|+s}\right] \tag{69}$$

$$\geq (1 + \epsilon(e))\frac{s}{|h| + s} - 1 \tag{70}$$

$$= (1 + \epsilon(e))\frac{\left\lceil \frac{|H|}{\epsilon(e)} \right\rceil + 1}{|h| + \left\lceil \frac{|H|}{\epsilon(e)} \right\rceil + 1} - 1 \tag{71}$$

$$> (1 + \epsilon(e))\frac{\frac{|h|}{\epsilon(e)}}{|h| + \frac{|h|}{\epsilon(e)}} - 1 = 0 \tag{72}$$

This means that after Alice follows e for $s = \left\lceil \frac{|H|}{\epsilon(e)} \right\rceil + 1$ steps, Bob's best response to $\alpha_{|h|+s}$ cannot be a suboptimal response to e, simply because e will have too much mass under the resulting empirical distribution. Bob will switch to some $b \in B(e)$ after no more than s steps, and will never change back so long as Alice continues to play e. Therefore, starting from any history $h \in \mathcal{H}$, for any $e \in E$, we have that $V_G(e, \phi^{\text{fp}}) = \mu_e$, concluding the proof. □

C. Proof of Lemma 5.2

Define the random variables c_s on $(\Omega, \mathcal{F})$ as

$$c_s(\omega) = \begin{cases} 0, & \exists t > s, a_t(\omega) \neq a_s(\omega) \\ 1, & \text{otherwise} \end{cases} \tag{73}$$

For each stage s, the variable c_s indicates whether Alice has converged to a fixed action by this stage. We need to ensure that each $c_s : \Omega \mapsto \Re$ is measurable w.r.t. $\mathcal{F}$. We also need to show that the $\lim_{s \to \infty} c_s$ is measurable. To see this, let $C_{s,t} \in \mathcal{F}$ be the event that Alice takes a fixed action from stage s to at least stage t. We have that $\{\omega \in \Omega : c_s(\omega) = 1\} = \bigcap_{t=s}^{\infty} C_{s,t}$, and $\{\omega \in \Omega : \lim_{s \to \infty} c_s(\omega) = 1\} = \bigcup_{s=1}^{\infty} \{\omega \in \Omega : c_s(\omega) = 1\}$. These are countable intersections and unions, and so $\{\omega \in \Omega : c_s(\omega) = 1\} \in \mathcal{F}$ and $\{\omega \in \Omega : \lim_{s \to \infty} c_s(\omega) = 1\} \in \mathcal{F}$. Therefore, each c_s, as well as $\lim_{s \to \infty} c_s$, is measurable w.r.t. $\mathcal{F}$.

Note that because $c_s \in \{0, 1\}$, the indicator functions $1_{\{c_s=0\}} = 1 - c_s$, and the indicator function $1_{\{\lim_{s \to \infty} c_s=0\}} = 1 - \lim_{s \to \infty} c_s$. We then have that

$$\gamma = \Pr\{\forall s, \exists t > s, a_t \neq a_s\} \tag{74}$$

$$= \Pr\left\{\lim_{s \to \infty} c_s = 0\right\} \tag{75}$$

$$= E\left[1_{\{\lim_{s \to \infty} c_s=0\}}\right] \tag{76}$$

$$= E\left[1 - \lim_{s \to \infty} c_s\right] \tag{77}$$

$$= E\left[\lim_{s \to \infty} 1 - c_s\right] \tag{78}$$

$$= \lim_{s \to \infty} E\left[1 - c_s\right] \tag{79}$$

$$= \lim_{s \to \infty} E\left[1_{\{c_s=0\}}\right] \tag{80}$$

$$= \lim_{s \to \infty} \Pr\{c_s = 0\}. \tag{81}$$

Since c_s is bounded, we can exchange the limit and expectation by the dominated convergence theorem (Bartle, 2014). This means that for any $\delta > 0$ there exists τ such that

$$\Pr\{c_\tau = 0\} \leq \gamma + \delta \tag{82}$$

Since $\{c_s = 0\} \Leftrightarrow \{\exists t > s, a_t \neq a_s\}$, this implies that the probability that learning has not converged by stage τ is less than to $\gamma + \delta$, and so this τ satisfies Equation 10 for the chosen value of δ. $\square$

D. Proof of Lemma 5.5

The proof is analogous to that of Lemma 5.2. First, recall from the previous proof that $\{\omega \in \Omega : \lim_{s \to \infty} c_s(\omega) = 1\} = C$ is measurable w.r.t. $\mathcal{F}$. We then note that for any $h \in \mathcal{H}$ with $\Pr\{h\} > 0$,

$$\Pr\{\forall s > |h|, \exists t > s, a_t \neq a_s \mid h \wedge \neg C\} = 1. \tag{83}$$

which comes from the definition of $\neg C$. For all $t \in \mathbb{N}$, we define a sequence of random variables d_s^t on $(\Omega, \mathcal{F})$ such that

$$d_s^t(\omega) = \begin{cases} 0 & \forall n \in [t, t+s], a_t(\omega) = a_n(\omega) \\ 1 & \text{otherwise} \end{cases} \tag{84}$$

where $d_s^t = 1$ indicates that Alice will switch actions within the next s stages after stage t. We then have that

$$1 = \Pr\{\forall s > |h|, \exists t > s, a_t \neq a_s \mid h \wedge \neg C\} \tag{85}$$

$$= \Pr\left\{\lim_{s \to \infty} d_s^{|h|} = 1 \mid h \wedge \neg C\right\} \tag{86}$$

$$= \lim_{s \to \infty} \Pr\left\{d_s^{|h|} = 1 \mid h \wedge \neg C\right\} \tag{87}$$

where once again we use the fact that each $d_s^{|h|}$ and $\lim_{s \to \infty} d_s^{|h|}$ are bounded, combined with the dominated convergence theorem. This means that for any $\delta > 0$, and any $h \in \mathcal{H}$ with $\Pr\{h\} > 0$, there exists σ such that

$$\Pr\left\{d_\sigma^{|h|} = 1 \mid h \wedge \neg C\right\} > 1 - \delta \tag{88}$$

Since $\{d_s^t = 1\} \Leftrightarrow \{\exists n \in [t, t+s], a_t \neq a_n\}$, this σ must satisfy Equation 30 for the chosen value of δ. $\square$