Parameter Identification for Fictitious Play Algorithm in Repeated Games

Hongcheng Dong[1,2] and Yifen Mu[2](✉)

[1] School of Mathematical Sciences, University of Chinese Academy of Sciences,
Beijing, China
donghongcheng@amss.ac.cn

[2] Key Lab of Systems and Control, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, People's Republic of China
mu@amss.ac.cn

Abstract. In the previous works [1] and [2], we solved the optimal strategy of the human player against a machine player who makes decisions based on Fictitious Play in infinitely repeated 2×2 games, in which the information is assumed to be complete and perfect. In this paper, we consider the problem of identification when the human player does not know the initial assessment of the machine player. In this scenario, we propose an identification algorithm for the human player and prove that the process of identification will end successfully in a finite time if the machine's payoff parameter is rational. When the machine's payoff parameter is irrational, the identification process will not end, which implies some advantage for the algorithm with irrational parameters.

Keywords: Repeated games · Algorithm identification · Fictitious Play · Dynamical game systems

1 Introduction

In this paper, we will study the problem of parameter identification in repeated games between a human player and a machine player which adopts some learning algorithms to take its action. To be specific, we will try to identify the initial assessment of the Fictitious Play algorithm in repeated 2×2 games. This is a simple and starting scenario for the problem of algorithm identification, which constitutes a necessary part of the evolution and control for dynamical game systems. With the development of Artificial Intelligence (AI), such games involving learning algorithms will become more and more common and important.

In the past decade, lots of algorithms have been developed to play games with different features, from the complete information 0-sum game like Go to the incomplete information 0-sum game like poker to the multi-player incomplete information stochastic games like the electronic game StarCraft [3–8]. These AIs are based on different learning algorithms to generate the near-optimal strategy in specific games which can perform very well and even beat the top human players.

These developments make AI algorithms more and more participating in people's life and work. Hence the game involving algorithms are becoming common

B. Xin et al. (Eds.): IWACIII 2023, CCIS 1931, pp. 270–282, 2024.
https://doi.org/10.1007/978-981-99-7590-7_22

and important, which are called algorithm games in the literature [9]. Early there have been many works analyzing the repeated games between symmetric algorithms. Researchers try to investigate whether the equilibrium (Nash Equilibrium, correlated equilibrium, etc) will arise as the long-run outcome when the players adopt the same algorithm to update their actions in the game. This topic has been studied extensively and still attracts attention of researchers in game theory [10–15], control theory [16,17] and computer sciences [18,19]. In fact, the convergence results provide theoretical basis for the training to optimal strategies in the construction of game AI we stated at the beginning.

On the other hand, from the point of opponent exploitation, we can provide a different perspective to understand the learning algorithms by considering the asymmetric algorithm game, i.e. the game between an algorithm (called a machine player) and a perfect opponent (called a human player). Such games also happen in many different situations. For example, many people play chess game or computer game with an AI for entertainment, the unmanned autonomous systems for different tasks would interact with each other or the human. Besides, algorithms may be forced to play a game by being attacked or fooled, which may cause surprising even serious consequences, see [20] and [21] in which scientists find that some tiny perturbations may cause the algorithms to make mistakes in classifying an image. This technical vulnerability may bring on attacks to medical learning systems, or the face recognition system designed for crime detection, which further lead to uneconomical or even dangerous consequences for the society [22–25].

Thus, in order to apply these algorithms in practice in a correct way, to understand and handle the system involving algorithms, to design better algorithms to play specific games, the analysis for the human-machine game system is necessary and urgent. However, to the best of our knowledge, research on such systems is not sufficient and related works are not much in the literature. Previously, [26,27] studies the optimal strategy against an opponent with finite memory and gives the theoretical results. Recently, [28,29] and [30] use myopic best response or look-ahead strategy to fight against the opponent which is approximated by Recurrent Neural Network (RNN). Also, [31] presents safe strategy and proposes an algorithm of exploiting sub-optimal opponents under the condition of ensuring safety, and [32] presents an exact algorithm in imperfect information games to exploit the opponent using the Dirichlet prior distribution.

In this paper we assume that the machine player uses the classical Fictitious Play (FP) algorithm to update its actions. FP algorithm was the first learning algorithm to achieve Nash equilibrium [33]. When the players adopt FP algorithm, convergence has been proved for repeated games with two players or zero-sum payoffs [11,34,35]. However, even for simple 3×3 general-sum game, the convergence does not hold [36]. This implies the complexity of the dynamical game systems driven by learning algorithms. So far, Fictitious Play and many variants have been studied, wherein stochastic FP [37] considers the perturbed payoff in the game and the players choose a distribution on the best response according to the private information about the perturbation, in weakened FP

[38] and generalised weakened FP [39] players take the $\epsilon-$best response as his action. Recently, [40] proposes the Full-Width Extensive Form Fictitious Self-Play (XSP) based on reinforcement learning and supervised learning for extensive form games, and [41] further proposes Neural Fictitious Self Play (NFSP) which uses neural networks to approximate the mapping of FP.

In [1] and [2], we have proved and solved the optimal strategy against the machine adopting FP under the assumption of complete and perfect information. In this paper, we will assume that some parameters in the algorithm is unknown to the human player and consider the identification problem. Specifically, we assume that the human player know that the machine adopts the FP algorithm but does not know the initial assessment of the algorithm. In order to get the near-optimal averaged utility over the infinite time, one natural idea for the human player is to identify the unknown parameters of the machine. Since the human can infer the inequality of the assessment parameters from the stage action of the machine, this seems very possible given enough probes. In this paper, we will give a simple and natural algorithm to identify the assessment-parameter in the FP algorithm and prove that the identification process will stop successfully in finite time if the machine's payoff parameter is a rational number. However, by an example, we will show that the identification process can not stop if the machine's payoff parameter is irrational. This finding implies some advantage of the learning algorithms with irrational parameters and may help design better algorithms.

The paper is organized as below: Sect. 2 gives the problem formulation; Sect. 3 gives the results when the machine's payoff parameter is rational and illustrates the case when the machine's payoff parameter is irrational by giving an example; Sect. 4 concludes the paper with some remarks and the future work.

2 Problem Formulation

Consider a 2×2 general-sum strategic-form game. Player 1 and Player 2 are called the machine player and the human player. The machine has two actions, denoted by A, B. The human has two actions, denoted by a, b. Thus there are 4 different possible outcomes (equally, the action profile) of the game: Aa, Ab, Ba, Bb. Given any outcome, the machine and the human have their individual utility $q_i, w_i, i = 1, 2, 3, 4$. We describe the game by the bi-matrix below.

Player 1	Player 2	
	a	b
A	q_1,w_1	q_2,w_2
B	q_3, w_3	q_4, w_4

The mixed strategy of the player is a probability distribution over the pure action set {A,B} or {a,b}.

Consider the repeated game. Denote the action of the machine and the human at time t by α_t^1, α_t^2. The machine player will choose its action α_t^1 according to the Fictitious Play (FP) algorithm, i.e.,

$$\alpha_t^1 = BR((\frac{\kappa_t(a)}{\kappa_t(a)+\kappa_t(b)}, \frac{\kappa_t(b)}{\kappa_t(a)+\kappa_t(b)})) \tag{1}$$

where the BR function denotes the best response of the machine player against his assessment $\kappa_t(a), \kappa_t(b)$ to the human's behavior and $\kappa_t(a), \kappa_t(b)$ are non-negative real numbers which are updated by

$$\kappa_t(i) = \kappa_{t-1}(i) + \begin{cases} 1, & if\, \alpha_{t-1}^2 = i; \\ 0, & if\, \alpha_{t-1}^2 \neq i. \end{cases} \tag{2}$$

where $i = a, b$. Here we let Player 1 choose A when both actions A, B are the best response of the machine player.

Obviously, once the initial assessments $\kappa_0(a), \kappa_0(b)$ are fixed, the updating rule of the machine is totally determined. Then how the system evolve will be determined by the human's action sequence. If the human takes his action sequence to be $\{\alpha_t^2\}, t = 1, 2, \ldots,$ then at each time t, the human will get an instantaneous utility $u_t = u_t(\alpha_t^1, \alpha_t^2) \in \{w_1, w_2, w_3, w_4\}$.

Define the averaged utility of the human over the infinite time to be

$$U_\infty = \limsup_{T\to\infty} \frac{\sum_{t=1}^T u_t(\alpha_t^1, \alpha_t^2)}{T}, \tag{3}$$

which always exists.

In [1] and [2], by assuming the complete and perfect information, we have solved the optimal strategy of the human player to get the optimal U_∞. Now we assume that the human does not know the initial assessment $(\kappa_0(a), \kappa_0(b))$, then how should the human do in order to get a bigger U_∞?

One natural idea for the human is to identify the initial parameter $(\kappa_0(a), \kappa_0(b)$. This seems possible since the human can get more information with the system running.

Before stating the related results, like we have done in the previous works, we rewrite the bi-matrix into a new one:

Player 1	Player 2	
	a	b
A	$q_3 + \Delta_1$,w_1	q_2,w_2
B	q_3, w_3	$q_2 + \Delta_2, w_4$

where $\Delta_1 > 0, \Delta_2 > 0$.

Then strategy updating rule of the machine is rewritten to be

$$\alpha_t^1 = \begin{cases} A, & if\, \Delta_1 \cdot \kappa_t(a) \geq \Delta_2 \cdot \kappa_t(b); \\ B, & otherwise. \end{cases}$$

3 The Identification for Parameters in the FP Algorithm

In [1] and [2], we have investigated the dynamical game systems in which the machine uses the FP algorithm to update its actions. We showed that the human's optimal strategy depends on the ratio $\frac{\Delta_2}{\Delta_1}$ and the long-run behavior of the system depends on $\frac{\Delta_2}{\Delta_1}$ being rational or irrational too.

By the explicit form of the human's optimal strategy [1,2], it is independent of the specific values of $\kappa_0(a)$ and $\kappa_0(b)$ but is solely dependent on the parameter

$$K_t = \Delta_1 \cdot \kappa_t(a) - \Delta_2 \cdot \kappa_t(b)$$

which can be computed by the initial assessment $(\kappa_0(a), \kappa_0(b))$ and the realized actions α_t^1, α_t^2. For example, if $w_2 > w_3 > max\{w_1, w_4\}$, the optimal strategy of the human is just the naive/myopic best response of the machine's action which can correctly predicted by the human. Denote the prediction of the human for the machine's action to be $\tilde{\alpha}_t^1$.

Now, assume that the machine's initial assessment $(\kappa_0(a), \kappa_0(b))$ is unknown to the human. Then for the human it is enough to identify the initial assessment parameter

$$K = \Delta_1 \cdot \kappa_0(a) - \Delta_2 \cdot \kappa_0(b)$$

in order to get his optimal strategy. This offers great convenience to the human compared to determining the precise values of $\kappa_0(a)$ and $\kappa_0(b)$.

On the other hand, according to Eq. (6), the machine takes actions according to an inequality of $(\kappa_0(a) + X_t(a), \kappa_0(b) + X_t(b))$, where $X_t(a)$ and $X_t(b)$ are the numbers of times at which the human player takes action a and action b up to time t (not included). Thus it is possible for the human to infer the feasible set of K from the machine's action.

Next we will give an algorithm to identify K. Since the system behavior is very different for rational and irrational $\frac{\Delta_2}{\Delta_1}$, we will also study the identification for rational and irrational $\frac{\Delta_2}{\Delta_1}$ respectively.

3.1 The Identification Algorithm for Assessment Parameter *K*

Now the goal of the human is to determine the value of the parameter $K = \Delta_1 \cdot \kappa_0(a) - \Delta_2 \cdot \kappa_0(b)$. We will take the game with the relationship $w_2 > w_3 > max\{w_1, w_4\}$ as a typical case to state the identification results. However, it is easy to see that the other cases share the same idea.

Denote the estimation of K by $\tilde{K}$. In the following, the estimation at each time in the identification process is denoted by $\tilde{K}_t$, and the subscript t is omitted when it does not lead to misunderstanding.

We give an identification algorithm as below:

Algorithm 1. Identify initial assessment parameter K

```
function F($\kappa_0(a), \kappa_0(b), M$)    ▷ Identify the initial evaluation under the game matrix M
    $K \leftarrow \Delta_1 \cdot \kappa_0(a) - \Delta_2 \cdot \kappa_0(b)$
    initial $\tilde{K}$
    for t do
        $\tilde{\alpha}_1(t) = BS(\tilde{K}, M, t, h)$    ▷ BS is the optimal strategy in [1] and [2]
        $\alpha_1(t) = BS(K, M, t, h)$
        if $\tilde{\alpha}_1(t) \neq \alpha_1(t)$ then    ▷ wrong forecast
            Update $\tilde{K}$    ▷ Update to a value that ensures the correct action before
        end if
        Update $h$    ▷ Update history action sequence
    end for
    return $\tilde{K}$    ▷ Output the final identification result
end function
```

By the identification algorithm, the estimation $\tilde{K}$ is updated as follows:

$$\tilde{K}_{t+1} = \begin{cases} \epsilon - \Delta_1 \cdot X_t(a) + \Delta_2 \cdot X_t(b), & if\, \alpha_t^1 = A, \tilde{\alpha}_t^1 = B, \\ -\epsilon - \Delta_1 \cdot X_t(a) + \Delta_2 \cdot X_t(b), & if\, \alpha_t^1 = B, \tilde{\alpha}_t^1 = A. \end{cases}$$

where $\epsilon > 0$ is small enough.

Then we have

Theorem 1. *Assume that $\frac{\Delta_2}{\Delta_1}$ is a rational number and the human player adopts the identification Algorithm 1 above. Then, for any initial identification value $\tilde{K}_0$, there exists a finite time t_f such that for all $t \geq t_f$, $\tilde{\alpha}_t^1 \equiv \alpha_t^1$, i.e., the human player can predict the machine's action correctly after t_f.*

Proof. When the human adopts the identification Algorithm 1, the evolution path of the system is definite, that is, the sequence $\{\alpha_t^1\}$, $\{\tilde{\alpha}_t^1\}$, $\{\alpha_t^2\}$, $X_t(a)$, $X_t(b)$ are determined. Denote $\eta_1 = \min_t\{K + f_t : K + f_t \geq 0\}$, $\eta_2 = \max_t\{K + f_t : K + f_t < 0\}$, where $f_t = \Delta_1 \cdot X_t(a) - \Delta_2 \cdot X_t(b))$.

We will prove this theorem in three steps.

Step 1: First, we prove that when $\tilde{K} \in [K - \eta_1, K + \eta_2)$, the human's prediction of the machine's action $\tilde{\alpha}_1^t$ can always be consistent with player 1's action α_1^t, i.e.,, $\tilde{\alpha}_t^1 \equiv \alpha_t^1$.

In this case, if $K + f_t \geq 0$, then $\tilde{K} + f_t \geq K - \eta_1 + f_t = K + f_t - \eta_1$, then from the definition of η_1 $\tilde{K} + f_t \geq 0$.

If $K + f_t < 0$, then $\tilde{K} + f_t < K + \eta_2 + f_t = K + f_t - \eta_2$, then from the definition of η_2, $\tilde{K} + f_t < 0$.

Step 2: Next, we prove that for any initial $\tilde{K}_0$, $\tilde{K}_t$ will enter $[K - \eta_1, K + \eta_2)$, and stop updating.

For the initial $\tilde{K}_0$, if $\tilde{K}_0 \in [K - \eta_1, K + \eta_2)$, then from the previous step, $\tilde{\alpha}_t^1 \equiv \alpha_t^1$, $\forall t \geq 0$. That is, the human will always predict correctly, so $\tilde{K}_t$ stops updating.

Suppose the human made a wrong prediction at time t_0, i.e., $\tilde{\alpha}^1_{t_0} \neq \alpha^1_{t_0}$. Without loss of generality, we can set $\alpha^1_{t_0} = B$, $\tilde{\alpha}^1_{t_0} = A$, which corresponds to $K + \Delta_1 \cdot X_{t_0}(a) - \Delta_2 \cdot X_{t_0}(b)) < 0$ and $\tilde{K}_{t_0} + \Delta_1 \cdot X_{t_0}(a) - \Delta_2 \cdot X_{t_0}(b)) \geq 0$.

Then according to the identification Algorithm 1,

$$\tilde{K}_{t_0+1} = -\epsilon - \Delta_1 \cdot X_{t_0}(a) + \Delta_2 \cdot X_{t_0}(b) = -\epsilon - f_{t_0}.$$

Then there are three situations to be discussed below.

(1) If $\tilde{\alpha}^1_t \equiv \alpha^1_t, \forall t \geq t_0 + 1$, i.e., the human has been predicting correctly after t_0, then the estimation stops updating.

(2) If the human still predicts wrongly after t_0, then denote $t_1 = \underset{t \geq t_0+1}{\arg\min}\{\tilde{\alpha}^1_t \neq \alpha^1_t\}$ to be the time of the next mistake. Then, in this case, at time $t_0+1, t_0+2, \ldots, t_1$, the human will not update the estimation, i.e., $\tilde{K}_{t_0+1} = \tilde{K}_{t_0+2} = \cdots = \tilde{K}_{t_1}$.

(2.1) If $\alpha^1_{t_1} = B$ and $\tilde{\alpha}^1_{t_1} = A$, which means $K + f_{t_1} < 0$ and $\tilde{K}_{t_1} + f_{t_1} \geq 0$, according to the identification Algorithm 1, $\tilde{K}_{t_1+1} = -\epsilon - f_{t_1}$.

In this case, we first prove that $f_{t_1} > f_{t_0}$. If not, then

$$\tilde{K}_{t_1} + f_{t_1} = \tilde{K}_{t_0+1} + f_{t_1} \leq \tilde{K}_{t_0+1} + f_{t_0} = -\epsilon < 0,$$

contradicts with $\tilde{K}_{t_1} + f_{t_1} \geq 0$. So it must hold $f_{t_1} > f_{t_0}$.

From $f_{t_1} > f_{t_0}$,

$$\tilde{K}_{t_1+1} = -\epsilon - f_{t_1} < -\epsilon - f_{t_0} = \tilde{K}_{t_0+1}.$$

And by calculation,

$$\tilde{K}_{t_0+1} - \tilde{K}_{t_1+1} = -f_{t_0} + f_{t_1} = -\Delta_1 \cdot (X_{t_0}(a) - X_{t_1}(a)) + \Delta_2 \cdot (X_{t_0}(b) - X_{t_1}(b)).$$

Define $\eta = \min_{m,n \in \mathrm{N}^+}\{|\Delta_1 \cdot m - \Delta_2 \cdot n| > 0\}$. By the rationality of $\frac{\Delta_2}{\Delta_1}$, η is a positive constant. So $\tilde{K}_{t_1+1} - \tilde{K}_{t_0+1} \leq -\eta$. That is, when the estimation $\tilde{K}$ is larger than K, the updated estimation will be smaller than the previous estimation by at least a positive constant.

(2.2) Assume that $\alpha^1_{t_1} = A$ and $\tilde{\alpha}^1_{t_1} = B$, that is, the prediction mistake at time t_1 is different from the prediction mistake at time t_0 and assume that there exists a future time $t_j, j \geq 2$ at which the prediction mistake is the same with the time t_0. Denote $t_j = \underset{t \geq t_0+1}{\arg\min}\{\alpha^1_t = B, \tilde{\alpha}^1_t = A\}$.

In this case, first of all, it holds that $K + f_{t_{j-1}} \geq 0$, i.e., $K \geq -f_{t_{j-1}}$. According to the identification Algorithm 1, $\tilde{K}_{t_{j-1}+1} = \epsilon - f_{t_{j-1}}$.

Meanwhile, at t_0, $\alpha^1_{t_0} = B$, which requires $K + f_{t_0} < 0$, so $K < -f_{t_0}$. Thus we get $-f_{t_{j-1}} < -f_{t_0}$, i.e.,, $f_{t_{j-1}} > f_{t_0}$. By the definition of η, it must hold $f_{t_{j-1}} \geq f_{t_0} + \eta$.

Since ϵ is small enough, $f_{t_{j-1}} > 2\epsilon + f_{t_0}$. Hence we have $\epsilon - f_{t_{j-1}} < -\epsilon - f_{t_0}$, that is, $\tilde{K}_{t_{j-1}+1} < \tilde{K}_{t_0+1}$.

On the other hand, at time t_j, $\alpha_t^1 = B$, $\tilde{\alpha}_t^1 = A$, which requires $K + f_{t_j} < 0$, $\tilde{K}_{t_j} + f_{t_j} \geq 0$. According to the identification Algorithm 1, it holds $\tilde{K}_{t_j+1} = -\epsilon - f_{t_j}$.

Now we aim to prove $f_{t_j} > f_{t_0}$. If not, it holds $f_{t_j} \leq f_{t_0}$. Then according to the inequality $\tilde{K}_{t_{j-1}+1} < \tilde{K}_{t_0+1}$ we just proved and the updating formula of $\tilde{K}_{t_0+1}$, we have

$$\tilde{K}_{t_j} + f_{t_j} = \tilde{K}_{t_{j-1}+1} + f_{t_j} < \tilde{K}_{t_0+1} + f_{t_0} = -\epsilon < 0,$$

which contradicts with $\tilde{K}_{t_j} + f_{t_j} \geq 0$. So we prove that $f_{t_j} > f_{t_0}$.

From $f_{t_j} > f_{t_0}$, according to the updating formula of $\tilde{K}_{t_j+1}$,

$$\tilde{K}_{t_j+1} = -\epsilon - f_{t_j} < -\epsilon - f_{t_0} = \tilde{K}_{t_0+1},$$

i.e., it holds $\tilde{K}_{t_j+1} < \tilde{K}_{t_0+1}$.

Then by the same way with in (2.1), we get $\tilde{K}_{t_0+1} - \tilde{K}_{t_j+1} \geq \eta$, which means that if the estimation value of the human is larger than K for more than once, i.e., the human will make the same prediction mistake at least twice, then there is a good property between the adjacently updated estimates, i.e.,, the updated estimation is smaller than the previous estimation by at least a positive constant.

(2.3) If $\alpha_{t_1}^1 = A$ and $\tilde{\alpha}_{t_1}^1 = B$, and there is no time $t_p > t_1$ making $\alpha_{t_p}^1 = B$, $\tilde{\alpha}_{t_p}^1 = A$, i.e., the human will never make the same mistake as at time t_0 after time t_0, then it is only necessary to analyze the prediction mistakes corresponding to $\alpha_t^1 = A$ and $\tilde{\alpha}_t^1 = B$. And this analysis is symmetric with the above.

To sum up, if the estimation is not "sufficiently" correct, then the human must make a mistake at some time, so the updated value of $\tilde{K}_t$ will move towards the correct direction at a speed greater than a positive constant until the estimation is "sufficiently" correct. And then the human will never make mistakes. Obviously this process will end after only a finite time.

That proves the theorem. ■

Remark 1: When $t \geq t_f$, although the human's prediction is always correct, $\tilde{K}$ and K can still be different. This is because that the FP algorithm only requires the inequality of K holds.

Remark 2: Through the proof of Theorem 1, it can be computed that after at most $\lceil \frac{\tilde{K}_{max} - \tilde{K}_{min}}{\eta} \rceil + 2$ updates, $\tilde{K}_t$ will be "sufficiently" correct. If the evaluation range $[\tilde{K}_{min}, \tilde{K}_{max}]$ about K is given at the initial time, then the initial estimation $\tilde{K}_0$ can be set to be any number in the interval.

Below we will give an example to illustrate how the identification is carried out.

Consider the following game:

Player 1	Player 2	
	a	b
A	2, 2	1, 5
B	0, 4	4, 3

where $\varDelta_1 = 2, \varDelta_2 = 3$.

Assuming the machine's initial assessment of the human is $\kappa_0(a) = 5$, $\kappa_0(b) = 4$, then $K = \varDelta_1 \cdot \kappa_0(a) - \varDelta_2 \cdot \kappa_0(b) = -2$. Let the initial estimation be $\tilde{K}_0 = 1$ and take $\epsilon = 0.1$. Then under the identification Algorithm 1, the evolution of the game system is as follows in Table 1, where the values in () represent the values of K and $\tilde{K}$ respectively at the current moment.

Table 1. The evolution process of system parameters under the Algorithm 1

t	$K + \varDelta_1 X_t(a) - \varDelta_2 X_t(b)$	$\widetilde{K}_t + \varDelta_1 X_t(a) - \varDelta_2 X_t(\text{ b})$	α_t^1	$\tilde{\alpha}_t^1$	α_t^2
0	$(-2) + 0$	$(1) + 0$	B	A	b
1	$(-2) - 3$	(-0.1) -3	B	B	a
2	$(-2) - 1$	$(-0.1) - 1$	B	B	a
3	$(-2) + 1$	$(-0.1) + 1$	B	A	b
4	$(-2) - 2$	(-1.1) -2	B	B	a
5	$(-2) + 0$	$(-1.1) + 0$	B	B	a
6	$(-2) + 2$	$(-1.1) + 2$	A	A	b
7	$(-2) - 1$	$(-1.1) - 1$	B	B	a
8	$(-2) + 1$	$(-1.1) + 1$	B	B	a
9	$(-2) + 3$	$(-1.1) + 3$	A	A	b
10	$(-2) + 0$	$(-1.1) + 0$	B	B	a
11	$(-2) + 2$	$(-1.1) + 2$	A	A	b
12	$(-2) - 1$	$(-1.1) - 1$	B	B	a
13	$(-2) + 1$	$(-1.1) + 1$	B	B	a
14	$(-2) + 3$	$(-1.1) + 3$	A	A	b

The results in Table 1 are represented by Figs. 1 and 2 as follows.

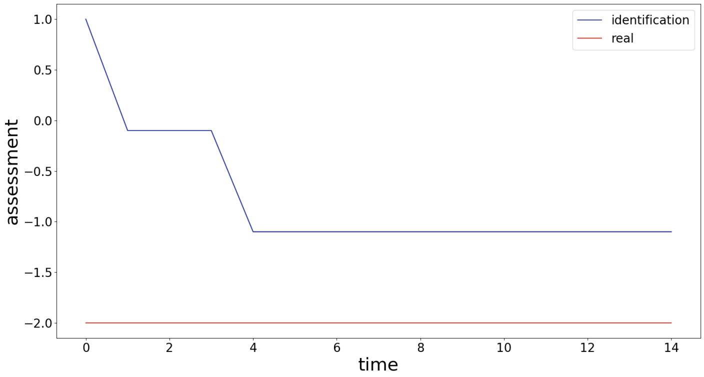

Fig. 1. The change of $\tilde{K}$ along time t

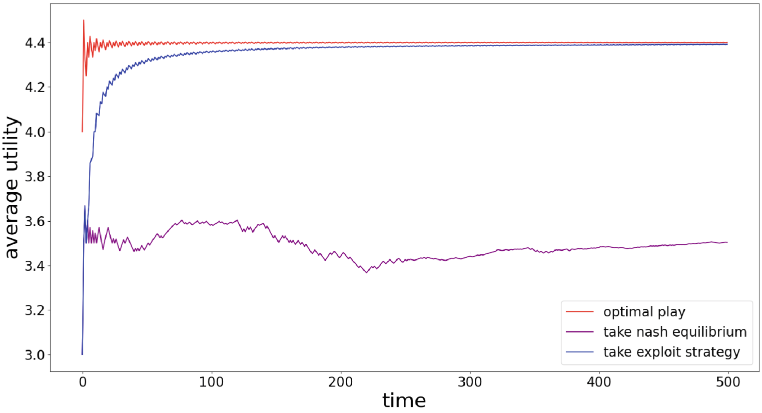

Fig. 2. The human's averaged utility along time t

As can be seen from the Table 1 and Figs. 1 and 2, if the prediction of the machine's action at time t is wrong, the estimation $\tilde{K}$ will be updated at time $t+1$. In this example, after $t \geq 5$, the predictions are accurate, implying that the estimation is "sufficiently" accurate. By "sufficient" accuracy, we find that the precise identification of $\tilde{K}$ is not necessary and it suffices for $\tilde{K}$ to approximate the value of K "closely".

Theorem 1 gives the result on identification for rational $\frac{\Delta_2}{\Delta_1}$. When $\frac{\Delta_2}{\Delta_1}$ is irrational, the situation will be different as shown in the next subsection.

3.2 The Identification for Irrational $\frac{\Delta_2}{\Delta_1}$

For the irrational $\frac{\Delta_2}{\Delta_1}$, consider the following game matrix:

Player 1	Player 2	
	a	b
A	$\sqrt{2}$, 2	1, 5
B	0, 4	4, 3

where $\Delta_1 = \sqrt{2}$, $\Delta_2 = 3$.

Assume Player 1's initial assessment of Player 2 is $\kappa_0(a) = 5$, $\kappa_0(b) = 4$, then $K = \Delta_1 \cdot \kappa_0(a) - \Delta_2 \cdot \kappa_0(b) = -2$. Let the initial estimate be $\tilde{K}_0 = 1$, take $\epsilon = 10^{-9}$.

Then the estimation $\tilde{K}_t$ changes with time as shown in Fig. 3 below where the vertical ordinate being 1 means that the estimation is updated at this moment.

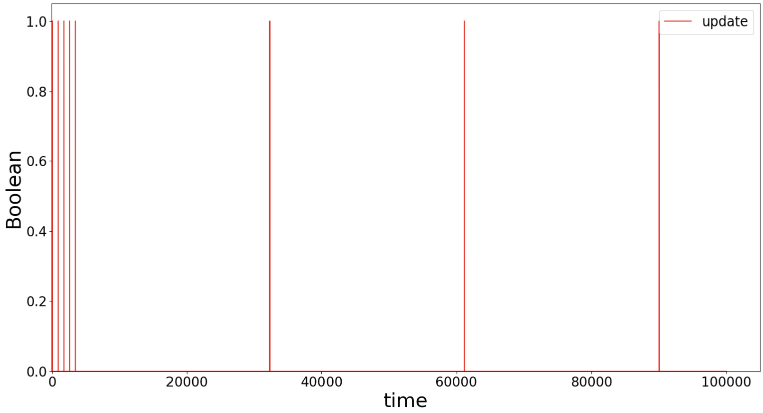

Fig. 3. Estimated value $\tilde{K}_t$ over time

From Fig. 3, we can see that for the irrational $\frac{\Delta_2}{\Delta_1}$, the identification will not end in any finite time. Along with time, the estimation error becomes more and more smaller. However, since in the FP algorithm, the parameter K is irrational, the inequality will never stop changing its signs. Thus for almost all the initial estimation and games (with measure 1), the identification process will never stop and the human will never get the total prediction of the machine. This might help us to design better algorithms.

4 Conclusions and Future Work

In this paper, we considered the repeated human-machine games where the machine uses the Fictitious Play algorithm to update its action at eat time. In the previous works [1] and [2], we solved the optimal strategy of the human player against the machine under assumption of complete and perfect information. In this paper, we assume that the human player does not know the initial assessment of the machine player and consider the identification problem of the human. We propose an identification algorithm for the human player and prove that the identification can end successfully in a finite time if the machine's payoff parameter is rational. When the machine's payoff parameter is irrational, the identification process will not end, which implies some advantage for the algorithm with irrational parameters. The results in this paper are rigorous and might shed some light on general games and algorithms.

This paper can be regarded as the first step to solve the problem of algorithm identification, which is a necessary part to exploit an algorithm in repeated games in the future application of some AI. The algorithm can also be regarded as an approximation of the real human behavior, thus the algorithm identification is the necessary step to find the pattern of the opponent's behavior. We will leave these general problems as future work.

Acknowledgement. This work was supported by the National Key Research and Development Program of China under grant No.2022YFA1004600, the Natural Science Foundation of China under Grant T2293770, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000, the Major Project on New Generation of Artificial Intelligence from the Ministry of Science and Technology (MOST) of China under Grant No. 2018AAA0101002.

References

1. Dong, H., Mu, Y.: The optimal strategy against fictitious Play in infinitely repeated games. In: Proceedings of the 41st Chinese Control Conference, pp. 6852–6857 (2022)
2. Dong, H., Mu, Y.F.: The optimal strategy against the opponent adopting fictitious play algorithm in infinitely repeated 2×2 games. SSRN Electron. J. (2022). https://doi.org/10.2139/ssrn.4201849
3. Silver, D., Huang, A., Maddison, C.J., et al.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
4. Silver, D., Hunert, T., Schrittwieser, J., et al.: A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science **362**(6419), 1140–1144 (2018)
5. Moravčík, M., Mchmid, M., Burch, N., et al.: DeepStack: expert-level artificial intelligence in heads-up no-limit poker. Science **356**(6337), 508–513 (2017)
6. Brown, N., Sandholm, T.: Superhuman AI for heads-up no-limit poker: libratus beats top professionals. Science **359**(6374), 418–424 (2017)
7. Brown, N., Sandholm, T.: Superhuman AI for multiplayer poker. Science **365**(6456), 885–890 (2019)
8. Vinyals, O., Babuschkin, I., Czarnecki, W.M., et al.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)
9. Bouzy, B., Métivier, M., Pellier, D.: Hedging algorithms and repeated matrix games. arXiv preprint arXiv:1810.06443 (2018)
10. Brown, G.W.: Some Notes on Computation of Games Solutions. RAND Corp., Santa Monica (1949)
11. Robinson, J.: An iterative method of solving a game. Ann. Math., 296–301 (1951)
12. Monderer, D., Sela, A.: A 2 × 2 game without the fictitious play property. Games Econ. Behav. **14**(1), 144–148 (1996)
13. Monderer, D., Shapley, L.S.: Fictitious play property for games with identical interests. J. Econ. Theory **68**(1), 258–265 (1996)
14. Christian, E., Valkanova, K.: Fictitious play in networks. Games Econ. Behav. **123**, 182–206 (2020)
15. Fudenberg, D., Drew, F., Levine, D.K., et al.: The Theory of Learning in Games. MIT press, Cambridge (1998)
16. Yuan, S., Guo, L.: Stochastic adaptive dynamical games. Sci China Math **46**, 1367–1382 (2016)
17. Hu, H.Y., Guo, L.: Non-cooperative stochastic adaptive multi-player games. Control Theory Appl. **35**(5) (2018)
18. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Machine Learning Proceedings, pp. 157–163. Morgan Kaufmann (1994)
19. Hu, J., Wellman, M.P.: Nash Q-learning for general-sum stochastic games. J. Mach. Learn. Res. **4**, 1039–1069 (2003)

20. Szegedy, C., et al.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (2014)
21. Nguyen, A., et al.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436 (2015)
22. Finlayson, S.G., et al.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
23. Sharif, M., et al.: Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1528–1540 (2016)
24. Shamma, J.S.: Game theory, learning, and control systems. Natl. Sci. Rev. **7**(7), 1118–1119 (2020)
25. Cao, M.: Merging game theory and control theory in the era of AI and autonomy. Natl. Sci. Rev. **7**(7), 1122–1124 (2020)
26. Mu, Y., Guo, L.: Towards a theory of game-based non-equilibrium control systems. J. Syst. Sci. Complex. **25**(2), 209–226 (2012)
27. Mu, Y., Guo, L.: Optimization and identification in a non-equilibrium dynamic game. In: The 48th IEEE Conference on Decision and Control, Shanghai, China, pp. 5750–5755 (2009)
28. Deng, X., et al.: Exploiting a no-regret opponent in repeated zero-sum games, personal communication
29. Tang, Z., Zhu, Y., Zhao, D., et al.: Enhanced rolling horizon evolution algorithm with opponent model learning. IEEE Trans. Games (2020)
30. Deng, Y., Schneider, J., Sivan, B.: Strategizing against no-regret learners. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
31. Ganzfried, S., Sandholm, T.: Safe opponent exploitation. ACM Trans. Econ. Comput. **3**(2), 1–28 (2015)
32. Ganzfried, S., Sun, Q.: Bayesian opponent exploitation in imperfect-information games. In: 2018 IEEE Conference on Computational Intelligence and Games, pp. 1–8. IEEE (2018)
33. Brown, G.W.: Iterative solution of games by fictitious play. Act. Anal. Prod. Allocat. **13**(1), 374–376 (1951)
34. Miyasawa, K.: On the convergence of learning processes in a 2 × 2 non-zero-person game, Technical Report Research Memorandum No. 33, Econometric Research Program, Princeton University
35. Sayin, M.O., Parise, F., Ozdaglar, A.: Fictitious play in zero-sum stochastic games. SIAM J. Control. Optim. **60**(4), 2095–2114 (2022)
36. Shapley, L.: Some topics in two-person games. Adv. Game Theory **52**, 1–29 (1964)
37. Fudenberg, D., Kreps, D.M.: Learning mixed equilibria. Games Econ. Behav. **5**(3), 320–367 (1993)
38. Van der Genugten, B.: A weakened form of fictitious play in two-person zero-sum games. Int. Game Theory Rev. **2**(04), 307–328 (2000)
39. Leslie, D.S., Collins, E.J.: Generalised weakened fictitious play. Games Econ. Behav. **56**(2), 285–298 (2006)
40. Heinrich, J., Lanctot, M., Silver, D.: Fictitious self-play in extensive-form games. In: International Conference on Machine Learning. PMLR (2015)
41. Heinrich, J., Silver, D.: Deep reinforcement learning from self-play in imperfect-information games. arXiv:1603.01121 (2016)