FROM WEAK LEARNING TO STRONG LEARNING IN FICTITIOUS PLAY TYPE ALGORITHMS

BRIAN SWENSON[†⋆], SOUMMYA KAR[†] AND JOÃO XAVIER[⋆*]

Abstract. The paper studies the highly prototypical Fictitious Play (FP) algorithm, as well as a broad class of learning processes based on best-response dynamics, that we refer to as FP-type algorithms. A well-known shortcoming of FP is that, while players may learn an equilibrium strategy in some abstract sense, there are no guarantees that the period-by-period strategies generated by the algorithm actually converge to equilibrium themselves. This issue is fundamentally related to the discontinuous nature of the best response correspondence and is inherited by many FP-type algorithms. Not only does it cause problems in the interpretation of such algorithms as a mechanism for economic and social learning, but it also greatly diminishes the practical value of these algorithms for use in distributed control. We refer to forms of learning in which players learn equilibria in some abstract sense only (to be defined more precisely in the paper) as weak learning, and we refer to forms of learning where players' period-by-period strategies converge to equilibrium as strong learning. An approach is presented for modifying an FP-type algorithm that achieves weak learning in order to construct a variant that achieves strong learning. Theoretical convergence results are proved.

Key words. game-theoretic learning, repeated play, fictitious play, strong convergence

1. Introduction. Fictitious Play (FP), introduced in [1], is one of the oldest and best-known game theoretic learning algorithms. FP has been shown to be an effective algorithm for distributed learning of Nash equilibria in various classes of games including two-player zero-sum games [2], generic $2 \times m$ games [3], supermodular games [4, 5], one-against-all games [6], and potential games [7, 8]. However, the manner in which players *learn* in FP is often unsatisfactory, especially in the context of distributed control.

In FP, players learn equilibrium strategies in the sense that the time-averaged empirical distribution of players' actions converges to the set of Nash equilibria — a form of learning known as *convergence in empirical distribution*. This notion of learning tends to be problematic when the limit set of a learning algorithm contains mixed-strategy equilibria. In particular, convergence of the time-averaged empirical distribution to a mixed-strategy equilibrium does not imply any form of convergence in players' period-by-period strategies or actions. In practice, players' period-by-period strategies tend to move in progressively longer and longer cycles around an equilibrium set—the time-averaged empirical distribution is driven to equilibrium, but the period-by-period strategies never approach the equilibrium set themselves.

In the context of repeated-play algorithms, we refer to convergence of the empirical distribution (or some function thereof) to an equilibrium set as weak convergence, and we refer to any form of learning involving weak convergence as weak learning. We refer to the convergence of players' period-by-period strategies to an equilibrium set

*The work was partially supported by the FCT project FCT [UID/EEA/50009/2013] through the Carnegie-Mellon/Portugal Program managed by ICTI from FCT and by FCT Grant CMU-PT/SIA/0026/2009, and was partially supported by NSF grant ECCS-1306128.

†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA (brianswe@andrew.cmu.edu and soummyak@andrew.cmu.edu).

⋆Institute for Systems and Robotics (ISR/IST), LARSyS, Instituto Superior Técnico, Portugal (jxavier@isr.ist.utl.pt).

as strong convergence, and we refer to any form of learning involving strong convergence as strong learning. Intuitively speaking, weak learning means that players learn an equilibrium strategy in some abstract sense (i.e., convergence in empirical distribution) but may never actually implement the strategy they are learning. In strong learning, not only do players *learn* an equilibrium strategy, but they also implement it.

FP is proven to achieve learning only in the weak sense, and thus no guarantees can be made regarding the convergence nor optimality of players period-by-period strategies. For example, Jordan [9] presents a continuum of games for which FP achieves weak learning, yet in all but a countable subset of games, the period-by-period strategies produced by FP never approach the game's unique equilibrium. As another example, Young [10] presents a 2×2 game in which FP achieves weak learning, but the period-by-period actions produced by FP achieve the lowest possible utility in every stage of the repeated play (see also Section 3.2).

Our first main contribution is the presentation of a simple variant of FP that converges strongly to equilibrium. In our strongly convergent variant of FP, players gradually and independently transition from using the FP best response rule to determine the next-iteration action, to using their current empirical distribution as a probability mass function from which they sample to determine the next-iteration action. We show that, for any game in which FP can be shown to converge weakly to equilibrium (and for which a certain robustness assumption holds—see **A.8**), our variant of FP will converge strongly to equilibrium.

One advantage of this approach is that it is readily applicable to more general FP-type learning algorithms. Our second (and more general) main contribution is a method for taking a weakly convergent FP-type learning algorithm, and constructing from it, a strongly convergent variant. We study a general class of FP-type algorithms and show that, so long as an algorithm achieves weak learning in a sufficiently robust sense (see **A.8**), then a strongly convergent variant of the algorithm can be constructed. As an example of how the general result may be applied, we consider three weakly convergent FP-type algorithms—classical FP, Generalized Weakened FP [11], and Empirical Centroid FP [12, 13]—and construct the strongly convergent variant of each.

1.1. Related Work. An overview of the topic of learning in games can be found in [10, 14]. Various problems associated with learning mixed-strategy equilibria in best-response-type learning algorithms (including FP-type algorithms) are discussed in [9]. In particular, the issue of weak convergence is considered, along with a discussion of some of the underlying mechanics that lead to weak convergence.

Many learning algorithms are designed to ensure that their limit points are pure-strategy equilibria [15–19]. Ensuring convergence to a pure strategy is a natural way of ensuring strong learning, since weak learning can generally only occur when the limit set contains mixed strategies.

In contrast, this paper studies a method of ensuring strong convergence when the limit set of the algorithm contains mixed strategies. The ability to (strongly) learn mixed equilibria is important for many reasons, the foremost being that, in finite games, the set of Nash equilibria (NE) is only guaranteed to be non-empty if mixed equilibria are considered. Mixed strategies play an important role when the learned strategy needs to be robust to uncertainty in opponent behavior or game structure, or secure against the actions of malicious players [6, 20–23]. With regards to FP in particular, it was recently shown in [24] that, for the class of near-potential games,

the limit set of the FP dynamics (weakly speaking) is a neighborhood of a mixed equilibrium.

Regret-testing algorithms [25], [26] achieve strong convergence to mixed-strategy equilibria in generic finite games. However, such algorithms operate on fundamentally different principles from FP-type algorithms—players implement a form of exhaustive search to coordinate on a NE strategy. Such algorithms tend to have slow convergence rates, especially when the number of players or available actions is large.

Stochastic FP (SFP)—introduced in [27]—was proposed as a learning mechanism that could (i) mitigate the problem of weak convergence to mixed equilibria in FP and (ii) provide a reasonable explanation for why real-world players might learn mixed-strategy equilibria. In SFP, the issue of weak convergence is addressed by smoothing each player's best response correspondence with the addition of small random shocks or perturbations. The stable points of SFP are not Nash equilibria, but rather Nash distributions. The set of Nash distributions converges to the set of Nash equilibria as the size of the perturbations goes to zero [27]. SFP has been shown to obtain strong convergence to the set of Nash distributions in various classes of games [8, 14, 28]. Moreover, if the perturbations are permitted to gradually decay throughout the course of the repeated play, then SFP converges to the set of NE [11].

In contrast to SFP, the present work does not consider the descriptive agenda of providing an explanation for why real-world learners might act according to a given behavior rule. Furthermore, we present a simple and intuitive procedure for modifying a variety of weakly convergent learning algorithms in order to obtain a strong convergent variant. From a technical perspective, the current work differs from SFP in that the best response correspondence is not directly smoothed in any way.

The work [11] by Leslie et al. studies a useful generalization of FP termed Generalized Weakened FP (GWFP). Among other contributions, the paper demonstrates that the convergence of FP is not affected by asymptotically decaying perturbations to players' best response sets. This result provides a cornerstone for our proofs by ensuring that FP (and GWFP) meet the critical robustness assumption **A.8**. We study a strongly convergent variant of GWFP in Section 6.2. Furthermore, [11] also presents a payoff-based, actor-critic learning algorithm based on GWFP that achieves strong learning. Our work differs from this in that we provide a general method for constructing a strongly convergent algorithm from a weakly convergent one in a setting where instantaneous payoffs information may or may not be available.

Our preliminary results on strong convergence in FP is found in [29]. The present work expands on [29] by considering algorithms beyond classical FP and establishing more general conditions under which convergence can be attained (in particular, see **A.1**–**A.3**). Furthermore, [29] contains a gap in reasoning in the proof of Lemma 2 which the present paper fills in.

The remainder of the paper is organized as follows. Section 2 sets up notation to be used in the subsequent development. Section 3 introduces classical FP and discusses the problem of weak convergence in classical FP. Section 4 presents the strongly convergent variant of classical FP and states the strong convergence theorem for classical FP. Section 5 presents the general notion of an FP-type algorithm, then presents the strongly convergent variant of an FP-type algorithm, states the general strong convergence result in the context of an FP-type algorithm, and presents the proof of the result. In Section 6, the general result is applied to prove strong convergence in classical FP, Generalized Weakened FP, and Empirical Centroid FP. Section

7 concludes the paper.

2. Preliminaries.

2.1. Setup and Notation. A game in normal form is represented by the triple $\Gamma := (N, (Y_i, u_i)_{i\in N})$, where $N = \{1, \ldots, n\}$ denotes the set of players, Y_i denotes the finite set of actions available to player i, and $u_i : \prod_{i\in N} Y_i \to \mathbb{R}$ denotes the utility function of player i. Denote by $Y := \prod_{i\in N} Y_i$ the joint action space.

In order to guarantee the existence of Nash equilibria it is necessary to consider the mixed extension of Γ in which players are permitted to play probabilistic strategies. Let $m_i := |Y_i|$ be the cardinality of the action space of player i, and let $\Delta_i := \{p \in \mathbb{R}^{m_i} : \sum_{k=1}^{m_i} p(k) = 1,\ p(k) \geq 0\ \forall k\}$ denote the set of mixed strategies available to player i—note that a mixed strategy is probability distribution over the action space of player i. Denote by $\Delta^n := \prod_{i\in N} \Delta_i$, the set of joint mixed strategies.

In this context, we often wish to retain the notion of playing a deterministic action. For this purpose, let $A_i := \{e_1, \ldots, e_{m_i}\}$ denote the set of "pure strategies" of player i, where e_j is the j-th cannonical vector containing a 1 at position j and zeros otherwise.

The mixed utility function of player i is given by $U_i(p) := \sum_{y\in Y} u_i(y)p_1(y)\ldots p_n(y)$, where $U_i : \Delta^n \to \mathbb{R}$. When convenient we sometimes write $U_i(p)$ as $U_i(p_i, p_{-i})$, where p_i denotes the mixed strategy of player i and p_{-i} denotes the mixed strategies of all other players. The set of Nash equilibria is given by $NE := \{p \in \Delta^n : U_i(p_i, p_{-i}) \geq U_i(p_i', p_{-i}),\ \forall p_i' \in \Delta_i,\ \forall i \in N\}$. Let

$$BR_i^\epsilon(p_{-i}) := \{a_i \in A_i : U(a_i, p_{-i}) \geq \max_{\alpha_i \in A_i} U(\alpha_i, p_{-i}) - \epsilon\} \tag{2.1}$$

be the i-th players set of ϵ-best responses to a strategy profile p_{-i} adopted by the other players. Note that in this definition we only consider pure-strategy ϵ-best responses. Denote by $v_i(p_{-i}) := \max_{p_i\in\Delta_i} U_i(p_i, p_{-i})$, the value obtained by playing a best response.

Throughout, we assume there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ rich enough to carry out the construction of the various random variables required in this paper. For a random object X defined on a measurable space $(\Omega, \mathcal{F})$, let $\sigma(X)$ denote the σ-algebra generated by X [30]. As a matter of convention, all equalities and inequalities involving random objects are to be interpreted almost surely (a.s.) with respect to the underlying probability measure, unless otherwise stated.

2.2. Repeated Play. Suppose players repeatedly face off in the game Γ. Denote by $t \in \{1, 2, \ldots\}$ a round of the repeated play. Let $\{a_i(t)\}_{t\geq 1}$ denote the sequence of actions taken by player i, where $a_i(t) \in A_i$, and let $\{a(t)\}_{t\geq 1}$, $a(t) = (a_1(t), \ldots, a_n(t))$ denote the sequence of joint actions.

Let $\{\mathcal{F}_t\}_{t\geq 1}$ be a filtration (sequence of σ-algebras) that contains the information available to players in round t of the repeated play. For $t \geq 1$ and $\alpha_i \in A_i$, let $g(\alpha_i,\ t) \in \mathbb{R}$ be an $\mathcal{F}_{t-1}$-measurable random variable with $g_i(\alpha_i,\ t) := \mathbb{P}(a_i(t) = \alpha_i|\mathcal{F}_{t-1})$, and let $g_i(t) \in \Delta_i$ be the vector with components $g_i(t) := (g_i(\alpha_1,\ t), \ldots, g_i(\alpha_{m_i},\ t))$, where m_i is the cardinality of A_i. We say $g_i(t)$ is the mixed strategy used by player i in round t, and we say $\{g_i(t)\}_{t\geq}$ is the sequence of period-by-period (mixed) strategies used by player i. The sequence of joint period-by-period strategies is given by $\{g(t)\}_{t\geq 1}$, $g(t) := (g_1(t), \ldots, g_n(t))$.

Denote by $q_i(t) \in \Delta_i$, the empirical distribution of player i. The precise manner

in which the empirical distribution[1] is formed will depend on the algorithm at hand. In general, $q_i(t)$ is formed as a function of the action history $\{a_i(s)\}_{s=1}^{t}$ and serves as a compact representation of the action history of player i up to and including the round t. The joint empirical distribution is given by $q(t) := (q_1(t), \ldots, q_n(t))$.

Unless otherwise stated, $d(\cdot,\ \cdot)$ denotes the standard Euclidean norm. For $m \geq 1$ and $S \subset \mathbb{R}^m$ define the distance from $p \in \mathbb{R}^m$ to $S \subset \mathbb{R}^m$ by $d(p,\ S) := \inf\{d(p,\ p') :\ p' \in S\}$. We say a repeated-play learning process converges *weakly* to equilibrium if for some map $f : \Delta^n \to \Delta^n$ there holds $d(f(q(t)),\ NE) \to 0$ as $t \to \infty$. In most cases in this paper, f will simply be the identity function. We say a repeated-play learning process converges *strongly*[2] to equilibrium if $d(g(t),\ NE) \to 0$ as $t \to \infty$. Note that weak learning implies that players *learn* an equilibrium strategy, but may never actually begin to implement the strategy that is being learned. On the other hand, in strong learning players both *learn* an equilibrium strategy, and implement the strategy that is being learned (see Section 3.2 for more details).

3. Fictitious Play.

3.1. Fictitious Play. Let

$$q_i(t) := \frac{1}{t}\sum_{s=1}^{t} a_i(s), \tag{3.1}$$

be the normalized histogram[3] of the actions of player i.

FP may be intuitively understood as follows. Players repeatedly face off in a stage game Γ. In any given stage of the game, players choose a next-stage action by assuming (perhaps incorrectly) that opponents are using stationary and independent strategies. Thus, in FP, players use the marginal empirical distribution of each opponent's past play, $q_i(t)$, as a prediction of the opponent's behavior in the upcoming round and choose a next-round strategy which is a best response against this prediction.

A sequence of actions $\{a(t)\}_{t\geq 1}$ such that[4]

$$a_i(t+1) \in BR_i(q_{-i}(t)),\ \forall i, \tag{3.2}$$

for all $t \geq 1$, is referred to as a *fictitious play process*. FP has been studied extensively to determine the classes of games for which it can be said to converge (weakly) to the set of Nash equilibria. Among other results, it has been shown that FP leads to weak learning in two-player zero-sum games [2], potential games [7], and generic $2 \times m$ games [3]. We summarize these results in the following theorem.

THEOREM 3.1. *Let $\Gamma = (N, \{u_i(\cdot)\}_{i\in N}, Y^n)$ be a two-player zero-sum game, potential game, or generic $2 \times m$ game, and let $\{a(t)\}_{t\geq 1}$ be a fictitious play process on Γ. Then $d(q(t),\ NE) \to 0$ as $t \to \infty$.*

3.2. Weak Convergence in Fictitious Play. The following example (see [10], p. 78), while fairly simple, clearly illustrates the phenomenon of weak convergence in

[1]The term *empirical distribution* is often used to refer explicitly to the time-averaged histogram of the action choices of some player i; i.e., $q_i(t) = \frac{1}{t}\sum_{s=1}^{t} a_i(s)$. Here, we allow for a broader definition that will permit interesting and useful algorithmic generalizations.

[2]The notion of strong convergence presented in this paper is comparable to the notions of "convergence in intended behavior" presented in [27] and "convergence in strategic intentions" given in [10].

[3]Recall that the actions $a_i(t) \in A_i$ are dirac distributions in the mixed-strategy space Δ_i.

[4]In all variants of FP discussed in this paper, the initial action $a_i(1)$ may be chosen arbitrarily for all i.

FP, and demonstrates why weak convergence can be a deeply unsatisfactory notion of learning.

	A	B
A	$\sqrt{2}, 1$	$0, 0$
B	$0, 0$	$1, \sqrt{2}$

Fig. 3.1

Consider the two-player asymmetric coordination game shown in Figure 3.1. The game has three Nash equilibria: both players play A, both players play B, and an asymmetric mixed-strategy Nash equilibrium. The game is a potential game [7] (in fact, an identical interests game [31]) and hence falls within the purview of Theorem 3.1—regardless of the initial conditions, players engaged in an FP process will learn an equilibrium in the weak sense that $d(q(t),\ NE) \to 0$ as $t \to \infty$.

Suppose that the players are engaged in an FP process on this game, and in the first round they miscoordinate their actions (e.g., one chooses A, and the other chooses B). Young [10] shows the somewhat counterintuitive result that the FP dynamics will in fact lead players to miscoordinate their action choices in every subsequent round of the learning process. Thus, despite the fact that $\lim_{t\to\infty} d(q(t),\ NE) = 0$, the players' realized action choices are extremely suboptimal—yielding the lowest possible utility in each round of play. Intuitively speaking, this phenomenon occurs when players' actions cycle in such a way as to drive the time-averaged empirical distribution to a mixed-strategy Nash equilibrium, yet player's period-by-period strategies never constitute (nor even approach) a Nash equilibrium themselves.

It may be said that in weak learning players "learn" a NE strategy in some abstract sense, but never actually implement the strategy they are learning. In strong learning, players not only learn a NE strategy, but they also physically implement the strategy that is being learned.

The following section presents a simple modification of FP that achieves strong learning; i.e., players' period-by-period strategies converge to equilibrium in addition to convergence of the empirical distributions.

4. Strong Convergence in Classical Fictitious Play. Consider a variant of FP in which the action for player i at time t is chosen by drawing a random sample from the mixed strategy (i.e., probability distribution) $g_i(t)$, where

$$g_i(t) \in BR_i(q_{-i}(t-1))\rho_i(t) + q_i(t-1)(1-\rho_i(t)), \tag{4.1}$$

$\rho_i(t) \in [0,1]$, and $\lim_{t\to\infty} \rho_i(t) = 0$. Intuitively, this is similar to the classical FP process (3.2), but rather than playing a deliberate best response each round, players gradually transition toward drawing their stage t action as a random sample from their own empirical distribution, $q_i(t)$.

The idea is that players will play a best response sufficiently often so that, per FP, the empirical distribution $q(t)$ will be driven toward equilibrium, as in Theorem 3.1. Then, since $\rho_i(t) \to 0$ as $t \to \infty$, the mixed strategy $g_i(t)$ tends towards $q_i(t)$, which is itself tending towards equilibrium. Informally, (4.1) captures the main idea of strongly convergent FP. A formal presentation of the algorithm is given below.

4.1. Strongly Convergent Variant of Classical FP. Consider a variant of FP in which the action for player i at time t is chosen according to the following randomized rule:

$$a_i(t) \sim g_i'(t) := \begin{cases} b_i(t-1), & \text{if } X_i(t) = 1, \\ q_i(t-1), & \text{otherwise,} \end{cases} \tag{4.2}$$

where $b_i(t-1) \in BR_i(q_{-i}(t-1))$, the notation $a_i(t) \sim g_i'(t)$ indicates that the action $a_i(t)$ is drawn as a random sample[5] from the probability mass function $g_i'(t)$, $X_i(t) \in \{0,1\}$ is a random variable, and $q_i(t)$ is the player's empirical distribution as defined in (4.4) below. Let $\mathcal{F}_t := \sigma(\{a(s), X_1(s), \ldots, X_n(s), b_1(s), \ldots, b_n(s)\}_{s \leq t})$, and note that $g_i'(t)$ is $\mathcal{F}_t$-measurable. Let

$$\rho_i(t) := \mathbb{P}(X_i(t) = 1 |\ \mathcal{F}_{t-1}),$$

and note that $\rho_i(t)$ is $\mathcal{F}_{t-1}$-measurable. Intuitively speaking, $\rho_i(t)$ represents the probability that player i deliberately chooses to play a best response strategy in round t given the history of play up through the previous round. We make the following assumptions regarding each player's probability of deliberately choosing a best response:

A. 1. $\lim_{t\to\infty} \rho_i(t) = 0,\ \forall i \in N,\ a.s.,$

A. 2. $\sum_{t\geq 1} \rho_i(t) = \infty,\ \forall i \in N,\ a.s.,$

A. 3. $\lim_{t\to\infty} \frac{\sum_{k=1}^{t} \rho_i(k)}{\sum_{k=1}^{t} \rho_j(k)} = 1,\ \forall i, j \in N,\ a.s.$

The first assumption ensures that players eventually transition towards playing their next-stage action as a sample from their empirical distribution rather than playing a deliberate best response. The second assumption ensures that, for each player, a deliberate best response is played infinitely often. The third assumption ensures that the number of deliberate best responses taken by each player remain relatively in sync.[6] In practice, players may choose their deliberate best responses completely asynchronously; for example, setting $\rho_i(t) = 1/t^r,\ \forall i$, with $r \in (0,1]$, results in (purely) independent sampling of deliberate best response rounds and secures **A.1**–**A.3**.

Let

$$\ell_i(t) := \sum_{k=1}^{t} X_i(k) \tag{4.3}$$

count the number of times player i has deliberately played a best response until and including round t. Note that $\ell_i(t)$ is $\mathcal{F}_t$-measurable. The empirical distribution $q_i(t)$ is defined recursively as[7]

$$q_i(t+1) = q_i(t) + \frac{1}{\ell_i(t+1)} \left(a_i(t+1) - q_i(t)\right) X_i(t+1). \tag{4.4}$$

Intuitively speaking, the empirical distribution (4.4) is updated only over rounds when a deliberate best response was played. Note that $q_i(t)$ is $\mathcal{F}_t$-measurable.[8]

[5] The action $a_i(t) \in A_i$ is technically a dirac distribution over the finite action space Y_i (see Section 2), and the mixed strategy $g_i'(t)$ is a probability distribution over Y_i. More precisely, the notation $a_i(t) \sim g_i'(t)$ means that an action $y_i(t)$ is drawn as a random sample from $g_i'(t)$ with $a_i(t) := \delta_{y_i(t)}(y_i)$, where $\delta_{y_i(t)}(y_i) = 1$ if $y_i = y_i(t)$ and $\delta_{y_i(t)}(y_i) = 0$ otherwise.

[6] Note that since $\rho_i(t)$ is only required to be $\mathcal{F}_{t-1}$-measurable, this parameter is in fact adaptively tunable. This is a feature of practical interest since it allows players to adjust their deliberate best response rates on the fly—possibly adapting to the (initially unknown) deliberate best response rates of others and to underlying process dynamics—in order to satisfy **A.1**–**A.3**.

[7] To initialize the process, let the action $a_i(1)$ be chosen arbitrarily, let $q_i(1) = a_i(1)$, and let $X_i(1) = 1$ for all i.

[8] Note that, (4.2) implicitly assumes that players have knowledge of the empirical distributions

Finally, let

$$g_i(t) := b_i(t-1)\rho_i(t) + q_i(t-1)(1-\rho_i(t)), \tag{4.5}$$

and note that $g_i(t)$ is $\mathcal{F}_{t-1}$ measurable.[9] More importantly, note that for every $\alpha_i \in A_i$, $g_i(\alpha_i, t) = \mathbb{P}(a_i(t) = \alpha_i |\ \mathcal{F}_{t-1})$, and thus $g_i(t)$ represents the mixed strategy (conditioned on past play) used by player i in round t. The joint mixed strategy used in round t is given by $g(t) := (g_1(t), \ldots, g_n(t))$.

We refer to a process where, for each player i, $a_i(t)$ is updated according to (4.2), $q_i(t)$ is updated according to (4.4), and $g_i(t)$ is updated according to (4.5) as the strongly convergent variant of (classical) FP (for reasons to be clear soon).

4.2. Strong Convergence in Classical FP: Main Result. The following result states that in the strongly convergent variant of FP, players' period-by-period mixed strategies converge to the set of Nash equilibria—i.e., strong learning is achieved.

COROLLARY 1. *Let Γ be a two-player zero-sum game, potential game, or generic $2 \times m$ game. Assume **A.1**–**A.3** hold. Then the strongly convergent variant of FP achieves strong learning in the sense that $\lim_{t\to\infty} d(g(t),\ NE) = 0$ almost surely.*

In order to prove the above result, we first study a more general notion of fictitious play and then prove the result as a corollary of the general theorem (see Theorem 5.1). Taking this general approach allows our strong convergence results to be be applied to other FP-type algorithms, e.g., Generalized Weakened FP (Section 6.2) and Empirical Centroid FP (Section 6.3). The proof of Corollary 1 is given in Section 6.1.

4.3. Simulation Example. In order to demonstrate the learning properties of strongly convergent FP, we simulated classical FP and strongly convergent FP in a simple two-player matching pennies game with utility functions as shown in Figure 4.1a. The game has a unique (symmetric) mixed-strategy equilibrium in which both players choose either action with probability 1/2. Figure 4.1b shows the period-by-period strategies generated by classical FP. Players' strategies are always pure and progress in continuously lengthening cycles. While the time-averaged empirical distribution is being driven to equilibrium, the period-by-period strategies clearly are not.

Figure 4.1c shows the period-by-period strategies generated by strongly convergent FP with $\rho(t) = t^{-.35}$. Players' period-by-period strategies are converging to the unique Nash equilibrium of the game.

Figure 4.1d shows the utility received by the realized joint action $a(t)$ in each round of repeated play for both learning algorithms. The received payoffs in classical FP cycle around the value of the game, while the received payoffs in strongly convergent FP converge to the value of the game.

One possible tradeoff in strongly convergent FP is that less frequent deliberate best response actions and less frequent updating of the empirical distribution (see

of opponents when computing a best response. This may be accomplished by assuming that players actions are accompanied with a "tag" indicating whether or not the played action was a deliberate best response. Alternatively, the information regarding $q_i(t)$ may tracked by the individual player i and disseminated by a gossip-type algorithm [12] or implicitly disseminated through a payoff-based scheme.

[9] To see this, note first that $q_i(t-1)$ and $\rho_i(t)$ have been shown to be $\mathcal{F}_{t-1}$ measurable. Furthermore, this implies that $BR_i(q_i(t-1))$ is $\mathcal{F}_{t-1}$-measurable. Lastly, by construction, $b_i(t) \in BR_i(q_i(t-1))$ is $\mathcal{F}_{t-1}$-measurable.

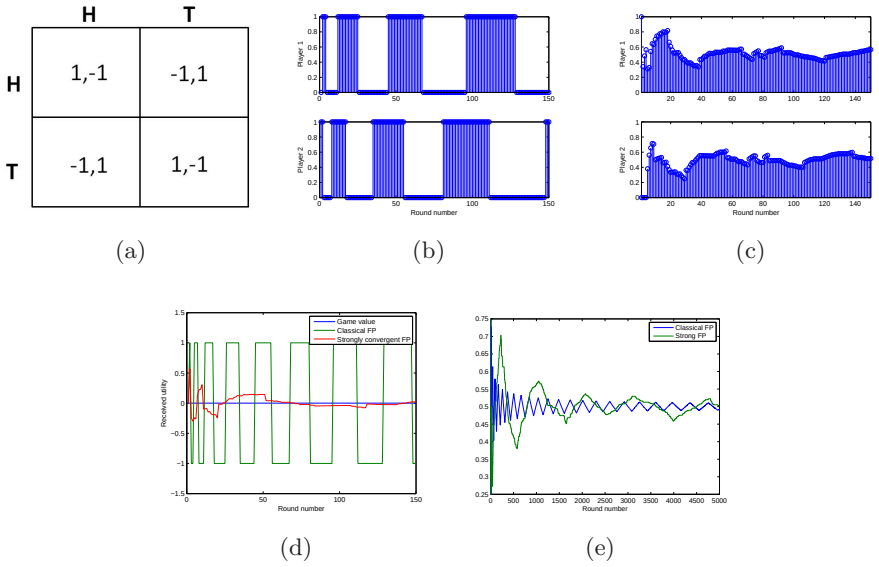

Fig. 4.1: 4.1a: Matching pennies payoff matrix, 4.1b: The probability of each player playing heads in round t using the classical FP algorithm, 4.1c: The probability of each player playing heads in round t using the strongly convergent FP algorithm, 4.1d: The received utility in round t given the realized action $a(t)$, 4.1e: The empirical distribution process of the action H (heads) for player 1 in both FP and strongly convergent FP.

(4.4)) may lead to a slow-down in convergence rate. The empirical distribution processes for player 1 in each algorithm is shown in Figure 4.1e with $\rho(t) = t^{-.35}$.

5. General Setup. In this section we study strong learning in FP-type algorithms —a class of algorithms that generalizes FP and includes many learning processes based on best-response dynamics.[10] In Section 5.1, we define the notion of an FP-type algorithm. In Section 5.2 we present some examples of an FP-type algorithm. In Section 5.3 we define the strongly convergent variant of an FP-type algorithm. In Section 5.4 we provide the general strong convergence result for an FP-type algorithm (see Theorem 5.1), and in Sections 5.5–5.7 we prove the general result.

5.1. FP-Type Algorithm. An FP-type algorithm generalizes classical FP in the following ways: (i) the notion of a player's empirical distribution is generalized, (ii) players are permitted to use a function of the empirical distribution (rather than use the empirical distribution itself) as a predictor of the next-round strategy of opponents, (iii) convergence to equilibrium may occur in terms of a function of the empirical distribution (rather than convergence to equilibrium of the empirical distribution itself), and (iv) limit sets other than the set of NE are permitted.

We define an FP-type algorithm as follows. Let players be engaged in repeated play of a stage game Γ. Let $a_i(t)$ represent the action of player i in round $t \in \{1, 2, \ldots\}$, and let $H_i(t) := \{a_i(s)\}_{s=1}^{t}$ represent the action history of player i up to and including round t.

[10] The class of FP-type algorithms proposed here is similar in spirit to the class of best-response-based algorithms considered in [9].

In classical FP, for each player i, the normalized histogram of the player's action choices (3.1) is used as a compact representation of the player's action history. In the general formulation of an FP-type algorithm, we still suppose that players track a compact representation of the action history, but we allow the compact representation to take on a fairly general form,[11] as stated in the following assumption:

A. 4. *The empirical distribution of player i is of the form $q_i(t) := f_i^q(H_i(t),\ t)$, where $f_i^q(\cdot,\ t) : \prod_{s=1}^t A_i \to \Delta_i$.* We make the following assumption regarding the sequence of functions $\{f_i^q(\cdot,\ t)\}_{t\geq 1}$ used to form the empirical distribution sequence of player i:

A. 5. *For any history sequence $\{H_i(t)\}_{t\geq 1}$ for player i, there holds* $\lim_{t\to\infty} \|f_i^q(H_i(t+1),\ t+1) - f_i^q(H_i(t),\ t)\| = 0$.

In particular, this implies that—regardless of the action history—there holds $\lim_{t\to\infty} \|q_i(t+1) - q_i(t)\| = 0$ for each player i. This fairly mild assumption captures the essential characteristics required for our asymptotic analysis, and may be seen as a generalization of classical FP where exact averaging of actions over time yields $\|q_i(t+1) - q_i(t)\| \leq \frac{1}{t}$ (see Section 5.2.1). Together, assumptions **A.4**–**A.5** allow us to consider a variety of FP inspired algorithms, including those with general step sizes [11] and those with more intricate history dependent rules such as derivative action [32].

In an FP-type algorithm, players form a prediction of the future behavior of opponents as a function of the current empirical distribution. Let $p_i(t)$ be player i's prediction of opponent strategies for the upcoming round $(t+1)$. We assume,

A. 6. *Player i's prediction $p_i(t)$ of opponent behavior is of the form $p_i(t) = f_i^p(q(t))$, where $f_i^p : \Delta^n \to \Delta_{-i}$ is a Lipschitz continuous, time-invariant function.*

We say a sequence of actions $\{a(t)\}_{t\geq 1}$ is an FP-type process if for all $i \in N$ and all $t \geq 1$, $a_i(t+1) \in BR_i^{\epsilon_t}(p_i(t))$, where $BR_i^{\varepsilon_t}(\cdot)$ is the ε_t-best response set (recall (2.1)), and $\{\epsilon_t\}_{t\geq 1}$ is a sequence satisfying $\lim_{t\to\infty} \epsilon_t = 0$.

In many variants of FP, including classical FP, learning occurs in the sense that $d(q(t),\ NE) \to 0$. We generalize this notion of learning by allowing for limit sets other than the set of NE and allowing for convergence in terms of a function of $q(t)$ rather than permitting convergence only in terms of $q(t)$ itself.

Let E be some target equilibrium set (not necessarily the set of NE). An FP-type process is said to learn elements of E if for each i there exists a function f_i^ξ satisfying:

A. 7. *The function $f_i^\xi : \Delta^n \to \Delta_i$ is Lipschitz continuous and time invariant,* and such that, for $\xi_i(t) := f_i^\xi(q(t))$ and $\xi(t) := (\xi_1(t), \ldots, \xi_n(t))$ there holds $\lim_{t\to 0} d(\xi(t),\ E) = 0$. We refer to $\xi(t)$ as the asymptotic learning distribution, and f_i^ξ as the convergence map of player i.

In general, we will denote an instance of an FP-type learning algorithm by $\Psi = (\{f_i^q(\cdot,\ t)\}_{t\geq 1}, f_i^p, f_i^\xi)_{i\in N}$. In order to construct a strongly convergent variant of Ψ we will require that Ψ obtain weak convergence in sufficiently robust sense as stated in the following assumption.

A. 8. *For the stage game Γ and equilibrium set E, the FP-type algorithm Ψ is such that for any sequence $(\epsilon_t)_{t\geq 1}$ satisfying $\lim_{t\to\infty} \epsilon_t = 0$, the FP-type algorithm Ψ*

[11] In most literature, the notion of an *empirical distribution* refers strictly to the time-averaged empirical histogram of a player's action choices, as in classical FP (3.1). However, as discussed in Section 2, we use the term empirical distribution more generally to refer to an arbitrarily formed (see **A.4**) distribution that a player uses to track information regarding opponents' empirical action histories. This abuse of terminology allows us to more naturally extend concepts to the general FP-type setting.

obtains weak convergence in the sense that $\lim_{t\to 0} d(\xi(t),\ E) = 0$.

The above assumption ensures that the FP-type algorithm is robust to asymptotically decaying perturbations in a player's best response set. When studying the strongly convergent variant of Ψ in the following section, the assumption **A.8** will serve to ensure that convergence of the process is not disrupted by minor asynchronies in the number of deliberate best responses taken by each player (i.e., minor disparities in (4.3)).

5.2. Examples.

5.2.1. Classical Fictitious Play. Classical FP (Section 3.1) fits the template of an FP-type algorithm with $q_i(t) = \frac{1}{t}\sum_{s=1}^{t} a_i(s)$. Note that $q_i(t)$ may be written in recursive form as: $q_i(t+1) = q_i(t) + 1/(t+1)\,(a_i(t+1) - q_i(t))$. Thus, $\|q_i(t+1) - q_i(t)\| \leq \frac{M_i}{t+1}$, where $M_i := \sup_{p_i', p_i'' \in \Delta_i} \|p_i' - p_i''\|$, and **A.5** is satisfied. The prediction map f_i^p is given by the identity function, and convergence map f_i^ξ also given by the identity function. The target equilibrium set is given by $E := NE$, the set of Nash equilibria.

5.2.2. Generalized Weakened Fictitious Play. Leslie et al. [11] study a useful generalization of FP, termed Generalized Weakened FP (GWFP), in which players are permitted to choose a suboptimal best response each round, so long as the degree of suboptimality decays asymptotically to zero, and in which step-size sequences other than $\{1/t\}_{t\geq 1}$ are permitted.

Formally, for $p_{-i} \in \Delta_{-i}$ and $\epsilon \geq 0$, let[12] $B\bar{R}_i^\epsilon(p_{-i}) := \{p_i \in \Delta_i : U_i(p_i, p_{-i}) \geq \max_{\alpha_i \in A_i} U_i(\alpha_i, p_{-i}) - \epsilon\}$, and for $p \in \Delta^n$, let $B\bar{R}^\epsilon(p) := (B\bar{R}_1^\epsilon(p_{-1}), \ldots, B\bar{R}_n^\epsilon(p_{-n}))$. A sequence $\{q(t)\}_{t\geq 1}$ is said to be a GWFP process if $q(t+1) \in (1-\gamma(t+1))q(t) + \gamma(t+1)(B\bar{R}^{\epsilon_t}(q(t)) + M_{t+1})$ with $\gamma(t) \to 0$ and $\epsilon_t \to 0$ as $t \to \infty$, $\sum_{t\geq 1}\gamma(t) = \infty$, and $\{M_t\}_{t\geq 1}$ is a deterministic (or stochastic) perturbation sequence satisfying $\lim_{t\to\infty} \sup_k \{\|\sum_{i=t}^{k-1} \gamma_{i+1} M_{i+1}\| : \sum_{i=t}^{k-1} \gamma_{i+1} \leq T\} = 0$ (a.s.).

We consider a special case of GWFP in which $M_t = 0,\ \forall t$ and the ϵ-best response set is restricted to the set of pure strategy ϵ-best responses. That is, we consider the subset of GWFP process such that $a(t+1) \in BR^{\epsilon_t}(q_{-i}(t))$, and,

$$q(t+1) = q(t) + \gamma(t+1)\,(a(t+1) - q(t))\,, \tag{5.1}$$

with $\epsilon_t \to 0$, and in a slight variation of terminology we refer to the sequence of actions $\{a(t)\}_{t\geq 1}$ satisfying the above as a GWFP process.

In the terminology of Section 5.1, GWFP fits the template of an FP-type algorithm with the empirical distribution $q_i(t)$ defined recursively as in (5.1) (where it is assumed that $\lim_{t\to\infty}\gamma(t) = 0$), the prediction map f_i^p given by the identity function for all i, and the convergence map f_i^ξ given by the identity function for all i, and the target equilibrium set is given by $E := NE$—the set of Nash equilibria.

5.2.3. Empirical Centroid Fictitious Play—Learning Consensus Equilibria. Empirical Centroid FP (ECFP) was conceived as a variant of FP suited to implementation in large-scale games [12, 13]. In ECFP, rather than tracking the empirical distribution of each individual opponent (as in FP), players track and respond

[12]The set $B\bar{R}_i^\epsilon(p_{-i})$ defined below differs from the set $BR_i^\epsilon(p_{-i})$ defined in the preliminaries in that $B\bar{R}_i^\epsilon(p_{-i})$ includes all *mixed* strategy best responses, whereas $BR_i^\epsilon(p_{-i})$ contains only the pure strategy best responses. The set $B\bar{R}_i^\epsilon(p_{-i})$ is used here in order to precisely define a GWFP process as given in [11], but the remainder of the paper focuses on the set $BR_i^\epsilon(p_{-i})$.

to only the centroid of the empirical distributions. In order to ensure the process is well defined the following assumption is made:

A. 9. *All players use the same strategy space.* Under this assumption, let the empirical distribution be defined by

$$q_i(t) := \frac{1}{t}\sum_{s=1}^{t} a_i(s), \tag{5.2}$$

and let the empirical centroid distribution be defined by $\bar{q}(t) := \frac{1}{n}\sum_{i\in N} q_i(t)$. We say a sequence of actions $\{a(t)\}_{t\geq 1}$ is an ECFP process if for all i and all $t \geq 1$,

$$a_i(t+1) \in BR_i^{\epsilon_t}(\bar{q}_{-i}(t)), \tag{5.3}$$

where $\bar{q}_{-i}(t) = (\bar{q}(t), \ldots, \bar{q}(t)) \in \prod_{j\neq i}\Delta_j$ is the $(n-1)$-tuple containing $(n-1)$ repeated copies of $\bar{q}(t)$, and $\{\epsilon_t\}_{t\geq 1}$ is a sequence satisfying $\lim_{t\to\infty}\epsilon_t = 0$.

In ECFP, players learn elements of the set of consensus Nash equilibria[13], defined by $C := \{p = (p_1, \ \ldots \ , p_n) \in NE : \ p_1 = p_2 = \ldots = p_n\}$, the subset of Nash equilibria in which all players use identical strategies (see [12] for more details). Define $\bar{q}^n(t) := (\bar{q}(t), \ldots, \bar{q}(t)) \in \Delta^n$ to be the n-tuple containing repeated copies of $\bar{q}(t)$; learning in ECFP takes place in the sense that $\lim_{t\to\infty} d(\bar{q}^n(t), \ C) = 0$.

In the terminology of Section 5.1, ECFP fits the template of an FP-type algorithm with the empirical distribution given by (5.2), the prediction map f_i^p given by $f_i^p(q(t)) := \left(\frac{1}{n}\sum_{j\in N} q_j(t), \ldots, \frac{1}{n}\sum_{j\in N} q_j(t)\right), \ \forall i$, where the right-hand side is a $(n-1)$-tuple containing repeated copies of $\bar{q}(t)$, and the convergence map given by $f_i^{\xi}(q(t)) := \frac{1}{n}\sum_{j=1}^{n} q_j(t), \ \forall i$. The target equilibrium set is given by $E := C$, the set of consensus Nash equilibria.

5.2.4. Empirical Centroid Fictitious Play—Learning Mean-Centric Equilibria. In this section we consider a slight modification of the ECFP algorithm presented in Section 5.2.3 that enables players to learn elements of an alternate (non-Nash) equilibrium set.

Let an ECFP action process be defined as in (5.3). Define the set of mean-centric equilibria by $MCE := \{p \in \Delta^n : \ U_i(p_i, \ \bar{p}_{-i}) \geq U_i(p_i', \ \bar{p}_{-i}) \ \forall p_i' \in \Delta_i, \ \forall i\}$. The set of MCE is neither a superset nor a subset of the NE—rather, it is a set of natural equilibrium points tailored to the ECFP dynamics [33]. The set of consensus Nash equilibria C (see Section 5.2.3) however, is contained in the set of MCE.

In ECFP, players learn elements of MCE in the sense that $\lim_{t\to\infty} d(q(t), \ MCE) = 0$. In the terminology of Section 5.1, this fits the template of an FP-type algorithm with $q_i(t)$ given by (5.2), f_i^p defined in the same way as in Section 5.2.3, the convergence map f_i^{ξ} given by the identity for all i, and the target equilibrium set given by $E := MCE$.

Note that the only difference between the ECFP algorithm discussed in the Section 5.2.3 and the ECFP algorithm discussed here is the choice of target equilibrium set E and convergence maps f_i^{ξ}.

5.3. Strongly Convergent Variant of an FP-type Algorithm. In this section we construct the strongly convergent variant of an FP-type learning algorithm.

[13]We assume here that the set of consensus Nash equilibria is non-empty. When revisiting ECFP in Section 6.3, we provide an assumption on the utility structure that ensures that the set is indeed non-empty.

The construction here is a generalization of that of Section 4.1 where we constructed the strongly convergent variant of classical FP.

Let $\Psi = (\{f_i^q(\cdot,\ t)\}_{t\geq 1},\ f_i^p,\ f_i^\xi)_{i\in N}$ be an FP-type learning algorithm. For each $i \in N$, let $\{X_i(t)\}_{t\geq 1}$ be a sequence of random variables with $X_i(t) \in \{0, 1\}$. Analogous to Section 4, $X_i(t) = 1$ will serve to indicate that player i took a deliberate best response in round t. Let

$$\ell_i(t) := \sum_{s=1}^{t} X_i(s) \tag{5.4}$$

count the number of deliberate best responses taken by player i through t.

In Section 4.1 the empirical distribution of player i, (4.4), is a time average taken only over rounds when player i took a deliberate best response. In order to generalize this notion to an FP-type algorithm, define the term

$$\tau_i(s) := \inf\{t : \ell_i(t) = s\}. \tag{5.5}$$

For $s \geq 1$, $\tau_i(s)$ indicates the round when player i took their s-th deliberate best response,[14] and the sequence $\{\tau_i(s)\}_{s\geq 1}$ gives the subsequence of rounds when player i took a deliberate best response. For $t \in \{1,\ 2, \ldots\}$ let $\bar{H}_i(t) := \{a_i(\tau_i(s)) :\ \tau_i(s) \leq t\}$ denote the action history of player i. Note that $\bar{H}(t)$ records only the history of actions that were taken as deliberate best responses. Let the empirical distribution of player i at time t be formed as

$$q_i(t) := f_i^q(\bar{H}_i(t),\ \ell_i(t)). \tag{5.6}$$

Let the asymptotic learning distribution (see **A.7** and subsequent discussion) be given by $\xi_i(t) := f_i^\xi(q(t))$ and $\xi(t) := (\xi_1(t), \ldots,\ \xi_i(t))$.

Let the action for player i in round $t \geq 2$ be chosen according to the random rule[15]

$$a_i(t) \sim g_i'(t) := \begin{cases} b_i(t-1), & \text{if } X_i(t) = 1, \\ \xi_i(t-1), & \text{otherwise,} \end{cases} \tag{5.7}$$

where $p_i(t-1) = f_i^p(q(t-1))$, and $b_i(t-1) \in BR_i^{\eta_t}(p_i(t-1))$, and assume:[16]

A. 10. *The sequence* $(\eta_t)_{t\geq 1}$ *associated with* $b_i(t)$ *of* (5.7) *is such that* $\lim_{t\to\infty} \eta_t = 0$.

Let $\mathcal{F}_t := \sigma(\{a(s), X_i(s), \ldots, X_n(s), b_1(s), \ldots, b_n(s)\}_{s\leq t})$. Let the probability that player i chooses a deliberate best response in round t conditioned on past events be given by $\rho_i(t) := \mathbb{P}(X_i(t) = 1|\mathcal{F}_{t-1})$, and assume **A.1**–**A.3** hold. Note that $q_i(t)$, $p_i(t)$, $\xi_i(t)$, and $g_i'(t)$ are $\mathcal{F}_t$–measurable and that by definition, $\rho_i(t)$ is $\mathcal{F}_{t-1}$–measurable.

Finally, let

$$g_i(t) := b_i(t-1)\rho_i(t) + \xi_i(t)(1 - \rho_i(t)). \tag{5.8}$$

Note that $g_i(t)$ is $\mathcal{F}_{t-1}$–measurable and that $g(\alpha_i, t) = \mathbb{P}(a_i(t) = \alpha_i|\mathcal{F}_{t-1})$; that is, $g_i(t)$ represents the mixed strategy in use by player i in round t (compare with (4.5)).

[14] Note that by (5.10), $\tau_i(s)$ is finite valued a.s. for any $s \in \{1,\ 2, \ldots\}$.

[15] To initialize the process, let the action $a_i(1)$ be chosen arbitrarily, let $X_i(1) = 1$, and let $\bar{H}(1) = a_i(1)$ for all i.

[16] Note that this assumption subsumes the more typical assumption that $\eta_t = 0,\ \forall t$. By making this more general assumption we are able to handle interesting scenarios that may arise in a practical implementation of the algorithm; e.g., players have some asymptotically decaying error in their knowledge of their utility function or knowledge of opponent's empirical distributions.

Let $g(t) := (g_1(t), \ldots, g_n(t))$ denote the joint mixed strategy in use at time t.

We refer to a process where, for each player i, $q_i(t)$ is updated according to (5.6), $a_i(t)$ is updated according to (5.7), and $g_i(t)$ is updated according to (5.8) as the strongly convergent variant of Ψ (for reasons to be clear soon—see Theorem 5.1). In Section 6 we will demonstrate applications of this in the context of the previous examples.

5.4. General Result. The following theorem provides the general result from which the strong convergence of various FP-type algorithms can be derived.

THEOREM 5.1. *Let Γ be a finite normal form game, let E be an equilibrium set, and let Ψ be an FP-type algorithm satisfying **A.4**–**A.8**. If the strongly convergent variant of Ψ satisfies **A.1**–**A.3** and **A.10** then it achieves strong learning in the sense that $\lim_{t\to\infty} d(g(t),\ E) = 0$, almost surely.*

We emphasize that in the above result players' period-by-period mixed strategies $g(t)$ are converging to equilibrium. In general, when seeking to construct the strongly convergent variant of some FP-type algorithm Ψ, the most challenging aspect of applying Theorem 5.1 is the verification that Ψ satisfies **A.8**. The remaining assumptions **A.4**–**A.7** are generally fairly trivial to verify. Assumptions **A.1**–**A.3** and **A.10** pertain to the manner in which the strongly convergent variant of Ψ is constructed and are not related to intrinsic properties of Ψ itself.

5.5. Some Additional Definitions. In order to prove Theorem 5.1 we will study the behavior of an underlying FP-type process that is embedded in the action, history, and empirical distribution processes produced by the strongly convergent variant of Ψ. In particular, for $i \in N$ and $s \in \{1, 2, \ldots\}$, let $\tau_i(s)$ be defined as in (5.5), and define the following terms: $\tilde{a}_i(s) := a_i(\tau_i(s))$, $\tilde{a}(s) := (\tilde{a}_1(s), \ldots, \tilde{a}_n(s))$, $\tilde{H}_i(s) := \bar{H}_i(\tau_i(s))$, $\tilde{q}_i(s) := q_i(\tau_i(s))$, $\tilde{q}(s) := (\tilde{q}_1(s), \ldots, \tilde{q}_n(s))$, $\tilde{p}_i(s) := f_i^p(\tilde{q}(s))$, $\tilde{\xi}(s) := (f_1^\xi(\tilde{q}(s)), \ldots, f_n^\xi(\tilde{q}(s)))$. The aforementioned terms (marked with a tilde) correspond to to the embedded FP-type process that we will study in the proof of Theorem 5.1. In particular, for each player i, the sequence $\{\tau_i(s)\}_{s\geq 1}$ denotes the subsequence of rounds when the player chose to play a deliberate best response. The sequence $\tilde{a}_i(s)_{s\geq 1}$ is the action sequence occurring along the subsequence of rounds when player i chose to play a deliberate best response. The sequence $\{\tilde{H}_i(s)\}_{s\geq 1}$ corresponds to the action history of player i along the same subsequence. The sequence $\{\tilde{q}_i(s)\}_{s\geq 1}$ corresponds to the empirical distribution of player i along the same subsequence; in particular, note that by Lemma 7.5 (see appendix), $\{\tilde{q}_i(s)\}_{s\geq 1}$ fits the format prescribed by **A.4** for the embedded FP-type process: $\tilde{q}_i(s) = f_i^q(\tilde{H}(s), s)$. Finally, the term $\tilde{\xi}(s)$ is the asymptotic learning distribution associated with the embedded FP-type process.

In studying the embedded FP-type process, it will be important to characterize the terms to which players are best responding. With this in mind, note that per (5.7), the action at time $\tau_i(s+1)$ (in the strongly convergent variant of Ψ) is chosen as $a_i(\tau_i(s+1)) \in BR_i^{\eta_{\tau_i(s+1)}}(p_i(\tau_i(s+1)-1))$. In order to translate this to the embedded FP-type process, define the following terms: $\hat{q}_j^i(s) := q_j(\tau_i(s+1)-1)$, $\hat{q}^i(s) := (q_1(\tau_i(s+1)-1), \ldots, q_n(\tau_i(s+1)-1))$ $\hat{p}_i(s) := f_i^p(\hat{q}^i(s))$, By construction, the $(s+1)$-th action of player i in the embedded FP-type process is chosen as,

$$\tilde{a}_i(s+1) \in BR_i^{\eta_{\tau_i(s+1)}}(\hat{p}_i(s)). \tag{5.9}$$

In the embedded FP-type process, the term $\tilde{q}_j(s)$ may be thought of as the 'true'

empirical distribution of player j. The term $\hat{q}_j^i(s)$ may be thought of as the estimate which player i maintains of $\tilde{q}_j(s)$, and the term $\hat{q}^i(s)$ (note the superscript) may be thought of as player i's estimate of the joint empirical distribution $\tilde{q}(s)$ at the time of player i's $(s+1)$-th best response. Finally, the term $\hat{p}_i(s)$ may be thought of as player i's prediction of opponents next-stage strategy given $\hat{q}^i(s)$; in particular, note that—in the embedded FP-type process—player i chooses their stage $(s+1)$ action (5.9) as an asymptotic best response to $\hat{p}_i(s)$.

5.6. Some Useful Properties. Let

$$\Omega' := \{\omega : \lim_{t\to\infty} \frac{\ell_i(t)}{\sum_{k=1}^t \rho_i(t)} = 1,\ \forall i\}.$$

By Lemma 7.6 (see appendix), there holds $\mathbb{P}(\Omega') = 1$. In proving Theorem 5.1 we will restrict attention to (sample path) realizations in Ω'.

Note that under assumption **A.2**, there holds $\{\omega : \lim_{t\to\infty} \ell_i(t) = \infty,\ \forall i\} \supset \Omega'$. By the equivalence $\{\omega : \lim_{t\to\infty} \ell_i(t) = \infty,\ \forall i\} = \{\omega : X_i(t) = 1 \text{ infinitely often } \forall i\}$, there holds $\{\omega : X_i(t) = 1 \text{ infinitely often } \forall i\} \supset \Omega'$. Therefore, by the definitions of ℓ_i and τ_i, there holds for any realization in Ω', $\lim_{t\to\infty} \ell_i(t) = \infty$, and

$$\tau_i(s) < \infty,\ \forall s \in \mathbb{N}, \tag{5.10}$$

$$\lim_{s\to\infty} \tau_i(s) = \infty. \tag{5.11}$$

These properties will be useful in the proof of Theorem 1. In particular, the proof will frequently make reference to $\tilde{q}_i(s)$, or $\tilde{a}_i(s)$ for arbitrary $s \in \mathbb{N}$—the property (5.10) ensures that such terms are well defined for any $\omega \in \Omega'$.

Note also that for any realization in Ω', for $i \in N$ and $s \in \{1, 2, \ldots\}$,

$$\ell_i(\tau_i(s)) = s, \tag{5.12}$$

and for $i \in N$ and $t \in \{1, 2, \ldots\}$

$$X_i(t) = 1 \implies \tau_i(\ell_i(t)) = t. \tag{5.13}$$

Furthermore, note that $X_i(t) = 0$ implies that $\ell_i(t) = \ell_i(t-1)$ and $\bar{H}_i(t) = \bar{H}_i(t-1)$, and in particular,

$$X_i(t) = 0 \implies q_i(t) = q_i(t-1). \tag{5.14}$$

These facts are readily verified by conferring with the definitions of τ_i, ℓ_i, and X_i.

5.7. Proof of Theorem 5.1. *Proof.* Since $\mathbb{P}(\Omega') = 1$ it is sufficient to show that the desired result holds for any $\omega \in \Omega'$. Henceforth, we restrict attention to realizations $\omega \in \Omega'$, and for ease of notation suppress the term ω when referring to random variables.

As a first step, we wish to show that $\lim_{s\to\infty} d(\tilde{\xi}(s),\ E) = 0$. We accomplish this by showing that there exists a sequence $\{\epsilon_s\}_{s\geq 1}$ such that $\lim_{s\to\infty} \epsilon_s = 0$ and $\tilde{a}_i(s+1) \in BR_i^{\epsilon_s}(\tilde{p}_i(s))$. By assumption **A.8**, it will then follow that $\lim_{s\to\infty} d(\tilde{\xi}(s),\ E) = 0$.

To that end, note that by Lemma 7.1 (see appendix), $\lim_{s\to\infty} |U_i(a_i(\tau_i(s+1)), p_i(\tau_i(s+1)-1)) - v_i(p_i(\tau_i(s+1)-1))| = 0,\ \forall i$, or equivalently by the definitions of $\tilde{a}(s)$ and $\hat{p}_i(s)$ (see Section 5.5),

$$\lim_{s\to\infty} |U_i(\tilde{a}_i(s+1)), \hat{p}_i(s)) - v_i(\hat{p}_i(s))| = 0,\ \forall i. \tag{5.15}$$

By Lemma 7.3 (see appendix), $\lim_{s\to\infty} \|\hat{q}^i(s) - \tilde{q}(s)\| = 0$. By **A.6**, it follows that $\lim_{s\to\infty} \|\hat{p}_i(s) - \tilde{p}_i(s)\| = 0$, which by the Lipschitz continuity of $U_i(\cdot)$ implies that

$\lim_{s\to\infty} |U_i(\alpha_i, \hat{p}_i(s)) - U_i(\alpha_i, \tilde{p}_i(s))| = 0,\ \forall \alpha_i \in A_i, \forall i$, and $\lim_{s\to\infty} |v_i(\hat{p}_i(s)) - v_i(\tilde{p}_i(s))| = 0, \forall i$. Returning to (5.15) we see that $\lim_{s\to\infty} |U_i(\tilde{a}_i(s+1)), \tilde{p}_i(s)) - v_i(\tilde{p}_i(s))| = 0,\ \forall i$, i.e., there exists a sequence $\{\epsilon_s\}_{s\geq 1}$ such that $\epsilon_s \to 0$ and $\tilde{a}_i(s+1) \in BR_i^{\epsilon_s}(\tilde{p}_i(s))$. It follows by **A.8** that

$$\lim_{s\to\infty} d(\tilde{\xi}(s),\ E) = 0. \tag{5.16}$$

We now proceed to show that $\lim_{t\to\infty} d(\xi(t),\ E) = 0$. Let $\varepsilon > 0$ be given. By Lemma 7.2 (see appendix) and assumption **A.7**, for each $i \in N$, there exists a random time $S_i > 0$ such that $\forall s \geq S_i$, $\|\xi(\tau_i(s)) - \tilde{\xi}(s)\| < \frac{\varepsilon}{2}$. Let $S^{'} = \max_i\{S_i\}$. By (5.16) there exists a random time $S^{''}$ such that $\forall s \geq S^{''}$, $d(\tilde{\xi}(s),\ E) < \frac{\varepsilon}{2}$. Let $S = \max\{S^{'}, S^{''}\}$. Then

$$d(\xi(\tau_i(s)),\ E) < \varepsilon,\ \forall i,\ \forall s \geq S. \tag{5.17}$$

Let $T = \max_i\{\tau_i(S)\}$. Note that for some i, $\xi(T) = \xi(\tau_i(S))$, and therefore by (5.17),

$$d(\xi(T),\ E) < \varepsilon. \tag{5.18}$$

Also note that for any $t_0 > T$, it holds that $\ell_i(t_0) \geq S$ (since $\ell_i(\tau_i(S)) = S$, and $\ell_i(t)$ is non-decreasing in t), and moreover

$$\begin{aligned} X_i(t_0) = 1 \text{ for some } i \implies q(t_0) = q(\tau_i(\ell_i(t_0))) \implies \xi(t_0) = \xi(\tau_i(\ell_i(t_0))), \\ X_i(t_0) = 0 \text{ for all } i \implies q(t_0) = q(t_0 - 1) \implies \xi(t_0) = \xi(t_0 - 1), \end{aligned} \tag{5.19}$$

where the first implication holds with with $\ell_i(t_0) \geq S$. In the above, the first line follows from (5.13), and the second line follows from (5.14). Consider $t \geq T$. If for some i, $X_i(t) = 1$, then by (5.19) and (5.17), $d(\xi(t),\ E) = d(\xi(\tau_i(\ell_i(t))),\ E) < \varepsilon$. Otherwise, if $X_i(t) = 0\ \forall i$, then $\xi(t) = \xi(t-1)$.

Iterate this argument m times until either (i) $X_i(t-m) = 1$ for some i, or (ii), $t - m = T$. In the case of (i), $d(\xi(t),\ E) = d(\xi(t-m),\ E) = d(\xi(\tau_i(\ell_i(t-m))),\ E) < \varepsilon$, where the inequality again follows from (5.17) and the fact that $t - m > T \implies \ell_i(t - m) \geq S$. In the case of (ii), $d(\xi(t),\ E) = d(\xi(T),\ E) < \varepsilon$, where the inequality follows from (5.18). Since $\varepsilon > 0$ was chosen arbitrarily, it follows that $\lim_{t\to\infty} d(\xi(t),\ E) = 0$.

Finally, we show that $\lim_{t\to\infty} d(g(t),\ E) = 0$. Note that by (5.8), $\|g_i(t) - \xi_i(t-1)\| \leq M_i\rho_i(t),\ \forall i$, where $M_i := \max_{p',p''\in\Delta_i} \|p' - p''\|$ is a constant. Invoking assumption **A.1** gives, $\lim_{t\to\infty} \|g_i(t) - \xi_i(t-1)\| = 0,\ \forall i$. Combining this with the fact that $\lim_{t\to\infty} d(\xi(t),\ E) = 0$ yields the desired result, $\lim_{t\to\infty} d(g(t),\ E) = 0$. □

6. Applications of the General Result. In this section we consider three different FP-type algorithms and study the strongly convergent variant of each. In each case, we prove strong convergence by showing that the FP-type algorithm fits the template of Theorem 5.1. Generally, the only non-trivial aspect of applying Theorem 5.1 will be to show that **A.8** is satisfied.

In Section 6.1 we consider classical FP. The fact that classical FP satisfies **A.8** was shown by Leslie et al. [11]. In Section 6.2 we consider GWFP—a generalization of FP proposed in [11]. Again, the crucial step of showing that GWFP satisfies **A.8** was shown in [11]. In Section 6.3 we consider a variant of FP termed ECFP. That ECFP satisfies **A.8** was shown in [34]. We emphasize that each of these algorithms is known to achieve weak learning in the sense that $d(\xi(t),\ E) \to 0$ as $t \to \infty$. Our contribution is to construct a variant where players also achieve learning in the strong

sense that period-by-period mixed strategies also converge to equilibrium.

6.1. Strong Convergence in Classical FP. We now prove Corollary 1 using the general convergence result of Theorem 5.1.

Proof. Classical FP fits the template of an FP-type algorithm with the empirical distribution given by $q_i(t) = \frac{1}{t}\sum_{s=1}^{t} a_i(s)$, the functions f_i^p and f_i^ξ given by the identity function for each i, and the best response perturbation given by $\epsilon_t = 0,\ \forall t$. To show that the strongly convergent variant of classical FP attains strong learning, it suffices to show that the assumptions of Theorem 5.1 are met.

To that end, note that **A.1**–**A.3** are satisfied by assumption, and **A.10** is trivially satisfied (with $\eta_t = 0,\ \forall t$). Furthermore, the empirical distribution sequence satisfies $\lim_{t\to\infty} \|q_i(t) - q_i(t-1)\| = 0$ (see Section 5.2.1), and hence **A.5** is satisfied. The functions f_i^p and f_i^ξ (each being the identity function) satisfy **A.6**–**A.7**. Therefore, it is sufficient to show that **A.8** is satisfied. But, for zero-sum games, potential games, and generic $2 \times m$ games this holds by [11], Corollary 5. □

6.2. Strong Convergence in Generalized Weakened FP. GWFP was introduced in Section 5.2.2, where it was shown to fit the template of an FP-type algorithm.

Since, by definition, a GWFP process allows players to choose an ϵ_t sub-optimal best response with $\epsilon_t \to 0$, the following result ([11], Corollary 5) guarantees a GWFP process satisfies **A.8** in the noted classes of games.

THEOREM 6.1. *Any generalized weakened fictitious play process will converge to the set of Nash equilibria in two-player zero-sum games, potential games, and generic* $2 \times m$ *games.*

To clarify the precise meaning of the convergence stated above as it relates to the present work, we emphasize that Theorem 6.1 implies that $\lim_{t\to\infty} d(q(t),\ NE) = 0$; i.e., the process converges weakly to equilibrium.

Let the strongly convergent variant of GWFP be constructed using the approach laid out in Section 5.3. The following Corollary to Theorem 5.1 states that the strongly convergent variant of a GWFP process will achieve strong learning.[17]

COROLLARY 2. *Let* Γ *be a two-player zero-sum game, potential game, or generic* $2 \times m$ *game. Let* Ψ *be an instance of GWFP. If the strongly convergent variant of* Ψ *satisfies* ***A.1***–***A3*** *and* ***A.10****, then it achieves strong learning in the sense that* $\lim_{t\to\infty} d(g(t),\ NE) = 0$.

Proof. It is sufficient to show that the conditions of Theorem 5.1 are met. Note that **A.1**–**A.3**, **A.10** hold by assumption. Furthermore, by definition, any GWFP process satisfies $\lim_{t\to\infty} \gamma(t) = 0$, and hence satisfies **A.5**. The functions f_i^p and f_i^ξ are given by the identity function for each i, and hence **A.6** and **A.7** hold. Thus, it suffices to show that **A.8** holds for the specified class of games—but, this follows from Theorem 6.1. □

6.3. Strong Convergence in Empirical Centroid FP. ECFP was introduced in Sections 5.2.3 and 5.2.4. It In order to study the asymptotic behavior of ECFP (in either of the above formats introduced in Sections 5.2.3 and 5.2.4) we make the following assumption regarding the structure of players' utility functions:

A. 11. *The players' utility functions are identical and permutation invariant. That is, for any* $i, j \in N$, $u_i(y) = u_j(y)$, *and* $u([y']_i, [y'']_j, y_{-(i,j)}) = u([y'']_i, [y']_j, y_{-(i,j)})$,

[17] It should be noted that classical FP may be seen as an instance of GWFP, and thus Corollary 1 may in fact be deduced as a corollary to Corollary 2. However, for clarity and continuity of presentation, the results regarding classical FP have been presented separately.

where, for any player $k \in N$, the notation $[y']_k$ indicates the action $y' \in Y_k$ being played by player k, and $y_{-(i,j)}$ denotes the set of actions being played by all players other than i and j.

We note that, under this assumption, the sets C and MCE are nonempty [12, 33]. The following theorem ([34], Theorem 1) specifies the manner in which players engaged in an ECFP process (weakly) learn elements of the sets C and MCE.

THEOREM 6.2. *Let $\{a(t)\}_{t\geq 1}$ be an ECFP process.*
*Assume Γ is such that **A.9** and **A.11** hold. Then players learn equilibrium strategies in the sense that (i) $\lim_{t\to\infty} d(\bar{q}^n(t),\ C) = 0$, and (ii) $\lim_{t\to\infty} d(q(t),\ MCE) = 0$.*

Note that case (i) above corresponds to ECFP with the convergence map f_i^ξ as given in Section 5.2.3, and case (ii) corresponds to the convergence map f_i^ξ given by the identity function (as in Section 5.2.4). Since, by definition, an ECFP process (5.3) allows players to choose actions from the ϵ_t-sub-optimal best response set with $\epsilon_t \to 0$, Theorem 6.2 ensures that ECFP satisfies **A.8**.

Let Ψ be an instance of ECFP as presented in either Section 5.2.3 or Section 5.2.4, and let the strongly convergent variant of Ψ be constructed using the approach laid out in Section 5.3. The following corollary to Theorem 5.1 states that players engaged in the strongly convergent variant of an ECFP process learn elements of C and MCE in the strong sense that players' period-by-period strategies converge to equilibrium.

COROLLARY 3. *(i) Let Ψ be an instance of ECFP with $f_i^\xi(q) = \frac{1}{n}\sum_j q_j,\ \forall i$ and assume Γ is such that **A.9** and **A.11** hold. If the strongly convergent variant of Ψ satisfies **A.1**–**A.3** and **A.10**, then it achieves strong learning in the sense that $\lim_{t\to 0} d(g(t),\ C) = 0$.*
*(ii) Let Ψ be an instance of ECFP with $f_i^\xi(q)$ given by the identity function for all i and assume Γ is such that **A.9** and **A.11** hold. If the strongly convergent variant of Ψ satisfies **A.1**–**A.3** and **A.10**, then it achieves strong learning in the sense that $\lim_{t\to 0} d(g(t),\ MCE) = 0$.*

Proof. Cases (i) and (ii) differ only in terms of the function $f_i^\xi(t)$ and target equilibrium set E. However, in both cases the function f_i^ξ satisfies **A.7**. It suffices to show the remaining conditions of Theorem 5.1 are satisfied. Henceforth we treat cases (i) and (ii) equivalently.

Note that **A.1**–**A.3** and **A.10** hold by assumption. The empirical distribution sequence satisfies $\|q_i(t) - q_i(t-1)\| \leq \frac{M_i}{t} \to 0$ as $t \to \infty$, where $M_i := \sup_{p',p''\in\Delta_i} \|p' - p''\|$, and hence **A.5** is satisfied. Note that the function $f_i^p(q) = \frac{1}{n}\sum_j q_j$ satisfies **A.6**. Finally, Theorem 6.2 shows that **A.8** is satisfied. ∎

7. Conclusions. An algorithm is said to achieve weak learning if players learn an equilibrium strategy in an abstract sense (see Section 2), but period-by-period strategies do not necessarily converge to equilibrium. An algorithm is said to achieve strong learning if (additionally) players' period-by-period strategies converge to equilibrium. Weak learning may be thought of as a form of learning where players *learn* a strategy in some abstract sense, but never begin to implement the strategy they are learning. On the other hand, in strong learning, not only do players *learn* a strategy, but they also physically implement the learned strategy through the course of the learning process.

Fictitious Play (FP) and its variants are known to exhibit weak learning but not necessarily strong learning. An approach was presented for taking a general FP-type algorithm that achieves weak learning, and constructing from it a strongly

convergent variant of the algorithm. General convergence results were proved and used to construct a strongly convergent variant of several example FP-type processes.

In order to apply the convergence results proved in this paper, it is necessary to ensure a candidate algorithm meets **A.8** (the other necessary assumptions are relatively trivial to verify). An interesting future research direction might be to investigate other FP-type algorithms (e.g., [32, 35]) and verify whether they meet the assumptions sufficient for construction of a strongly convergent variant.

Appendix.

7.1. Some Useful Inequalities. We consider some useful inequalities related to the strongly convergent variant of an FP-type algorithm. We restrict attention to realizations $\omega \in \Omega'$. Let $\{q_i(t)\}_{t\geq 1}$ be given by (5.6). By **A.5** there exists a sequence $\gamma(t)$ such that $\lim_{t\to\infty} \gamma(t) = 0$, and for each $i \in N$,

$$\|q_i(t+1) - q_i(t)\| \leq M_i \gamma(\ell_i(t)), \tag{7.1}$$

where $M_i := \sup_{q',q''\in\Delta_i} \|q' - q''\|$. Similarly, there holds for any integer $s > 0$,

$$\|\tilde{q}(s+1) - \tilde{q}(s)\| \leq M\gamma(s), \tag{7.2}$$

where $M := \sup_{q',q''\in\Delta^n} \|q' - q''\|$. More generally, for any integers $s_1, s_2 > 0$, if **A.5** holds then,

$$\|\tilde{q}(s_1) - \tilde{q}(s_2)\| \leq M \sum_{s=\min\{s_1,s_2\}}^{\max\{s_1,s_2\}-1} \gamma(s) \leq |s_1 - s_2|B, \tag{7.3}$$

where $0 < B < \infty$ is such that $\sup_t \gamma(t) \leq B/M$.

7.2. Intermediate Results.

LEMMA 7.1. *Let $\tau_i(s)$ be defined as in section 5.5, and assume **A.10** holds. Then for any realization in Ω' there holds, $\lim_{s\to\infty} |U_i(a_i(\tau_i(s)), p_i(\tau_i(s)-1)) - v_i(p_i(\tau_i(s)-1))| = 0$, $\forall i$.*

Proof. Let $s \in \mathbb{N}$. Note that by definition $\tau_i(s) := \inf\{t : \ell_i(t) = s\}$ and $\ell_i(t) := \sum_{k=1}^{t} X_i(k)$, thus $X_i(\tau_i(s)) = 1$. By (5.7) this implies $a_i(\tau_i(s)) = b_i(\tau_i(s)) \in BR_i^{\eta_{\tau_i(s)}}(p_i(\tau_i(s)-1))$, which implies $|U_i(a_i(\tau_i(s)), p_i(\tau_i(s)-1)) - v_i(p_i(\tau_i(s)-1))| \leq \eta_{\tau_i(s)}$. By **A.10**, $\eta_t \to 0$ as $t \to \infty$, and moreover, by (5.11), $\tau_i(s) \to \infty$ as $s \to \infty$. Thus $\eta_{\tau_i(s)} \to 0$ as $s \to \infty$, and the claim holds. □

LEMMA 7.2. *Let $i, j \in N$, let $\tau_i(s)$ and $\tilde{q}_j(s)$ be defined as in Section 5.5, and assume **A.2**–**A.3** hold. Then for any realization in Ω', $\lim_{s\to\infty} \|q_j(\tau_i(s)) - \tilde{q}_j(s)\| = 0$.*

Proof. Note that for any $t \in \mathbb{N}$, $q_j(t) = q_j(\tau_j(\ell_j(t))) = \tilde{q}_j(\ell_j(t))$, where the first equality follows from Lemma 7.4, and the second equality follows from the definition of $\tilde{q}_i(s)$. Hence,

$$\begin{aligned}\|q_j(\tau_i(s)) - \tilde{q}_j(s)\| = \|\tilde{q}_j(\ell_j(\tau_i(s))) - \tilde{q}_j(s)\| &= \|\tilde{q}_j(\ell_j(\tau_i(s))) - \tilde{q}_j(\ell_i(\tau_i(s)))\| \\ &\leq |\ell_j(\tau_i(s)) - \ell_i(\tau_i(s))|B,\end{aligned}$$

where the first equality follows from the previous statement, and the second equality follows from the fact that $\ell_i(\tau_i(s)) = s$ (see (5.12)), and the final inequality follows from (7.3). Thus, it suffices to show that

$$\lim_{s\to\infty} |\ell_j(\tau_i(s)) - \ell_i(\tau_i(s))| = 0. \tag{7.4}$$

For convenience in notation let $h_i(t) := \sum_{m=1}^{t} \rho_i(m)$. By Lemma 7.6 and the definition of Ω' there holds for any $k \in N$, $\lim_{t\to\infty} \frac{\ell_k(t)}{h_k(t)} = 1$. By assumption **A.3**, for any $k \in N$, $\lim_{t\to\infty} (h_k(t)/(h_i(t)) = 1$. Hence, for any $k \in N$,

$$\lim_{t\to\infty} \frac{\ell_k(t)}{h_i(t)} = \lim_{t\to\infty} \frac{\ell_k(t)}{h_k(t)} \frac{h_k(t)}{h_i(t)} = 1. \tag{7.5}$$

Returning attention to (7.4) and recalling that by (5.11), $\lim_{s\to\infty} \tau_i(s) = \infty$ on Ω', we have,

$$\limsup_{s\to\infty}|\ell_j(\tau_i(s))-\ell_i(\tau_i(s))|\le\limsup_{t\to\infty}|\ell_j(t)-\ell_i(t)|=\limsup_{t\to\infty}\left|\frac{\ell_j(t)}{h_i(t)}h_i(t)-\frac{\ell_i(t)}{h_i(t)}h_i(t)\right|$$
$$=\limsup_{t\to\infty}|h_i(t)-h_i(t)|=0,$$

where the transition to the last line follows from application of (7.5). Thus, (7.4) is verified, and the desired result holds. □

LEMMA 7.3. *Let $i,j\in N$, let $\hat{q}_j^i(s)$ and $\tilde{q}_j(s)$ be defined as in Section 5.5, and assume **A.2**–**A.3** hold. Then for any realization in Ω' there holds* $\lim_{s\to\infty}\|\hat{q}_j^i(s)-\tilde{q}_j(s)\|=0$.

Proof. Recall that by definition, $\hat{q}_j^i(s)=q_j(\tau_i(s+1)-1)$; our objective then is to show that $\lim_{s\to\infty}\|q_j(\tau_i(s+1)-1)-\tilde{q}_j(s)\|=0$. By Lemma 7.2, $\lim_{s\to\infty}\|q_j(\tau_i(s))-\tilde{q}_j(s)\|=0$. By (7.2) and **A.5** there holds, $\lim_{s\to\infty}\|\tilde{q}_j(s+1)-\tilde{q}_j(s)\|=0$. Combining this with the previous statement,

$$\lim_{s\to\infty}\|q_j(\tau_i(s+1))-\tilde{q}_j(s)\|=0. \tag{7.6}$$

Recalling (7.1), there holds,

$$\limsup_{s\to\infty}\|q_j(\tau_i(s+1)-1)-q_j(\tau_i(s+1))\|\le\limsup_{s\to\infty}M_j\gamma(\ell_j(\tau_i(s+1)))=0, \tag{7.7}$$

where the equality holds since $\lim_{s\to\infty}\ell_j(\tau_i(s))=\infty$ on Ω', and by **A.5**, $\lim_{s\to\infty}\gamma(s)=0$.

Consider now the quantity of interest,

$$\|q_j(\tau_i(s+1)-1)-\tilde{q}_j(s)\|\le\|q_j(\tau_i(s+1)-1)-q_j(\tau_i(s+1))\|+\|q_j(\tau_i(s+1))-\tilde{q}_j(s)\|.$$

The first term on the right hand side (RHS) goes to zero by (7.7) and the second term on the RHS goes to zero by (7.6). Thus, $\lim_{s\to\infty}\|q_j(\tau_i(s+1)-1)-\tilde{q}_j(s)\|=0$, and the claim holds. □

LEMMA 7.4. *Let $i\in N$, let $q_i(\cdot)$ be as defined in* (5.6)*, let $\ell_i(\cdot)$ be as defined in* (5.4)*, and let $\tau_i(\cdot)$ be as defined in* (5.5)*. Then for every realization in Ω' and any $t\in\{1,2,\ldots\}$ there holds $q_i(\tau_i(\ell_i(t)))=q_i(t)$.*

Proof. Let $t_0:=\tau_i(\ell_i(t))=\inf\{t':\ell_i(t')=\ell_i(t)\}$, where the second equality follows from the definition of $\tau_i(\cdot)$. Note that $t_0\le t$ and by definition of t_0, there holds $\tau_i(\ell_i(t_0))=t_0$, and hence $q_i(\tau_i(\ell_i(t_0)))=q_i(t_0)$. Furthermore, by the definition of t_0, for $t_0\le t'\le t$, there holds $\ell_i(t)=\ell_i(t')=\ell_i(t_0)$, and hence $\tau_i(\ell_i(t))=\tau_i(\ell_i(t_0))$. Moreover, the fact that $\ell_i(t)=\ell_i(t')=\ell_i(t_0)$ implies by definition of $\ell_i(\cdot)$ that $X_i(t')=0$ for $t_0<t'\le t$ (if such a t' exists). Thus, by (5.14) there holds $q_i(t)=q_i(t')=q_i(t_0)$ for $t_0\le t'\le t$, and in particular $q_i(t)=q_i(t_0)$. Combining this with the facts that $q_i(\tau_i(\ell_i(t_0)))=q_i(t_0)$ and $\tau_i(\ell_i(t))=\tau_i(\ell_i(t_0))$ yields the desired result. □

LEMMA 7.5. *Let $\Psi=(\{f_i^q(\cdot,t)\}_{t\ge1},f_i^p,f_i^\xi)_{i\in N}$ be an FP-type algorithm, and let the strongly convergent variant of Ψ be constructed as in Section 5.3. Let $\tilde{a}(s)$, $\tilde{H}(s)$, and $\tilde{q}_i(s)$ be as defined in Section 5.5. Then for every realization in Ω', and for $s\ge1$, $\tilde{q}_i(s)=f_i^q(\tilde{H}(s),s)$.*

Proof. For $s\ge1$, note that $\tilde{q}_i(s)=q_i(\tau_i(s))=f_i^q(\bar{H}_i(\tau_i(s)),\ell_i(\tau_i(s)))=f_i^q(\tilde{H}(s),s)$, where the first equality follows from the definition of $\tilde{q}_i(s)$ in Section 5.5, the second follows from **A.4**, and the third follows from the definition of $\tilde{H}_i(s)$ in Section 5.5 and (5.12). □

LEMMA 7.6. *Let $\{X(t)\}_{t\ge1}$ be $0-1$ Bernoulli random variables, let $\ell(t):=\sum_{k=1}^t X(k)$ be the associated counting process, let $\mathcal{G}_t:=\sigma(\{X(k)\}_{k=1}^t)$, and let $\rho(t)=\mathbb{P}(X(t)=1|\mathcal{G}_{t-1})$. Assume $\sum_{t\ge1}\rho(t)=\infty$. Then there holds, $\lim_{t\to\infty}(\ell(t))\,/\,\left(\sum_{k=1}^t\rho(t)\right)=1$, a.s.*

Proof. The result follows via Levi's extension of the Borel-Cantelli Lemmas, [30] p.124. □

REFERENCES

[1] G. W. Brown. *"Iterative Solutions of Games by Fictitious Play" In Activity Analysis of Production and Allocation.* Wiley, New York, 1951.

[2] J. Robinson. An iterative method of solving a game. *Ann. Math.*, 54(2):296–301, 1951.

[3] U. Berger. Fictitious play in $2 \times n$ games. *Journal of Economic Theory*, 120(2):139–154, 2005.
[4] P. Milgrom and J. Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 58(6):1255–1277, 1990.
[5] U. Berger. Learning in games with strategic complementarities revisited. *Journal of Economic Theory*, 143(1):292–301, 2008.
[6] A. Sela and D. Herreiner. Fictitious play in coordination games. *International Journal of Game Theory*, 28(2):189–197, 1999.
[7] D. Monderer and L. Shapley. Potential Games. *Games and Econ. Behav.*, 14(1):124–143, 1996.
[8] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control and Optim.*, 44(1):328–348, 2005.
[9] J. S. Jordan. Three problems in learning mixed-strategy Nash equilibria. *Games and Econ. Behav.*, 5(3):368–386, 1993.
[10] H. P. Young. *Strategic learning and its limits*, volume 2002. Oxford University Press, 2004.
[11] D. S. Leslie and E. J. Collins. Generalised weakened fictitious play. *Games and Econ. Behav.*, 56(2):285–298, 2006.
[12] B. Swenson, S. Kar, and J. Xavier. Empirical centroid fictitious play: an approach for distributed learning in multi-agent games. Accepted for publication in IEEE Transactions on Signal Processing, http://arxiv.org/abs/1304.4577, 2012.
[13] B. Swenson, S. Kar, and J. Xavier. Distributed learning in large-scale multi-agent games: A modified fictitious play approach. In *46th Asilomar Conference on Signals, Systems, and Computers*, pages 1490 – 1495, Pacifc Grove, CA, USA, 2012.
[14] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*, volume 2. MIT press, 1998.
[15] J. R. Marden, G. Arslan, and J. S. Shamma. Joint strategy fictitious play with inertia for potential games. *IEEE Trans. Automat. Contr.*, 54(2):208–220, 2009.
[16] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma. Payoff based dynamics for multiplayer weakly acyclic games. *SIAM J. Control and Optim.*, 48(1):373–396, 2009.
[17] G. C. Chasparis, A. Arapostathis, and J. S. Shamma. Aspiration learning in coordination games. *SIAM J. Control and Optim.*, 51(1):465–490, 2013.
[18] B. Pradelski and H. P. Young. Learning efficient Nash equilibria in distributed systems. *Games and Econ. Behav.*, 75(2):882–879, 2012.
[19] J. R. Marden and J. S. Shamma. Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation. *Games and Econ. Behav.*, 75(2):788–808, 2012.
[20] S. Rass and B. Rainer. Numerical computation of multi-goal security strategies. In *Decision and Game Theory for Security*, pages 118–133. Springer, 2014.
[21] M. Voorneveld. Pareto-optimal security strategies as minimax strategies of a standard matrix game. *J. Optimiz. Theory App.*, 102(1):203–210, 1999.
[22] T. Alpcan and T. Basar. *Network Security: A Decision and Game-Theoretic Approach*. Cambridge University Press, 2010.
[23] K. Dabcevic, A. Betancourt, L. Marcenaro, and C. S. Regazzoni. A fictitious play-based game-theoretical approach to alleviating jamming attacks for cognitive radios. In *Acoust. Speech, Signal Proc. (ICASSP), 2014 IEEE Int. Conf. on*, pages 8158–8162. IEEE, 2014.
[24] O. Candogan, A. Ozdaglar, and P. A. Parrilo. Dynamics in near-potential games. *Games and Econ. Behav.*, 82:66–90, 2013.
[25] D. P. Foster and H. P. Young. Regret testing: A simple payoff-based procedure for learning Nash equilibrium. *University of Pennsylvania and Johns Hopkins University (mimeo)*, 2003.
[26] F. Germano and G. Lugosi. Global Nash convergence of Foster and Young's regret testing. *Games and Econ. Behav.*, 60(1):135–154, 2007.
[27] D. Fudenberg. Learning mixed equilibria. *Games and Econ. Behav.*, 5(3):320–367, 1993.
[28] J. Hofbauer and W. H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.
[29] B. Swenson, S. Kar, and J. Xavier. Strong convergence to mixed equilibria in fictitious play. In *Information Sciences and Systems, 48th Annual Conference on*, pages 1–6. IEEE, 2014.
[30] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.
[31] D. Monderer and L. S. Shapley. Fictitious play property for games with identical interests. *Journal of Economic Theory*, 68(1):258–265, 1996.
[32] G. Arslan and J. S. Shamma. Distributed convergence to Nash equilibria with local utility measurements. In *Proc. of the 43rd IEEE Conf. on Decision and Control*, volume 2, pages 1538 – 1543, 2004.
[33] B. Swenson, S. Kar, and J. Xavier. Mean-centric equilibrium: An equilibrium concept for learning in large-scale games. In *IEEE Glob. Conf. Signal Inf. Process.*, pages 571–574,

2013.

[34] B. Swenson, S. Kar, and J. Xavier. On robustness properties in empirical centroid fictitious play. Submitted for conference publication, http://arxiv.org/abs/1504.00391, 2015.

[35] J. S. Shamma and G. Arslan. Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Trans. Automat. Contr.*, 50(3):312–327, 2005.