Sequential targeted optimality as a new criterion for teaching and following in repeated games

Max Knobbout
Utrecht University
Dept. of Computer Science
The Netherlands
mknobbout@gmail.com

Gerard A.W. Vreeswijk
Utrecht University
Dept. of Computer Science
The Netherlands
gv@cs.uu.nl

ABSTRACT

In infinitely repeated games, the act of teaching an outcome to our adversaries can be beneficial to reach coordination, as well as allowing us to 'steer' adversaries to outcomes that are more beneficial to us. Teaching works well against followers, agents that are willing to go along with the proposal, but can lead to miscoordination otherwise. In the context of infinitely repeated games there is, as of yet, no clear formalism that tries to capture and combine these behaviours into a unified view in order to reach a solution of a game. In this paper, we propose such a formalism in the form of an algorithmic criterion, which uses the concept of targeted learning. As we will argue, this criterion can be a beneficial criterion to adopt in order to reach coordination. Afterwards we propose an algorithm that adheres to our criterion that is able to teach pure strategy Nash Equilibria to a broad class of opponents in a broad class of games and is able to follow otherwise, as well as able to perform well in self-play.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent system*

General Terms

Algorithms, Theory

Keywords

Game Theory, Implicit Cooperation, Coordination, Teaching

1. INTRODUCTION

In the area of multiagent learning, game theory is an important tool to model the interaction between agents that arises. In order to establish and sustain coordination in a repeated game, the agents need to achieve a mutual beneficial outcome. In a setting where the agents are not pre-coordinated and have no explicit way of communication (only by observing actions/outcomes) this quickly becomes a complex scenario. From this perspective, the act of proposing (or forcing) an outcome to our adversaries makes sense, which we will informally describe as 'teaching' behaviour. On the other hand we have 'following' behaviour, which can be understood as the act of going along with such a proposal. Teaching behaviour does not only make sense in order to reach coordination, but often adopting the role of a teacher allows us to 'steer' followers to outcomes that are more beneficial to us. However, it can lead to miscoordination if multiple agents try to teach different outcomes of the game. Without an external designation of these roles, it can be hard to decide whether to take on the role of a teacher or a follower. In the context of infinitely repeated games there is, as of yet, no clear formalism that tries to capture and combine these behaviours into a unified view in order to reach a solution of a game. A reason for this could be the fact that the distinction between teaching behaviour on one hand, and following behaviour on the other, is often not so clear-cut as one might presume. In order for the reader to place this observation into perspective, the next section discusses different lines of previous work related to this work in which either (1) intuitively (a combination of) teaching and following behaviour occurs, but the authors do not explicitly mention this or in which (2) the authors mention the existence of (one of) these behaviours but do not provide a formalism or a unified view.

Cite as: Sequential targeted optimality as a new criterion for teaching and following in repeated games, Max Knobbout and Gerard A.W. Vreeswijk, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011),* Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 517-524.

2. PREVIOUS WORK

In order to make the distinction between teaching and following behaviour, one can try and identify the nature of teaching behaviour. In [5], the authors mention the concept of teaching (or leading) in repeated games, which is introduced in the form of two strategies. These strategies, named Godfather and Bully, can be used to induce good performance from 'followers'. Bully assumes it has first mover advantage, and optimizes its payoff assuming that the other player is a follower. Godfather (a generalization of Tit for Tat) uses the threat of security level to maintain a mutually beneficial outcome. Intuitively, these strategies can indeed be understood as teacher strategies, but the authors do not explicitly mention why this is the case. In [3], the authors argue that the main difference between teaching strategies as opposed to following strategies is the fact that teacher strategies also take into account the payoff of the opponent. We believe that this notion is insufficient, since arbitrary mixed strategies can also be considered teaching strategies: they force the opponent into a way of play by reducing the setting to a Markov problem.

Another approach can be to try to identify follower strategies in order to make the distinction between the two. Fol-

lower strategies can be intuitively understood as strategies that condition on the way of play of the opponent. Typical examples are model-based learners, such as Fictitious Play and Rational Learning. However, Godfather also conditions on the way of play of the opponent, so it is a natural question to ask why it is not a following strategy. In [5], the authors argue that reinforcement learning algorithms like Q-learners can be considered followers. However, we believe that Q-learners capture a bit of both: with a high learning rate they are able to quickly adapt (follow), while a low learning rate ensures that the agent stays committed to a way of play (teach). Even more so, the difference between teaching and following quickly becomes a grey area when we consider strategies that use a multitude of strategies, like no-regret learners.

We also have related research in which intuitively the combination of teaching and following behaviour occurs, but often the authors never seem to mention it. One example in which the authors do mention this can be found in [3], in which the authors use a variant of godfather that uses both teacher and follower utility together with the notion of guilt to determine the length of the punishment phase. Here guilt is the extra reward the opponent has accumulated by deviating from the target solution. The problem with this approach is that guilt has little scientific basis and it is often very unclear if the opponent should remain guilty in particular cases. Moreover, the algorithm can lead to very unpredictable and complex behaviour, which is hard for the opponent to predict this.

In other research we have that authors never mention the existence of teaching and following behaviour, even though their approach does intuitively seem to exhibit it to some degree. In [1], the authors use the WoLF principle (Win or Learn Fast) to extend the basic gradient ascent IGA algorithm. The WoLF principle states that if the player is winning, the algorithm should use a lower learning rate in the case if it is losing. Adopting a low learning rate can be seen as unwillingness to change your strategy, hence teaching behaviour, while adopting a high learning rate can be seen as follower behaviour. Another example of such research can be found in [2], where the authors propose the AWESOME (adapt when everybody is stationary, otherwise move to equilibrium) algorithm which is able to play a best response against stationary opponents in n-action n-player games, and is also able to converge to a Nash equilibrium in self-play. This algorithm does exactly what the name implies, except the other way around: It starts out with the assumption that the opponents are equilibrium players, and thus plays their part of the pre-computed equilibrium strategy (teaching). If this hypothesis later on is refuted, it then proceeds to assume the opponent is stationary and adapts accordingly by playing a best-response to the empirical frequency of play (following). But again, the point about teaching and following behaviour is not explicitly made, the algorithm merely works this way to ensure the above mentioned properties. In a multitude of papers found in [6], [7] and [8], the authors propose a criterion that states that an algorithm should achieve a close to optimal payoff against certain classes of opponents with high probability. It is then possible to use this criterion to demand a best response value against a multitude of opponents, which can lead to interesting step-wise teaching and following behaviour. For example in [6], the algorithm (1) first considers that the opponent is stationary (and plays a best response), (2) afterwards considers that the opponent is a follower (and plays a mixed strategy variant of Bully) and (3) if no considerations can be made, concludes that the best we can do is follow (and plays Fictitious Play). But again, the point about teaching and following is not explicitly made. As we will motivate later, we believe that this approach, using beliefs about our opponent to decide whether to teach or to follow, is a suitable approach for our problem.

Figure 1: Non-teaching game

	Left	Right
Top	**1**,0	**0**,0
Bottom	**0**,0	**1**,0

Lastly, there is economic research about the subject which, for the purpose of this section, should not be omitted. In this approach (market) leaders and followers make up the complex dynamics of a system that arises. The point of departure is the model, such as the Stackelberg leadership model, where leader and follower are defined by the game itself and distinguished by the first-mover advantage. An important question here is whether or not the notion of teaching and following can completely exist outside the model(game)-level. This is actually questionable, as we will demonstrate by the following example. Consider the game shown in Figure 1, which we have chosen to name the "non-teaching game". Now consider that we are the row player, and our opponent is the column player, and we are playing an infinitely repeated game where we want to maximize our average returned payoff. In this game, the opponent is indifferent about all possible outcomes of the game. Forcing outcomes by leading (Bully) or retaliating (Godfather) is impossible in this game, since the opponent does not prefer any outcome over another. The security value for this game is 0.5 by adopting the mixed strategy (0.5,0.5) (the security value is the value the agent can guarantee regardless of the opponent by playing purely defensively). Arguably, the best strategy to adopt in this game is to start out with this defensive strategy and afterwards play a best response based on the frequency of play of the opponent (for example in the case he plays Left more than Right), or in other words following behaviour. These types of games are evidence that teaching might not be feasible in all games.

3. BASIC CONCEPTS

In the previous section, we saw that the distinction between teaching and following behaviour is not so clear cut as one might suspect. In this section we will provide the reader with a criterion that tries to capture the essence of teaching and following. In our setting we consider infinitely repeated, complete, perfect information normal-form games (complete and perfect information implies that all agents have knowledge about the payoffs of the game and the actions that have taken place), with 2 players and n actions. These games are defined in the normal game-theoretic sense, that is to say they consist of a finite set of players, a finite set of actions for each player and a real valued payoff function that maps for each player an action profile to a real

number. The goal is to maximize the average reward the player receives. The n player extension might be interesting for possible future research, but for now we do not want to complicate matters too much.

During our approach we will sometimes stop and deliberate on two important aspects of teaching and following, which is the *what* ("what to teach?") and the *when* ("when to teach and when to follow?"). Here we can already partly state the "what", namely we should teach something that is within the capabilities of our opponent. This idea to get a best response value against strategies that belong to a certain class of strategies is discussed by Shoham and Brown in [9, pp. 222–223], where they discuss the concept of targeted learning and use a criterion named (efficient) *targeted optimality*. The following definition is similar, except that we replaced the somewhat vague notion of 'class of opponents' to a set of strategies, which can be any subset of the full strategy set available.

Definition 1. Given a (finite or infinite) strategy set S, a strategy is said to be *targeted optimal* if it holds that for any choice of $\epsilon > 0$ and $\delta > 0$ there should exist a number of rounds τ, polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, such that for every number of rounds $t \geq \tau$ the strategy against an arbitrary strategy $\sigma \in S$ achieves average payoff of at least $V_{BR}(\sigma) - \epsilon$ with probability $1 - \delta$, where $V_{BR}(\sigma)$ is the value of the best response given that the opponent plays σ. If, during run-time, for a choice of ϵ and δ the average payoff when playing our strategy remains ϵ-close to the best response value for every number of rounds $t \geq \tau$, where τ is defined as previous, we say that the property of optimality is *maintained*.

Notice that (ϵ, δ)-optimality is quite a weak notion of optimality. To explain this choice in the context of teaching and following, the choice of ϵ can be explained by the fact that sometimes we need room to identify whether or not the opponent can be taught. The latter choice can be explained by the fact that we can never be certain whether or not the opponent actually belongs to the target class. Here, δ can be seen as a 'measure of stubbornness' to determine when to abandon our hopes to achieve an average payoff ϵ-close to the best response value (namely when we are certain enough that the opponent does not belong to the target class). We believe that this measure is something which we need when it comes to teaching. In the next part of this section, definition 1 will be used to define a novel criterion that tries to capture teaching and following behaviour into a unified view.

3.1 A new criterion for teaching and following

The first step in the construction of our criterion is to use the notion of targeted optimality and create a new notion in which it is applied sequentially. We propose this new criterion as *sequentially targeted optimality* (we drop the 'efficient' adjective to keep the criterion name more compact).

Definition 2. A strategy σ is said to be *sequentially targeted optimal* given strategy sets S^p and S^s if it holds that this strategy first deploys a strategy, referred to as σ^p, and σ^p should be targeted optimal given strategy set S^p. If for a choice of ϵ and δ during run-time the property of optimality is not maintained (either because (1) the strategy of the opponent indeed belongs to S^p but with probability δ we have not achieved an average payoff ϵ-close to the best response value or (2) the strategy of the opponent does not belong to S^p), then our strategy should deploy another strategy, referred to as σ^s, and σ^s should be targeted optimal given S^s. If a strategy is sequentially targeted optimal with respect to S^p and S^s, we refer to the first deployed strategy σ^p as the primary strategy, and the second deployed strategy σ^s as the secondary strategy.

The reason that we also applied the weaker notion of (ϵ, δ)-optimality to the secondary strategy is simply because we want to have room for an algorithm to also adhere to other criteria (and not just one criterion which overrules any other possible criterion). Notice that this criterion already states some of the aspects of teaching and following. It states the "what": we try to achieve the best possible payoff (or at least arbitrary close to) given that we condition on the opponent. It also (partly) states the "when": first we could have a period in which we try to 'teach' the opponent, and if that fails, we could have a period in which we try to 'follow'. However, if we just use an arbitrary primary strategy set and secondary strategy set to create a sequential targeted optimal strategy, this resulting strategy can definitely not be labelled a teaching- and following strategy in all cases. This is because we have not laid any restrictions on these strategy sets and because we have to show that teaching can indeed be beneficial. However, formalizing a notion of teaching and following strategies is problematic, since it is often a grey area as we saw in section 2. To overcome this problem, we will try to define when a sequential targeted optimal strategy is a sequentially teaching-following strategy as a whole, without defining its specific parts S^p and S^s. To do this, we will first introduce the notion of self-teachability.

Definition 3. A strategy σ is *self-teachable* if it is sequentially targeted optimal given S^p and S^s, using primary strategy σ^p and secondary strategy σ^s, if it holds that $\sigma^p \in S^s$ and $\sigma^s \in S^p$.

Loosely speaking, a strategy is self-teachable if we are able to 'follow' (and get our desired best response value) on the strategy which we use to 'teach' and we are able to 'teach' (and get our desired best response value) on the strategy which we use to 'follow'. Thus, if a strategy is self-teachable it contains some sort of symmetry within the different strategies that are deployed. Using this notion of 'symmetry', we propose a novel criterion that tries to capture both teaching and following behaviour, which is given by the *sequential teaching-following* criterion in the next definition.

Definition 4. A strategy is said to be a *sequential teaching-following strategy* if it is self-teachable in a set of games G (that is, it achieves the property of self-teachability in all these games) using strategy sets S^p and S^s and if it holds that in all games belonging to G, the guaranteed best response value of playing against a strategy from S^p is at least as high as the guaranteed best response value of playing against a strategy from S^s:

$$\min_{\sigma \in S^p} V_{BR}(\sigma) \geq \min_{\sigma' \in S^s} V_{BR}(\sigma')$$

If a strategy is a sequential teaching-following strategy, we refer to the primary strategy as the teacher strategy and the secondary strategy as the follower strategy.

This criterion states when a sequential targeted optimal strategy is a sequential teaching-following strategy without

Figure 2: Matching pennies

	Left	Right
Top	**-1**,1	**1**,-1
Bottom	**1**,-1	**-1**,1

explicitly specifying its parts S^p and S^s and it states a certain beneficialness which is restricted to a set of games. The beneficialness is stated in terms of payoff guarantees (and not for example in terms of maximum payoff or expected payoff), because minimum payoff is an important concept in repeated games to identify enforceable outcomes. The restriction to a set of games is because we already talked about the feasibility of teaching: not all games are suited for teaching. It also allows us to play around more with the concept, since we can form sequential teaching-following strategies that use non-mixed strategies like Bully and Godfather as teaching strategies for example, without necessarily consorting to mixed variants. This is because in some games, the only equilibrium strategies are mixed. One such well known example is the matching pennies game shown in Figure 2. Moreover, there is nothing restricting anyone to drop the requirement by creating a sequential teaching-following strategy that conditions over every game. We believe that this notion captures the essence of teaching and following: here teaching and following are defined as behaviours that are able to coordinate together (both players are able to get a best response) and they can be separated by the fact that teacher behaviour has a certain beneficialness to it.

The symmetry we demanded in our previous definition of teaching-following strategies may seem overly restrictive, since we demanded a 2-way interaction: teaching should be good against following and vice versa. However, this demand not only serves as a way to distinguish teaching from following strategy, but also to ensure certain beneficial properties in self-play.

PROPOSITION 1. *When using a sequential teaching-following strategy in self-play, if it is the case that one player maintains its teacher strategy $\sigma^p \in S^s$ while the other maintains his follower strategy $\sigma^s \in S^p$, then both players converge to a Nash equilibrium.*

PROOF. Since the strategy σ^p is targeted optimal given strategy set S^p for any arbitrary choice of $\epsilon > 0$, and strategy σ^s is targeted optimal given strategy set S^s for any arbitrary choice of $\epsilon' > 0$, we know that the first player will achieve for any ϵ an average payoff ϵ-close to $V_{BR}(\sigma^s)$ while the second player will achieve for any ϵ' a pay ϵ'-close to $V_{BR}(\sigma^p)$. This means that, given an arbitrary ϵ and ϵ', it holds that for the first player there are no strategies available such that more than ϵ expected payoff can be gained and for the second player there are no strategies available such that more than ϵ' expected payoff can be gained. Thus both players can not gain more than $\max(\epsilon, \epsilon')$ by deviating unilaterally, which implies a $\max(\epsilon, \epsilon')$-Nash equilibrium. Since the players maintain their strategies, we can let $\epsilon \to 0$ and $\epsilon' \to 0$, and thus $\max(\epsilon, \epsilon') \to 0$, which means that in the limit the players converge to a Nash equilibrium. □

This proposition is important when we want to show when a specific teaching-following strategy converges to a Nash equilibrium in self-play. As we will see later, in order to guarantee convergence to a Nash equilibrium in self-play we also need to consider the case in which both the players maintain their teaching strategy (if possible) and the case in which both players maintain their following strategy.

The teaching-following criterion we supplied tried to incorporate intuitive aspects of teaching and following, such as the "what" and the "when". Based on the criterion, it can be argued that in infinitely repeated games, it can be beneficial to first try to teach an outcome that allows us to receive a greater guaranteed outcome. This is especially the case for conservative agents that care more about payoff guarantees than payoff maximization. Many known strategies can be extended to have a teaching phase, so there is not really anything to lose given that the game is not finite. If the rate of convergence plays a role, the criterion also states that the properties should be achieved in efficient time. Moreover, as we will see later on with our algorithm, combining two strategies with the use of the criterion will cause the resulting strategy to maintain many of the properties of the original strategies. In other words, the criterion not only tries to capture the essence of teaching and following, but it is also a beneficial criterion for algorithms to adhere to. Moreover, it allows authors to create strategies in terms of 'weaknesses': what works good against what in which situations? In the next section we will create an algorithm that adheres to our proposed criterion.

4. IMPLEMENTATION

In this section, we will first look at the teaching and following component of our algorithm individually and afterwards we will combine them to create an algorithm that is both able to teach and follow in repeated games by adhering to our teaching-following criterion.

4.1 Teaching strategy

For the teaching part of our strategy, we will use a variant of Bully. We already saw that intuitively this strategy is indeed a teaching strategy, since it assumes it has Stackelberg leader advantage. On the other hand, Bully does not work well in all games, in particular games that require mixed equilibria. In the long run this will imply that our strategy is not able to teach beneficial outcomes in all games.

The idea is that Bully, in some games, works specifically well against opponents that are willing 'to go along with the proposal', such as learning rules that play a best response to the distribution of play. As it turns out, the class of strategies that are 'susceptible' to Bully is very broad and covers many examples found in literature. We refer to these strategies as *pure consistent* strategies, which is a superclass of the consistent strategies defined in [4]. The difference is that pure consistent strategies should achieve a best response against pure strategies, in stead of arbitrary mixed strategies in the case of consistent strategies. We also extend the definition with the notion of a polynomial rate of convergence, which will play a role in the next proposition.

Definition 5. A strategy is said to be *ϵ-pure consistent* if there exists a T such that against any pure strategy σ_{-i} and for any $t > T$ the strategy achieves a payoff ϵ-close to $V_{BR}(\sigma_{-i})$ with probability $1 - \epsilon$. A strategy is *pure consistent* if it is ϵ-pure consistent for every positive ϵ and is said to have a *polynomial rate of convergence* if T is polynomial in $\frac{1}{\epsilon}$.

It can easily be shown that any consistent strategy (like Fictitious play), universal consistent strategy (like no-regret learners) and rational strategy (mentioned in [1], not to be confused with the economical definition of rationality) are pure consistent, as well as countless more strategies. The reason for this is because a pure strategy is very easy to learn for the opponent. This is again one of the beautiful aspects of teaching and following: if the message we are trying to teach is simple, the class we can target is much larger than in the case in which we are trying to teach a more complex message.

As our teacher strategy, we use a modified version of Bully. This is because Bully is not well defined in cases in which our opponent is indifferent about several outcomes. To cope with this, we define our teacher value and action in the following way:

Definition 6. The teacher value, $V_{teacher}$, is defined as:

$$V_{teacher} = \max_i V_i(i, j_i^*)$$

where

$$j_i^* = \operatorname*{argmin}_{j \in J_i} V_i(i, j)$$

and

$$J_i = \{\, a \mid V_{-i}(i, a) = \max_j V_{-i}(i, j) \,\}$$

In short, $V_{teacher}$ is defined as the best possible payoff the agent can guarantee by assuming it has first-mover advantage and by assuming that the opponent plays a best response to this pure strategy which is least beneficial to us. The action belonging to $V_{teacher}$ is defined as $a_{teacher}$.

Observe that $V_{teacher}$ is indeed a best response value against an arbitrary pure consistent opponent (notice that it can still be considered a best response value in repeated games if the opponent is (universally) consistent, as long as we are teaching a feasible and enforceable outcome; more on this observation later). However, if it is the case that $V_{teacher} < V_{Maximin}$ (recall for example the matching pennies game in Figure 2), it is arguably better to play our (possibly mixed) Maximin strategy. As we will see later, this will not pose a problem since the notion of teaching-following can be restricted to a set of games. The proof that we will use is unique in the sense that it does not rely on probability bounds to show a probability dependent payoff guarantee. This is because our opponent is using a learning/adaptive strategy (which cannot be simply captured by a Random variable). However, observe that if we play a pure strategy against a pure consistent strategy, the strategy we play also seems to show 'consistent behaviour'. This idea will be the basis of the upcoming proof, in which we will show targeted optimality against the set of pure consistent strategies by adopting the strategy in which we repeatedly play $a_{teacher}$.

PROPOSITION 2. *For any choice of $\epsilon > 0$ and $\delta > 0$ against an opponent that uses a pure consistent strategy σ_{-i} with a polynomial convergence rate, there exists a finite T, polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, such that playing $a_{teacher}$ repeatedly will for any $t > T$ result in an average payoff of at least $V_{teacher} - \epsilon$ with probability $1 - \delta$ against this opponent.*

PROOF. We will first show that for any given value of ϵ, there exists an $\epsilon' > 0$, such that if it is the case that our opponent with probability equal or greater than $1 - \epsilon'$ receives an average payoff ϵ'-close to his optimal payoff, we receive an average payoff ϵ-close to $V_{teacher}$. Since our opponent has a polynomial rate of convergence, we use a polynomial function $T_{-i}(\frac{1}{\epsilon'})$ to denote the actual time steps needed to achieve the property of pure consistency. Let p_i and p_{-i} be the payoff belonging to the action profile $(a_{teacher}, BR(a_{teacher}))$. Without loss of generality, we consider that there is another action profile in the vector, (a_i, a_{-i}) with payoff p'_i and p'_{-i} respectively such that p'_i is the worst payoff in the vector for our agent and p'_{-i} the (second) best for the other agent. Let's also consider that $p_i > p'_i + \epsilon$, since otherwise any combination of actions by the opponent would guarantee that the average payoff we receive is larger or equal than $V_{teacher} - \epsilon$. Similarly we have that $p_{-i} > p'_{-i}$, since by definition of $a_{teacher}$ we have that any action with payoff equal to p_{-i} will net our agent a payoff of at least $V_{teacher}$. For every possible ϵ, the worst-case candidate h to violate the property is playing k proportion $(a_{teacher}, BR(a_{teacher}))$ and $(1-k)$ proportion (a_i, a_{-i}) such that it holds that our opponent still receives an average payoff ϵ'-close to his optimal payoff. Since in this case $V_{teacher} = p_i$ and $V_{BR}(a_{teacher}) = p_{-i}$, we have to find an ϵ' such that the proportion k is high enough such that:

$$k * p_{-i} + (1 - k) * p'_{-i} + \epsilon' \geq p_{-i}$$

implies that the following also holds:

$$k * p_i + (1 - k) * p'_i + \epsilon \geq p_i$$

Solving for ϵ', we see that

$$\epsilon' \leq \epsilon * \kappa$$

where

$$\kappa = \left(\frac{p_{-i} - p'_{-i}}{p_i - p'_i} \right)$$

Since we know that $p_i > p'_i$, $p_{-i} > p'_{-i}$ and $\epsilon > 0$, this outcome is strictly positive. Thus for ϵ' any value in the interval $(0...b]$, where $b = \epsilon * \kappa$ guarantees that if our opponent (with probability $1 - \epsilon'$) receives a payoff ϵ'-close to his optimal payoff then our agent receives a payoff ϵ-close to $V_{teacher}$. Notice that this happens after $T_{-i}(\frac{1}{\epsilon'})$ iterations.

The second step in our proof is to observe that this result is general enough to apply to any game, since we can just drop the assumption that p'_i and p'_{-i} belong to the same payoff profile. It is not hard to see that fixating the proportion that $(a_{teacher}, BR(a_{teacher}))$ is played in combination with an arbitrary action profile allows us to find a larger value for ϵ' than in the case of repeatedly getting the worst possible payoff for our agent and the second best for the other agent. In other words, this is the largest possible range we can find for ϵ' that is small enough to ensure the property. Moreover, we can make the observation that it also holds that for every later iteration than $T_{-i}(\frac{1}{\epsilon'})$, the average payoff will not decrease. For a small enough value of ϵ' for the opponent (namely small enough such that there exists no other payoff in the payoff vector that is smaller than $\max_{a \in A_2} V_{-i}(a_{teacher}, a)$ and larger or equal than $\max_{a \in A_2} V_{-i}(a_{teacher}, a) - \epsilon'$) the opponent can do no better to maintain or increase the proportion k in which $BR(a_{teacher})$ is played. Thus, for a small enough value of ϵ for our agent, the proportion in which we receive $V_{teacher}$ is also maintained or increased. Since in the above

proof the calculation for ϵ' was based on achieving the worst possible payoff in the remaining proportion of rounds, it is impossible that our average payoff also drops lower; it is enough that the proportion in which $V_{teacher}$ is achieved remains constant or increases.

The final step is to prove the proposition. Using the earlier defined function T_{-i} and our found value for κ, we see that after $T_{-i}\left(\max(\frac{1}{\delta}, \frac{1}{\kappa * \epsilon})\right)$ time steps, we receive for any later time step an average payoff ϵ-close to $V_{teacher}$ with probability $1-\delta$. Since T_{-i} is a polynomial function, we also achieve this polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ (notice that κ is just a game-constant). □

This proof concludes the teaching part of our strategy and enables us to move on to the following strategy.

4.2 Following strategy

For our following strategy, we have a number of possibilities, since many strategies achieve a best response value against pure strategies within polynomial time (for example Fictitious Play). However, we have chosen to select the AWESOME strategy to fill in this role, which is discussed in [2]. We stress that for the sake of understanding the message we are trying to convey in this paper no thorough understanding of AWESOME is required. The most important aspect of AWESOME is the fact that it has two key properties, namely AWESOME (1) converges to a Nash equilibrium in self-play, which, as we will prove later, cause our sequential teaching-following strategy to converge as well; and (2) converges to a best-response against arbitrary stationary opponents. The resulting teaching-following strategy will (more or less) also have this property. Unfortunately, proving targeted optimality against pure strategies when using AWESOME is not so easy as it may seem, and requires knowledge of valid schedules and the specific steps taken in the algorithm. Moreover, the exact amount of rounds needed in which we acquire targeted optimality is not relevant in the case of our algorithm, since we will play AWESOME for the rest of the game once we adopt it. Thus instead of giving the full proof, we give a brief proof outline.

PROPOSITION 3. *When using the AWESOME algorithm, for any $\delta > 0$ and $\epsilon > 0$, there exists a number of rounds τ, polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, such that for any number of rounds $t \geq \tau$ the strategy against an arbitrary pure strategy σ achieves average payoff of at least $V_{BR}(\sigma) - \epsilon$ with probability $1-\delta$, where $V_{BR}(\sigma)$ is the value of the best response against σ.*

The proof is heavily based on the fact that the observed distribution of play of the opponent is identical to the true distribution of play (contrary to mixed strategies). After every restart, AWESOME will first consider that the opponent is an equilibrium player. This hypothesis is refuted after a fixed amount of rounds, based on the monotonically decreasing closeness parameter (belonging to the schedule) that denotes the maximum allowed distance between distributions in the equilibrium playing phase, and it is based on the distance between the pure strategy distribution and the equilibrium strategy distribution. Afterwards AWESOME will consider that the opponents are stationary, which we will refer to as the stationary playing phase. First AWESOME will play a random action that either is a best response or not. In the first case, AWESOME will play this action for the rest of the game since it will never switch actions and thus the algorithm will never restart on behalf of itself nor the opponent. If this is not a best response, we will eventually switch actions after a fixed amount of rounds based on the number of players, the maximum number of actions, the payoff difference between our best and worst outcome in the game and our monotonically decreasing closeness parameter (belonging to the schedule) that denotes the maximum allowed distance between distributions in the stationary playing phase. If this function decreases fast enough, we will restart the algorithm. Using this information, we can find the number of restarts (and thus eventually the number of iterations) needed to ensure targeted optimality against pure strategies.

Using this proof we can immediately see that AWESOME is not only targeted optimality given the class of pure strategies, but also pure consistent. This property implies that by Definition 3 our eventual algorithm will be self-teachable.

4.3 Algorithm

The combination of the teacher and follower strategy gives us a new strategy that is able to teach pure strategy outcomes to adversaries that are willing to go along with this (pure consistent strategies) and is able to follow otherwise with a strategy we targeted in the teaching phase (in this case AWESOME). Observe that, as we will show later, the games in which this resulting strategy may work does not include games in which a mixed equilibrium strategy is required. The resulting algorithm is shown in 'Algorithm 1'. The input parameter $\langle(\epsilon^p, \delta^p), (\epsilon^s, \delta^s)\rangle$ should always be the same for any sequential targeted optimal algorithm: it contains a pair of ϵ and δ values for both the primary and secondary strategy. These parameters, as previously discussed, depict the closeness of the average payoff required and the probability that this will be reached. As we have seen, the lower the values, the longer the teaching/following process will take. The meaning of the κ variable can be found in Proposition 2 and the function T_{-i} is a polynomial function that estimates the rate of convergence of the opponent, and can effectively limit the target class to slow or fast learners (notice that we cannot make the teaching phase too short, since we also have to retain the self-teachability criterion). We again see a beautiful aspect of teaching arise: if the opponent is a slow learner, we might stop on teaching our opponent prematurely. Since the best response value against an arbitrary pure strategy is $V_{Minimax'}$, where $V_{Minimax'}$ is the pure strategy Minimax value, we know that this strategy is a teaching-following for all games in which $V_{teacher} \geq V_{Minimax'}$ (observe that $V_{Minimax'} \geq V_{Maximin}$, which settles our earlier conncern that repeatedly playing $a_{teacher}$ is not a best response in games in which $V_{teacher} < V_{Maximin}$

Algorithm 1 Sequential teaching-following strategy

Require: $\langle(\epsilon^p, \delta^p), (\epsilon^s, \delta^s)\rangle$
Ensure: $\epsilon^p > 0$, $\delta^p > 0$, $\epsilon^s > 0$, $\delta^s > 0$
1: $t \leftarrow 0$
2: **while** $(t < T_{-i}\left(\max(\frac{1}{\delta^p}, \frac{1}{\kappa * \epsilon^p})\right) \vee (\text{AvgPayoff} \geq V_{teacher} - \epsilon^p)$ **do**
3: $playaction(a_{teacher})$
4: $t \leftarrow t + 1$
5: **end while**
6: $playstrategy$(AWESOME)

Figure 3: Battle of the sexes

	Left	Right
Top	**3**,1	**0**,0
Bottom	**0**,0	**1**,3

Figure 4: Stackelberg game

	Left	Right
Top	**1**,0	**3**,2
Bottom	**2**,1	**4**,0

such as the matching pennies game shown in Figure 2). From a game-theoretic viewpoint, this result also makes perfect sense, since in this case we are indeed teaching a feasible and enforceable outcome, which then in turn can constitute a repeated Nash equilibrium as justified by the Folk theorem (for readers unfamiliar with this observation, we refer to [9, pp. 151-153] where this is very well explained). This observation can be used to prove convergence to a Nash equilibrium in self-play.

PROPOSITION 4. *In infinitely repeated games, our teaching-following algorithm, restricted to its set of games, will necessarily converge to a Nash equilibrium in self-play if it holds that ϵ_i^p and ϵ_{-i}^p are sufficiently small.*

PROOF. First let us define what 'sufficiently small' means: the values for ϵ_i^p and ϵ_{-i}^p are sufficiently small if for both players it holds that there exists no other payoff-profile in the payoff matrix for which both players receive a payoff of at least $V_{teacher} - \epsilon^p$. Notice that this is not a big restriction, since we can just compute this and pick such a small value for ϵ^p accordingly.
We distinguish the following 3 cases in self-play:

1. Both players maintain their primary strategy σ^p. This happens when both agents coincidently achieve an ϵ^p-close best response value while making false assumptions about their opponent. However, our demand for the values of ϵ^p ensure that we are indeed teaching $V_{teacher}$ and not settling on another payoff profile. Since we know that this outcome is both feasible and enforceable in our set of games, we know that we are playing a repeated Nash equilibrium.

2. Both players achieve their best response value when one player uses primary strategy σ^p while the other uses secondary strategy σ^s. Since both players are playing a best response to each other in these games, we know that σ^p is targeted optimal given σ^s and vice versa, which implies by Proposition 1 a Nash equilibrium.

3. Both players maintain their secondary strategy σ^s for the rest of the game. Convergence to a Nash equilibrium in this specific case is proven in [2].

□

Moreover, our algorithm more or less retains all the properties of AWESOME. For example, it can be easily shown that if the strategy of the opponent converges to a stationary strategy, our algorithm will converge to a best-response given this stationary strategy or we will achieve an average payoff ϵ^p-close to $V_{teacher}$.

Our teaching-following strategy enables us to teach a repeated Nash equilibrium which provably can be learned by a very broad class of opponents (contrary to just playing AWESOME) in efficient time and allows us to switch if the former fails. On top of the beneficial theoretical properties of our algorithm, we believe we can make our discussion of our algorithm even more convincing by looking at some specific games.

The following games are examples in which our algorithm is able to perform particularly well.

1. In the battle of the sexes game, shown in Figure 3, our algorithm is able to teach ('force') the (most) beneficial outcome of 3 to follower strategies that are willing to go along, while other strategies that are able to coordinate might reach a point on the Pareto boundary which is less beneficial (such as 1).

2. Our algorithm is able to signal repeated Nash equilibrium outcomes that are easy to learn by the opponent and can ensure greater payoff than the equilibrium of the stage game. This is the case with the Stackelberg game shown in 4 (with 'Stackelberg game' we do not mean the formal definition, but rather we refer to [9, p. 200] where they use this name to distinguish a particular simultaneous action Cournot game). In this particular game, our sequential teaching-following strategy is able to teach the outcome that will give our agent a payoff of 3, where as the equilibrium strategy of the stage game gives us a lower payoff of 2.

This section was mainly concerned with presenting an algorithm that is able to teach and follow with the use of our proposed criterion. In the next section we will take a step back to take a look at our criterion again, which will open the way for some general discussion.

5. GENERAL DISCUSSION

In this paper we used the notion of sequential targeted optimality to create a teaching-following criterion as a way of capturing both teaching and following behaviour in repeated games. However, some choices we made during the construction of our criterion could be made differently. An important choice we made was when we defined the notion of self-teachability. The only demand we had is that teaching and following behaviour are able to coordinate together, and that the teacher strategy sets itself apart from the follower strategy in terms of payoff guarantee in certain games. This definition can potentially imply that in some games what we understand as a 'teaching' strategy can conversely function as a 'following' strategy in other games. Since this definition still fully captures the coordination aspect of teaching and following this is not really a problem, but admittedly there might be something more to the broad meaning of a teaching strategy and the broad meaning of a following strategy. Another choice immediately becomes apparent when we define the beneficialness of the teaching part over the following part. We used payoff guarantees to define this beneficialness, which makes sense from the viewpoint of a conservative agent. On the other hand, expected payoff or maximal payoff guarantees also make sense when we consider for example greedy agents or risk-taking agents. We made this choice mainly because a minimal payoff guarantee allows us

to identify cases in which playing a strategy will necessarily lead to an enforceable outcome. But again we stress that this was nothing more than a choice.

Another important point of discussion is the fact that the notion is restricted to a set of games. By showing that in some games teaching strategies (other than our Maximin strategy) are not really feasible, we tried to make the point more clear that we really need this restriction. However, this restriction also has its problems. For example: what does it mean that a strategy is restricted to a set of games? Does it mean that the strategy is useless in other games? We have not really give an interpretation to this restriction. It becomes even more troublesome when the payoff matrices are not known. When do we know which strategy to use? We stress that this was never our intention to define; we are merely interested in defining the set of games in which 'it makes sense' to use such a strategy. The exact interpretation of this restriction is up to the creator of the strategy.

We also made a choice with the switching criterion in our definition of sequentially targeted optimality. As shown in [6], by smart use of the probability factors δ we can devise an algorithm that is targeted optimal simultaneously given different classes of opponents, instead of sequentially in our algorithm. If we would allow simultaneous optimization, it could lead to a potentially different definition of teaching and following.

As a final point of discussion, we note that our definition of sequentially teaching-following was not concerned with safety and convergence to Nash equilibria in self-play (although we have given conditions in which this can happen). We note that the latter is the least of our worries, since in a teacher and follower setting one might be less concerned about self-play. It is questionable why we even need to perform well given that we face ourselves, given that we are only concerned whether or not our opponent is a follower. The first point, a safety condition, is arguably more important. Any strategy should be safe to use, else we can just play our security strategy instead. However, we did not feel the need to include this in our criterion; this can simply be a separate criterion instead when devising a sequential teaching-following strategy.

6. FUTURE RESEARCH

There are many possibilities for future research. First of all, we would really like to see a sequential teaching-following strategy that uses Bully extended to the set of mixed strategies as its teaching strategy and for example AWESOME as its following strategy. This strategy targets in its teaching phase the set of consistent opponents (and not necessarily the *pure* consistent opponents) and its following phase the set of stationary opponents (and not necessarily pure strategies). However, we note that proving targeted optimality for AWESOME against stationary opponents can be tricky as it requires manipulation of many probability factors.

Another possible point of departure is to extend the notion of teaching-following to n-player games. In this particular case, we have to take into account the fact that our opponents might belong to different classes. The notion of targeted optimality has to be extended to cope with this fact. As shown in [8], checking if multiple opponents belong to a single class also becomes quite tricky, but is definitely an interesting direction to go in.

In this paper, our focus was on teaching and following in a sequential way. But it might be perfectly possible to teach and follow in different ways (such as periodic). This could be a direction for possible future research. For example, dropping sequentially optimality in favour of simultaneous optimality might cause interesting behaviour. In this case, if a solution of the game is reached, the agent still needs to worry about the fact whether or not the opponent might belong to a different target class. This might open up the way to new insights concerning the subject.

Another quite different point of departure is to investigate the exact nature of teaching and following. We used the self-teachability criterion, but we also mentioned in the introduction that teacher and follower strategies also have 'certain properties' that allow us to identify them as such (for example Bully and Godfather can be reasonably understood as teacher strategies). The challenge becomes to devise a formal notion of when a strategy is a teaching strategy and when a strategy is a following strategy.

A last point for possible future research we like to discuss is in settings where the payoff matrices (initially) are not known. If the payoff matrix of the adversary stays hidden throughout, it can be troublesome for teaching strategies, since (arguably) they rely heavily on the payoff matrix of the opponent. In these settings, it might be interesting to investigate how teaching and following can still arise.

7. REFERENCES

[1] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[2] V. Conitzer and T. Sandholm. AWESOME: A General Multiagent Learning Algorithm that Converges in Self-Play and Learns a Best Response against Stationary Opponents. In *Proceedings of the 20th International Conference on Machine Learning*, pages 83–90, 2006.

[3] J. W. Crandall. Learning to teach and follow in repeated games. In *AAAI Workshop on Multiagent Learning*, July 2005.

[4] D. Fudenberg and D. K. Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, 1995.

[5] M. L. Littman and P. Stone. Implicit negotiation in repeated games. In *Proceedings of The Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, pages 393–404, 2001.

[6] R. Powers and Y. Shoham. New criteria and a new algorithm for learning in multi-agent systems. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 17, 2004.

[7] R. Powers and Y. Shoham. Learning against opponents with bounded memory. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 817–822, 2005.

[8] R. Powers, Y. Shoham, and T. Vu. A general criterion and an algorithmic framework for learning in multi-agent systems. In *Machine Learning*, volume 67, pages 45–76, 2007.

[9] Y. Shoham and L. Brown. *Multiagent Systems: Algorithmic, Game-Theoretic and Logical Foundations*. Cambridge University Press, 2009.