Journal of Computational Social Science (2021) 4:839–850
https://doi.org/10.1007/s42001-021-00109-9

RESEARCH ARTICLE

Algorithms may not learn to play a unique Nash equilibrium

Takako Fujiwara-Greve[1] · **Carsten Krabbe Nielsen**[2]

Received: 25 March 2020 / Accepted: 1 March 2021 / Published online: 14 March 2021

Abstract

There is a widespread hope that, in the near future, algorithms become so sophisticated that "solutions" to most problems are found by machines. In this note, we throw some doubts on this expectation by showing the following impossibility result: given a set of finite-memory, finite-iteration algorithms, a continuum of games exist, whose unique and strict Nash equilibrium cannot be reached from a large set of initial states. A Nash equilibrium is a social solution to conflicts of interest, and hence finite algorithms should not be always relied upon for social problems. Our result also shows how to construct games to deceive a given set of algorithms to be trapped in a cycle without a Nash equilibrium.

Keywords Algorithm · Learning · Nash equilibrium · Impossibility

Introduction

This note addresses two issues. First, we ask whether algorithms/machines can reliably "solve" social problems. We do this by investigating if algorithms as players can learn to reach a Nash equilibrium of any (finite) game, as they play the game many times and adjust the way to play the game.[1] Nash equilibrium is an action profile from which no one wants to deviate unilaterally and is interpreted as a social solution to play the game. Since nowadays, there is a widespread hope that sophisticated algorithms can help humans sort out complex problems of social interactions (for example, organizing

[1] It is well-known that Lemke–Howson algorithm [11] can find a Nash equilibrium of any finite game, although the time it takes can be exponential (Savani and Stengel [16]). Our motivation is completely different from this line of research.

✉ Takako Fujiwara-Greve
takakofg@econ.keio.ac.jp

[1] Department of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan

[2] Dipartimento di Economia e Finanza, Catholic University of Milan, Via Necchi 5, 20123 Milan, Italy

traffic efficiently), it is important to investigate if algorithms can learn to play a Nash equilibrium.

Second, we reconsider the learning problem framework in game theory from a practical point of view. Many convergence results (e.g., Kandori et al. [8], Young [20], and Hurkens [7]) are based on a model such that, the game is given first and then it is asked if there is a learning process that converges to a Nash equilibrium or a set of similarly stable states (see also Gilli [6] for a more general approach to learning). We think that a practical framework has the opposite structure: starting from a fixed learning mechanism (such as installing learning algorithms in cars like the ones proposed by Koh et al. [10]), we need to ask whether the action profile sequence generated by the learning mechanism reaches a reasonable solution in any actual game (traffic situations that may occur). There is only limited literature using this approach (e.g., Milgrom and Roberts [13] and Selten [17]), and therefore we contribute to the literature by further advancing the approach.

Algorithms are necessarily finite in two senses: their memory capacity and their number of iterative reasoning are bounded. Hence, it is easy to imagine that a fixed set of algorithms cannot deal with an arbitrary game to find a Nash equilibrium, even if it is unique and strict. We prove this impossibility in a clear, simple setup. Our result also shows how to construct games to deceive a given set of algorithms to be trapped in a cycle without a Nash equilibrium.

Basic model

Nash equilibrium

Let $G = (A_1, A_2, u_1, u_2)$ be a two-person, normal-form game, where A_i is a finite set of *actions* (sometimes called "pure actions") of player $i \in \{1, 2\}$ and $u_i : A_1 \times A_2 \to \mathbb{R}$ is the *payoff function* of player i, i.e., player i wants to maximize the value of u_i. For any finite set X, let $\Delta(X)$ be the set of all probability distributions over X. The support of a probability distribution $q \in \Delta(X)$ is written as $supp(q)$. For each $i \in \{1, 2\}$, an element in the set $\Delta(A_i)$ can be interpreted in two ways: it is a probabilistic choice of actions by player i (called a "mixed action" by player i) or it is a *belief* over possible actions of i from player j's point of view (where $j \neq i$). A pure action is a degenerate mixed action, and hence $A_i \subset \Delta(A_i)$. Denote the expected payoff function by $Eu_i : \Delta(A_1) \times \Delta(A_2) \to \mathbb{R}$. Since A_i is finite, $\Delta(A_i)$ is compact with the usual topology on $\mathbb{R}^{|A_i|}$ and is convex.

For each player $i \in \{1, 2\}$ and each probability distribution $\sigma_j \in \Delta(A_j)$ by the opponent ($j \neq i$), define the set of *best responses* (in mixed actions) by player i to σ_j as

$$BR_i(\sigma_j) = \{\sigma_i \in \Delta(A_i) \mid Eu_i(\sigma_i, \sigma_j) \geqq Eu_i(\alpha, \sigma_j) \ \ \forall \alpha \in \Delta(A_i)\}.$$

Because Eu_i is continuous and $\Delta(A_i)$ is compact, $BR_i(\sigma_j) \neq \emptyset$ for any $\sigma_j \in \Delta(A_j)$.

For each player $i \in \{1, 2\}$ and any set $\Sigma_j \subset \Delta(A_j)$ (possible mixed actions by the opponent player j, or the set of beliefs held by player i), define the set of *pure actions* that can be played as a part of a best response to some (mixed) action in Σ_j:

Table 1 A game with a unique, strict Nash equilibrium

Player 1\2	a	b	c	d	e
A	7, 0	0, 1	0, 0	0, 0	−1, −1
B	0, 0	7, 0	0, 1	0, 0	−1, −1
C	0, 0	0, 0	7, 0	0, 1	−1, −1
D	0, 1	0, 0	0, 0	7, 0	−1, −1
E	2, 0	2, 0	2, 0	2, 0	3, 3

$$b_i(\Sigma_j) := \{a_i \in A_i \mid \exists \sigma_j \in \Sigma_j,\ \exists \sigma_i \in BR_i(\sigma_j) \text{ such that } a_i \in supp(\sigma_i)\}.$$

This is also nonempty for any nonempty Σ_j.

Definition 1 A mixed action profile $(\sigma_1^*, \sigma_2^*) \in \Delta(A_1) \times \Delta(A_2)$ is a *Nash equilibrium* if, for each player $i \in \{1, 2\}$, $\sigma_i^* \in BR_i(\sigma_j^*)$.

Definition 2 A mixed action profile $(\sigma_1^*, \sigma_2^*) \in \Delta(A_1) \times \Delta(A_2)$ is a *strict Nash equilibrium* if, for each player $i \in \{1, 2\}$, $\{\sigma_i^*\} = BR_i(\sigma_j^*)$.

Remark 1 A strict Nash equilibrium is a pure-action profile.

Remark 1 is a well-known result.

Definition 3 (Basu and Weibull [1]) A nonempty product set $C_1 \times C_2 \subset A_1 \times A_2$ is *closed under rational behavior* (a curb set) if $b_1(\Delta(C_2)) \times b_2(\Delta(C_1)) \subset C_1 \times C_2$.

That is, for *any* belief over the opponent's actions within C_j, the pure best response is contained in C_i. A curb set is a weaker stability concept than a Nash equilibrium.

Remark 2 (Basu and Weibull [1]) A strict Nash equilibrium is a curb set.

Motivating example

Consider a dynamic game of two players, player 1 and 2, over a discrete-time horizon $t = 1, 2, \ldots$. In each period t, they play the "component game" of Table 1, which is a two-person, normal-form game. The component game has the unique, strict Nash equilibrium of (E, e). However, we show that a dynamic action choice process may get stuck in a non-curb set $\{A, B, C, D\} \times \{a, b, c, d\}$.

Note that

$$BR_1(a) = \{A\},\ BR_1(b) = \{B\},\ BR_1(c) = \{C\},\ BR_1(d) = \{D\}; \tag{1}$$

$$BR_2(A) = \{b\},\ BR_2(B) = \{c\},\ BR_2(C) = \{d\},\ BR_2(D) = \{a\}. \tag{2}$$

Assume that players have up to two-period memory and the following algorithms (or *behavior rules*), which map observations to actions, are feasible for the players.

- Inertia algorithm: play the same action as the previous period.
- Cournot algorithm: play a best response to the opponent's previous period action.
- S2-algorithm: play a best response to an opponent using the Cournot algorithm. (That is to play a best response to a best response by the opponent to your previous period action.)
- M2-algorithms: play a best response to a probability distribution over the possible actions by the opponent using either the Inertia algorithm or the Cournot algorithm.
- M3-algorithms: play a best response to a probability distribution over the possible actions by the opponent using one of the Inertia algorithm, the Cournot algorithm, and the S2-algorithm.

Note that the M2- and the M3-algorithms are classes of algorithms since the probability distribution can vary. Suppose that the initial action combination was within $\{A, B, C, D\} \times \{a, b, c, d\}$. We first claim that, if players use one of the above algorithms throughout the time, no player can play action E or e. Consider player 1 (she). Since Inertia's case is obvious, suppose that she uses the Cournot algorithm. Her action in $t = 2$ belongs to $\{A, B, C, D\}$ because player 2's first period action was within $\{a, b, c, d\}$ and her best responses are as shown in (1). If she uses the S2-algorithm, her action in $t = 2$ is again within $\{A, B, C, D\}$, because player 2, using the Cournot algorithm, would play within $\{a, b, c, d\}$ by the same logic and (2). If she uses one of the M2- or M3-algorithms, she puts a positive probability on at most three different actions by player 2 in $t = 2$. Let x, y, and $1 - x - y$ be the probabilities of three different actions in $\{a, b, c, d\}$. Notice that

$$\min_{x+y \leqq 1} \max \{7x, 7y, 7(1-x-y)\} = \frac{7}{3} > 2.$$

Hence, action E is not a best response for any (x, y), and player 1's actions are contained in $\{A, B, C, D\}$ in $t = 2$. The logic for player 2 is similar, and this continues for $t = 3, 4, \ldots$.

Next, consider that players try to learn the best algorithm within the above five classes by changing the algorithms over time, possibly based on observations and the expected payoff for the next period. For example, suppose that player 1 was using an M2-algorithm with probability 0.5 on Inertia and 0.5 on the Cournot algorithm by player 2. If she observed (A, a) in $t = 1$ and (A, b) in $t = 2$, she may put probability 1 on the event that player 2 is using the Cournot algorithm and switch to the S2-algorithm to choose her action in $t = 3$. Alternatively, she may increase the probability on the Cournot algorithm only slightly and use a different M2-algorithm. We note that such a switching rule for algorithms (which we call a "meta-rule") can

be an algorithm. However, recall that in any period, as long as players use one of the above algorithms, they would not play an action outside of $\{A, B, C, D\}$ or $\{a, b, c, d\}$. Hence, any meta-rule would not lead to the Nash equilibrium, either.

We can generalize the logic of this example to dynamic games played by two populations and to any finite-memory, finite-iteration algorithms.

Dynamic model

We describe a simple dynamic model that is sufficient to arrive at the impossibility result in "Impossibility of learning" section. Many extensions are possible without affecting the result and are discussed in "Extensions" section.

Consider a two-player (component) game $G = (A_1, A_2, u_1, u_2)$ and a population which is partitioned into two non-empty classes V_1 and V_2, corresponding to the potential player(s) for player 1 and 2. The time horizon is discrete and written as $t = 1, 2, \ldots$. In each period, one player from each group is randomly chosen to play G. When V_i's are singletons, the model is a learning model of two, fixed players as in "Motivating example" section. When V_i's have many members, this is a social learning model of two groups. In each period, each player in that period chooses a pure action based on a *behavior rule*, a function that maps information regarding the component game and its past action profiles to an available action.

There are many well-known behavior rules. One is the Cournot algorithm described in "Motivating example" section.[2] Another well-known behavior rule is the fictitious play (e.g., Brown [2], Fudenberg and Kreps [4], and Fudenberg and Takahashi [5]) which chooses a pure action best response to the observed frequency of actions by the opponent group.[3] Each of the level-k-rules in the level-k theory (e.g., Nagel [15], Stahl [18] and Mohlin [14]) is also a behavior rule[4]: the level-0 rule is to choose all actions with equal probability, the level-1 rule best responds to the choice of the level-0 opponent and so on. Notice that these behavior rules are well-defined without the knowledge of the component game G. A behavior rule can be interpreted as a way of reasoning and can be an algorithm.

We allow players in any group to hold different behavior rules and to change behavior rules over time. The latter case includes rule-learning (e.g., Stahl [18] and [19]) and hypothesis testing (e.g., Foster and Young [3]), that is, in each period, a behavior rule (which can be an algorithm) for each player is determined by a *meta-rule* (which can be an algorithm as well) on how to adjust behavior rules over time.

[2] In order to always choose a pure-action best response, some tie-breaking rule must be added. This caveat applies to all behavior rules in the following, but our impossibility result is independent of the tie-breaking rules.

[3] The standard fictitious play rule uses the entire history to compute the "observed frequency" and thus requires unlimited memory. However, we can modify the definition of the "observed frequency" to allow bounded memory.

[4] The standard model of the level-k theory is for a single population model with a symmetric component game.

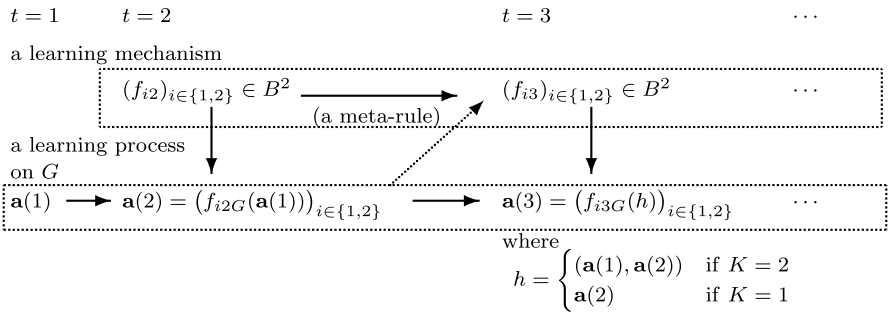

Fig. 1 Learning mechanism and its generated learning process

Let B be the set of feasible behavior rules (the contents are specified in the next subsection). A *learning mechanism*[5] is a sequence of behavior rules of the two groups $\{(f_{it})_{i\in\{1,2\}}\}_{t=2}^{\infty}$ in B^2. For our result, we do not need to specify the meta-rule structure underlying a learning mechanism. For example, a deterministic dynamic (e.g., the Cournot dynamic) can be interpreted as a degenerate learning mechanism such that $f_{it} = f$ for any $i \in \{1,2\}$ and any t.

When the component game G is given and the maximal memory capacity $K \in \mathbb{Z}_{++}$ for all players is given, the functional form of f_{itG} is induced from f_{it}. For each period $t = 2, 3, \ldots$, $H_t := [A_1 \times A_2]^{\min\{K,t-1\}}$ is the set of possible *histories* that a player remembers/collects. A behavior rule for a player in group V_i in period $t = 2, 3, \ldots$ is a function $f_{itG} : H_t \to A_i$. Since f_{itG} can choose the same action for a subset of histories, this formulation includes the case that the memory capacity is less than K. The standard Cournot dynamic, given G, is $f_{itG}(h) = b_i(a_j(t-1))$ for each $t = 2, 3, \ldots$ where $a_j(t-1)$ is the previous period action by player j in the observed history h, for any $K \geqq 1$.

A learning mechanism $\{(f_{it})_{i\in\{1,2\}}\}_{t=2}^{\infty}$, a component game G, a memory capacity K, and an initial action profile $\mathbf{a}(1)$ of G generate an infinite sequence $\{\mathbf{a}(t)\}_{t=1}^{\infty}$ of action profiles (a *learning process*) on G as illustrated in Fig. 1. The dashed arrow indicates that the underlying meta-rule may or may not use the information regarding the history.[6] Figure 1 resembles a stimulus-response model: a learning mechanism responds to the stimulus of G and the initial action profile to generate an action profile $\mathbf{a}(2)$ in $t = 2$, which becomes a part of the stimulus for $t = 3$ and so on.

Behavior rules

We focus on "rational or justifiable" behavior rules in the sense that the rule prescribes (i) a best response to some belief or (ii) a previously chosen action. Such behaviors are predominant among humans (e.g., Stahl [18], and Kneeland [9]). When designing an algorithm, we also want it to have these properties. To encompass both human learning mechanisms and algorithmic learning mechanisms, we use the word "players" (instead of algorithms) as the actors below. We explain the principle of each class of behavior rules in words first and then give formula.

[5] Since we allow learning of how to choose a behavior rule, it is not a simple "learning process".

[6] For an illustration of a rule-learning model, see Fig. 1 of Stahl [18].

For any G, any $i \in \{1,2\}$, any $t = 2,3,\ldots$, and any observed history $h = (\mathbf{a}(t-k),\ldots,\mathbf{a}(t-1)) \in H_t$ in the past $k = \min\{K, t-1\}$ periods, let

$$A_i \mid_h = \{a_i \in A_i \mid \exists \tau \in \{t-k,\ldots,t-1\}; a_i = a_i(\tau)\}$$

be the set of actions played by player i in the history h.

Simple behavior rules

In this subsection, we define simple behavior rules in the sense that the players choose actions based on a simple belief that the current opponent is one-step less sophisticated than yourself (i.e., the opponent uses a behavior rule in the one-step lower S-class) or based on no belief (S0-class).

S0-rules (conservative behavior rules):
An S0-rule chooses an action within the observed actions of one's own group. If $K \geqq 2$, then each S0-rule corresponds to a way to choose one out of K observations. If $K = 1$, then there is a unique S0-rule, which chooses the same action as the previous period. This is the Inertia rule in "Motivating example" section.

When G and t are given, an S0-rule of group V_i is a function $s_{i0} : H_t \to A_i$ such that

$$s_{i0}(h) \in A_i \mid_h, \ \forall h \in H_t.$$

S1-rules (adaptive behavior rules):
An S1-rule chooses a best response to an opponent using an S0-rule. Players of this type are often called "adaptive players" (e.g., Milgrom and Roberts [13]). The Cournot behavior rule is an S1-rule with $K = 1$.

When G and t are given, an S1-rule of population V_i is a function $s_{i1} : H_t \to A_i$ such that

$$s_{i1}(h) \in b_i(A_j \mid_h), \ \forall h \in H_t,$$

since the actions by S0-players in the other group are contained in $A_j \mid_h$.[7]

S2-rules (one-step forward-looking rules):
An S2-rule plays a best response to an opponent using an S1-rule. Since actions of such an opponent are contained in $b_j(A_i \mid_h)$, an S2-rule of population V_i at period $t = 2,3,\ldots$ is a function $s_{i2} : H_t \to A_i$ such that

$$s_{i2}(h) \in b_i(b_j(A_i \mid_h)), \ \forall h \in H_t.$$

This rule uses a forward-looking reasoning in the sense that the opponent is believed to react to the past actions in your group using some adaptive behavior rule. To

[7] We can allow an S1-player to choose a mixed-action best response $\sigma_i \in BR_i(a_j)$ for some $a_j \in A_j \mid_h$. This, however, complicates the analysis without affecting our impossibility result.

implement an S2-rule, one needs to know the opponents' payoff function. Stahl [18] provides evidence that human subjects can compute a-few-times iterated best responses. Selten [17] considers a rule similar to an S2-rule called anticipatory strategies.

S3-rules (two-step forward-looking rules):

An S3-rule plays a best response to an opponent using an S2-rule; for any $t = 2, 3, \ldots$,

$$s_{i3}(h) \in b_i(b_j(b_i(A_j \mid_h))),\ \forall h \in H_t.$$

We can iteratively define Sn-rules for $n = 4, 5, \ldots$. The index number of the rules indicates the iteration of best responses. The iterative definition is very similar to that of the level-k theory, except that the above iteration starts with a pure-action belief, while the standard level-k theory starts with a uniform distribution belief. The next class of mixed-belief behavior rules allows probabilistic beliefs.

Mixed-belief behavior rules

The Sn-rules (for $n > 0$) are based on a belief that the opponent uses an Sn-1-rule. More generally, players may have a probabilistic belief that the opponent uses up to Sn-1-rules of iterative reasoning.

M2-rules (behavior rules incorporating S0- and S1-rules by the opponent):

An M2-rule plays a best response to *some* probability distribution over S0-and S1-rules by the opponent. The functional form is, for any $t = 2, 3, \ldots$,

$$m_{i2}(h) \in b_i\big(\Delta(A_j \mid_h \cup\, b_j(A_i \mid_h))\big),\ \forall h \in H_t.$$

In other words, if a player believes that the next opponent's possible behavior rules are contained within S0- and S1-rules, then his reaction falls in the group of M2-behavior rules. The iteration of best responses is up to twice. A degenerate rule which puts weight 1 on the belief that the opponent chooses an action in the set $A_j \mid_h$ is an S1-rule, and another extreme rule which puts weight 1 on the belief that the opponent chooses an action in the set $b_j(A_i \mid_h)$ is an S2-rule.

M3-rules (behavior rules incorporating S0- and M2-rules by the opponent):

Play a best response to some probability distribution over the use of S0- and M2-rules (of various weights) by the opponent, i.e., for any $t = 2, 3, \ldots$,

$$m_{i3}(h) \in b_i\Big(\Delta\big(A_j \mid_h \cup\, b_j\big(\Delta(A_i \mid_h \cup b_i(A_j \mid_h))\big)\big)\Big),\ \forall h \in H_t.$$

Higher level Mn-rules are iteratively defined and include all lower level Mk-rules and Sk-rules as special-weight cases except S0-rules.

Table 2 Cyclic games with a unique, strict Nash equilibrium

Player 1\2	1	2	3	⋯	m	$m+1$	$m+2$
1	$x, 0$	$0, x$	$0, 0$	⋯	$0, 0$	$0, 0$	$-1, -1$
2	$0, 0$	$x, 0$	$0, x$	⋯	$0, 0$	$0, 0$	$-1, -1$
3	$0, 0$	$0, 0$	$x, 0$	⋯	$0, 0$	$0, 0$	$-1, -1$
⋮	⋯	⋯	⋯	⋯	⋯	⋯	⋯
m	$0, 0$	$0, 0$	$0, 0$	⋯	$x, 0$	$0, x$	$-1, -1$
$m+1$	$0, x$	$0, 0$	$0, 0$	⋯	$0, 0$	$x, 0$	$-1, -1$
$m+2$	$y, 0$	$y, 0$	$y, 0$	⋯	$y, 0$	$y, 0$	z, z

Definition 4 For each $n = 2, 3, \ldots$, an *n-sophisticated learning mechanism* is an infinite (possibly stochastic) sequence of behavior rules of the two groups $\{(f_{it})_{i\in\{1,2\}}\}_{t=2}^{\infty}$ within B, which is a subset of S0-rules and Mn-rules.

For example, a 2-sophisticated learning mechanism is a sequence within the S0-, S1-, S2-, and M2-rules, or equivalently, within S0-rules and M2-rules.

Impossibility of learning

Proposition 1 *For any finite $K \geqq 1$, $n \geqq 2$, and any n-sophisticated learning mechanism with memory capacity K, there exist a continuum of component games such that each G has a unique and strict Nash equilibrium but the generated sequence of action profiles on G cannot reach the Nash equilibrium from a non-singleton set of initial action profiles.*

Proof Let $m = K + n$ and consider the class of component games of the form shown in Table 2, where $x > 0$, $z > 0$ and $x/m > y > x/(m+1)$.

Note that, for any $M = 1, 2, \ldots,$

$$\min_{p \in \Delta^M} \max \{p_1, \ldots, p_M\} = \frac{1}{M}, \tag{3}$$

where $\Delta^M := \{p \in [0,1]^M \mid \sum_{k=1}^{M} p_k = 1\}$ is the $M-1$-dimensional simplex.

Consider an arbitrary probability distribution σ_j over $\{1, \ldots, m+1\}$ with m or less actions in the support. If the opponent j is expected to use σ_j, (3) implies that any pure action $k \leqq m+1$ of player i has the expected payoff of at least x/m while the pure action $m+2$ gets only y or -1. Hence the inequality $x/m > y$ implies that the pure-action best responses of player i belong to $\{1, \ldots, m+1\}$.

Also, the game of Table 2 has a unique and strict Nash equilibrium $(m+2, m+2)$. This is because the inequality $y > x/(m+1)$ and (3) imply that there are beliefs with $m+1$ actions in the support whose pure-action best responses lie outside of $\{1, \ldots, m+1\}$ for player 1. Therefore the product set $\{1, \ldots, m+1\} \times \{1, \ldots, m+1\}$ is not a curb set.

Since any n-sophisticated learning mechanism with K period memory has a belief with the support containing at most $m = K + n$ different actions of the opponents, if the initial action profile is in the non-curb set $\{1, \ldots, m+1\} \times \{1, \ldots, m+1\}$, then it cannot leave this set and reach the unique and strict Nash equilibrium $(m+2, m+2)$. Finally, notice that the set of payoff parameters that has the above property has a positive measure. □

The essence of the proof is that it is possible to construct a component game with a very large cycle of best responses so that players with limited memory and iterative reasoning cannot form a belief that rationalizes an action outside of the cycle. Note also that the unique and strict Nash equilibrium is efficient when $z > x$, i.e., there is no action profile that makes any player better off. Then it is the way to play this game, but finite-algorithms may never find it.

The impossibility result of course hinges on the fact that there is no "sufficiently wide experimentations" among actions or beliefs, as assumed in Foster and Young [3], Matros [12], and Mohlin [14]. However, we do allow randomness in choosing a behavior rule.

Concluding remarks

Extensions

We can weaken many of the assumptions without changing the result. First, a wide variety of informational structures can be allowed. The impossibility result does not change, even if players sample among K-period histories with or without recall (e.g., as in the model of Young [20]), because such samplings do not enlarge the set of iterative best responses. Second, players need not play a pure action each period. They can play a mixed best response to their beliefs, i.e., they can randomize over the actions in the best response set.

We did not consider behavior rules that assume that the opponent is as smart as yourself. To play a best response to the action by such an opponent, one needs to find a fixed point of the mutual best responses. This is essentially asking players (algorithms) to find a Nash equilibrium of a "restricted" game, where the set of available actions is the support of the starting belief. Even if we extend the model in this way, the impossibility result still holds, because, if the support of the starting belief is limited, players may not be able to find a way out of a cycle.

Practical impossibility and future directions

A possible remedy to the impossibility result is to install large volatility in actions (e.g., make the learning mechanism choose every available action with a small probability). This will take the process away from non-curb sets. However, we face another "practical impossibility": since algorithms are made to conduct routine tasks efficiently, installing sufficient randomness to find a Nash equilibrium in any game is at odds with this goal. Imagine that a self-learning, car-navigation system has some built-in random actions. To avoid any possible cycle with another navigated car, such randomness may be useful. But, for the day-to-day activities, random actions delay the time to reach the destination.

Acknowledgements The authors are grateful to an anonymous referee for useful comments.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Basu, K., & Weibull, J. W. (1991). Strategy subsets closed under rational behavior. *Economics Letters, 36*(2), 141–146.
2. Brown, G. W. (1951). Iterative solutions of games by fictitious play. In T. C, Koopmans (Ed.), *Activity analysis of production and allocation* (pp. 374–376). New York: Wiley.
3. Foster, D., & Young, P. (2003). Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior, 45*(1), 73–96.
4. Fudenberg, D., & Kreps, D. (1993). Learning mixed equilibria. *Games and Economic Behavior, 5*(3), 320–367.
5. Fudenberg, D., & Takahashi, S. (2011). Heterogeneous beliefs and local information in stochastic fictitious play. *Games and Economic Behavior, 71*(1), 100–120.
6. Gilli, M. (2001). A general approach to rational learning in games. *Bulletin of Economic Research, 53*(4), 275–303.
7. Hurkens, S. (1995). Learning by forgetful players. *Games and Economic Behavior, 11*(2), 304–329.
8. Kandori, M., Mailath, G., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica, 61*(1), 29–56.
9. Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica, 83*(5), 2065–2079.
10. Koh, S., Zhou, B., Fang, H., Yang, P., Yang, A., Yang, Q., et al. (2020). Real-time deep reinforcement learning based vehicle nagivation. *Applied Soft Computing, 96,* 106694. https://doi.org/10.1016/j.asoc.2020.106694.
11. Lemke, C., & Howson, J. (1964). Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics, 12*(2), 413–423.
12. Matros, A. (2003). Clever agents in adaptive learning. *Journal of Economic Theory, 111*(1), 110–124.
13. Milgrom, P., & Roberts, D. J. (1991). Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior, 3*(1), 82–100.
14. Mohlin, E. (2012). Evolution of theories of mind. *Games and Economic Behavior, 75*(1), 299–318.
15. Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85*(5), 1313–1326.
16. Savani, R., & von Stengel, B. (2006). Hard-to-solve bimatrix games. *Econometrica, 74*(2), 397–429.
17. Selten, R. (1991). Anticipatory learning in 2 person games. In R. Selten (Ed.), *Game equilibrium models I* (pp. 98–154). Berlin: Springer Verlag.

18. Stahl, D. (1996). Boundedly rational rule learning in a guessing game. *Games and Economic Behavior, 16*(2), 303–330.
19. Stahl, D. (2000). Rule learning in symmetric normal-form games: Theory and evidence. *Games and Economic Behavior, 32*(1), 105–138.
20. Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*(1), 57–84.