Distributed Convergence to Nash Equilibria with Local Utility Measurements

Gurdal Arslan and Jeff S. Shamma

Abstract— We consider multiplayer repeated matrix games in which several players seek to increase their individual rewards by updating their strategies based on limited information. One body of work assumes that players can measure the actions of other players, but do not have access to the utility functions of other players. In this case, well known strategy update mechanisms such as Fictitious Play (FP) and Gradient Play (GP) provide convergence to Nash equilibria in certain special classes of games. Recent work by the authors introduced "dynamic" versions of FP and GP, where players use derivative action to process and respond to the information available to them. These mechanisms, called *derivative action FP* and *derivative Action GP*, lead to behavior converging to Nash equilibria in a significantly larger set of games than standard FP and GP provide. In this paper, we consider the case where players *do not* have access to opposing actions. As before, players do not have access to opposing player utility functions. Furthermore, a player's access to its own utility function is restricted to the *measured* utility at each round of the repeated game—structural parameters of its own utility remain unknown. Our main result is to show that derivative action FP and GP can be adapted to the utility measurement case to yield the same dynamics (in continuous-time and up to a coordinate transformation) as though players could measure other player actions. The transformation holds for both two-player games as well as in multiplayer games with a specific utility structure. The implication is that many of the stability and convergence properties obtained under derivative action FP and GP can be extended to the utility measurement case.

I. OVERVIEW

There is a substantial body of literature on the topic of learning in games and the related topic of evolutionary games. This includes several recent monographs [1], [2], [3], [4], [5]. At issue in much of this work is understanding the limiting behavior of interacting players that adapt their strategies given incomplete information. Of particular concern is whether player strategies will converge to a Nash equilibrium. In this regard, many strategy update mechanisms have been analyzed and a variety of convergence—and non-convergence—results have been obtained.

Our particular concern in this paper is how one may overcome non-convergence properties that are exhibited by a broad class of strategy update mechanisms. To motivate this problem, consider a multiplayer repeated matrix game in which players have access to the actions taken by other players. Players *do not* have access to the strategies that generated these actions. Nor do players have access to

the utility functions of other players. Let us presume that each player keeps a running histogram of the actions of the other players. Call these running averages "empirical frequencies". The paper [6] showed that if players use strategies that are functions of the *current value* of the empirical frequencies, then convergence to a (mixed) Nash equilibrium cannot occur. The result strongly relies on utility functions not being shared among players. This non-convergence result is reminiscent of earlier results, such as [7], that established non-convergence for certain special classes of strategy update mechanisms.

Recent work by the authors [8], [9], showed that it is possible to overcome this lack of convergence by processing the empirical frequencies in a "dynamic" manner, i.e., by allowing strategies to depend on the evolution of the empirical frequencies, and not just their current values. From a control theory perspective, this is akin to using dynamic feedback versus static feedback. These papers showed that the use of "derivative action" can lead to Nash equilibria in situations that other strategic update mechanisms can not. Derivative action can be interpreted as a device for approximate anticipation of opposing player moves. Indeed, such an interpretation was taken in [10], which established convergence to Nash equilibrium in zero-sum games that used forecasts based on averaging over intervals.

The papers [8], [9] considered a *continuous-time* version of repeated games and assumed that players can measure the actions of other players. This paper analyzes the discrete-time case and uses dynamical systems methods of stochastic approximation as in [11], [12] to establish positive probabilities of convergence to Nash equilibria in *discrete-time* under certain conditions. These conditions admit the possibility of establishing convergence in the derivative action case when convergence was impossible in other approaches. This paper goes on to consider the more restrictive case where players can only measure the reward received at each stage. Players do not have access to the actions of other players, nor do they have access to the structural parameters of their own utility functions. In this more restrictive setting, the paper presents derivative action versions of utility measurement processing and again establishes positive probability of convergence to Nash equilibria in discrete-time under certain conditions.

Notation

— For $i \in \{1, 2, \ldots, n\}$, $-i$ denotes the complementary set $\{1, \ldots, i-1, i+1, \ldots, n\}$.

— Boldface $\mathbf{1}$ denotes the vector $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathcal{R}^n$.

G. Arslan is with the Department of Electrical Engineering, University of Hawaii at Manoa, 440 Holmes Hall, 2540 Dole Street, Honolulu, HI 96822, gurdal@hawaii.edu

J. S. Shamma is with the Department of Mechanical and Aerospace Engineering, University of California, 37-146 Engineering IV, Los Angeles, CA 90095-1597, shamma@seas.ucla.edu

— $\Delta(n)$ denotes the simplex in $\mathcal{R}^n$, i.e.,

$$\left\{ s \in \mathcal{R}^n \mid s \geq 0 \text{ componentwise, and } \mathbf{1}^T s = 1 \right\}.$$

— $\text{Int}(\Delta(n))$ denotes the set of interior points of a simplex, i.e., $s > 0$ componentwise.
— For $\varepsilon > 0$, $\Delta_\varepsilon(n)$ denotes the set

$$\left\{ s \in \mathcal{R}^n \mid s \geq \varepsilon \text{ componentwise, and } \mathbf{1}^T s = 1 \right\}.$$

— $\Pi_K : \mathcal{R}^n \to K$ denotes the projection to the convex $K \subset \mathcal{R}^n$,

$$\Pi_K[x] = \arg \min_{s \in K} |x - s|,$$

where $|\cdot|$ denotes the usual 2-norm in $\mathcal{R}^n$.
— x^i denotes the ith component of the vector x. The convention will be to reserve subscripts for indexing the players of a game.
— $\mathbf{v}^i \in \Delta(n)$ denotes the ith vertex of the simplex $\Delta(n)$, i.e., the vector whose i^{th} term equals 1 and remaining terms equal 0.
— $\mathcal{H} : \text{Int}(\Delta(n)) \to \mathcal{R}$ denotes the entropy function

$$\mathcal{H}(s) = -s^T \log(s).$$

— $\sigma : \mathcal{R}^n \to \Delta(n)$ denotes the "logit" or "soft-max" function

$$(\sigma(x))^i = \frac{e^{x^i}}{e^{x^1} + \cdots + e^{x^n}}.$$

II. SETUP

This section outlines our framework and notation for learning in games. A suitable reference is [1].

A. Matrix Games and Smoothed Best Response

We consider a multiplayer game with players $\mathcal{P}_1, ..., \mathcal{P}_{n_\text{P}}$, where n_P is the number of players. In the non-repeated (one-shot) game, each player, $\mathcal{P}_i$, generates a random action, $a_i \in \{1, ..., m_i\}$, according to the player's *strategy*, p_i, which is a probability distribution in $\Delta(m_i)$. Each player receives a real-valued reward according to its *utility function* $U_i(a)$, which is evaluated on the total action profile $a = (a_1, ..., a_{n_\text{P}})$. These utility functions may be extended to the product space of probability distributions in the usual way by identifying

$$U_i(p) \overset{\text{def}}{=} \mathbf{E}_p[U_i(a)],$$

where $p = (p_1, \ldots, p_{n_\text{P}})$ denotes the total strategy profile.

Define the "smoothed" utility function,

$$\mathcal{U}_i(p) = \mathbf{E}_p[U_i(a)] + \tau \mathcal{H}(p_i),$$

The entropy term $\tau \mathcal{H}(p_i)$ may be viewed as a τ-weighted reward for randomization. Other interpretations, including connections to information theory, are discussed in [13].

We will use p_{-i} to denote the collection of strategies of players *other than* player $\mathcal{P}_i$, i.e.,

$$p_{-i} = (p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_{n_\text{P}}).$$

With this notation, we will sometimes write a strategy profile p as (p_i, p_{-i}). Similarly, we may write $U_i(p)$ as $U_i(p_i, p_{-i})$ and $\mathcal{U}_i(p)$ as $\mathcal{U}_i(p_i, p_{-i})$,

Using the above notation, a strategy profile p^* is called a *Nash equilibrium* if, for all $i \in \{1, ..., n_\text{P}\}$,

$$\mathcal{U}_i(p_i^*, p_{-i}^*) \geq \mathcal{U}_i(p_i, p_{-i}^*), \ \forall \ p_i \in \Delta(m_i).$$

For $\tau = 0$, a Nash equilibrium is *strict* if the above holds with strict inequalities, and a Nash equilibrium is *completely mixed* if $p_i^* \in \text{Int}(\Delta(m_i))$ for all $i \in \{1, ..., n_\text{P}\}$.

Define the *best response* function as

$$\beta_i(p_{-i}) = \arg \max_{s \in \Delta(m_i)} \mathcal{U}_i(s, p_{-i}).$$

In case $\tau = 0$, then $\beta_i(\cdot)$ may be multi-valued. In the smoothed case, $\tau > 0$, then the best response function may be written explicitly as follows. Define $G_i(p_{-i})$ as the m_i-dimensional vector

$$G_i(p_{-i}) = \begin{pmatrix} U_i(\mathbf{v}^1, p_{-i}) \\ \vdots \\ U_i(\mathbf{v}^{m_i}, p_{-i}) \end{pmatrix}.$$

The jth component of $G_i(p_{-i})$ may be interpreted as the expected reward to player $\mathcal{P}_i$ when using action j given that other players use strategies p_{-i}. The vector $G_i(\cdot)$ is also the gradient

$$G_i(p_{-i}) = \nabla_{p_i} U_i(p_i, p_{-i}).$$

In terms of $G_i(\cdot)$, the (single-valued) best response function is given by the logit or soft-max function (see Notation)

$$\beta_i(p_{-i}) = \sigma\left(\frac{1}{\tau} G_i(p_{-i})\right).$$

B. Repeated Matrix Games with Restricted Information

Suppose now that the game is sequentially repeated over stages $k \in \{0, 1, 2, \ldots\}$. At each stage, k, player $\mathcal{P}_i$ uses its *current* strategy, $p_i(k)$, to generate its current action, $a_i(k)$. Again, each player receives a reward, $U_i(a(k))$, according to its utility function evaluated on the total current action profile.

Player strategies, $p_i(k)$, are updated, or adapted, at each stage according to the information available to player $\mathcal{P}_i$ over times $\{0, \ldots, k-1\}$. Two commonly investigated informational assumptions are:

- At each stage, player $\mathcal{P}_i$ can observe the actions of other players, $a_{-i}(k)$, and knows the structural form of its own utility function, $U_i(\cdot)$.

  -or-

- Player $\mathcal{P}_i$ can only measure the realized value, $U_i(a(k))$, of its own utility.

The second assumption is a more stringent restriction of information, and this scenario will be the ultimate focus of this paper.

C. Fictitious Play (FP) and Gradient Play (GP)

Define the *empirical frequency*, $q_i(\cdot)$, to be the running averages of the actions of players, i.e.,

$$q_i(k+1) = q_i(k) + \frac{1}{k+1}(\mathbf{v}^{a_i(k)} - q_i(k)),$$

where actions, $a_i(k)$, are generated as random outcomes to the evolving strategies, $p_i(k)$.

In the scenario where player actions are public knowledge, then empirical frequencies are also public knowledge, and hence these can be used as part of processes that define the strategies $p_i(k)$. We will review two such processes.

The first process is smooth *fictitious play* (FP). Using $\tau > 0$, a players's strategy is the best response to the observed empirical frequencies, i.e.,

$$p_i(k) = \beta_i(q_{-i}(k)).$$

The second process is *gradient play* (GP), in which a player's strategy is

$$p_i(k) = \Pi_{\Delta_\varepsilon}[q_i(k) + G_i(q_{-i}(k))],$$

for some small $\varepsilon > 0$. The interpretation of gradient play is that the strategy is updated according to the evolving gradient of the non-smoothed ($\tau = 0$) utility. In order to impose some level of "exploration", the projection is to Δ_ε which lies in the interior of the original simplex. Such exploration will be used to obtain the desired discrete-time convergence properties from continuous-time analysis.

III. REVIEW OF DERIVATIVE ACTION PLAY

In this section, we first review the continuous-time "derivative action" versions of FP and GP introduced in [8], [9]. We will go on to define discrete-time versions of derivative action FP and GP and establish certain probabilistic convergence properties that are derived from a combination of the local stability analysis in [9] and results from the dynamical systems method of stochastic approximation (e.g., [11], [12]). These tools show that one can infer certain probabilistic convergence properties of stochastic discrete-time iterations by analyzing appropriate deterministic continuous-time equations.

The continuous-time version of FP is

$$\dot{q}_i = -q_i + \beta_i(q_{-i}), \tag{1}$$

It is straightforward to see that the only stationary points of (1) are Nash equilibria of the smoothed matrix game. The continuous-time version of GP is

$$\dot{q}_i = -q_i + \Pi_{\Delta_\varepsilon}[q_i + G_i(q_{-i})]. \tag{2}$$

Assume that all Nash equilibria of the non-smoothed matrix game are either strict or completely mixed. Then for sufficiently small ε, the stationary points of (2) are either 1) the original completely mixed Nash equilibria or 2) vertices of Δ_ε that are of order ε distance from the original strict Nash equilibria.

A. Derivative Action FP

An interpretation of continuous-time FP is that the state-variable, $q_i(t)$, evolves as a low-pass filtered version of the continuous-time strategy, i.e., $p_i(t) = \beta_i(q_{-i}(t))$. Derivative action FP seeks to exploit the *derivative*, $\dot{q}_{-i}(t)$, in defining a player's strategy, $p_i(t)$. Ideally, this would take the form

$$p_i(t) = \beta_i(q_{-i}(t) + \gamma \dot{q}_{-i}(t)).$$

Derivative action may be viewed as an anticipatory best response, since

$$\beta_i(q_{-i}(t) + \gamma \dot{q}_{-i}(t)) \approx \beta_i(q_{-i}(t + \gamma))$$

In case $\gamma = 1$, derivative action also has the interpretation of an attempt to "invert" the low-pass filter dynamics that map strategies into empirical frequencies.

The non-idealized case recognizes that exact derivative measurements are not available. Accordingly, references [8], [9], use the following approximate differentiator implementation,

$$\begin{aligned}
\dot{q}_i &= -q_i + \beta_i(q_{-i} + \gamma \dot{r}_{-i}) \\
\dot{r}_i &= \lambda(q_i - r_i)
\end{aligned} \tag{3}$$

with $\lambda > 0$. The intent is that for large λ, r_i closely tracks q_i, and so $\dot{r}_i$ may be a good approximation for $\dot{q}_i$. It turns out such intuition need not hold, because the ability to reconstruct the derivative $\dot{q}_i$ depends on the *second* derivative magnitude $\ddot{q}_i$. The implication is that the asymptotic convergence of the idealized situation might not be "recovered" using an approximate differentiator implementation.

We now state a result that characterizes under what conditions the approximate differentiator implementation (3) maintains local asymptotic stability. An important implication is that approximate differentiator FP can lead to a Nash equilibrium even when standard FP fails to converge.

Define

$$q(t) = (q_1(t), ..., q_{n_P}(t)), \quad r(t) = (r_1(t), ..., r_{n_P}(t)).$$

Clearly, (q^*, q^*) is a stationary point of the dynamics (3) if and only if q^* is a Nash equilibrium of the smoothed matrix game. Since $q_i(t)$ and $r_i(t)$ are probability distributions for all $t \geq 0$ (assuming of course $q_i(0)$ and $r_i(0)$ are probability distributions), we can write the deviation $(q(t) - q^*, r(t) - q^*)$ as $\mathcal{N}\delta x$, for some δx, where $\mathcal{N}$ is a block diagonal matrix with each block being an orthonormal matrix whose columns span the null space of a row vector $\mathbf{1}^T$ of appropriate dimension. Linearizing the dynamics of $(q(t), r(t))$ around (q^*, q^*) results in

$$\frac{d}{dt}\delta x = \begin{pmatrix} -I + (1 + \gamma\lambda)\mathcal{D} & -\gamma\lambda\mathcal{D} \\ \lambda I & -\lambda I \end{pmatrix} \delta x, \tag{4}$$

for some matrix $\mathcal{D}$ with $-I + \mathcal{D}$ being the Jacobian matrix of the linearization of standard FP (1) around q^*. The following result from [9] establishes that the linearized

dynamics (4) can be locally stable with a suitable derivative gain γ when the linearization of standard FP is unstable.

Theorem 3.1 ([9]): Consider a multiplayer game with a Nash equilibrium p^* under derivative action FP described by (3). Assume that $\mathcal{D}$ in (4) is non-singular. Let $a_i + jb_i$ denote the eigenvalues of $-I + \mathcal{D}$. The linearization (4) is asymptotically stable for large $\lambda > 0$ if and only if

1) $\max_i a_i < \frac{1-\gamma}{\gamma}$, *if* $\max_i a_i < 0$;
2) $\max_i \frac{a_i}{a_i^2 + b_i^2} < \frac{\gamma}{1-\gamma} < \frac{1}{\max_i a_i}$, *if* $\max_i a_i \geq 0$.

Condition 1 in Theorem 3.1 implies that the linearization of standard FP is asymptotically stable. In this case, any $0 < \gamma < 1$ renders the derivative action FP linearization (4) stable. Condition 2 implies that derivative action FP may have a stable linearization in situations where standard FP does not.

B. Derivative Action GP

In the spirit of the prior modification of FP, we now define a derivative action version of GP. First, note that the gradient functions $G_i(p_{-i})$ can be extended beyond the product space of probability distributions to a domain that includes all $\mathcal{R}^n$ (of appropriate dimension). In this case, the gradient loses its "expected value" interpretation. With this extension, derivative action GP is defined as

$$\dot{q}_i = -q_i + \Pi_{\Delta_\varepsilon}[q_i + \gamma G_i(q_{-i} + \dot{r}_{-i})]$$
$$\dot{r}_i = \lambda(q_i - r_i). \qquad (5)$$

The following result is a straightforward generalization of a similar result stated in [9].

Theorem 3.2: Consider a multiplayer game under derivative action GP described by (5) with a completely mixed Nash equilibrium p^* satisfying $p_i^* > \varepsilon \mathbf{1}$ (element-by-element). Assume that the Jacobian matrix $\tilde{\mathcal{M}}$ of the linearization of standard GP (2) around p^* is non-singular. Let $a_i + jb_i$ denote the eigenvalues of $\tilde{\mathcal{M}}$. The linearization of (5) around (p^*, p^*) is asymptotically stable for large $\lambda > 0$ if and only if

$$\max_i \{a_i/(a_i^2 + b_i^2)\} < \gamma < 1/\max_i\{a_i\}.$$

Note that the trace of the Jacobian matrix $\tilde{\mathcal{M}}$ of the linearization of standard GP (2) vanishes, and so no completely mixed equilibrium can be asymptotically stable under standard GP (2). Theorem 3.2 shows that the use of derivative action in GP renders a mixed Nash equilibrium locally asymptotically stable in a large class of games.

C. Discrete Time Algorithms and Positive Probabilities of Convergence

The stability results presented in the previous section revealed that continuous-time approximate derivative action FP may render a Nash equilibrium locally stable even though the same Nash equilibrium may be unstable under standard FP. Similar statements hold for GP. The implication of such local stability for the corresponding discrete-time dynamics is convergence to Nash equilibrium with positive

probability provided that an attainability condition is satisfied, as established in [11], [14]. It turns out that the randomization induced by either the entropy terms in derivative action FP ($\tau > 0$) or the exploration parameter in derivative action GP ($\varepsilon > 0$) will assure the attainability condition in discrete-time versions of approximate derivative action FP and GP, respectively.

Define

$$q_i(k+1) = q_i(k) + \frac{1}{k+1}(\mathbf{v}^{a_i(k)} - q_i(k)),$$
$$r_i(k+1) = r_i(k) + \frac{\lambda}{k+1}(q_i(k) - r_i(k)), \qquad (6)$$

where actions, $a_i(k)$, are generated as random outcomes to the evolving strategies, $p_i(k)$. The discrete-time approximate derivative action FP strategy is

$$p_i(k) = \beta_i(q_{-i}(k) + \gamma\lambda(q_{-i}(k) - r_{-i}(k))). \qquad (7)$$

whereas the approximate derivative action gradient play strategy is

$$p_i(k) = \Pi_{\Delta_\varepsilon}[q_i(k) + \gamma G_i(q_{-i}(k) + \lambda(q_{-i}(k) - r_{-i}(k)))]. \qquad (8)$$

The following theorem is a direct consequence of Proposition 7.5 of [11].

Theorem 3.3:

1) *Consider a multiplayer game under discrete-time approximate derivative action FP, described by (6) and (7), with a Nash equilibrium p^*. Let γ and λ satisfy the stability conditions of Theorem 3.1. Then the random sequence $(q_i(k), r_i(k))$ converges to (p^*, p^*) with non-zero probability.*

2) *Consider a multiplayer game under discrete-time approximate derivative action GP, described by (6) and (8), with a completely mixed Nash equilibrium p^* satisfying $p_i^* > \varepsilon \mathbf{1}$ (element-by-element). Let γ and λ satisfy the stability conditions of Theorem 3.2. Then the random sequence $(q_i(k), r_i(k))$ converges to (p^*, p^*) with non-zero probability.*

IV. DERIVATIVE ACTION ON UTILITY MEASUREMENTS

A. Utility Measurement Processing

We now analyze the case where players do *not* have access to each other's actions. Rather, at each stage, a player only measures the reward received at that stage. Players are not even aware of which players influence their reward. Rather than track empirical frequencies, the "bookkeeping" done by each player is an estimate of the average reward obtained when using a specific action.

More precisely, the utility measurement framework proceeds as follows. At stage k, player $\mathcal{P}_i$ plays an action $a_i(k) \in \{1, \ldots, m_i\}$, according to the current strategy $p_i(k)$ and receives the reward $U_i(a(k))$, where $a(k) = (a_1(k), \ldots, a_{n_p}(k))$ is the overall action profile. Upon observing $U_i(a(k))$, player $\mathcal{P}_i$ updates an estimate of the

average utility received for using $a_i(k)$ as follows (see [15]):

$$
\overline{U}_i^\ell(k+1) =
\begin{cases}
\overline{U}_i^\ell(k) + \frac{1}{(k+1)p_i^\ell(k)}(U_i(a_i(k)) - \overline{U}_i^\ell(k)), & \text{if } a_i(k) = \ell; \\
\overline{U}_i^\ell(k), & \text{otherwise}
\end{cases} \quad (9)
$$

where $p_i^\ell(k)$ is the ℓth component of $p_i(k)$, i.e., $\text{Prob}[a_i(k) = \ell]$, and $\overline{U}_i^\ell(k)$ represents player $\mathcal{P}_i$'s estimate of the average reward over time for using action $\ell \in \{1, \ldots, m_i\}$.

In anticipation of "derivative action" on utility measurements, we also define

$$
\overline{W}_i(k+1) = \overline{W}_i(k) + \frac{\lambda}{k+1}(\overline{U}_i(k) - \overline{W}_i(k)), \quad (10)
$$

where $\overline{U}_i(k) = (\overline{U}_i^1(k), \ldots, \overline{U}_i^{m_i}(k))$ and $\lambda > 0$.

We will investigate two forms of utility measurement processing.

The first is *utility measurement derivative action FP*. At time k, the strategy of player $\mathcal{P}_i$ is

$$
p_i(k) = \sigma\left(\frac{1}{\tau}\left(\overline{U}_i(k) + \gamma\lambda(\overline{U}_i(k) - \overline{W}_i(k))\right)\right), \quad (11)
$$

for some $\gamma > 0$ and $\tau > 0$.

The second is *utility measurement derivative action GP*. At time k, the strategy of player $\mathcal{P}_i$ is

$$
p_i(k) = \Pi_{\Delta_\varepsilon}[q_i(k) + \gamma(\overline{U}_i(k) + \lambda(\overline{U}_i(k) - \overline{W}_i(k)))], \quad (12)
$$

for some small exploration rate $\varepsilon > 0$, and where

$$
q_i(k+1) = q_i(k) + \frac{1}{k+1}(\mathbf{v}^{a_i(k)} - q_i(k)). \quad (13)
$$

Note that in utility measurement derivative action GP, each player computes the empirical frequencies of *its own* actions, but still does not observe the actions of other players.

B. Special Case: Pairwise Structured Utility Functions

We will now show that a under certain utility function structure, the resulting ODE analysis of utility measurement processing leads to the same equations (up to a change of coordinates) as the case where players could construct empirical frequencies. The implication is that the utility measurement versions inherit similar probabilistic convergence properties as their empirical frequency measurement counterparts.

Assumption 4.1: The (non-smoothed) utility functions $U_i(p_i, p_{-i})$ have the form

$$
U_i(p_i, p_{-i}) = \sum_{j \neq i} p_i^T M_{ij} p_j,
$$

for matrices M_{ij}.

Assumption 4.1 imposes a "pairwise" structure in the utility functions, i.e., the total utility is the sum of pairwise interactions with other players. In the case of two players, Assumption 4.1 is satisfied trivially.

1) Utility Measurement Derivative Action FP Analysis: The following theorem states that under the structure of Assumption 4.1, utility measurement derivative action FP inherits the same convergence properties as its "action measurement" counterpart.

Theorem 4.1: Let p^* be a Nash equilibrium of a smoothed ($\tau > 0$) mutliplayer game with a utility structure as in Assumption 4.1. Let $p(k) = (p_1(k), \ldots, p_{n_{\mathrm{P}}}(k))$ be the strategy profile generated by utility measurement derivative action FP, described by (9), (10), and (11). If (p^*, p^*) is a locally asymptotically stable equilibrium of (3), then $\text{Prob}\left[\lim_{k \to \infty} p(k) = p^*\right] > 0$.

The necessary and sufficient conditions for the local asymptotic stability of (p^*, p^*) of (3) for large λ are provided in Theorem 3.1. We note that *global* asymptotic stability would imply almost sure convergence to the Nash equilibrium [11].

The proof of Theorem 4.1 relies on showing that the differential equations suggested by Proposition 7.5 of [11] are identical, up to a coordinate transformation, to those of (3).

Towards this end, we first compute

$$
\mathbf{E}\left[\overline{U}_i^\ell(k+1) - \overline{U}_i^\ell(k)|\overline{U}(k), \overline{W}(k)\right]
$$
$$
= \frac{1}{k+1}\left(\sum_{a:a_i=\ell} U_i(a)\prod_{j \neq i} p_j^{a_j}(k) - \overline{U}_i^\ell(k)\right)
$$

and

$$
\mathbf{E}\left[\overline{W}_i(k+1) - \overline{W}_i(k)|\overline{U}(k), \overline{W}(k)\right]
$$
$$
= \frac{\lambda}{k+1}(\overline{U}_i(k) - \overline{W}_i(k)),
$$

where

$$
\overline{U}(k) = (\overline{U}_1(k), \ldots, \overline{U}_{n_{\mathrm{P}}}(k)),
$$
$$
\overline{W}(k) = (\overline{W}_1(k), \ldots, \overline{W}_{n_{\mathrm{P}}}(k)).
$$

These lead to the differential equations (for $\ell \in \{1, \ldots, m_i\}$)

$$
\dot{\overline{U}}_i^\ell = -\overline{U}_i^\ell + \sum_{a:a_i=\ell} U_i(a)\prod_{j \neq i}\sigma^{a_j}\left(\frac{1}{\tau}(\overline{U}_j + \gamma\dot{\overline{W}}_j)\right),
$$
$$
\dot{\overline{W}}_i = \lambda(\overline{U}_i - \overline{W}_i). \quad (14)
$$

The stationary points of (14) satisfy

$$
\overline{U}_i^* = \overline{W}_i^* =
\begin{pmatrix}
\sum_{a:a_i=1} U_i(a)\prod_{j \neq i}\sigma^{a_j}(\frac{1}{\tau}\overline{U}_j^*) \\
\vdots \\
\sum_{a:a_i=m_i} U_i(a)\prod_{j \neq i}\sigma^{a_j}(\frac{1}{\tau}\overline{U}_j^*)
\end{pmatrix}.
$$

Therefore, the corresponding stationary strategies, defined as $p_i^* = \sigma(\overline{U}_i^*/\tau)$, satisfy

$$
p_i^* = \sigma\left(\frac{1}{\tau}\left(
\begin{matrix}
\sum_{a:a_i=1} U_i(a)\prod_{j \neq i}(p_j^*)^{a_j} \\
\vdots \\
\sum_{a:a_i=m_i} U_i(a)\prod_{j \neq i}(p_j^*)^{a_j}
\end{matrix}
\right)\right),
$$

which corresponds to a Nash equilibrium of the smoothed game.

Under Assumption 4.1, it is possible to simplify (14) to

$$\dot{\overline{U}}_i = -\overline{U}_i + \sum_{j \neq i} M_{ij} \sigma \left(\frac{1}{\tau} (\overline{U}_j + \gamma \lambda (\overline{U}_j - \overline{W}_j)) \right)$$
$$\dot{\overline{W}}_i = \lambda (\overline{U}_i - \overline{W}_i) \quad (15)$$

Assumption 4.1 also implies that (3) may be written as

$$\dot{q}_i = -q_i + \sigma \left(\frac{1}{\tau} \sum_{j \neq i} M_{ij} (q_j + \gamma \lambda (q_j - r_j)) \right)$$
$$\dot{r}_i = \lambda (q_i - r_i). \quad (16)$$

Local asymptotic stability of (16) now implies local asymptotic of (15) through the identification

$$\overline{U}_i \leftrightarrow \sum_{j \neq i} M_{ij} q_j \quad \text{and} \quad \overline{W}_i \leftrightarrow \sum_{j \neq i} M_{ij} r_j.$$

2) Utility Measurement Derivative Action GP Analysis: The following theorem states that under the structure of Assumption 4.1, utility measurement derivative action GP inherits the same convergence properties as its "action measurement" counterpart. This theorem is analogous to Theorem 4.1.

Theorem 4.2: Let p^* be a Nash equilibrium of a non-smoothed $(\tau = 0)$ multiplayer game with a utility structure as in Assumption 4.1. Let $p(k) = (p_1(k), ..., p_{n_p}(k))$ be the strategy profile generated by utility measurement derivative action GP, described by (9), (10), (12), and (13). Assume that p^* is completely mixed and satisfies $p_i^* > \varepsilon \mathbf{1}$. If (p^*, p^*) is a locally asymptotically stable equilibrium of (5), then $\text{Prob}\left[\lim_{k \to \infty} p(k) = p^*\right] > 0$.

Necessary and sufficient conditions for the local asymptotic stability of (p^*, p^*) of (5) are provided in Theorem 3.2. As before, global asymptotic stability would imply almost sure convergence to the Nash equilibrium [11].

As before, the proof of Theorem 4.2 relies on analyzing the differential equations suggested by Proposition 7.5 of [11], which are

$$\dot{q}_i = p_i - q_i$$
$$\dot{\overline{U}}_i^\ell = -\overline{U}_i^\ell + \sum_{a:a_i=\ell} U_i(a) \Pi_{j \neq i} p_i^{a_j}$$
$$\dot{\overline{W}}_i = \lambda (\overline{U}_i - \overline{W}_i), \quad (17)$$

where

$$p_i = \Pi_{\Delta_\varepsilon} [q_i + \gamma(\overline{U}_i + \lambda(\overline{U}_i - \overline{W}_i))].$$

Note that the only empirical frequencies used by a player are its own.

Using Assumption 4.1 leads to the simplification,

$$\dot{q}_i = p_i - q_i$$
$$\dot{\overline{U}}_i = -\overline{U}_i + \sum_{j \neq i} M_{ij} p_j$$
$$\dot{\overline{W}}_i = \lambda (\overline{U}_i - \overline{W}_i) \quad (18)$$

One can show that local asymptotic stability of (5) implies local asymptotic stability of (18). The argument is more involved than a immediate identification, because (18) is of higher order. However, the dimensionality of (18) is effectively reduced because the quantity $\sum_{j \neq i} M_{ij} q_j - \overline{U}_i$ decays exponentially.

V. CONCLUDING REMARKS

This paper has shown how multiple players using only local utility measurements can evolve towards a mixed strategy Nash equilibrium in a repeated game setting through the use of derivative action. Two (of several) unresolved issues in this framework are 1) whether the local stability of derivative action is actually globally attractive in the case of a unique Nash equilibrium and 2) if other simple mechanisms complementary to derivative action can lead to similar or stronger convergence results.

REFERENCES

[1] D. Fudenberg and D. Levine, *The Theory of Learning in Games.* Cambridge, MA: MIT Press, 1998.
[2] J. Hofbauer and K. Sygmund, *Evolutionary Games and Population Dynamics.* Cambridge University Press, 1998.
[3] L. Samuelson, *Evolutionary Games and Equilibrium Selection.* Cambridge, MA: MIT Press, 1997.
[4] H. P. Young, *Individual Strategy and Social Structure.* Princeton, NJ: Princeton University Press, 1998.
[5] J. Weibull, *Evolutionary Game Theory.* Cambridge, MA: MIT Press, 1995.
[6] S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, vol. **93**, no. 5, pp. 1830–1836, 2003.
[7] V. Crawford, "Learning behavior and mixed strategy Nash equilibria," *Journal of Economic Behavior and Organization*, vol. **6**, pp. 69–78, 1985.
[8] J. Shamma and G. Arslan, "A feedback stabilization approach to fictitious play," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, 2003, pp. 4140–4145.
[9] ——, "Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria," 2003, accepted for publication in *IEEE Transactions on Automatic Control*, http://www.seas.ucla.edu/~shamma/papers.html.
[10] J. Conlisk, "Adaptation in games: Two solutions to the Crawford puzzle," *Journal of Economic Behavior and Organization*, vol. **22**, pp. 25–50, 1993.
[11] M. Benaim, "Dynamics of stochastic approximation algorithms," in *Seminaire de Probabilites XXXIII*, J. Azema *et al.*, Eds. Springer-Verlag Lecture Notes in Mathematics, 1999, vol. 1709, pp. 1–68.
[12] H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications.* Springer-Verlag, 1997.
[13] D. H. Wolpert, "Information Theory – The Bridge Connecting Bounded Rational Game Theory and Statistical Physics," 2004, http://arxiv.org/PS_cache/cond-mat/pdf/0402/0402508.pdf.
[14] M. Benaim and M. Hirsch, "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games," *Games and Economic Behavior*, vol. **29**, pp. 36–72, 1999.
[15] D. Fudenberg and D. Levine, "Consistency and cautious fictitious play," *Journal of Economic Dynamics & Control*, vol. **19**, pp. 1065–1089, 1995.