Learning in games with quantized payoff observations

Kyriakos Lotidis, Panayotis Mertikopoulos, Nicholas Bambos

Learning in Games with Quantized Payoff Observations

Kyriakos Lotidis[§], Panayotis Mertikopoulos[◇] and Nicholas Bambos[†]

***Abstract*— This paper investigates the impact of feedback quantization on multi-agent learning. In particular, we analyze the equilibrium convergence properties of the well-known "follow the regularized leader" (FTRL) class of algorithms when players can only observe a quantized (and possibly noisy) version of their payoffs. In this information-constrained setting, we show that coarser quantization triggers a qualitative shift in the convergence behavior of FTRL schemes. Specifically, if the quantization error lies below a threshold value (which depends *only* on the underlying game and not on the level of uncertainty entering the process or the specific FTRL variant under study), then (*i*) FTRL is attracted to the game's strict Nash equilibria with arbitrarily high probability; and (*ii*) the algorithm's asymptotic rate of convergence remains the same as in the non-quantized case. Otherwise, for larger quantization levels, these convergence properties are lost altogether: players may fail to learn anything beyond their initial state, even with full information on their payoff vectors. This is in contrast to the impact of quantization in continuous optimization problems, where the quality of the obtained solution degrades smoothly with the quantization level.**

I. Introduction

In the implementation of distributed learning and control systems, observations and feedback often need to be quantized down to the bit-resolution allowed by the sensing/sampling and data communication rates. This is driven by various design pressures, including sensing/sampling and communication bandwidth constraints, as well as computation, memory, and power limitations. In particular, such challenges are ubiquitous in current and emerging distributed systems (like the Internet of Things or edge/mobile computing), where edge devices must often contend with granular, reduced-precision data and measurements. For example, a mobile device may only be able to measure the quality of its downlink channel up to a relatively low precision and then request setting the downlink transmitter power (which also affects other devices via interference) based on low-rate feedback. Likewise, an edge computing node may only be able to receive a low-bit representation of the data of a control-plane application and must then process and resubmit this data using some low-bit encoding.

Reduced-precision settings of this type can be modeled efficiently by assuming that, in addition to any random factors affecting the process, observable quantities are also *quantized*, reflecting the granularity of the measurement / communication process. With this in mind, our paper examines *quantized multi-agent learning processes* that unfold as follows:

1) At each stage $n = 1, 2, \dots$, every participating agent selects an action from some finite set.
2) Each agent's reward is determined by their chosen action and that of all other participating agents.
3) Agents observe a noisy quantized version of their rewards, they update their actions, and the process repeats.

In terms of the agents' learning dynamics – i.e., the way that they update their actions – we consider the widely studied *"follow the regularized leader"* (FTRL) class of algorithms, as introduced by [1] and containing as special cases the seminal multiplicative/exponential weights algorithm of [2]–[4], as well as the standard projection dynamics of [5], [6]. Within this setting, we aim to address the following questions:

1) What is the *impact of quantization* on the learning process relative to the non-quantized case?
2) Is there *robust deterioration* – i.e., graceful degradation, as opposed to abrupt collapse – of the outcome of the learning process as the coarseness of the quantization increases?

Related work: The literature on learning in games has traditionally focused on identifying when a learning process converges to equilibrium – locally or globally. In this regard, a widely known result is that the empirical frequency of play under no-regret learning converges to the game's set of *coarse correlated equilibria* [7], [8]. However, since this set may contain highly undesirable, dominated strategies [9], this convergence result typically needs to be refined.

On that account, a very large body of work has focused on the sharper question of convergence to a *Nash equilibrium* (NE), i.e., a state from which no player has an incentive to deviate unilaterally. This question is much more difficult and only partial results are known: as a representative (but otherwise incomplete) list of relevant results, [10]–[13] established the convergence of an "adjusted" variant of FTRL to approximate Nash equilibria in potential, $2 \times m$, and $2 \times \cdots \times 2$ games. This convergence was established under the assumption that players receive perfect realizations of their in-game payoffs – i.e., there are no observation or measurement errors, random or otherwise. More recently, and under similar feedback assumptions, [14] showed that, in any generic game, strict Nash equilibria – i.e., Nash equilibria where each player has a unique best response – are *precisely* the states that are stable and attracting under the (unadjusted) dynamics of FTRL in discrete time.

The algorithmic stability and convergence results discussed above were achieved via the use of an *importance-weighted estimator* (IWE) which provides a counterfactual surrogate for the payoff that a player would have obtained from an action that

[§]Dept. of Management Science & Engineering, Stanford University, klotidis@stanford.edu
[◇]CNRS, Univ. Grenoble Alpes, panayotis.mertikopoulos@imag.fr
[†]Dept. of Electrical Engineering and Management Science & Engineering, Stanford University, bambos@stanford.edu

they did not actually pick. The key property of this estimator is that its bias can be balanced against its variance so as to yield progressively more accurate payoff predictions with only a mild deterioration in precision. In turn, this "asymptotic unbiasedness" property plays a major role in the convergence results discussed above because it allows players to eventually gravitate towards actions that yield consistently better payoffs against the "mean field" of the other players' actions.

However, this crucial property is lost the moment quantization enters the picture: the granularity of the players' payoff observations can *never* become finer than the quantization gap of their feedback / measurement mechanism, so any learning process would not be able to resolve this gap either. Indeed, *any* payoff estimator must contend with a persistent bias that disallows the resolution of payoffs corresponding to nearby mixed strategies – e.g., playing $(1/3, 1/3, 1/3)$ versus $(1/3, 1/3 - \varepsilon, 1/3 + \varepsilon)$ in Rock-Paper-Scissors for sufficiently small ε. As a result, any learning process that relies on gradual changes in the players' mixed strategies – like FTRL and its variants – would seem unable to make consistent progress towards a Nash equilibrium, even if starting relatively close.

Our contributions: Our analysis paints a different account of the above. First, if the quantization error does not exceed a certain threshold value, we show that FTRL with quantized feedback – dubbed "*follow the quantized leader*" (FTQL) for short – continues to identify strict Nash equilibria *with perfect accuracy*, despite the *persistent bias* induced by the quantization process. More precisely, we show that strict Nash equilibria are locally stable and attracting with arbitrarily high probability under FTQL, just as in the case of FTRL with *perfect* payoff-based feedback. Second, we derive a series of sharp convergence rate estimates for FTQL which echo the convergence speed of FTRL with *non-quantized* feedback as derived recently in [15]. Specifically, *despite the quantization*, the convergence rate of FTQL differs from its non-quantized variant only by a multiplicative constant, showing that the algorithm's asymptotic rate of convergence remains otherwise unimpeded by the coarseness of the quantization scheme (as long as this coarseness does not exceed the critical level beyond which learning is impossible).

Importantly, this quantization threshold depends *only* on the underlying game and is otherwise independent of the level of uncertainty involved and/or the specific FTQL variant in play. Beyond this threshold, the learning landscape changes abruptly and dramatically. In particular, for larger values of the quantization gap, the convergence properties of FTQL are lost altogether: players may fail to learn anything beyond their initial state, even with full information on their mixed payoff vectors (even in simple 2×2 common interest games that are otherwise easy to learn).

This behavior comes in stark contrast to the impact of quantization in continuous optimization where the quality of the obtained solution degrades gracefully with the quantization gap [16]–[18]. This suggests a fundamental shift in design principles when dealing with game-theoretic problems as above: the robust deterioration observed in the discretization of continuous optimization problems is no longer present, and the quantization granularity has to be tuned judiciously as a function of the agents' interactions.

II. Background and motivation

A. Games in normal form

Throughout this paper, we consider normal form games with a finite number of players and a finite number of actions per player. More precisely, we posit that each player, indexed by $i \in \mathcal{N} = \{1, \dots, N\}$, has a finite set of *actions* – or *pure strategies* – $\alpha_i \in \mathcal{A}_i$ and a *payoff function* $u_i : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} \coloneqq \prod_{i \in \mathcal{N}} \mathcal{A}_i$ denotes the set of all possible action profiles $\alpha = (\alpha_1, \dots, \alpha_N)$. Players can mix their strategies, i.e., play a probability distribution $x_i \in \mathcal{X}_i \coloneqq \Delta(\mathcal{A}_i)$ over their pure strategies, and we write $x = (x_1, \dots, x_N) \in \mathcal{X} \coloneqq \prod_{i \in \mathcal{N}} \mathcal{X}_i$ for the associated *mixed strategy profile*. For notational convenience, we will also write $\mathcal{Y}_i \coloneqq \mathbb{R}^{\mathcal{A}_i}$ and $\mathcal{Y} \coloneqq \prod_{i \in \mathcal{N}} \mathcal{Y}_i$, for the space of payoff vectors of player $i \in \mathcal{N}$ and the ensemble thereof.

Given a mixed strategy profile $x \in \mathcal{X}$, we will use the standard shorthand $x = (x_i; x_{-i})$ to keep track of the mixed strategy profile x_{-i} of all players other than i, and we further define

(i) The *expected payoff* of player i under x:

$$u_i(x) = u_i(x_i, x_{-i}) \tag{1}$$

(ii) The *mixed payoff vector* of player i under x:

$$v_i(x) = (u_i(\alpha_i; x_{-i}))_{\alpha_i \in \mathcal{A}_i} \tag{2}$$

In words, $v_i(x) \in \mathcal{Y}_i$ simply collects the expected payoffs $v_{i\alpha_i}(x) \coloneqq u_i(\alpha_i; x_{-i})$, $\alpha_i \in \mathcal{A}_i$, that player $i \in \mathcal{N}$ would have obtained by playing $\alpha_i \in \mathcal{A}_i$ against the mixed strategy profile x_{-i} of all other players. Then, aggregating over all players $i \in \mathcal{N}$, we will also write $v(x) = (v_1(x), \dots, v_N(x)) \in \mathcal{Y}$ for the ensemble of the players' mixed payoff vectors.

In terms of solution concepts, we say that a strategy profile x^* is a *Nash equilibrium* (NE) if no player has an incentive to unilaterally deviate from it, i.e.,

$$u_i(x^*) \geq u_i(x_i; x^*_{-i}) \quad \forall x_i \in \mathcal{X}_i, \forall i \in \mathcal{N}. \tag{NE}$$

Finally, we say that x^* is a *strict Nash equilibrium* if the inequality in (NE) is strict for all $x_i \neq x^*_i$, $i \in \mathcal{N}$, i.e., if any deviation from x^*_i results to a strictly worse payoff for the deviating player $i \in \mathcal{N}$. It is straightforward to verify that a strict equilibrium $x^* \in \mathcal{X}$ is also *pure* in the sense that each player assigns positive probability only to a single pure strategy.

B. Quantization: definitions and impact on learning

1) Basics of quantization: As we discussed in the introduction, our paper concerns models of repeated play where all observable quantities – the players' payoffs, the associated vectors, etc. – are subject to rounding and/or precision cutoffs. To formalize this, let $\ell > 0$ be the *quantization error* of the players' observation / measurement device, and let $\mathcal{R} \colon \mathbb{R} \to \ell\mathbb{Z} \equiv \{\dots, -2\ell, -\ell, 0, \ell, 2\ell, \dots\}$ be the associated *quantization operator* which reduces to the identity on $\ell\mathbb{Z}$, and which maps any real number $x \in \mathbb{R}$ to an integer multiple $\mathcal{R}(x) \in \ell\mathbb{Z}$ of

ℓ such that $|x - \mathcal{R}(x)| \leq \ell/2$ for all $x \in \mathbb{R}$. For example, the "floor" operation $x \mapsto \lfloor x \rfloor$ has quantization error $\ell = 2$ (since $\sup_x |x - \lfloor x \rfloor| = 1$), whereas the "round half away from zero" (or "commercial rounding") operation $\mathcal{R}(x) = \text{sgn}(x)\lfloor |x| + 1/2 \rfloor$ in Python and Java has a quantization error of $\ell = 1$.

Vectorizing this construction in the obvious way, we will write $\mathcal{R}(v) := (\mathcal{R}(v_k))_{k=1,\dots,d}$ for an arbitrary vector $v \in \mathbb{R}^d$. Then, by construction, $\mathcal{R}$ reduces to the identity on $(\ell\mathbb{Z})^d$ and we have

$$\|\mathcal{R}(v) - v\|_\infty \leq \ell/2 \quad \text{for all } v \in \mathbb{R}^d. \tag{3}$$

Unless explicitly mentioned otherwise, we will not assume a specific quantization operator in the sequel, and we will state our results only as a function of the quantization error ℓ.

2) The impact on learning: To motivate the analysis to come, we provide below two examples where the process of quantization can lead to significant challenges in multi-agent learning, even in the simplest case where players observe their full (quantized) payoff vectors.

For concreteness, we will present our examples in the context of the well-known *exponential* (or *multiplicative*) *weights* (EW) algorithm [2]–[4] which, in our setting, can be written as

$$\begin{aligned} X_{i\alpha_i,n} &\propto \exp(Y_{i\alpha_i,n}) \\ Y_{i,n+1} &= Y_{i,n} + \gamma_n V_{i,n} \end{aligned} \tag{EW}$$

where (i) $X_{i,n} \in \mathcal{X}_i$ is the mixed strategy of player $i \in \mathcal{N}$ at the n-th stage of the process; (ii) $V_{i,n}$ is an approximation of the player's mixed payoff vector $v_i(X_n)$ which we discuss in detail below; (iii) $Y_{i,n} \in \mathcal{Y}_i$ is an auxiliary "score vector" that aggregates payoff information (so $Y_{i\alpha_i,n}$ indicates the propensity of player i to employ the pure strategy $\alpha_i \in \mathcal{A}_i$); and (iv) $\gamma_n > 0$ is a "learning rate" (or step-size) parameter that controls the weight with which new information enters the algorithm.

With all this in hand, the examples that follow are intended to highlight two critical issues: *a*) the evolution of (EW) when $V_{i,n}$ is obtained by rounding $v_i(X_n)$ at different precision cutoffs; and *b*) the difference between learning in Γ with quantized feedback versus learning with *non-quantized* feedback in a quantized version $\mathcal{R}(\Gamma)$ of the original game.

Example 1 (The role of the quantization error)**.** Consider a two-player common-interest game with $\mathcal{A}_1 = \{a_1, a_2\}, \mathcal{A}_2 = \{b_1, b_2\}$, and rewards given by the following payoff matrix:

Player 1/2	b_1	b_2
a_1	99.1	100.9
a_2	100.9	99.1

Clearly, (a_1, b_2) is a strict Nash equilibrium of the game.

We now examine the case where, at each round $n = 1, 2, \dots$, both players observe their quantized mixed payoff vectors $V_{i,n} = \mathcal{R}(v(X_n))$, $i = 1, 2$, and subsequently update their strategies according to (EW). The specific quantization schemes we consider are as follows:

(i) "*Round to closest even away from zero*" ($\ell = 2$): this scheme maps x to the closest even integer and resolves ties by moving away from 0, i.e., $\mathcal{R}(x) = 2\text{sgn}(x)\lfloor |x|/2 + 1/2 \rfloor$. Then, for any initial mixed strategy profile $X_1 \in \mathcal{X}$ that assigns positive probability to all actions, all coordinates of $v(1)$ will lie in the interval $(99.1, 100.9)$, so every entry of $\mathcal{R}(v(1))$ will in turn be equal to 100. We thus conclude that all coordinates of Y_n will be increased by the same amount in the iterative step $Y_n \leftarrow Y_{n+1}$. Since this constant increase disappears under the normalization step in (EW), we readily obtain $X_n = X_1$ for all $n = 1, 2, \dots$, i.e., the players' strategy profile remains unchanged for all time in this learning model.[1]

(ii) "*Round half away from zero*" ($\ell = 1$): as discussed above, this scheme maps x to the closest integer and resolves ties by moving away from 0, i.e., $\mathcal{R}(x) = \text{sgn}(x)\lfloor |x| + 1/2 \rfloor$. For simplicity, we assume that the learning rate is constant, say $\gamma_n = 1, \forall n$. Now, taking $(X_{a_1,1}, X_{a_2,1}) = (0.8, 0.2)$ and $(X_{b_1,1}, X_{b_2,1}) = (0.2, 0.8)$, we readily obtain $\mathcal{R}(v_{a_1}(X_1)) = \mathcal{R}(v_{b_2}(X_1)) = 101$, and $\mathcal{R}(v_{a_2}(X_1)) = \mathcal{R}(v_{b_1}(X_1)) = 99$. As a result, the corresponding score differences for $n = 1$ satisfy: $Y_{a_1,n+1} - Y_{a_2,n+1} > Y_{a_1,n} - Y_{a_2,n}$ and $Y_{b_2,n+1} - Y_{b_1,n+1} > Y_{b_2,n} - Y_{b_1,n}$ from which we readily get $X_{a_1,n+1} > X_{a_1,n}$ and $X_{b_2,n+1} > X_{b_2,n}$ Therefore, inductively we have that $v_{a_1}(X_n)$ and $v_{b_2}(X_n)$ increase as n grows, while $v_{a_2}(X_n)$ and $v_{b_1}(X_n)$ decrease. We thus obtain $\mathcal{R}(v_{a_1}(X_n)) = \mathcal{R}(v_{b_2}(X_n)) = 101$, and $\mathcal{R}(v_{a_2}(X_n)) = \mathcal{R}(v_{b_1}(X_n)) = 99$ for all n. Hence:

$$\begin{aligned} Y_{a_1,n+1} - Y_{a_2,n+1} &= Y_{a_1,n} - Y_{a_2,n} + 2 \\ &= Y_{a_1,1} - Y_{a_2,1} + 2n \end{aligned} \tag{4}$$

from which we readily get

$$\frac{\exp(Y_{a_1,n+1})}{\exp(Y_{a_2,n+1})} = \frac{\exp(Y_{a_1,1})}{\exp(Y_{a_2,1})} \exp(2n) \tag{5}$$

Taking $n \to \infty$ we obtain $(X_{a_1,n}, X_{a_2,n}) \to (1, 0)$ as $n \to \infty$, and, likewise: $(X_{b_1,n}, X_{b_2,n}) \to (0, 1)$ as $n \to \infty$. We thus conclude that X_n converges to a strict Nash equilibrium.

From the above, we see that for two different quantization lengths, the learning process may exhibit a completely different behavior: in (i) it remains static throughout the execution of the algorithm, whereas in (ii) X_n converges to a strict Nash equilibrium of the underlying game. As we will see later, there is a threshold value ℓ associated with the minimum payoff differences, where this transition is sharp. §

Example 2 (Learning with quantized feedback vs. learning in the quantized game)**.** This example is intended to highlight the difference between learning in Γ with quantized feedback versus learning with perfect feedback in a quantized version $\mathcal{R}(\Gamma)$ of Γ. As before, suppose there are two players, 1 and 2, with action spaces $\mathcal{A}_1 = \{a_1, a_2\}$ and $\mathcal{A}_2 = \{b_1, b_2\}$ respectively, and let $\mathcal{R}(x) = \text{sgn}(x) \cdot \lfloor |x| + 1/2 \rfloor$. The payoff matrix of the original game Γ, along with the quantized version of it, is shown below:

Player 1\2	b_1	b_2
a_1	$0.04 \xrightarrow{\mathcal{R}} 0$	$0.8 \xrightarrow{\mathcal{R}} 1$
a_2	$0.8 \xrightarrow{\mathcal{R}} 1$	$0.04 \xrightarrow{\mathcal{R}} 0$

[1] Note here that the precise values of the game are not important: we would obtain the same result if we replaced $\{99.1, 100.9\}$ with $\{99 + \varepsilon, 101 - \varepsilon\}$ for any $\varepsilon > 0$ sufficiently small.

We denote by X_n, Y_n and $\tilde{X}_n, \tilde{Y}_n$ the sequences of states generated by (EW) on Γ and $\mathcal{R}(\Gamma)$ respectively. Moreover, we assume for concreteness that $\gamma_n = 1$ for all n, and $X_1 = \tilde{X}_1$ with $(X_{a_1,1}, X_{a_2,1}) = (0.6, 0.4)$ and $(X_{b_1,1}, X_{b_2,1}) = (0.4, 0.6)$. So, the different procedures are as follows:

(i) "Γ *with quantized feedback*": In this setting, players observe $V_n = \mathcal{R}(v(X_n))$ at the n-th stage. By the initial conditions, we obtain $V_1 = 0$, which means that all coordinates of the score vector remain unchanged, so, inductively, we get $Y_n = Y_1$ and hence $X_n = X_1$ for all stages, i.e., the learning process does not evolve.

(ii) "$\mathcal{R}(\Gamma)$ *without quantization*": Unlike the previous setting, the players observe the full payoff vector of $\mathcal{R}(\Gamma)$, i.e., $V_n = \mathbb{E}_{\tilde{\alpha}_n \sim \tilde{X}_n}[\mathcal{R}(v(\tilde{\alpha}_n))]$. By the initial conditions, we have $(V_{a_1,1}, V_{a_2,1}) = (0.6, 0.4)$ and $(V_{b_1,1}, V_{b_2,1}) = (0.4, 0.6)$. Therefore, the corresponding score differences for $n = 1$ satisfy: $\tilde{Y}_{a_1,n+1} - \tilde{Y}_{a_2,n+1} > \tilde{Y}_{a_1,n} - \tilde{Y}_{a_2,n}$ and $\tilde{Y}_{b_2,n+1} - \tilde{Y}_{b_1,n+1} > \tilde{Y}_{b_2,n} - \tilde{Y}_{b_1,n}$. With a similar reasoning as in Example 1(ii), we have: $X_{a_1,n+1} > X_{a_1,n}$ and $X_{b_2,n+1} > X_{b_2,n}$. Iterating over n, we get: $(X_{a_1,n}, X_{a_2,n}) \to (1, 0)$ and $(X_{b_1,n}, X_{b_2,n}) \to (0, 1)$ i.e., X_n converges to a strict equilibrium of $\mathcal{R}(\Gamma)$.

The above shows a remarkable difference in behavior: in the case of Γ with quantized feedback, players learn nothing beyond their initial state; by contrast, learning with *perfect* feedback in the quantized game $\mathcal{R}(\Gamma)$ converges to the strict Nash equilibrium (a_1, b_2). This serves to highlight the fact that learning with quantized feedback cannot be compared to learning in a quantized game: the players' end behavior is drastically different in the two cases. §

III. The learning model

We now proceed to describe our general model for learning with quantized feedback; for ease of reference, we will refer to this scheme as *follow the quantized leader* (FTQL).

Viewed abstractly, our model is based on the standard FTRL template [1] run with quantized (and possibly noisy) payoff observations as follows:

$$\begin{aligned} X_{i,n} &= Q_i(Y_{i,n}) \\ Y_{i,n+1} &= Y_{i,n} + \gamma_n V_{i,n} \end{aligned} \tag{FTQL}$$

In more detail, the defining elements of (FTQL) are (i) the approximate payoff vectors $V_{i,n} \in \mathcal{Y}_i$ which are reconstructed from the players' payoff observations; and (ii) the players' "choice maps" $Q_i \colon \mathcal{Y}_i \to \mathcal{X}_i$ which determine each player's mixed strategy $X_{i,n} \in \mathcal{X}_i$ as a function of the "aggregate payoff" variables $Y_{i,n} \in \mathcal{Y}_i$. In the rest of this section, we describe both of these elements in detail; for a pseudocode implementation of the method, see also Algorithm 1 below.

1) The feedback process: The vanilla version of FTRL assumes that each player $i \in \mathcal{N}$ observes the full (mixed) payoff vector $V_{i,n} \leftarrow v_i(X_n)$ in order to update their individual score vector $Y_{i,n}$ at each stage n. However, in our model, we only assume that players observe a quantized – and possibly noisy – version of their in-game, realized payoffs. Specifically, if $\hat{\alpha}_{i,n} \in \mathcal{A}_i$ denotes the action (pure strategy) chosen by the i-th player at stage n, we assume that each player receives as feedback the quantized reward

$$\hat{u}_{i,n} = \mathcal{R}[u_i(\hat{\alpha}_{i,n}; \hat{\alpha}_{-i,n}) + \xi_{i,n}] \tag{6}$$

where $\mathcal{R}$ is a quantization operator with gap ℓ (cf. Section II) and $\xi_{i,n} \in \mathbb{R}$, $n = 1, 2, \dots$, is a random, zero-mean error capturing all sources of uncertainty in the process. Specifically, letting $\mathcal{F}_n$ denote the history (natural filtration) of X_n, we will make the following statistical assumptions for ξ_n:

$$\text{Zero-mean:} \quad \mathbb{E}[\xi_{i,n} \,|\, \mathcal{F}_n] = 0 \tag{7a}$$

$$\text{Finite variance:} \quad \mathbb{E}[|\xi_{i,n}|^2 \,|\, \mathcal{F}_n] \leq \sigma^2 \tag{7b}$$

i.e., ξ_n is an L^2-bounded martingale difference sequence relative to the history of play up to stage n (inclusive).

To reconstruct their payoff vectors from the quantized feedback model (6), we further assume that players employ the *importance-weighted estimator* (IWE)

$$V_{i\alpha_i,n} = \frac{\mathbb{1}\{\hat{\alpha}_{i,n} = \alpha_i\}}{\hat{X}_{i\alpha_i,n}} \hat{u}_{i,n} \tag{IWE}$$

where $\hat{X}_{i\alpha_i,n} = (1 - \varepsilon_n) X_{i\alpha_i,n} + \varepsilon_n / |\mathcal{A}_i|$ denotes the probability with which player i selects action $\alpha_i \in \mathcal{A}_i$ at stage n given the mixed strategy profile $X_{i,n} \in \mathcal{X}_i$ and an "*explicit exploration*" parameter $\varepsilon_n > 0$. The role of this parameter will be discussed in detail in the next section.

2) The players' choice maps: As mentioned above, the second defining element of (FTQL) is the players' choice map $Q_i \colon \mathcal{Y}_i \to \mathcal{X}_i$ whose role is to translate the "aggregate score" vectors $Y_{i,n} \in \mathcal{Y}_i$ into mixed strategies $X_{i,n} = Q_i(Y_{i,n}) \in \mathcal{Y}_i$. This choice map is in turn defined as a regularized best response of the form

$$Q_i(y_i) = \arg\max_{x_i \in \mathcal{X}_i} \{\langle y_i, x_i \rangle - h_i(x_i)\} \tag{8}$$

where $h_i \colon \mathcal{X}_i \to \mathbb{R}$ denotes the method's namesake "regularizer function".

For concreteness, we will focus on a class of decomposable regularizers of the form $h_i(x_i) = \sum_{\alpha_i \in \mathcal{A}_i} \theta_i(x_{i\alpha_i})$ where the "kernel function" $\theta_i \colon [0, 1] \to \mathbb{R}$ is (i) continuous on $[0, 1]$; (ii) twice differentiable on $(0, 1]$; and (iii) strongly convex, i.e., $\inf_{t \in (0,1]} \theta_i''(z) > 0$. Two standard examples of such functions are:

1) The *entropic regularizer* $\theta_i(z) = z \log z$: a standard calculation shows that the induced choice map is $Q_i(y_i) = (\exp(y_{i\alpha_i}))_{\alpha_i \in \mathcal{A}_i} / \sum_{\alpha_i \in \mathcal{A}_i} \exp(y_{i\alpha_i})$ which leads to the exponential (or multiplicative) weights update template (EW) of Section II.

2) The *Euclidean regularizer* $\theta_i(z) = z^2/2$: trivially, the induced choice map is the closest point projection $Q_i(y_i) = \arg\min_{x_i \in \mathcal{X}_i} \|y_i - x_i\|_2$, and the induced scheme is the projection dynamics [5], [6].

An important distinction between these regularizers is that $\theta_i'(0^+) = -\infty$ for the entropic regularizer while $\theta_i'(0^+)$ is finite for the Euclidean one. Regularizers that have the former behavior are called *steep* and have the property that the induced mirror map is interior-valued; regularizers with the latter

Algorithm 1: Follow the quantized leader (FTQL)

Initialize: Y_1
for $n = 1, 2, \dots$ **do**
  $X_{i,n} \leftarrow Q_i(Y_{i,n})$
  Update sampling strategy: $\hat{X}_{i\alpha_i,n} \leftarrow (1-\varepsilon_n)X_{i\alpha_i,n} + \frac{\varepsilon_n}{|\mathcal{A}_i|}$
  Sample $\hat{\alpha}_n \sim \hat{X}_n$
  Observe realized payoff: $\hat{u}_{i,n} \leftarrow \mathcal{R}(u_i(\alpha_i; \alpha_{-i}) + \xi_{i,n})$
  Estimate payoff vector through (IWE):

$$V_{i\alpha_i,n} \leftarrow \frac{\mathbb{1}\{\hat{\alpha}_{i,n} = \alpha_i\}}{\hat{X}_{i\alpha_i,n}} \hat{u}_{i,n}$$

  Update score vectors: $Y_{i,n+1} \leftarrow Y_{i,n} + \gamma_n V_{i,n}$
end for

behavior are called *non-steep* and have surjective mirror maps [19]. This behavior is captured by the *rate function*

$$\phi_i(y) = \begin{cases} 0 & \text{if } y \leq \theta_i'(0^+) \\ 1 & \text{if } y \geq \theta_i'(1^-) \\ (\theta_i')^{-1}(y) & \text{otherwise} \end{cases} \tag{9}$$

As we shall see below, this rate function plays a crucial role in determining the rate of convergence of (FTQL).

IV. Analysis and results

We are now in a position to proceed with the convergence analysis of the quantized learning scheme (FTQL). The first thing to note is that a finite game may admit several Nash equilibria – an odd number generically – so it is not reasonable to expect a global convergence result that applies to all games. For this reason, we will focus below on states that are *locally* stable and attracting:

Definition 1. Let X_n, $n = 1, 2, \dots$, be the sequence of mixed strategy profiles generated by (FTQL). We then say that $x^* \in \mathcal{X}$ is:

1) *Stochastically stable* if, for every confidence level $\delta > 0$ and every neighborhood $\mathcal{U}$ of x^* in $\mathcal{X}$, there exists a neighborhood $\mathcal{U}_1$ of x^* in $\mathcal{X}$ such that

$$\mathbb{P}(X_n \in \mathcal{U} \text{ for all } n \,|\, X_1 \in \mathcal{U}_1) \geq 1 - \delta. \tag{10}$$

2) *Attracting* if, for every confidence level $\delta > 0$, there exists a neighborhood $\mathcal{U}_1$ of x^* in $\mathcal{X}$ such that

$$\mathbb{P}(X_n \to x^* \text{ as } n \to \infty \,|\, X_1 \in \mathcal{U}_1) \geq 1 - \delta. \tag{11}$$

3) *Stochastically asymptotically stable* if it is stochastically stable and attracting.

Informally, the above states that x^* is stochastically stable if every trajectory X_n of (FTQL) that starts sufficiently close to x^* remains nearby with arbitrarily high probability; in addition, if X_n converges to x^* as well, then x^* is stochastically *asymptotically* stable [20], [21]. On that account, states that are (stochastically) asymptotically stable under (FTQL) are the only states that can be considered as viable, stable outcomes of the learning process.

In the context of FTRL with perfect, *non-quantized* payoff observations, it is known that a state is stochastically asymptotically stable *if and only if* it is a *strict* Nash equilibria of Γ [14]. With this in mind, and given that the advent of quantization can only worsen the attraction properties of any given point (cf. the relevant discussion in Section II), we will exclusively focus below on the asymptotic stability and attraction properties of strict Nash equilibria under (FTQL).

In this regard, our main result can be summarized along the following two axes:

1) If the quantization error ℓ is smaller than a threshold value ℓ^* that depends only on the underlying game, every strict Nash equilibrium of Γ is stochastically asymptotically stable under (FTQL).
2) Conditioned on the above, convergence to a strict equilibrium $x^* \in \mathcal{X}$ occurs at a rate of $\|X_n - x^*\|_1 \leq \phi\Big(-\Theta\Big(\sum_{k=1}^n \gamma_k\Big)\Big)$, where ϕ is the rate function (9).

The idea of our proof is to find a set of suitable initial conditions for the quantized version of $v(X_n)$ to remain in the interior of the *normal cone* $\mathrm{NC}(x^*)$ of $\mathcal{X}$ at x^* throughout the execution of the algorithm. For this, we need to delve into the geometry of $\mathrm{NC}(x^*)$ and find the limitations in the quantization legnth ℓ that guarantee that X_n will be contained in the desired region.

We start with the following lemma that gives a specific description of the *normal cone* at a vertex x^* of the polytope $\mathcal{X}$.

Lemma 1. *Let x^* be of the form $(e_{1\alpha_1^*}, \dots, e_{N\alpha_N^*})$, where $e_{i\alpha_i^*} \in \mathbb{R}^{|\mathcal{A}_i|}$ a standard basis vector. Then the normal cone of $\mathcal{X}$ at x^* can be expressed as:*

$$\mathrm{NC}(x^*) = \{w \in \mathcal{Y} : w_{i\alpha_i} - w_{i\alpha_i^*} \leq 0, \forall i \in \mathcal{N}, \alpha_i \in \mathcal{A}_i\} \tag{12}$$

Proof. We have that $\mathcal{X} = \{x \in \mathbb{R}^{|\mathcal{A}|} : \sum_{\alpha_i \in \mathcal{A}_i} x_{i\alpha_i} = 1, x_{i\alpha_i} \geq 0, \forall \alpha_i \in \mathcal{A}_i, i \in \mathcal{N}\}$, for $|\mathcal{A}| = |\mathcal{A}_1| + \cdots + |\mathcal{A}_N|$. We can equivalently write it in standard form, as:

$$\mathcal{X} = \{x \in \mathbb{R}^{|\mathcal{A}|} : Cx = e, x_{i\alpha_i} \geq 0, \forall \alpha_i \in \mathcal{A}_i, i \in \mathcal{N}\} \tag{13}$$

where C is a $N \times |\mathcal{A}|$ matrix whose i-th row of C is $c_i^T = (0, \dots, 0, 1, \dots, 1, 0, \dots, 0)$, with ones in positions $(|\mathcal{A}_1| + \cdots + |\mathcal{A}_{i-1}| + 1), \dots, (|\mathcal{A}_1| + \cdots + |\mathcal{A}_i|)$. Then, every vertex $\overline{x}$ of $\mathcal{X}$ is of the form: $\overline{x}_{i\overline{\alpha}_i} = 1$ for some $\overline{\alpha}_i \in \mathcal{A}_i$ and $\overline{x}_{i\alpha_i} = 0, \forall \alpha_i \neq \overline{\alpha}_i \in \mathcal{A}_i, \forall i \in \mathcal{N}$. Hence, x^* is an extreme point of the bounded polytope $\mathcal{X}$ and the set of adjacent vertices of x^* is the set $\mathcal{Z} = \{x^* - e_{i\alpha_i^*} + e_{i\alpha_i} : \alpha_i \in \mathcal{A}_i, i \in \mathcal{N}\}$. Now, let $\mathcal{C} = \{w : \langle w, z - x^* \rangle \leq 0, \forall z \in \mathcal{Z}\}$. The tangent cone of $\mathcal{X}$ at x^* equals to the closure of the cone of feasible directions at x^*, and since $\mathcal{X}$ is a convex polytope, we get: $\mathrm{TC}(x^*) = \mathrm{cone}(\{z - x^* : z \in \mathcal{Z}\})$. Since $\mathrm{NC}(x^*) = (\mathrm{TC}(x^*))^\circ := \{w : \langle w, x \rangle \leq 0, \forall x \in \mathrm{TC}(x^*)\}$, it remains to show that $\mathcal{C} = \mathrm{NC}(x^*)$. Clearly, $\mathrm{NC}(x^*) \subset \mathcal{C}$, since $z - x^* \in \mathrm{TC}(x^*), \forall z \in \mathcal{Z}$. For the opposite direction, take $w \in \mathcal{C}$. Then, for $x \in \mathrm{TC}(x^*)$, i.e. $x = \sum_{j=1}^k \lambda_j (z_j - x^*)$ for $z_j \in \mathcal{Z}, \lambda_j \geq 0$, we have $\langle w, x \rangle \leq 0$, since $\langle w, z_j - x^* \rangle \leq 0, \forall z_j \in \mathcal{Z}$. This shows that $w \in \mathrm{NC}(x^*) \implies \mathcal{C} \subset \mathrm{NC}(x^*)$, and our proof is complete. ∎

Given this representation of the normal cone at the vertices of $\mathcal{X}$, we can derive several geometric properties of strict Nash equilibria. Informally, the next lemma states that the payoff vector $v(x^*)$ at a strict equilibrium x^* belongs to the interior of $\mathrm{NC}(x^*)$, and also gives the distance from the cone's boundary.

Lemma 2. *Let $x^* = (\alpha_1^*, \dots, \alpha_N^*) \in \mathcal{X}$ be a strict Nash equilibrium and let d^* be defined as per (16).*

(a) If $\ell \leq d^*$, then $\mathbb{B}(v(x^*), \frac{\ell}{2}) \subseteq \mathrm{NC}(x^*)$, where $\mathbb{B}$ is with respect to $\|\cdot\|_\infty$

(b) If $\ell \leq \frac{d^*}{m}$ for $m \in \mathbb{N}$ and $d = d^* - m\ell$, then for any $w \in \mathbb{B}(v(x^*), \frac{d}{2})$, we have: $w_{i\alpha_i} - w_{i\alpha_i^*} + m\ell \leq 0$, for any $\alpha_i \in \mathcal{A}_i, i \in \mathcal{N}$.

Proof. (a) Let $w \in \mathbb{B}(v(x^*), \frac{\ell}{2})$. We have: $|w_{i\alpha_i} - v_{i\alpha_i}(x^*)| \leq \frac{\ell}{2}, \forall \alpha_i \in \mathcal{A}_i, i \in \mathcal{N}$ Then, for any $z_{i\alpha_i} \in \mathcal{Z}$, we have:

$$\begin{aligned} w_{i\alpha_i} - w_{i\alpha_i^*} &\leq v_{i\alpha_i}(x^*) - v_{i\alpha_i^*}(x^*) + \ell \\ &= u_i(\alpha_i; \alpha_{-i}^*) - u_i(\alpha_i^*; \alpha_{-i}^*) + \ell \leq 0 \end{aligned}$$

from which we conclude that $w \in \mathrm{NC}(x^*)$.

(b) Let $w \in \mathbb{B}(v(x^*), \frac{\ell}{2})$ and $\tilde{w}$ defined as follows:

$$\tilde{w}_{i\alpha_i} = \begin{cases} w_{i\alpha_i} - \frac{m\ell}{2} & \text{if } \alpha_i = \alpha_i^* \\ w_{i\alpha_i} + \frac{m\ell}{2} & \text{otherwise} \end{cases} \tag{14}$$

Then, $\|\tilde{w} - w\|_\infty \leq \frac{m\ell}{2}$, and hence we have:

$$\|\tilde{w} - v(x^*)\|_\infty \leq \|\tilde{w} - w\|_\infty + \|w - v(x^*)\|_\infty \leq \frac{d^*}{2} \tag{15}$$

from which we get that $\tilde{w} \in \mathbb{B}(v(x^*), \frac{d^*}{2})$, i.e., $\tilde{w} \in \mathrm{NC}(x^*)$ due to part *(a)*. Therefore, we conclude that for any $i \in \mathcal{N}, \alpha_i \in \mathcal{A}_i$: $\tilde{w}_{i\alpha_i} - \tilde{w}_{i\alpha_i^*} \leq 0 \implies w_{i\alpha_i} - w_{i\alpha_i^*} + m\ell \leq 0$ ∎

Now, we are ready to state and prove our main theorem.

Theorem 1. *Let $x^* = (\alpha_1^*, \dots, \alpha_N^*)$ be a strict Nash equilibrium of Γ and let*

$$d^* = \min_{i \in \mathcal{N}} \min_{\alpha_i \in \mathcal{A}_i \setminus \{\alpha_i^*\}} \{u_i(\alpha_i^*; \alpha_{-i}^*) - u_i(\alpha_i; \alpha_{-i}^*)\} \tag{16}$$

denote the minimum payoff difference incurred by a unilateral off-equilibrium deviation. Assume further that (FTQL) is run with quantization error $\ell < d^*/3$ and step-size and exploration parameters such that

$$\sum_{n=1}^{\infty} \gamma_n = \infty, \quad \sum_{n=1}^{\infty} \gamma_n \varepsilon_n < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{\gamma_n^2}{\varepsilon_n^2} < \infty. \tag{17}$$

Then x^* is stochastically asymptotically stable and, for all trajectories converging to x^*, we have

$$\|X_{i,n} - x_i^*\|_1 \leq 2 \sum_{\alpha_i \neq \alpha_i^*} \phi_i(-c\tau_n + o(\tau_n)) \tag{18}$$

where $\tau_n = \sum_{k=1}^{n} \gamma_k$ and $c \in (0, d^* - 3\ell)$.

Proof. Since (FTQL) updates the score vector Y_n at each stage n, we need a connection between the variables in the dual space, $\mathcal{Y}$, and the ones in the primal, $\mathcal{X}$. This connection is conveniently expressed through the so-called *score-dominant sets* [14]. Formally, [14] shows that for any $\varepsilon > 0$, there exist $M_{i,\varepsilon}$ for all $i \in \mathcal{N}$ so that

$$\prod_{i \in \mathcal{N}} Q_i(\mathcal{W}_i(M_{i,\varepsilon})) \subseteq \mathcal{U}_\varepsilon \tag{19}$$

where: $\mathcal{W}_i(M_{i,\varepsilon}) = \{Y_i : Y_{i,\alpha_i^*} - Y_{i,\alpha_i} > M_{i,\varepsilon}, \forall \alpha_i \neq \alpha_i^*\}$, and

$$\mathcal{U}_\varepsilon = \{x \in \mathcal{X} : x_{i\alpha_i^*} > 1 - \varepsilon, \forall i \in \mathcal{N}\}. \tag{20}$$

Therefore, our goal in the sequel will be to find a set of initial conditions so that the corresponding score differences $Y_{i\alpha_i^*,n} - Y_{i\alpha_i,n}$, stay large enough throughout the stages of the algorithm for all players $i \in \mathcal{N}$.

***Stochastic stability**:* To begin with, fix a confidence level $\delta > 0$, and let $\mathcal{U}$ be a neighborhood of x^* in $\mathcal{X}$. Invoking Lemma 2 with $m = 3$, we see that any $w \in \mathbb{B}(v(x^*), \frac{d}{2})$ satisfies $w_{i\alpha_i} - w_{i\alpha_i^*} + 3\ell < 0$. By continuity of v, there exist a neighborhood $\overline{\mathcal{U}}$ of x^* and $c > 0$ such that $\overline{\mathcal{U}} \subseteq \mathcal{U}$ and $v_{i\alpha_i}(x) - v_{i\alpha_i^*}(x) + 3\ell \leq -c$, for $x \in \overline{\mathcal{U}}$. Then, by (19),(20), there exist $\varepsilon_0 > 0$, M_{i,ε_0}, for all $i \in \mathcal{N}$ such that:

(a) $\mathcal{U}_{\varepsilon_0} \subseteq \overline{\mathcal{U}} \subseteq \mathcal{U}$

(b) $\prod_{i \in \mathcal{N}} Q_i(\mathcal{W}_i(M_{i,\varepsilon_0})) \subseteq \mathcal{U}_{\varepsilon_0}$

For our analysis, we decompose the approximate payoff vector V_n in components as follows

$$V_{i\alpha_i,n} = \mathbb{E}[\mathcal{R}(v_{i\alpha_i}(\alpha_n) + \xi_{i,n}) \,|\, \mathcal{F}_n] + U_{i\alpha_i,n} + b_{i\alpha_i,n} \tag{21}$$

where $v_{i\alpha_i}(\alpha_n) = u_i(\alpha_i; \alpha_{-i,n})$ and:

1) $U_{i\alpha_i,n} := V_{i\alpha_i,n} - \mathbb{E}[\mathcal{R}(v_{i\alpha_i}(\hat{\alpha}_n) + \xi_{i,n}) \,|\, \mathcal{F}_n]$ is a zero-mean error process.

2) $b_{i\alpha_i,n} := \mathbb{E}[\mathcal{R}(v_{i\alpha_i}(\hat{\alpha}_n) + \xi_{i,n}) \,|\, \mathcal{F}_n] - \mathbb{E}[\mathcal{R}(v_{i\alpha_i}(\alpha_n) + \xi_{i,n}) \,|\, \mathcal{F}_n]$ is a systematic (non-zero-mean) error process due to (a) quantization; and (b) sampling from $\hat{X}_n$ instead of X_n.

It is important to highlight that previous techniques of [15] and references therein can no longer be applied to this setting, because the bias term $b_{i,n}$ is *not* diminishing, but *persistent* in all stages of the (FTQL) due to the quantization error. By comparison, all previous analyses require that any bias entering a learning algorithm vanish appropriately in the long run.

To proceed, we denote by $\tilde{V}_{i,n} := \mathbb{E}[\mathcal{R}(v_i(\alpha_n) + \xi_{i,n} \cdot e) \,|\, \mathcal{F}_n]$ where e is a vector of ones of appropriate dimension, and by $\tilde{b}_{i\alpha_i,n} := v_{i\alpha_i}(\hat{X}_n) - v_{i\alpha_i}(X_n)$.

Claim 1. *The following inequalities hold: $\|\tilde{V}_{i,n} - v_i(X_n)\|_\infty \leq \frac{\ell}{2}$ and $\|b_{i,n} - \tilde{b}_{i,n}\|_\infty \leq \ell$*

Claim 2. $\mathbb{E}[\|\tilde{b}_n\|_* \,|\, \mathcal{F}_n] = O(\varepsilon_n)$ *and* $\mathbb{E}[\|U_n\|_*^2 \,|\, \mathcal{F}_n] = O(1/\varepsilon_n^2)$

Claim 1 follows from (3) and some algebraic derivations, while Claim 2 holds due to Lipschitz continuity of $v(\cdot)$, compactness of $\mathcal{X}$ and $\mathcal{L}^2$-boundedness of ξ. The proofs are omitted due to lack of space.

Then, dropping the index i for convenience, the general score-difference relation at stage n, skipping some algebraic manipulations due to lack of space, and using Claim 2 is

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} = Y_{\alpha,1} - Y_{\alpha^*,1} + \sum_{k=1}^{n} \gamma_k (\tilde{V}_{\alpha,k} - \tilde{V}_{\alpha^*,k})$$

$$+\sum_{k=1}^{n}\gamma_k(b_{\alpha,n}-b_{\alpha^*,n})+\sum_{k=1}^{n}\gamma_k(U_{\alpha,n}-U_{\alpha^*,n})$$
$$\leq Y_{\alpha,1}-Y_{\alpha^*,1}+\sum_{k=1}^{n}\gamma_k(v_\alpha(X_k)-v_{\alpha^*}(X_k)+3\ell)$$
$$+\sum_{k=1}^{n}\gamma_k(\tilde{b}_{\alpha,k}-\tilde{b}_{\alpha^*,k})+\sum_{k=1}^{n}\gamma_k(U_{\alpha,k}-U_{\alpha^*,k}) \tag{22}$$

We will first bound the term $R_n := \sum_{k=1}^{n}\gamma_k(U_{\alpha,k}-U_{\alpha^*,k})$ for all $n\in\mathbb{N}$. Then, R_n is a martingale, as $\mathbb{E}[|R_n|\,|\,\mathcal{F}_n]<\infty$ and $\mathbb{E}[R_{n+1}\,|\,\mathcal{F}_n]=R_n, \forall n$. Moreover, for $K_1>0$, whose value will be determined later, we define the sequence of events $D_{n,K_1}=\{\sup_{k\leq n}|U_{\alpha,k}-U_{\alpha^*,k}|\geq K_1\}$ and $D_{K_1}=\{\sup_{k\geq 1}|U_{\alpha,k}-U_{\alpha^*,k}|\geq K_1\}$. By Doob's maximal inequality (Theorem 2.4, [22]), we have $\mathbb{P}(D_{n,K_1})\leq\frac{\mathbb{E}[R_n^2]}{K_1^2}$. Furthermore, we have that $\mathbb{E}[R_n^2]=\sum_{k=1}^{n}\gamma_k^2\,\mathbb{E}[(U_{\alpha,k}-U_{\alpha^*,k})^2]$ since $\mathbb{E}[U_{\alpha_1,k}U_{\alpha_2,m}]=\mathbb{E}[\mathbb{E}[U_{\alpha_1,k}U_{\alpha_2,m}\,|\,\mathcal{F}_k]]=\mathbb{E}[U_{\alpha_1,k}\,\mathbb{E}[U_{\alpha_2,m}\,|\,\mathcal{F}_k]]=0$ as $U_{\alpha_1,k}$ is $\mathcal{F}_k$-measurable for $k<m$, $\alpha_1,\alpha_2\in\{\alpha,\alpha^*\}$ and $\mathbb{E}[U_{\alpha_2,m}\,|\,\mathcal{F}_k]=0$. Moreover, $\mathbb{E}[(U_{\alpha,k}-U_{\alpha^*,k})^2]\leq 2\,\mathbb{E}[\|U_k\|_*^2]=O(1/\varepsilon_k^2)$ by Claim 2, and, therefore, there exists a constant C_1 such that $\mathbb{E}[R_n^2]\leq C_1\sum_{k=0}^{n}\gamma_k^2/\varepsilon_k^2$. By taking $n\to\infty$ we get $\mathbb{P}(D_{K_1})\leq\frac{C_1\sum_{k=1}^{\infty}\gamma_k^2/\varepsilon_k^2}{K_1^2}<\infty$ so, for $K_1=\sqrt{2C_1\sum_{k=1}^{\infty}\gamma_k^2/\varepsilon_k^2\,\delta}$ we get

$$\mathbb{P}(D_{K_1})\leq\delta/2 \tag{23}$$

Now, for the term $\sum_{k=0}^{n}\gamma_k(\tilde{b}_{\alpha,k}-\tilde{b}_{\alpha^*,k})$, we have that: $\left|\sum_{k=1}^{n}\gamma_k(\tilde{b}_{\alpha,k}-\tilde{b}_{\alpha^*,k})\right|\leq\sum_{k=1}^{n}\gamma_k|\tilde{b}_{\alpha,k}-\tilde{b}_{\alpha^*,k}|\leq 2\sum_{k=1}^{n}\gamma_k\|\tilde{b}_k\|_*$. Defining $S_n:=2\sum_{k=1}^{n}\gamma_k\|\tilde{b}_k\|_*$ it is easy to see that S_n is a submartingale, as $\mathbb{E}[|S_n|]<\infty$ and $\mathbb{E}[S_{n+1}\,|\,\mathcal{F}_n]\geq S_n, \forall n$. With similar arguments as before, defining $E_{K_2}=\{\sup_{k\geq 1}S_k\geq K_2\}$ and using Doob's maximal inequality, we get that $\mathbb{P}(E_{K_2})\leq(C_2/K_2)\sum_{k=0}^{\infty}\gamma_k\varepsilon_k<\infty$. Setting $K_2=2(C_2/\delta)\sum_{k=0}^{\infty}\gamma_k\varepsilon_k$ we get

$$\mathbb{P}(E_{K_2})\leq\delta/2 \tag{24}$$

Hence, by the union bound, we get: $\mathbb{P}(D_{K_1}\cup E_{K_2})\leq\delta$. Setting $M>M_{\varepsilon_0}+K_1+K_2$, we get that if $Y_1\in\mathcal{W}(M)$, i.e. $Y_{\alpha,1}-Y_{\alpha^*,1}<-M$ then:

$$Y_{\alpha,n+1}-Y_{\alpha^*,n+1}\leq -M+K_1+K_2<-M_{\varepsilon_0} \tag{25}$$

on the event $(D_{K_1}\cup E_{K_2})^c$, from which we get that $X_{n+1}\in\mathcal{U}_{\varepsilon_0}$, i.e. $X_{n+1}\in\mathcal{U}$ with probability at least $1-\delta$. Therefore, we conclude that x^* is stochastically stable.

***Stochastic asymptotic stability**:* From the previous analysis, on the event $(D_{K_1}\cup E_{K_2})^c$ we have that: $Y_{\alpha,n+1}-Y_{\alpha^*,n+1}<-M_{\varepsilon_0}-c\sum_{k=1}^{n}\gamma_k$ Sending $n\to\infty$, we get that $Y_{\alpha,n+1}-Y_{\alpha^*,n+1}\to-\infty$, from which we have that for all $\tilde{M}>0, Y_k\in\mathcal{W}(\tilde{M})$ eventually. Hence, for all $\tilde{\varepsilon}>0, X_k\in\mathcal{U}_{\tilde{\varepsilon}}$ eventually, from which we get that $X_k\to x^*$ as $k\to\infty$.

***Rates of convergence**:* Finally, to establish the rate of convergence of (FTQL), let $\tau_n:=\sum_{k=1}^{n}\gamma_k$ for all n. Since $R_n=\sum_{k=1}^{n}\gamma_k(U_{\alpha,k}-U_{\alpha^*,k})$ is a martingale, $\lim_{n\to\infty}\tau_n=\infty$ and $\sum_{k=1}^{\infty}\tau_k^{-2}\,\mathbb{E}[|\gamma_k(U_{\alpha,k}-U_{\alpha^*,k})|^2\,|\,\mathcal{F}_k]\leq C_1\sum_{k=1}^{\infty}\tau_k^{-2}\gamma_k^2/\varepsilon_k^2<\infty$, we have, by Strong law of large numbers for martingales (Theorem 2.18, [22]), that $\frac{R_n}{\tau_n}\to 0$ a.s. as $n\to\infty$.

Moreover, since S_n is a nonnegative submartingale with $\mathbb{E}[S_n]$ bounded for all n, by Doob's submartingale convergence theorem (Theorem 2.5, [22]), we obtain that there exist a random variable S_∞ with $\mathbb{E}[|S_\infty|]<\infty$ and $S_n\to S_\infty$ a.s. as $n\to\infty$. Since $S_\infty\in\mathcal{L}^1$, we get that $S_\infty<\infty$ a.s., and, so, $\frac{S_n}{\tau_n}\to 0$ as $n\to\infty$.

Since $X_n\to x^*$ as $n\to\infty$, we have that $X_n\in\overline{\mathcal{U}}$ eventually, i.e. there exists $n_0\in\mathbb{N}$ such that $X_n\in\overline{\mathcal{U}}$ for all $n\geq n_0$, from which we get $v_{i\alpha_i}(x)-v_{i\alpha_i^*}(x)+3\ell\leq -c$. Hence, for $n\geq n_0$, after some calculations omitted due to lack of space, we obtain $Y_{\alpha,n+1}-Y_{\alpha^*,n+1}\leq -c\tau_n+o(\tau_n)$. Therefore, we have:

$$\theta'(X_{\alpha,n+1})\leq\theta'(X_{\alpha^*,n+1})-c\tau_n+o(\tau_n)\leq\theta'(1)-c\tau_n+o(\tau_n) \tag{26}$$

from which we conclude that: $X_{\alpha,n+1}\leq\phi(-c\tau_n+o(\tau_n))$. Aggregating over all strategies $\alpha\in\mathcal{A}, \alpha\neq\alpha^*$ we have:

$$\|x^*-X_{n+1}\|_1\leq 2\sum_{\alpha\neq\alpha^*}\phi(-c\tau_n+o(\tau_n)) \tag{27}$$

■

The qualitative behavior of the dynamics as a function of the quantization length is shown in Fig. 1. Some corollaries and remarks are in order. We begin by discussing the possible schedules for the algorithm's step-size and exploration parameters: here, a standard choice is to take $\gamma_n\propto 1/n^p$ and $\varepsilon_n\propto 1/n^q$, with $p,q\geq 0$ chosen so as to satisfy (17). A straightforward verification gives the conditions

$$p\leq 1,\quad p+q>1\quad\text{and}\quad 2p-2q>1 \tag{28}$$

which, in turn, provide the following explicit guarantee:

Corollary 1. *Suppose that* (FTQL) *is run with assumptions as in Theorem 1 and with step-size and exploration parameters satisfying* (28)*. Then, for all $p>3/4$,* (FTQL) *achieves*

$$\|X_{i,n}-x_i^*\|_1\leq 2\sum\nolimits_{\alpha_i\neq\alpha_i^*}\phi_i\Big(-\Theta(n^{1-p})\Big). \tag{29}$$

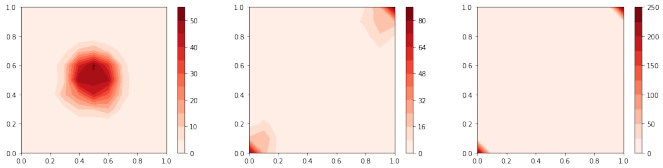
Proof. From (28) it is easy to see that $p\in(3/4,1]$. For $p=1$, $\tau_n=\Theta(\log n)$, so (FTQL) achieves faster convergence if $p<1$. Since for $p\in(3/4,1)$ we have that $\tau_n=\Theta(n^{1-p})$, the result is immediate from (27). ■

Remark. In the absence of quantization, FTRL achieves a convergence rate of $\phi(-\sum_{k=1}^{n}\gamma_k)$ [15] with $p\in[0,1]$. In our case, if the assumption for ξ is strengthened to almost sure boundedness (or sub-Gaussian increments), we can likewise relax the step-size requirements and achieve the rate (29) for *any* $p\in[0,1]$. §

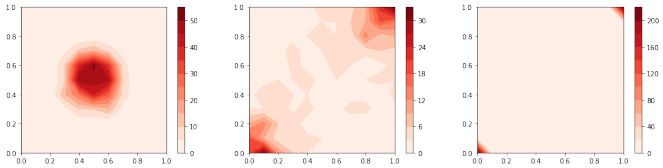
Our next result concerns the rate of convergence of (FTQL) for different choices of the mirror map Q as defined in (8):

Corollary 2. *Suppose that* (FTQL) *is run with assumptions as in Corollary 1. Then:*

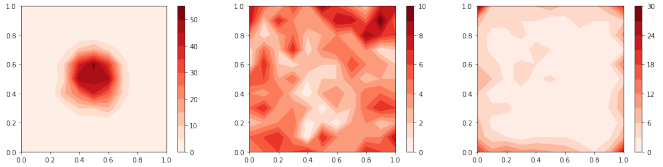
1) *The exponential* (or *multiplicative*) *weights variant of the algorithm* ($\theta(z)=z\log z$) *achieves convergence to strict Nash equilibria at a rate of:* $\|X_{i,n}-x_i^*\|_1\leq 2\sum_{\alpha_i\neq\alpha_i^*}\exp\Big(-\Theta(n^{1-p})\Big)$.



(a) $\ell = 0$



(b) $\ell = 1.5$



(c) $\ell = 4$

Fig. 1: Density 'heat-maps' of the temporal evolution of (FTQL) at times 0 (left), 50 (middle), 2000 (right) in 500 instances (sampled as $Y_1 \sim U[0,1]^4$) on a 2×2 symmetric game with $u(a_1,b_1) = u(a_2,b_2) = 5.1$, $u(a_1,b_2) = u(a_2,b_1) = 2.4$, $p = 0.75$, $q = 0.25$, and $\xi_n \sim U[-0.1, 0.1]$ for three quantization errors: $\ell = 0$, 1.5, 4. Both (a_1,b_1) and (a_2,b_2) are pure Nash equilibria. In each plot, the horizontal axis corresponds to x_{1a_1} and the vertical one to x_{2b_1}. We observe that for small quantization errors ($\ell = 0$, 1.5), the instances converge to the Nash equilibria (a_1,b_1) and (a_2,b_2); however, for $\ell = 1.5$, the convergence is slower. On the contrary, for large quantization length ($\ell = 4$), the behavior of the system is unpredictable and all the pure strategy profiles become attractors. Note the *gradual "disintegration" of convergence* with larger quantization error.

*2) The Euclidean variant of the algorithm ($\theta(z) = z^2/2$) achieves convergence to strict Nash equilibria at a **finite** number of iterations.*

Proof. 1) Since $\theta(x) = x\log x$, we have $\phi(x) = \exp(x-1)$. So, we obtain $\phi_i(-\Theta(n^{1-p})) = \exp(-\Theta(n^{1-p}))$. Then, the result is immediate.

2) Since $X_{i,n} \geq 0$, θ' increasing, (26) becomes

$$\theta'(0) \leq \theta'(X_{\alpha,n+1}) \leq \theta'(1) - c\tau_n + o(\tau_n) \tag{30}$$

and, since $\lim_n \tau_n = \infty$, we obtain $\tau_n \geq \frac{1}{c}(\theta'(1) - \theta'(0) + o(\tau_n))$ for large n. Combining the above inequalities, we get $\theta'(0) \leq \theta'(X_{\alpha,n+1}) \leq \theta'(0)$, from which we conclude that $X_{\alpha,n+1} = 0$ for sufficiently large n, as per our claim. ∎

V. Concluding remarks

Our results show that the impact of quantization on learning in games is somewhat different than what one would perhaps expect: instead of a gradual deterioration of the quality of learning as the quantization error increases, we see that (FTQL) continues to identify strict Nash equilibria *perfectly* if the quantization is not too coarse, and the rate of convergence is as in the non-quantized case. We find this property particularly appealing, as it shows that, if the feedback process is quantized judiciously, we can achieve significant gains in terms of memory storage and bandwidth expenditures *without* compromising the quality of learning.

References

[1] S. Shalev-Shwartz and Y. Singer, "Convex repeated games and Fenchel duality," in *NIPS' 06: Proceedings of the 19th Annual Conference on Neural Information Processing Systems*. MIT Press, 2006, pp. 1265–1272.

[2] V. G. Vovk, "Aggregating strategies," in *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*, 1990, pp. 371–383.

[3] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, no. 2, pp. 212–261, 1994.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

[5] D. Friedman, "Evolutionary games in economics," *Econometrica*, vol. 59, no. 3, pp. 637–666, 1991.

[6] E. Hopkins, "Learning, matching, and aggregation," *Games and Economic Behavior*, vol. 26, pp. 79–110, 1999.

[7] H. Moulin and J.-P. Vial, "Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon," *International Journal of Game Theory*, vol. 7, pp. 201–221, 1978.

[8] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, September 2000.

[9] Y. Viossat and A. Zapechelnyuk, "No-regret dynamics and fictitious play," *Journal of Economic Theory*, vol. 148, no. 2, pp. 825–842, March 2013.

[10] M. Bravo, "An adjusted payoff-based procedure for normal form games," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1469–1483, November 2016.

[11] R. Cominetti, E. Melo, and S. Sorin, "A payoff-based learning procedure and its application to traffic games," *Games and Economic Behavior*, vol. 70, no. 1, pp. 71–83, 2010.

[12] D. S. Leslie and E. J. Collins, "Convergent multiple-timescales reinforcement learning algorithms in normal form games," *The Annals of Applied Probability*, vol. 13, no. 4, pp. 1231–1251, November 2003.

[13] ——, "Individual Q-learning in normal form games," *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 495–514, 2005.

[14] A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos, "Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information," in *COLT '21: Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

[15] ——, "The convergence rate of regularized learning in games: From bandits and uncertainty to optimism and beyond," https://hal.inria.fr/hal-03357715/, 2021.

[16] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, ser. Applied Optimization. Kluwer Academic Publishers, 2004, no. 87.

[17] Y. Nesterov and A. S. Nemirovski, *Interior Point Polynomial Methods in Convex programming*. SIAM Publications, 1994.

[18] A. Juditsky, A. S. Nemirovski, and C. Tauvel, "Solving variational inequalities with stochastic mirror-prox algorithm," *Stochastic Systems*, vol. 1, no. 1, pp. 17–58, 2011.

[19] P. Mertikopoulos and W. H. Sandholm, "Learning in games via reinforcement and regularization," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1297–1324, November 2016.

[20] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, 2nd ed., ser. Stochastic Modelling and Applied Probability. Berlin: Springer-Verlag, 2012, no. 66.

[21] R. C. R. Robinson, *An Introduction to Dynamical Systems: Continuous and Discrete*, 2nd ed. Providence, RI: American Mathematical Society, 2012.

[22] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*, ser. Probability and Mathematical Statistics. New York: Academic Press, 1980.