Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information

Angeliki Giannou, Emmanouil Vasileios Vlatakis-Gkaragkounis, Panayotis Mertikopoulos

Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information

Angeliki Giannou GIANNOUANGELIKI@GMAIL.COM
National Technical University of Athens
Emmanouil-Vasileios Vlatakis-Gkaragkounis EMVLATAKIS@CS.COLUMBIA.COM
Columbia University
Panayotis Mertikopoulos PANAYOTIS.MERTIKOPOULOS@IMAG.FR
Univ. Grenoble Alpes, CNRS, Inria & Criteo AI Lab

Abstract

In this paper, we examine the Nash equilibrium convergence properties of no-regret learning in general N-player games. For concreteness, we focus on the archetypal *"follow the regularized leader"* (FTRL) family of algorithms, and we consider the full spectrum of uncertainty that the players may encounter – from noisy, oracle-based feedback, to bandit, payoff-based information. In this general context, we establish a comprehensive equivalence between the stability of a Nash equilibrium and its support: *a Nash equilibrium is stable and attracting with arbitrarily high probability if and only if it is strict* (i.e., each equilibrium strategy has a unique best response). This equivalence extends existing continuous-time versions of the "folk theorem" of evolutionary game theory to a bona fide algorithmic learning setting, and it provides a clear refinement criterion for the prediction of the day-to-day behavior of no-regret learning in games.

1. Introduction

The prototypical framework for online learning in games can be summarized as follows:

1. At each stage of the process, every participating agent chooses an action from some finite set.
2. All agents receive a reward based on the actions of all other players and their individual payoff functions (assumed a priori unknown).
3. The players record their rewards and any other feedback generated during the payoff phase, and the process repeats.

This multi-agent framework has both important similarities and major differences with *single-agent* online learning. Indeed, if we isolate a single, focal player and abstract away all others, we essentially recover a multi-armed bandit (MAB) problem – stochastic or adversarial, depending on the assumptions for the non-focal players [3, 4]. In this case, the most widely used figure of merit is the agent's *regret*, i.e., the difference between the agent's cumulative payoff and that of the best fixed action in hindsight. Accordingly, much of the literature on online learning has focused on deriving regret bounds that are min-max optimal, both in terms of the horizon T of the process, as well as the number of actions A available to the focal player.

On the other hand, from a game-theoretic standpoint, the main question that arises is whether players eventually settle on an equilibrium profile from which no player has an incentive to deviate. In this regard, a "folk" result states that the empirical frequency of play under no-regret play

converges to the game's set of *coarse correlated equilibria* (CCE) [13, 15]. However, there are two key caveats with this result. First, CCE are considerably weaker than Nash equilibria, to the extent that they fail even the most basic postulates of rationalizability [9]: as was shown by Viossat and Zapechelnyuk [41], CCE may be supported *exclusively* on *strictly dominated* strategies, even in simple, symmetric two-player games. Second, the convergence of the empirical mean does not carry any tangible guarantees for the players' day-to-day behavior: under this type of convergence, the player's best payoff over time could be close to that of a Nash equilibrium, but the players might otherwise be spending arbitrarily long periods of time on dominated strategies.

The above is just a well-known example of the convergence failures of no-regret learning in games with a possibly exotic equilibrium structure. More to the point, even when the underlying game admits a *unique* Nash equilibrium, recent works have shown that no-regret algorithms – such as the popular multiplicative weights update (MWU) method – could still lead to chaotic [5, 30, 31] or Poincaré recurrent / cycling behavior [20, 27, 29]. From a convergence viewpoint, all these results can be seen as instances of a much more general impossibility result at play: *there are no uncoupled dynamics leading to Nash equilibrium in* all *games* [Hart and Mas-Colell, 16].[1] Since no-regret dynamics are by definition unilateral, they are *a fortiori* uncoupled, so this result shatters any hope of obtaining a universal Nash equilibrium convergence result for the players' day-to-day behavior.

Our contributions. In view of the above, a critical question that arises is the following: *Is there a class of Nash equilibria that consistently attract no-regret processes?* Conversely, *are all Nash equilibria equally likely to emerge as outcomes of a no-regret learning process?*

To address these questions in as general a setting as possible, we focus on the *"follow the regularized leader"* (FTRL) family of algorithms: this is arguably the most widely used class of dynamics for no-regret learning in games, and it includes as special cases the seminal multiplicative weights / EXP3 algorithms [1, 35, 36]. In terms of feedback, we also consider a flexible, context-agnostic template in which players are only assumed to have access to an inexact model of their payoff vectors at a given stage. This model for the players' feedback covers a broad range of modeling assumptions, such as (*a*) the case where players can retroactively compute – or otherwise observe – their full payoff vectors (e.g., as in routing games); and (*b*) the *bandit* case, where players only observe their in-game payoffs and have no other information on the game being played.

The range of modeling assumptions covered by our framework is quite extensive, so one would likewise expect different, context-specific answers to these questions – presumably with equilibria becoming "less stable" as information becomes "more scarce". This expectation is justified by the behavior of no-regret learning in single-agent environments: there, the type of information available to the learner has a dramatic effect on the achieved regret minimization rate. Nevertheless, we show that this conjecture is *false:* as far as the algorithms' equilibrium convergence properties are concerned, the learning dynamics described above are all *equivalent*.

In more detail, we show that all FTRL algorithms under study enjoy the following properties:

a) *Strict Nash equilibria are* stochastically *asymptotically stable* – i.e., they are stable and attracting with arbitrarily high probability.

b) *Only strict Nash equilibria have this property:* mixed Nash equilibria supported on more than one strategies are inherently unstable from a learning viewpoint.

1. "Uncoupled" means here that each player's update rule does not depend explicitly on the payoffs of other players.

We are not aware of a similar result in the literature at this level of generality (i.e., including models with bandit feedback), and we believe that this equivalence represents an important refinement criterion for the prediction of the day-to-day behavior of no-regret learners in the face of uncertainty and lack of perfect information.

Related work. To put our contributions in the proper context, we provide below an account of relevant works in the literature, classified along the two directions of our main result: "strictness $\Longrightarrow$ stability" and "stability $\Longrightarrow$ strictness".

I. Strictness $\Longrightarrow$ Stability. Analyzing the convergence of game-theoretic learning dynamics has generated a vast corpus of literature that is impossible to survey here. Nonetheless, an emerging theme in this literature is the focus on specific classes of games (such as potential games or 2^N games). As a purely indicative – and highly incomplete – list, we cite here the works of Leslie and Collins [25] and Leslie [24], Cominetti et al. [7], Kleinberg et al. [22], Coucheney et al. [8], Syrgkanis et al. [39], and Cohen et al. [6], who provide a range of equilibrium convergence results in potential, 2^N, and (λ, μ)-smooth games, under different feedback assumptions – from payoff vector observations [22, 39] to bandit [6, 7, 24, 25]. By contrast, our focus is determining the stochastic stability of a class of *equilibria* – not *games*.

As far as we are aware, the only comparable results in this literature concern an idealized continuous-time, deterministic, full-information version of our setting, which is common in applications to population biology and evolutionary game theory. In this context, building on earlier results on the replicator dynamics [18, 43], the authors of [27] showed that strict Nash equilibria are asymptotically stable under the continuous-time dynamics of FTRL. However, we stress here again that these results only concern continuous-time, deterministic dynamical systems with an inherent full-information assumption; we are not aware of a result providing convergence to strict Nash equilibria with bandit feedback.

II. Stability $\Longrightarrow$ Strictness. In the converse direction, a related result in the literature on evolutionary games is that only strict Nash equilibria are asymptotically stable under the (multi-population) replicator dynamics [19, 34, 43], a continuous-time, deterministic dynamical system which can be seen as the "mean-field" limit of the exponential weights algorithm [20, 33, 38]. In a much more recent paper [10], this implication was extended to the dynamics of FTRL, but always in a deterministic, full-information, continuous-time setting. In this regard, our results are aligned with [10]; however, other than this high-level conceptual link, there is no precise connection, either at the level of implications or at the level of proofs. Specifically, the analysis of [10] relies crucially on volume-conservation arguments that are neither applicable nor relevant in a discrete-time stochastic setting – where the various processes involved could jump around stochastically without any regard for volume contraction or expansion.

Proof techniques. Learning with partial information is an inherently stochastic process, so our results are also stochastic in nature – hence the requirement for asymptotic stability with arbitrarily high probability. This constitutes a major point of departure from continuous-time models of learning [10, 27], so our proof techniques are also radically different as a result. The principal challenge in our proof of stability of strict Nash equilibria comes in controlling the aggregation of error terms with possibly unbounded variance (coming from inverse propensity scoring of bandit-type observations). Because of this, stochastic approximation techniques that have been used to

show convergence with L^2-bounded feedback [28] cannot be applied in this setting; we achieve this control by applying a sharp version of the Doob-Kolmogorov maximal inequality to control equilibrium deviations with high probability. In the converse direction, the crucial argument in the proof of the *instability* of mixed equilibria is a direct probabilistic estimate which leverages a non-degeneracy argument for the noise entering the process; we are not aware of other works using a similar technique.

2. Preliminaries

The stage game. Throughout this work we will focus on normal form games with a finite number of players and a finite number of actions per player. Formally, such a game is defined as a tuple $\Gamma = \Gamma(\mathcal{N}, \mathcal{A}, u)$ with the following primitives:

- A finite set of *players* – or *agents* – indexed by $i \in \mathcal{N} = \{1, \dots, N\}$.
- A finite set of *actions* – or *pure strategies* – indexed by $\alpha_i \in \mathcal{A}_i = \{1, \dots, A_i\}$, $i \in \mathcal{N}$. Players can also play *mixed strategies*, which represent probability distributions $x_i \in \mathcal{X}_i := \Delta(\mathcal{A}_i)$; in this case, we will write $x_{i\alpha_i}$ for the probability that player $i \in \mathcal{N}$ selects $\alpha_i \in \mathcal{A}_i$. Aggregating over all players, we will also write $x = (x_1, \dots, x_N)$ for the players' *mixed strategy profile* and $\mathcal{X} := \prod_i \mathcal{X}_i$ for the set thereof. Finally, when we want to focus on the strategy (or action) of a particular player $i \in \mathcal{N}$, we will use the shorthand $(x_i; x_{-i}) := (x_1, \dots, x_i, \dots, x_N)$ – and, similarly, $(\alpha_i; \alpha_{-i})$ for pure strategies.
- An ensemble of *payoff functions* $u_i \colon \mathcal{A} \to \mathbb{R}$ where $\mathcal{A} := \prod_i \mathcal{A}_i$ is the space of all pure strategy profiles. The expected payoff of player i in a mixed strategy profile $x \in \mathcal{X}$ is then given by

$$u_i(x) \equiv u_i(x_i; x_{-i}) = \sum_{\alpha_1 \in \mathcal{A}_1} \cdots \sum_{\alpha_N \in \mathcal{A}_N} u_i(\alpha_1, \dots, \alpha_N) \cdot x_{1,\alpha_1} \cdots x_{N,\alpha_N} \tag{1}$$

  where $u_i(\alpha_1, \dots, \alpha_N)$ is the payoff of player i in the action profile $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathcal{A}$.

For posterity, we will also write $v_{i\alpha_i}(x) = u_i(\alpha_i; x_{-i})$ for the payoff that player i would have gotten by playing $\alpha_i \in \mathcal{A}_i$ against the mixed strategy profile x_{-i} of all other players. In this way, the *mixed payoff vector* of the i-th player will be

$$v_i(x) = (v_{i\alpha_i}(x))_{\alpha_i \in \mathcal{A}_i} \tag{2}$$

and we will write $v(x) = (v_1(x), \dots, v_N(x))$ for the ensemble thereof. For notational convenience, we will also set $\mathcal{Y}_i = \mathbb{R}^{\mathcal{A}_i}$ and $\mathcal{Y} = \prod_i \mathcal{Y}_i$ for the space of payoff vectors and profiles respectively. Finally, in a slight abuse of notation, we will identify α_i with the mixed strategy that assigns all probability to α_i, and we will denote the corresponding *pure payoff vector* as $v_i(\alpha) = (u_i(\alpha_i; \alpha_{-i}))_{\alpha_i \in \mathcal{A}_i}$. The distinction between pure and mixed payoff vectors will become important later on, when we discuss the information at each player's disposal.

Nash equilibrium. In this general context, the most widely used solution concept is that of a Nash equilibrium, i.e., a mixed strategy profile that discourages unilateral deviations. Formally, x^* is a *Nash equilibrium* of Γ if

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

The set of pure strategies supported at the equilibrium component $x_i^* \in \mathcal{X}_i$ of each player will be denoted by $\operatorname{supp}(x_i^*) = \{\alpha_i \in \mathcal{A}_i : x_{i\alpha_i}^* > 0\}$. Accordingly, Nash equilibria can be equivalently characterized by means of the variational inequality

$$v_{i\alpha_i^*}(x^*) \geq v_{i\alpha_i}(x^*) \quad \text{for all } \alpha_i^* \in \operatorname{supp}(x_i^*) \text{ and all } \alpha_i \in \mathcal{A}_i,\ i \in \mathcal{N}. \tag{3}$$

The above characterization gives rise to the following classification of Nash equilibria:

- x^* is a *pure equilibrium* if $\operatorname{supp}(x_i^*)$ only contains a single strategy for all $i \in \mathcal{N}$.
- x^* is a *mixed equilibrium* in any other case; in particular, if $\operatorname{supp}(x_i^*) = \mathcal{A}_i$ for all $i \in \mathcal{N}$, we say that x^* is *fully mixed*.

By definition, pure equilibria correspond to vertices of $\mathcal{X}$, fully mixed equilibria lie in the relative interior $\operatorname{ri}(\mathcal{X})$ of $\mathcal{X}$, and, more generally, mixed equilibria lie in the relative interior of the face of the simplex spanned by the support of each player's equilibrium component. A further distinction between Nash equilibria that is inherited by the inequality (3) is as follows: if (3) holds as a strict inequality for all $\alpha_i \in \mathcal{A}_i \setminus \operatorname{supp}(x_i^*)$, $i \in \mathcal{N}$, the equilibrium in question is said to be *quasi-strict* [11]. Quasi-strict equilibria have the defining property that all pure best responses are played with positive probability; it is also well known that all Nash equilibria in all but a measure-zero set of games are quasi-strict. For this reason, the property of having a quasi-strict equilibrium is generic, and games that enjoy this property are called themselves *generic* [Specifically, the set of games with Nash equilibria that are not quasi-strict is *meager* in the Baire category sense.]

We stress here that quasi-strict equilibria could be either mixed or pure: for example, the equilibrium of Matching Pennies is fully mixed and quasi-strict, whereas the equilibrium of the Prisoner's dilemma is quasi-strict and pure. In this last case, when a quasi-strict equilibrium is pure, it will be called *strict*: any deviation from an equilibrium strategy results in a strictly worse payoff.

3. Regret minimization and regularized learning

A key requirement in the context of online learning is the minimization of the players' regret, i.e., the cumulative payoff difference between each player's chosen action and the best possible action in hindsight over a given horizon of play T. Formally, given a sequence of play $X_n \in \mathcal{X}$, $n = 0, 1, \dots$, the (external) *regret* of player $i \in \mathcal{N}$ is defined as

$$\operatorname{Reg}_i(T) = \max_{x_i \in \mathcal{X}_i} \sum_{n=0}^{T} [u_i(x_i; X_{-i,n}) - u_i(X_{i,n}; X_{-i,n})] \tag{4}$$

and we will say that player i has no regret if $\operatorname{Reg}_i(T) = o(T)$.

One of the most widely used online learning schemes to achieve this requirement is the so-called *"follow the regularized leader"* (FTRL) family of algorithms [35, 36]. Heuristically, at each stage of the learning process, FTRL prescribes a mixed strategy that maximizes the player's (perceived) cumulative payoff modulo a regularization penalty whose role is to "smooth out" the transition between strategies during play. Formally, this leads to the round-by-round recursive rule

$$\begin{aligned} X_{i,n} &= Q_i(Y_{i,n}) \\ Y_{i,n+1} &= Y_{i,n} + \gamma_n \hat{v}_{i,n} \end{aligned} \tag{FTRL}$$

where $Q_i \colon \mathcal{Y}_i \to \mathcal{X}_i$ denotes the "choice map" of player $i \in \mathcal{N}$, $\gamma_n > 0$ is a "learning rate" parameter such that $\sum_n \gamma_n = \infty$, and $\hat{v}_{i,n}$ is a "payoff signal" that provides an estimate for the mixed payoffs of player i at stage n. We discuss each of these components in detail below.

3.1. The feedback model

Depending on the specific framework at play, the modeling details concerning the feedback received by the players may vary wildly. For example, when modeling congestion in a city, it is reasonable to assume that commuters can estimate the time it would have taken them to get to their destination via a different route – e.g., by means of a GPS service or an app like GoogleMaps or Waze. By contrast, in applications of online learning to auctions and online advertising, it is not clear how a player could estimate the payoff of actions they did not play.

To account for as broad a range of feedback models as possible, we will take a context-agnostic approach and assume that each player receives a "black-box" model of their payoff vector of the form

$$\hat{v}_n = v(X_n) + \xi_n \tag{5}$$

for some abstract error process $\xi_n = (\xi_{i,n})_{i\in\mathcal{N}}$. To differentiate between random (zero-mean) and systematic (non-zero-mean) errors, we will further decompose ξ_n as $\xi_n = Z_n + b_n$, where

$$b_n = \mathbb{E}[\xi_n \,|\, \mathcal{F}_n] \quad \text{and} \quad \mathbb{E}[Z_n \,|\, \mathcal{F}_n] = 0 \tag{6}$$

with $\mathcal{F}_n$ denoting the history of X_n up to stage n (inclusive) [2]. We may then characterize the input signal $\hat{v}_n$ by means of the following statistics:

$$a) \quad \textit{Bias:} \qquad \mathbb{E}[\|b_n\|_* \,|\, \mathcal{F}_n] \leq B_n \tag{7a}$$

$$b) \quad \textit{Mean square:} \qquad \mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] \leq M_n^2 \tag{7b}$$

In the above, B_n and M_n represent deterministic bounds on the bias and variance of the feedback signal $\hat{v}_n$. For concreteness, we will also make the following blanket assumptions:

(A1) *Bias control:* $\lim_{n\to\infty} B_n = 0$ and $\sum_n \gamma_n B_n < \infty$.

(A2) *Variance control:* $\sum_n \gamma_n^2 M_n^2 < \infty$.

(A3) *Generic observation errors at equilibrium:* For every mixed Nash equilibrium x^* of Γ and for all $n = 0, 1, \dots$, there exists a player $i \in \mathcal{N}$ and strategies $a, b \in \operatorname{supp}(x_i^*)$ such that

$$\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta \,|\, \mathcal{F}_n) > 0 \quad \text{for all sufficiently small } \beta > 0. \tag{8}$$

The formulation of these hypotheses has been kept intentionally abstract because we have not made any modeling assumptions for how the players' payoff signals are generated. In this regard, they are to be construed as an "inexact model" that allows for a wide variety of settings; as an application, we illustrate below how these assumptions are verified in two widely used learning frameworks.

Model 1 (Oracle-based feedback)**.** *Assume that each player chooses an action based on a given mixed strategy. Then, once this procedure has been completed, an oracle reveals to each player the payoffs corresponding to their pure strategies given the other players' chosen strategies* (*in the congestion example, this oracle could be Waze or a GPS device*)*. Formally, at each round n, every player $i \in \mathcal{N}$ picks an action $\alpha_{i,n} \in \mathcal{A}_i$ based on $X_{i,n} \in \mathcal{X}_i$ and observes the pure payoff vector $v_i(\alpha_n) \equiv (u_i(\alpha_i; \alpha_{-i,n}))_{\alpha_i\in\mathcal{A}_i}$. Then the player's feedback signal is $\hat{v}_{i,n} = v_i(\alpha_n)$, which is a special case of the model* (5) *with $Z_n = v(X_n) - v(\alpha_n)$ and $b_n = 0$. In more detail, we have:*

2. Of course, since the feedback signal is generated only *after* the player chooses a strategy, $\hat{v}_n$ is not $\mathcal{F}_n$-measurable in general.

- (A1) *is trivial because* $b_n = 0$.
- (A2) *is satisfied as long as* $\sum_n \gamma_n^2 < \infty$ (*because* $\sup_n \mathbb{E}[\|\hat{v}_n\|_*^2] \leq \max_x \|v(x)\|_*^2 < \infty$).
- (A3) *is proved in Appendix B.*

Model 2 (Payoff-based feedback)**.** *Assume that each player picks an action based on some mixed strategy as above; however, players now only observe their realized payoffs* $u_i(\alpha_{i,n}; \alpha_{-i,n})$*. This is the standard model for multi-armed bandits [3, 4], and it is also known as the "bandit feedback" setting. In this case, players can estimate their payoff vectors by means of the importance-weighted estimator:*

$$\hat{v}_{i\alpha_i,n} = \frac{\mathbb{1}\{\alpha_{i,n} = \alpha_i\}}{\hat{X}_{i\alpha_{i,n}}} u_i(\alpha_n) \tag{IWE}$$

where $\hat{X}_{i,n} = (1 - \varepsilon_n) X_{i,n} + \varepsilon_n / |\mathcal{A}_i|$ *is the mixed strategy of the* i*-th player at stage* n*. Compared to* $X_{i,n}$*, the player's actual sampling strategy is recalibrated by an explicit exploration parameter* $\varepsilon_n \to 0$ *whose role is to stabilize the learning process by controlling the variance of* (IWE)*. The idea is that even if a strategy has zero probability to be chosen under* X_n*, it will still be sampled with positive probability thanks to the mixing factor* ε_n*.*

A standard calculation (that we defer to Appendix B) shows that (IWE) *can be recast in the general form* (5) *with* $B_n = \mathcal{O}(\varepsilon_n)$ *and* $M_n^2 = \mathcal{O}(1/\varepsilon_n)$*. We then have:*

- (A1) *is satisfied as long as* $\varepsilon_n \to 0$ *and* $\sum_n \gamma_n \varepsilon_n < \infty$.
- (A2) *is satisfied as long as* $\sum_n \gamma_n^2 / \varepsilon_n < \infty$.
- (A3) *is proved in Appendix B.*

Remark. The above conditions for the method's learning rate and exploration parameters can be achieved by using schedules of the form $\gamma_n \propto 1/n^p$ and $\varepsilon_n \propto 1/n^q$ with $p + q > 1$ and $2p - q > 1$. A popular choice is $p = 2/3 + \delta$ and $q = 1/3 + \delta$ for some arbitrarily small $\delta > 0$ – or $\delta = 0$ and including an extra logarithmic factor, cf. [37] and references therein.

3.2. Regularization

The second component of the FTRL method is the players' "choice map" $Q_i \colon \mathcal{Y}_i \to \mathcal{X}_i$. Because the players' score variables $Y_{i,n}$ essentially represent an estimate of each strategy's cumulative payoff over time, Q_i is defined as a "regularized" version of the best-response correspondence $y_i \mapsto \arg\max_{x_i \in \mathcal{X}_i} \{\langle y_i, x_i \rangle\}$ (the regularization being necessary to avoid prematurely committing to a strategy). On that account, we will consider *regularized best responses* of the general form

$$Q_i(y_i) = \arg\max_{x_i \in \mathcal{X}_i} \{\langle y_i, x_i \rangle - h_i(x_i)\}. \tag{9}$$

In the above, each player's *regularizer* $h_i \colon \mathcal{X}_i \to \mathbb{R}$ is defined as $h_i(x_i) = \sum_{\alpha_i \in \mathcal{A}_i} \theta_i(x_i)$ for some "kernel function" $\theta_i \colon [0,1] \to \mathbb{R}$ with the following properties: (*i*) θ_i is *continuous* on $[0,1]$; (*ii*) C^2-smooth on $(0,1]$; and (*iii*) $\inf_{[0,1]} \theta_i'' > 0$. Of course, different regularizers give rise to different instances of (FTRL); for concreteness, we present below two prototypical examples thereof.

Example 1 (Multiplicative/Exponential weights update)**.** *A popular choice of regularizer is the* (*negative*) *entropy* $h_i(x) = \sum_i x_i \log x_i$*, which leads to the* logit choice *map* $\Lambda_i(y) = \exp(y_i) / \sum_j \exp(y_j)$ *and the algorithm known as* multiplicative weights update (*MWU*)*, cf. [1, 2, 26, 35, 42].*

Example 2 (Euclidean projection)**.** *Another popular regularizer is the quadratic penalty* $h_i(x) = \sum_i {x_i}^2/2$*, which yields the* payoff projection *choice map* $\Pi_i(y) = \arg\min_{x\in\Delta}\|y-x\|^2$*, cf. [23, 45].*

4. Analysis and Results

To understand the long-run behavior of (FTRL), we will focus on the following overarching question: *Which Nash equilibria hold convergence and stability properties and how are these properties affected by the uncertainty in the players' feedback model?*

We provide the technical groundwork for our answers in Section 4.1 below; subsequently, we state our results in Section 4.2, and present the technical analysis in Section 5.

4.1. Asymptotic Stability

The first thing to note in this general context is that a game may admit several Nash equilibria, both mixed and pure. As a result, global convergence to an equilibrium from all initializations is not possible; for this reason, we will focus on the notion of (*stochastic*) *asymptotic stability* [18, 21, 34]. Heuristically, an equilibrium is *stochastically stable* if any sequence of play that begins close enough to the equilibrium in question, remains close enough with high probability; in addition, if the sequence of play eventually converges to said equilibrium, then we say that it is stochastically asymptotically stable. Formally, we have the following definition.

Definition 1. *Fix some arbitrary confidence level* $\delta > 0$*. Then* $x^* \in \mathcal{X}$ *is said to be*

1. ***Stochastically stable*** *if, for every neighborhood* U *of* x^* *in* $\mathcal{X}$*, there exists a neighborhood* U_0 *of* x^* *such that whenever* $X_0 = Q(Y_0) \in U_0$*, we have*

$$\mathbb{P}(X_n \in U \text{ for all } n = 0, 1, \dots) \geq 1 - \delta \tag{10}$$

   whenever $X_0 = Q(Y_0) \in U_0$*.*

2. ***Attracting*** *if there exists a neighborhood* U_0 *of* x^* *such that*

$$\mathbb{P}(\lim_{n\to\infty} X_n = x^*) \geq 1 - \delta \tag{11}$$

   whenever $X_0 = Q(Y_0) \in U_0$*.*

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

Definition 1 will be the mainstay of our analysis and results, so some remarks are in order.

Remark 1. A first intricate detail in the above definition is the high probability requirement: indeed, under uncertainty, a single unlucky estimation of the players' payoff vector could drive X_n away from any neighborhood of x^*, possibly never to return. In this regard, local stability results cannot be expected to hold with probability 1, hence the requirement to hold with some arbitrary confidence level in the definition above.

Remark 2. Another remark worth making is the requirement $X_0 = Q(Y_0) \in U_0$ that indicates that some strategies in $\mathcal{X}$ are not admissible as initial states. Going back to the two archetypal examples of (FTRL), Examples 1 and 2, there is a dichotomy in the properties of the corresponding mirror maps. On the one hand, the kernel of the Euclidean/quadratic regularizer is differentiable on all of $[0, 1]$. On the other hand, the derivative of the kernel of the negative Shannon-entropy goes to $-\infty$ as x goes to 0. This means that in the latter the boundaries are off the limits and inevitably some initial conditions do not belong in $\operatorname{im} Q$. We discuss this dichotomy extensively in Appendix A.2.

4.2. Main Results

We are now in a position to state our main results. The informal version is as follows.

Main Theorem. Suppose that Assumptions (A1)–(A3) hold. Then:
x^* is a strict Nash equilibrium $\iff$ x^* is stochastically asymptotically stable under (FTRL)

Formally, we get the following precise statements and corollaries for the specific feedback models described in Section 3.1.

Theorem 1. *Let $x^* \in \mathcal{X}$ be a strict Nash equilibrium of Γ. If (FTRL) is run with inexact payoff feedback satisfying Assumptions (A1) and (A2), then x^* is stochastically asymptotically stable.*

Theorem 2. *Let x^* be a mixed Nash equilibrium of Γ. If (FTRL) is run with inexact payoff feedback satisfying assumption (A3), then x^* is not stochastically asymptotically stable.*

Corollary 1. *Suppose that (FTRL) is run in a generic game with oracle-based feedback as in Model 1 and a sufficiently small step-size γ_n with $\sum_n \gamma_n^2 < \infty$. Then, a Nash equilibrium is stochastically asymprotically stable if and only if it is strict.*

Corollary 2. *Suppose that (FTRL) is run in a generic game with bandit feedback as in Model 2 and sufficiently small step-size and explicit exploration paramters with $\sum_n \gamma_n^2/\varepsilon_n < \infty$, $\sum_n \gamma_n \varepsilon_n < \infty$. Then, a Nash equilibrium is stochastically asymprotically stable if and only if it is strict.*

These results – and, in particular, the implications for the bandit case – provide a learning justification to the abundance of arguments that have been made in the refinement literature against selecting mixed Nash equilibria [9, 11, 40]. In the rest of our paper, we present an outline of the main proof ideas and defer the details to the appendix.

5. Our Techniques

5.1. The Stochastic Asymptotic Stability of Strict Nash Equilibria

At a high level, the standard tool in FTRL dynamics for questions pertaining to asymptotic stability of strict Nash equilibria is the construction of a potential – or *Lyapunov* – function. However, the analysis and the underlying structural results are considerably more involved when we shift from the continuous dynamics to discrete algorithms and more importantly in a stochastic framework with incomplete feedback information. Still, to build intuition we first recall the continuous and deterministic analogue.

The continuous-time case. In prior work [14, 17, 44], multiple instantiations of Bregman functions, like the KL-divergence have been employed as a potent tool for understanding *replicator & population dynamics*, which are the continuous analogues of MWU/EW (Example 1). Unfortunately, Bregman functions are insufficient to cover the full spectrum of regularizers studied in this work. This limitation has been sidesteped in [27] by exploiting the information of the dual space $\mathcal{Y}$ of the payoff scores, via the Fenchel coupling:

$$F_h(x, y) = h(x) + h^*(y) - \langle y, x \rangle \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \tag{12}$$

where $h^* : \mathcal{Y} \to \mathbb{R}$ is the convex conjugate of h: $h^*(y) = \sup_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}$. Indeed, $F_h(x^*, y) \geq 0$ where equality holds if and only if $x^* = Q(y)$ (Proposition A.4). Therefore, for the continuous FTRL dynamics $\dot{y}(t) = v(x(t)), x(t) = Q(y(t))$, it remains to show that the time derivative of the Lyapunov-candidate-function $L_{x^*}(y(t)) = F_h(x^*, y(t))$ is negative. This last key ingredient for the strict Nash equilibria is derived by their *variational stability* property. Formally, a point x^* is *variationally stable* if there exists a neighborhood U of x^* such that

$$\langle v(x), x - x^*\rangle \leq 0 \text{ for all } x \in U \tag{VS}$$

with equality if and only if $x = x^*$. Roughly speaking, this property states that the payoff vectors are pointing "towards" the equilibrium in question since in a neighborhood of x^*, it strictly dominates over all other strategies. Thus by applying the chain rule, (VS) implies that $\mathrm{d}L_{x^*}(y(t))/\mathrm{d}t \leq 0$ [3]. Given their usefulness also in the discrete time stochastic case, we present all the aforementioned properties in detail in the paper's supplement (Appendix A.1-A.5).

The discrete time. The core elements of the continuous time proof do not trivially extend to the discrete time case. Even though we are not able to show that $(F_h(x^*, Y_k))_{k=1}^{\infty}$ is a decreasing sequence, due to the discretization and the uncertainty involved, we prove that $F_h(x^*, Y_k) \to 0$. This immediately implies that FTRL algorithm converges to x^*, since from Proposition A.4 $F_h(x^*, Y_k) \geq \frac{1}{2K_h}\|x^* - X_k\|$.

To exploit again the Fenchel coupling as a Lyapunov function, successive differences have to be taken among $F_h(x^*, Y_{n+1}), \dots, F_h(x^*, Y_0)$. In contrast to the continuous time analysis, since the chain rule no longer applies, we can only do a second order Taylor expansion of the Fenchel coupling. Additionally, let us recall that in our stochastic feedback model, the payoff vector $\hat{v}_n = v(X_n) + Z_n + b_n$ including possibly either random zero-mean noise or systematic biased noise. Combining Proposition A.4, definition of $\hat{v}_n$ and (FTRL), we can create the following upper-bound of Fenchel coupling at each round:

$$F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_0) + \sum_{k=0}^{n} \gamma_k(\mathsf{drift}_k + \mathsf{noise}_k + \mathsf{bias}_k) + \frac{1}{2K_h}\sum_{k=0}^{n}\gamma_k^2\|\hat{v}_k\|_*^2 \tag{$\star$}$$

where $\mathsf{drift}_k = \langle v(X_k), X_k - x^*\rangle, \mathsf{noise}_k = \langle Z_k, X_k - x^*\rangle, \mathsf{bias}_k = \langle b_k, X_k - x^*\rangle$ are the related terms with the drift of the actual payoff, the zero-mean noise and the bias correspondingly. When X_n lies in a variationally stable region U_{VS} of x^*, the first-order term of drift_k, which also appears in the continuous time, corresponds actually to the negative "drift" of the variational stability which attracts Fenchel coupling to zero.

Having settled the basic framework, we split the proof sketch of Theorem 1 into two parts: *stochastic stability* & *convergence*. Our analysis relies heavily on tools from the convex analysis and martingale limit theory to control the influence of the stochastic terms in the aforementioned bound.

Step 1: *Stability*. Let $U_\varepsilon = \{x : D_h(x^*, x) < \varepsilon\}$ and $U_\varepsilon^* = \{y \in \mathcal{Y} : F_h(x^*, y) < \varepsilon\}$ be the $\varepsilon-$sublevel sets of Bregman function and Fenchel coupling respectively. Our first observation is that for all "natural" decomposable regularizers, it holds the so-called "reciprocity condition" (Propositions A.1 and A.5): essentially, this posits that U_ε and $Q(U_\varepsilon^*)$ are neighborhoods of x^* in $\mathcal{X}$. Additionally, since $F_h(x^*, y) = D_h(x^*, x)$ whenever $Q(y) = x$ and $\mathrm{supp}(x)$ contains $\mathrm{supp}(x^*)$, from

3. Analytically, $\frac{\mathrm{d}L_{x^*}(y(t))}{\mathrm{d}t} = \frac{\mathrm{d}h^*(y(t))}{\mathrm{d}t} - \langle\dot{y}(t), x^*\rangle = \langle\dot{y}(t), \nabla h^*(y)\rangle - \langle\dot{y}(t), x^*\rangle = \langle v(x(t)), x(t) - x^*\rangle \leq 0.$

Proposition A.4, it holds that $Q(U_\varepsilon^*) \subseteq U_\varepsilon$ and $Q^{-1}(U_\varepsilon) = U_\varepsilon^*$. Thus, we conclude that whenever $y \in U_\varepsilon^*$, $x = Q(y) \in U_\varepsilon$.

To proceed, fix a confidence level δ and ε sufficiently small such that (VS) holds for all $x \in U_\varepsilon$. Using Doob's maximal inequalities for (sub)martingales (Theorems 5 and 6) we can prove that with probability at least $1-\delta$, (a) $\{\sum_{k=0}^n \gamma_k \mathsf{noise}_k\}$, (b) $\{\sum_{k=0}^n \gamma_k \mathsf{bias}_k\}$ and (c) $\{\frac{1}{2K_h}\sum_{k=0}^n \gamma_k^2 \|\hat{v}_k\|_*^2\}$ are less than $\varepsilon/4$ for all $n \geq 0$. For concision, we defer the full proof to the supplement of the paper in Appendix A.7. For the rest of this part, we condition on this event and rewrite ($\star$) as $F_h(x^*, Y_{n+1}) < \sum_{k=0}^n \gamma_k \mathsf{drift}_k + \varepsilon$.

Following the definition of stability (Definition 1), we prove inductively that if X_0 belongs a smaller neighborhood, namely if $X_0 \in U_{\varepsilon/4} \cap \operatorname{im} Q$, then X_n never escapes U_ε, $X_n \in U_\varepsilon$ for all $n \geq 0$.

- Induction Basis/Hypothesis: Since $X_0 \in U_{\varepsilon/4} \cap \operatorname{im} Q$, apparently $F_h(x^*, Y_0) < \varepsilon/4$ and $X_0 \in U_\varepsilon$. Assume that $X_k \in U_\varepsilon$ for all $0 \leq k \leq n$.
- Induction Step: We will prove that $Y_{n+1} \in U_\varepsilon^*$ and consequently $X_{n+1} \in U_\varepsilon$. Since U_ε is a neighborhood of x^* in which (VS) holds we have that $\mathsf{drift}_k \leq 0$ for all $0 \leq k \leq n$. Consequently $F_h(x^*, Y_{n+1}) < \varepsilon$ which implies that $Y_{n+1} \in U_\varepsilon^*$ or equivalently $X_{n+1} \in U_\varepsilon$.

Step 2: *Convergence*. A tandem combination of stochastic Lyapunov and variational stability is the following lemma:

Lemma 1 (Informal statement of Lemma A.1)**.** *Let $x^* \in \mathcal{A}$ be a strict Nash equilibrium. If X_n does not exit a neighborhood R of x^*, in which variational stability holds, then there exists a subsequence X_{n_k} of X_n that converges to x^* almost surely.*

Indeed, if X_n is entrapped in a variationally stable region U_ε of x^* without converging to x^*, we can show that $\sum_{k=0}^\infty \gamma_k \mathsf{drift}_k \to -\infty$, while comparatively by the law of the large numbers for martingales (Theorem 3), the contribution of (a),(b),(c) is negligible. Thus, in limit ($\star$) implies that $0 \leq \liminf F_h(x^*, Y_n) \leq -\infty$, which is a contradiction.

Our final ingredient to complete the proof is that $(F_h(x^*, Y_k))_{k=1}^\infty$ behaves like an almost supermartingale when it is entrapped in a variationally stable region U_ε of x^*. So, by convergence theorem for (sub)-martingales (Theorem 4), $(F_h(x^*, Y_k))_{k=1}^\infty$ actually converges to a random finite variable. Inevitably though, $\liminf_{n\to\infty} F_h(x^*, Y_n) = \lim_{n\to\infty} F_h(x^*, Y_n) = 0$ and by Proposition A.4, $Q(Y_n) = X_n \to x^*$.

5.2. The Stochastic Instability of Mixed Nash Equilibria

For the proof of Theorem 2, it is worth mentioning that in this case stability fails for any choice of step-size. We start by focusing on the assumption of non-degeneracy (A3) of theorem's statement.

- From a game-theoretic perspective, (A3) actually demands that with non-zero probability, when players receive the payoffs corresponding to pure strategy profiles, there exists at least one player for whom at least two strategies of the equilibrium have distinct payoff signal. Note that if for each player, the payoffs corresponding to two different strategies of $\operatorname{supp}(x^*)$ were all equal [4] immediately implies a non-generic game with pure Nash equilibria.
- To illustrate this assumption in our generic feedback model, suppose that this error term ξ_n is standard normal random noise ξ_n. Indeed, the requirement of (A3) is satisfied since $\mathbb{P}(|v_{i,a}(X_n)+$

4. when all other players' also employ strategies of the equilibrium

$\xi_{ia,n} - v_{i,b}(X_n) - \xi_{ib,n}| \geq 1/|\mathcal{N}|) > 1 - \mathcal{O}\left(\exp(-1/|\mathcal{N}|^2)\right)$. Such kind of property can be derived actually for any per-coordinate independent noise since actually the event of two independent coordinates to be exactly equal has zero measure.

For the bandit models Model 1, Model 2 of the previous section, we show that (A3) is satisfied in Corollaries 3 and 4 of Appendix B.

Moving on to the proof of Theorem 2, we start our analysis by connecting the difference of the payoff signal between two pure strategies, with the difference of the changes in the output of the regularizers' kernels, θ_i:

Lemma 2 (Informal Statement of Lemma B.2)**.** *Let $X_{i,n}$ be the sequence of play in (FTRL) i.e., $X_{i,n} = Q(Y_{i,n}) \in \mathcal{X}_i$ of player $i \in \mathcal{N}$; and for some round $n \geq 0$ let $a, b \in \operatorname{supp}(X_{i,n})$ be two pure strategies of player $i \in \mathcal{N}$. Then it holds:*

$$\left(\theta_i'(X_{ia,n+1}) - \theta_i'(X_{ia,n})\right) - \left(\theta_i'(X_{ib,n+1}) - \theta_i'(X_{ib,n})\right) = \gamma_n(\hat{v}_{ia,n} - \hat{v}_{ib,n})$$

To proceed with the proof of Theorem 2 assume ad absurdum that a mixed Nash equilibrium x^* is stochastically asymptotically stable. Since x^* is mixed, there exist $a, b \in \operatorname{supp}(x^*)$. Second, the stochastic stability implies that for all $\varepsilon, \delta > 0$ if X_0 belongs to an initial neighborhood U_ε, then $\|X_n - x^*\| < \varepsilon$ for all $n \geq 0$, with probability at least $1 - \delta$. Third, by the triangle inequality for two consecutive instances of the sequence of play $X_{i,n}, X_{i,n+1}$ for any player $i \in \mathcal{N}$ it holds:

$$|X_{ia,n+1} - X_{ia,n}| + |X_{ib,n+1} - X_{ib,n}| < \mathcal{O}(\varepsilon) \text{ with probability } 1 - \delta \tag{13}$$

Consider ε sufficiently small, such that the probabilities of the strategies that belong to the support of the equilibrium are bounded away from 0, for all the points of the neighborhood. Since θ_i is continuously differentiable in $(0, 1]$, the differences described in Lemma 2 are bounded from $\mathcal{O}(\varepsilon)$ due to Eq. (13). Thus, if the sequence of play X_n is contained to an $\varepsilon-$neighborhood of x^*, then the difference of the feedback, for any player $i \in \mathcal{N}$, to two strategies of the equilibrium is $\mathcal{O}(\varepsilon/\gamma_n)$ with probability at least $1 - \delta$:

$$\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| = \mathcal{O}(\varepsilon/\gamma_n) \mid \mathcal{F}_n) \geq 1 - \delta$$

However, from assumption (A3) for a fixed round n and some player $i \in \mathcal{N}$, there exist $\beta, \pi > 0$ such that: $\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta \mid \mathcal{F}_n) = \pi > 0$. Thus by choosing $\varepsilon = \mathcal{O}(\beta\gamma_n)$ and $\delta = \pi/2$, we obtain a contradiction and our proof is complete.

6. Discussion

The equivalence between strict Nash equilibria and stable attracting states of feedback-limited (FTRL) implies that any equilibrium that exhibits payoff-indiffirence between different strategies is inherently unstable. This fragility has already been remarked from an epistemic viewpoint [40], and our results provide a complementary justification based on realistic models of learning.

In the converse direction, the generality of the feedback models considered also provides a template for proving stochastic asymptotic stability results in more demanding learning environments. A particular case of interest arises in online ad auctions where payoffs are observed with delay (or are dropped completely): depending on the delay, the estimation of the player's payoff could exhibit a bias relative to the sampling strategy, and our generic conditions provide an estimate of how large the delays can be before convergence breaks down. This opens the door to an array of fruitful research directions that we intend to pursue in the future.

References

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert 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.

[3] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[4] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[5] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *COLT '19: Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

[6] Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[7] Roberto Cominetti, Emerson Melo, and Sylvain Sorin. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83, 2010.

[8] Pierre Coucheney, Bruno Gaujal, and Panayotis Mertikopoulos. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*, 40(3):611–633, August 2015.

[9] Eddie Dekel and Drew Fudenberg. Rational behavior with payoff uncertainty. *Journal of Economic Theory*, 52: 243–267, 1990.

[10] Lampros Flokas, Emmanouil Vasileios Vlatakis-Gkaragkounis, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[11] Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, 1991.

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

[13] James Hannan. Approximation to Bayes risk in repeated play. In Melvin Dresher, Albert William Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games, Volume III*, volume 39 of *Annals of Mathematics Studies*, pages 97–139. Princeton University Press, Princeton, NJ, 1957.

[14] Marc Harper. Escort evolutionary game theory. *Physica D: Nonlinear Phenomena*, 240(18):1411–1415, September 2011.

[15] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, September 2000.

[16] Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

[17] Edwin Hewitt and Karl Stromberg. *Real and abstract analysis*. Graduate Texts in Mathematics. Springer-Verlag, New York, NY, 1975.

[18] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK, 1998.

[19] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, July 2003.

[20] Josef Hofbauer, Sylvain Sorin, and Yannick Viossat. Time average replicator and best reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, May 2009.

[21] Rafail Z. Khasminskii. *Stochastic Stability of Differential Equations*. Number 66 in Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2 edition, 2012.

[22] Robert David Kleinberg, Georgios Piliouras, and Éva Tardos. Load balancing without regret in the bulletin board model. *Distributed Computing*, 24(1):21–29, 2011.

[23] Ratul Lahkar and William H. Sandholm. The projection dynamic and the geometry of population games. *Games and Economic Behavior*, 64:565–590, 2008.

[24] David S. Leslie. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, August 2006.

[25] David S. Leslie and E. J. Collins. Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

[26] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108 (2):212–261, 1994.

[27] Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, November 2016.

[28] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

[29] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[30] Barnabé Monnot and Georgios Piliouras. Limits and limitations of no-regret learning in games. *The Knowledge Engineering Review*, 32, 2017.

[31] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[32] Ralph Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*, volume 317 of *A Series of Comprehensive Studies in Mathematics*. Springer-Verlag, Berlin, 1998.

[33] Aldo Rustichini. Optimal properties of stimulus-response learning models. *Games and Economic Behavior*, 29 (1-2):244–273, 1999.

[34] William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, Cambridge, MA, 2010.

[35] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[36] Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and Fenchel duality. In *NIPS' 06: Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pages 1265–1272. MIT Press, 2006.

[37] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2): 1–286, November 2019.

[38] Sylvain Sorin. Exponential weight algorithm in continuous time. *Mathematical Programming*, 116(1):513–528, 2009.

[39] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, pages 2989–2997, 2015.

[40] Eric van Damme. *Stability and perfection of Nash equilibria*. Springer-Verlag, Berlin, 1987.

[41] Yannick Viossat and Andriy Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, March 2013.

[42] Vladimir G. Vovk. Aggregating strategies. In *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*, pages 371–383, 1990.

[43] Jörgen W. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.

[44] Steven Weinberg. *Quantum Field Theory*. Cambridge University Press, Cambridge, UK, 2000.

[45] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML '03: Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A. Proof of stability of strict Nash equilibria

Looking at the continuous analogues of FTRL algorithms, the standard methodology leverages potential-Lyapunov arguments. However, in the discrete case multiple intrinsic challenges arise especially in the presence of uncertainty. In the prototypical example of MWU/EW (Example 1) the standard potential function is the Kullback-Leibler divergence. Thus, a natural candidate for the generalization from MWU/EW to any FTRL algorithm would be the Bregman divergence. Unfortunately, Bregman divergence does not always capture the actual behavior of FTRL algorithms in the boundary of the simplex. Hence, in order to trace the information entailed in the dual space, where FTRL algorithms truly evolve, we leverage the Fenchel coupling. In the first three subsections we present the main properties of Bregman divergence (Appendix A.1), the dichotomy among different regularizers (Appendix A.2) and then Fenchel coupling and its properties (Appendix A.4). Then we explore a structural property of strict Nash equilibria, namely variational stability (Appendix A.5). In the last section before we present our proofs we introduce some notions from martingales limit theory. With these last two sections we have established all the necessary machinery to bound the influence of the uncertainty in the behavior of the algorithm and finally prove the stability result.

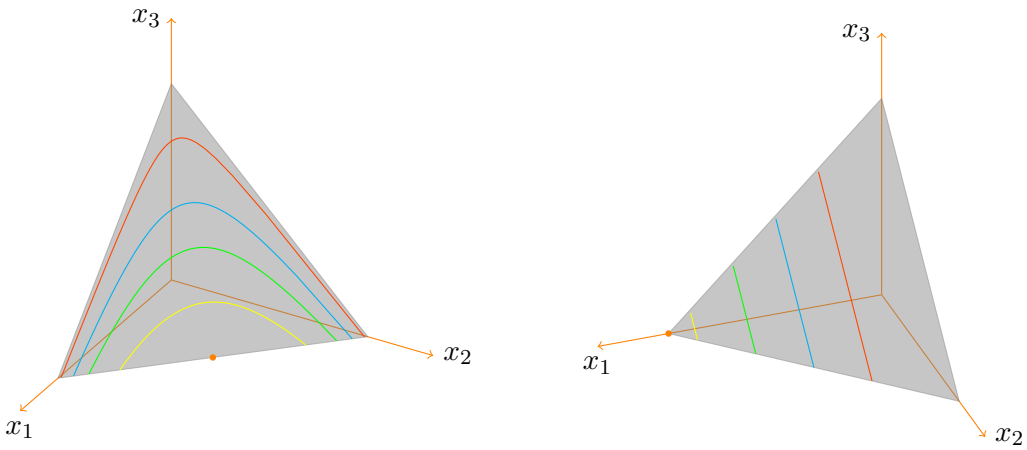

Figure 1: The level sets of KL-divergence

A.1. Bregman divergence

Bregman divergence provides a way to measure the distance of two points that belong to the simplex. Its properties render it a useful tool to prove convergence results. Below we state its definition and prove these properties that would be crucial in the establishment of our proof. Given a fixed point $p \in \mathcal{X}$ then the Bregman divergence of a function h is defined for all points $x \in \mathcal{X}$ as

$$D_h(p,x) = h(p) - h(x) - h'(x; p-x) \text{ for all } p, x \in \mathcal{X} \tag{A.1}$$

where $h'(x; p-x)$ is the one-sided derivative

$$h'(x; p-x) \equiv \lim_{t \to 0^+} t^{-1}[h(x + t(p-x)) - h(x)] \tag{A.2}$$

Notice that this definition of the Bregman divergence permits to work also with points on the boundary. It is possible that the limit of D_h attains the value of $+\infty$ if $h'(x; p-x) = -\infty$, as $x \to p$,

where p is a point of the boundary. However, the condition below ensures that this is not the case.

$$D_h(p;x) \to 0 \text{ whenever } x \to p \tag{Reciprocity}$$

This is known as the reciprocity condition. What this property actually means is that the sublevel sets of $D(p,\cdot)$ are neighborhoods of p. This is illustrated in Fig. 1, when the function employed is the negative Shannon-entropy and the induced Bregman divergnce the Kullback–Leibler divergence. Notice that for most decomposable functions h, this property holds. Below we present a proof of this statement.

Proposition A.1. *If $h(x) = \sum_i \theta(x_i)$, with θ having the properties described in* (reguralizer's properties) *and furthermore it holds that $\theta'(x) = o(1/x)$ for x close to 0, then $D_h(p;x) \to 0$ whenever $x \to p$ for all $x, p \in \mathcal{X}$.*

Proof. It is sufficient to prove that $\lim_{x\to 0}(\theta(0) - \theta(x) - \theta'(x)(0-x)) = 0$. The difference of the first two terms is obviously gives zero. Now, for the last term notice that if $\theta'(x) = o(1/x)$ for x close to 0, then $\lim_{x\to 0} x\theta'(x) = 0$ and the proof is completed. □

Additionally, Bregman divergence satisfies the properties described below.

Proposition A.2. *Let h be a K-strongly convex function defined on the simplex $\mathcal{X} = \Delta(\mathcal{A})$, that has the properties described in reguralizer's properties and let Δ_p be the union of the relative interiors of the faces of $\mathcal{X}$ that contain p i.e.,*

$$\Delta_p = \{x \in \mathcal{X} : \operatorname{supp}(p) \subseteq \operatorname{supp}(x)\} = \{x \in \mathcal{X} : x_a > 0 \text{ whenever } p_a > 0\} \tag{A.3}$$

Then

1. *$D_h(p,x) < \infty$ whenever $x \in \Delta_p$.*
2. *$D_h(p,x) \geq 0$ for all $x \in \mathcal{X}$, with equality if and only if $p = x$, more particularly*

$$D_h(p,x) \geq \frac{1}{2}K\|x-p\|^2 \text{ for all } x \in \mathcal{X} \tag{A.4}$$

Proof. For the first part, if $x \in \Delta_p$ then $h(x + t(x-p))$ is finite and smooth in a neighborhood of 0 and thus $D(p,x)$ is also finite.
The second part of the proposition, let $z = x - p$ then strong convexity yields

$$\begin{aligned} h(x+tz) &\leq th(p) + (1-t)h(x) - \frac{1}{2}Kt(1-t)\|x-p\|^2 \\ t^{-1}(h(x+tz) - h(x)) &\leq h(p) - h(x) - \frac{1}{2}(1-t)K\|x-p\|^2 \\ h(p) - h(x) - t^{-1}(h(x+tz) - h(x)) &\geq \frac{1}{2}(1-t)K\|x-p\|^2 \end{aligned}$$

And by taking $t \to 0$, we obtain the result. □

We mention at this point that from (reguralizer's properties), since for each $i \in \mathcal{N}$: $\inf_{\in[0,1]} \theta_i'' > 0$, there exists $K_i > 0$ such that for all $x, y \in [0,1]$ and $t \in [0,1]$

$$\theta_i(tx + (1-t)y) \leq t\theta_i(x) + (1-t)\theta_i(y) - \frac{K_i}{2}t(1-t)|x-y|^2 \tag{A.5}$$

In all the proofs h symbolizes the aggregate function of all the regularizers i.e., $h(x) = \sum_i h_i(x_i)$, with strong convexity parameter $K \equiv \min_i K_i$.

A.2. Steep vs non-steep

In this section we elaborate in detail the dichotomy of the properties of different regularizers mentioned in Remark 2. As we mentioned players may have different regularizers h_i employed in their choice maps $Q_i(y) = \arg\max_{x\in\mathcal{X}_i}\{\langle x, y\rangle - h_i(x)\}$. Depending on the regularizer chosen, FTRL dynamics may differ significantly. To formally express this difference, it is convenient to consider that h is an extended-real valued function $h : \mathcal{V} \to \mathbb{R}\cup\{\infty\}$ with value ∞ outside of the simplex $\mathcal{X}$. Then the subdifferential of h at $x \in \mathcal{V}$ is defined as:

$$\partial h(x) = \{y \in \mathcal{V}^* : h(x') \geq h(x) + \langle y, x' - x\rangle \; \forall x' \in \mathcal{V}\} \tag{A.6}$$

If $\partial h(x)$ is nonempty, then h is called subdifferentiable at $x \in \mathcal{X}$. When $x \in \mathrm{ri}(\mathcal{X})$ then $\partial h(x)$ is always non-empty or $\mathrm{ri}(\mathcal{X}) \subseteq \mathrm{dom}\,\partial h \equiv \{x \in \mathcal{X} : \partial h(x) \neq \emptyset\}$. Notice that when the gradient of h exists, then its subgradient always contains it. With these in mind, we present a typical separation between the different regularizers. On the one hand, *steep* regularizers like the negative Shannon-entropy become infinitely steep as x approaches the boundary or $\|\nabla h(x)\| \to \infty$. On the other hand, *non-steep* are everywhere differentiable, like the Euclidean, allowing the sequence of play to transfer between the different faces of the simplex. In the dual space of payoffs, steepness implies that the choice map is not surjective (since it cannot map all payoff vectors to points of the boundary), it is however injective (it maps a payoff vector plus a multiple of $(1, 1, \ldots, 1)$ to the same strategy). Non-steep regularizers give rise to surjective maps, which are not injective, not even up to a multiple of $(1, 1, \ldots, 1)$, to the boundary. Focusing on the more simple case of decomposable regularizers, the kernel of a steep one is differentiable on $(0, 1]$ while for non-steep the kernel is differentiable in all of $[0, 1]$. As a result, when a steep regularizer is employed the mirror map $Q : \mathcal{Y} \to \mathcal{X}$ cannot return any point of the boundary. In other words, the points of the boundary are infeasible not only as initial conditions but also as part of the sequence of play.

Remark 3. This dichotomy is important for our analysis since we study the stochastic asymptotic stability of Nash equilibria, which may lie on the boundary, and we seek a neighborhood of initial conditions such that the equilibrium to be stable and attracting. Thus, instead of demanding the existence of a neighborhood U of an equilibrium x^*, such that whenever $X_0 \in U$, x^* is stable and attracting; we demand the existence of a neighborhood U of x^* such that whenever $X_0 \in U \cap \mathrm{im}\, Q$ then x^* is stable and attracting.

A.3. Polar cone

The notion of the polar cone is tightly connected with the notion of duality. Given a finite dimensional vector space $\mathcal{V}$, a convex set $\mathcal{C} \subseteq \mathcal{V}$ and a point $x \in \mathcal{C}$ the tangent cone $\mathrm{TC}_{\mathcal{C}}(x)$ is the closure of the set of all rays emanating from x and intersecting $\mathcal{C}$ in at least one other point. The dual of the tangent cone is the polar cone $\mathrm{PC}_{\mathcal{C}}(x) = \{y \in \mathcal{V}^* : \langle y, z\rangle \leq 0 \text{ for all } z \in \mathrm{TC}_{\mathcal{C}}(x)\}$.
When the under consideration convex set is the simplex of the players' strategies, the polar cone corresponding to the boundary differs significantly from the one corresponding to the interior. Formally, the polar cone at a point x of the simplex is

$$\mathrm{PC}(x) = \{y \in \mathcal{Y} : y_a \geq y_b \text{ for all } a, b \in \mathcal{A}\}^{5} \tag{A.7}$$

5. It is always $y_a = y_b$ whenever $a, b \in \mathrm{supp}(x)$.

An illustration of this is depicted in Fig. 3. When (FTRL) is run, the notion of the polar cone emerges from the choice map $Q : \mathcal{Y} \to \mathcal{X}$, connecting the primal space of the strategies with the dual space of the payoffs. The proposition below presents this exact connection.

Proposition A.3. *Let h be a strong convex regularizer that satisfies the properties described in reguralizer's properties and let $Q : \mathcal{Y} \to \mathcal{X}$ be the induced choice map then*

1. $x = Q(y) \Leftrightarrow y \in \partial h(x)$
2. $\partial h(x) = \nabla h(x) + \mathrm{PC}(x)$ *for all* $x \in \mathcal{X}$.

A.4. Fenchel coupling

Even though Bregman divergence is a useful tool, (FTRL) evolves in the dual space of payoffs. Thus dually to the above the Fenchel coupling[6] is defined, $F_h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

$$F_h(p, y) = h(p) + h^*(y) - \langle y, p \rangle \text{ for all } p \in \mathcal{X}, y \in \mathcal{Y} \tag{A.8}$$

where $h^* : \mathcal{Y} \to \mathbb{R}$ is the convex conjugate of h: $h^*(y) = \sup_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$. The fenchel conjugate is differentiable on $\mathcal{Y}$ and it holds that

$$\nabla h^*(y) = Q(y) \text{ for all } y \in \mathcal{Y} \tag{A.9}$$

Fenchel coupling is also a measure that connects the primal with the dual space. As we mentioned above, (FTRL) evolves in the dual space and thus we use Fenchel coupling to trace its convergence properties. As the next proposition states, whenever Fenchel coupling $F(p, y)$ is bounded from above so does $\|Q(y) - p\|$. This proposition in its entity, is critical for our proof, since we first need to find a neighborhood U of attractness (See Definition 1). For this step, Bregman divergence is necessary in order to define the aforementioned neighborhood since $\|Q(y) - p\| < c$ for some constant c is not necessarily a neighborhood of p (See Appendix A.2).

Proposition A.4. *Let h be a K-strongly convex function on $\mathcal{X}$ and has the propertied described in reguralizer's properties. Let $p \in \mathcal{X}$, then*

1. $F_h(p, y) \geq \frac{1}{2} K \|Q(y) - p\|^2$ *for all* $y \in \mathcal{Y}$ *and whenever* $F_h(p, y) \to 0$, $Q(y) \to p$.
2. $F_h(p, y) = D_h(p, x)$ *whenever* $Q(y) = x$ *and* $x \in \Delta_p$.
3. $F_h(p, y') \leq F_h(p, y) + \langle y' - y, Q(y) - p \rangle + \frac{1}{2K} \|y' - y\|_*^2$.

Remark 4. Notice that the first part of the proposition is not implied by the second one, since it is possible that $\operatorname{im} Q = \operatorname{dom} \partial h$ is not always contained in Δ_p (see Appendix A.2).

Proof. For the first part, let $x = Q(y)$ then $h^*(y) = \langle y, x \rangle - h(x)$

$$F_h(p, y) = h(p) - h(x) - \langle y, p - x \rangle \tag{A.10}$$

Since $y \in \partial h(x)$ (Proposition A.3), it is

$$h(x + t(p - x)) \geq h(x) + t\langle y, p - x \rangle \tag{A.11}$$

6. The term is due to [27].

and by strong convexity of h, we have

$$h(x + t(p - x)) \leq th(p) + (1 - t)h(x) - \frac{1}{2}Kt(1 - t)\|p - x\|^2 \tag{A.12}$$

Thus by combining (A.11),(A.12) and taking $t \to 0$ we get

$$F_h(p, y) \geq h(p) - h(x) - h(p) + h(x) + \frac{K}{2}\|p - x\|^2 \geq \frac{K}{2}\|p - x\|^2 \tag{A.13}$$

For the second part of the proposition, notice that $x + t(p - x)$ lies in the relative interior of some face of $\mathcal{X}$ for t in a neighborhood of 0 and thus $h(x + t(p - x))$ is smooth and finite. So, h admits a two-sided derivative along $x - p$ and since $y \in \partial h(x)$, $\langle y, p - x\rangle = h'(x; p - x)$ and our claim naturally follows.
Finally for the last part of the proposition, we have

$$\begin{aligned} F_h(p, y') &= h(p) + h^*(y') - \langle y', p\rangle \\ &\leq h(p) + h^*(y) + \langle y' - y, \nabla h^*(y)\rangle + \frac{1}{2K}\|y' - y\|_*^2 - \langle y', p\rangle \\ &= F_h(p, y) + \langle y' - y, Q(y) - p\rangle + \frac{1}{2K}\|y' - y\|_*^2 \end{aligned}$$

where the second inequality follows from the fact that h^* is $1/K$ strongly smooth [32]. □

In terms of Fenchel coupling our reciprocity assumption can be written as

$$F_h(p, y) \to 0 \text{ whenever } Q(y) \to p \tag{Reciprocity}$$

Again for most of h decomposable, the assumption is turned into a property as we prove below.

Proposition A.5. *If $h(x) = \sum_i \theta(x_i)$, with θ having the properties described in* (reguralizer's properties) *and furthermore it holds that $\theta'(x) = o(1/x)$ for x close to 0, then $F_h(p, y) \to 0$ whenever $Q(y) \to p$ for all $p \in \mathcal{X}$.*

Proof. Again it is sufficient to prove that whenever $Q(y) = x \to 0$ then $F_h(p, y) \to 0$. Notice that from Proposition A.4 $F_h(p, y) = D_h(p, x)$ whenever $x = Q(y)$ and $x \in \Delta_p$. Thus by Proposition A.1 $Q(y) = x \to 0$ implies that $F_h(p, y) \to 0$. □

A.5. Variational stability

Definition 2 (Variational stability)**.** *A point $x^* \in \mathcal{X}$ is said to be* variationally stable *if there exists neighborhood U of x^* such that*

$$\langle v(x), x - x^*\rangle \leq 0 \text{ for all } x \in U \tag{VS}$$

with equation if and only if $x = x^*$.

What this property actually states is that in a neighborhood of x^*, it strictly dominates over all other strategies. Interestingly, strict Nash equilibria hold this property:

Proposition A.6. *For finite games in normal form, the following are equivalent:*

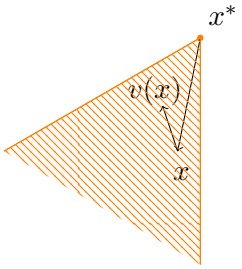

Figure 2: (VS) states that the payoff vectors are pointing "towards" the equilibrium

i) x^* *is a strict Nash equilibrium.*

ii) $\langle v(x^*), z\rangle \leq 0$ *for all* $z \in \mathrm{TC}(x^*)$ *with equality if and only if z=0.*

iii) x^* *is variationally stable.*

Proof. We will first prove that $i) \Rightarrow ii) \Rightarrow iii) \Rightarrow i)$.
$i) \Rightarrow ii)$ Since x^* is a Nash equilibrium by definition it holds for each player i that

$$\langle v(x^*), x - x^*\rangle \leq 0 \text{ for all } x \in \mathcal{X} \tag{A.14}$$

For the strict part of the inequality, by definition of strict Nash equilibria it holds that $\langle v_i(x^*), x_i - x_i^*\rangle < 0$ whenever $x_i \neq x_i^*$ and thus

$$\langle v(x^*), z\rangle = \sum_{i=1}^{N} \langle v_i(x^*), x_i - x_i^*\rangle < 0 \text{ if } x_i \neq x_i^* \text{ for some } i \text{ or } z \neq 0 \tag{A.15}$$

$ii) \Rightarrow iii)$ By definition of the polar cone, we have that $v(x^*)$ belongs to the interior of $\mathrm{PC}(x^*)$[7]. Thus by continuity there exists some neighborhood of x^* such that $v(x)$ also belongs to the polar cone of $\mathrm{PC}(x^*)$ or x^* is variationally stable.
$iii) \Rightarrow i)$ Assume now that x^* is variationally stable but not strict, then there exist for some player i a,$b \in \mathcal{A}_i$ such that $u_i(a; x^*_{-i}) = u_i(b; x^*_{-i})$. Then for $x_i = x_i^* + \lambda(e_a - e_b)$ and $x_{-i} = x^*_{-i}$ we have

$$\langle v(x^*), x - x^*\rangle = \langle v_i(x^*), \lambda(e_a - e_b)\rangle = 0 \tag{A.16}$$

which is a contradiction. □

A.6. Martingale limit theory

Our analysis leverages tools from martingale limit theory. Below we first present a simple fact for the reader to keep in mind, followed by the main theorems that we utilize in the main body of our proofs

Fact 1. *Let* $R_n = \sum_{k=1}^{n} r_k$*, where* r_k *is a positive random variable for all* $k = 0, 1, \ldots$ *attached to the filtration* $\mathcal{F}_{k-1}$*. Then* R_n *is a submartingale.*

We begin with the strong law of large numbers for martingale difference sequences:

7. Indeed if it belonged to the boundary then the equality in $ii)$ would not hold only for $z = 0$.

Theorem 3. *Let $R_n = \sum_{k=1}^{n} r_k$ be a martingale with respect to an underlying stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n=1}^{\infty}, \mathbb{P})$ and let $(\tau_n)_{n=1}^{\infty}$ be a nondecreasing sequence of positive numbers with $\lim_{n\to\infty} \tau_n = \infty$. If $\sum_{n=1}^{\infty} \tau_n^{-p} \mathbb{E}[|r_n|^p \,|\, \mathcal{F}_{n-1}] < \infty$ for some $p \in [1,2]$ almost surely, then*

$$\lim_{n\to\infty} \tau_n^{-1} R_n = 0 \text{ almost surely} \tag{A.17}$$

The second important result for our analysis is Doob's martingale convergence theorem:

Theorem 4. *If R_n is a submartingale that is bounded in L_1 (i.e., $\sup_n \mathbb{E}[|R_n|] < \infty$), R_n converges almost surely to a random variable R with $\mathbb{E}[R] < \infty$.*

Finally, we use the known as Doob's maximal inequality and one of its variants, presented below:

Theorem 5. *Let R_n be a non-negative submartingale and fix some $\varepsilon > 0$. Then:*

$$\mathbb{P}(\sup_n R_n \geq \varepsilon) \leq \frac{\mathbb{E}[R_n]}{\varepsilon} \tag{A.18}$$

Theorem 6. *Let R_n be a martingale and fix some $\varepsilon > 0$. Then:*

$$\mathbb{P}(\sup_n |R_n| \geq \varepsilon) \leq \frac{\mathbb{E}[R_n^2]}{\varepsilon^2} \tag{A.19}$$

Proofs of all these results can be found in [12].

A.7. Deferred Proof of Theorem 1

In the following preliminary result, we focus on the case of (FTRL) with payoff feedback as described in Section 3.1 and we show that if x^* is a *strict* Nash equilibrium, there exists a subsequence of $(X_n)_{n=0}^{\infty}$ that converges to it. In order to achieve this convergence result, it is necessary to assume that the sequence $(X_n)_{n=0}^{\infty}$ is contained in a neighborhood of x^*, in which (VS) holds. Here, we outline the basic steps below:

Step 0: By contradiction, assume that there exists a neighborhood, in which X_n is not contained for all sufficiently large n and assume without loss of generality that holds for all $n = 0, 1, \ldots$.

Step 1: We start by showing that the terms of the RHS of the third property described in Proposition A.4 are converging almost surely to finite values, except for one. This term, which is a consequence of x^* being variational stable, goes to $-\infty$ as $n \to \infty$.

Step 2: The next crucial observation is that the Fenchel coupling is bounded from below by 0, thanks to the first property in Proposition A.4, which gives us the contradiction.

Remark. For the interested reader, the assumption (A2), $\sum_n \gamma_n^2 M_n^2 < \infty$, that we use in the preliminary lemma and in Theorem 1 could be relaxed by using the Hölder inequality to $\sum_n \gamma_n^{1+q/2} M_n^q < \infty$ for any $q \in [2, \infty)$.

Lemma A.1. *Let $x^* \in \mathcal{A}$ be a strict Nash equilibrium. If (FTRL) is run with payoff feedback of the type (5), that satisfies (A1)-(A2) and the sequence of play $(X_n)_{n=0}^{\infty}$ does not exit a neighborhood $\mathcal{R}$ of x^*, in which variational stability holds, then there exists a subsequence X_{n_k} of X_n that converges to x^* almost surely.*

Proof. Suppose that there exists a neighborhood $U \subseteq \mathcal{R}$ of x^* , such that $X_n \notin U$ for all large enough n. Assume without loss of generality that this is true for all $n \geq 0$. Since variational stability holds in R, we have

$$\langle v(x), x - x^* \rangle < 0 \text{ for all } x \in \mathcal{R},\ x \neq x^* \tag{A.20}$$

Furthermore, from Proposition A.4 we have that for each round n:

$$F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_n) + \gamma_n \langle \hat{v}_n, X_n - x^* \rangle + \frac{1}{2K}\gamma_n^2 \|\hat{v}_n\|_*^2 \tag{A.21}$$

By applying the above inequality for all rounds from $1, ..., n$ and creating the telescopic sum we get

$$F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_0) + \sum_{k=0}^{n} \gamma_k \langle \hat{v}_k, X_k - x^* \rangle + \frac{1}{2K} \sum_{k=0}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2 \tag{A.22}$$

Remember that for the payoff vector holds that

$$\hat{v}_n = v(X_n) + Z_n + b_n$$

We now rewrite (A.22)

$$\begin{aligned} F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_0) &+ \sum_{k=0}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle + \sum_{k=0}^{n} \gamma_k \langle Z_k, X_k - x^* \rangle \\ &+ \sum_{k=0}^{n} \gamma_k \langle b_k, X_k - x^* \rangle + \frac{1}{2K} \sum_{k=0}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2 \end{aligned} \tag{A.23}$$

Let $\tau_n = \sum_{k=0}^{n} \gamma_k$ then

$$\begin{aligned} F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_0) &+ \sum_{k=0}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle + \tau_n \left(\frac{\sum_{k=0}^{n} \gamma_k \langle b_k, X_k - x^* \rangle}{\tau_n} \right) \\ &+ \tau_n \left(\frac{\sum_{k=0}^{n} \gamma_k \langle Z_k, X_k - x^* \rangle}{\tau_n} + \frac{\frac{1}{2K} \sum_{k=0}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2}{\tau_n} \right) \end{aligned} \tag{A.24}$$

We focus on the asymptotic behavior of each particular term of the previous inequality. We remind that $\mathcal{F}_n$ denotes the history of X_n up to stage n (inclusive) and thus the feedback signal, $\hat{v}_n$ is not $\mathcal{F}_n$-measurable in general.

- Let $R_n = \sum_{k=0}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2$. Then

$$\mathbb{E}[R_n] \leq \sum_{k=0}^{n} \gamma_k^2\, \mathbb{E}[\|\hat{v}_k\|_*^2] = \sum_{k=0}^{n} \gamma_k^2\, \mathbb{E}[\mathbb{E}[\|\hat{v}_k\|_*^2 \,|\, \mathcal{F}_k]] \leq \sum_{k=0}^{n} \gamma_k^2 M_k^2 < \infty \tag{A.25}$$

  where $\sum_{k=0}^{n} \gamma_k^2 M_k^2$ is finite by assumption (A2). Hence by Fact 1 and (A.25) R_n is an L_1 bounded submartingale while Doob's convergence theorem (Theorem 4) shows that almost surely

$$\lim_{n \to \infty} \tau_n^{-1} R_n = 0 \tag{A.26}$$

- Let $S_n = \sum_{k=0}^{n} \gamma_k \langle Z_k, X_k - x^* \rangle$ and $\psi_k = \gamma_k \langle Z_k, X_k - x^* \rangle$. For the expected value of ψ_n we have

$$\mathbb{E}[\psi_n \,|\, \mathcal{F}_n] = \gamma_n \langle \mathbb{E}[Z_n \,|\, \mathcal{F}_n], X_n - x^* \rangle = 0 \tag{A.27}$$

and so S_n is a martingale since $\mathbb{E}[S_n \,|\, \mathcal{F}_n] = S_{n-1}$. Moreover, for the expectation of the absolute value of ψ_n, Cauchy-Schwarz inequality implies

$$\begin{aligned}
\mathbb{E}[|\psi_n|^2 \,|\, \mathcal{F}_n] &\leq \gamma_n^2 \, \mathbb{E}[\|Z_n\|_*^2 \|X_n - x^*\|^2 \,|\, \mathcal{F}_n] && \text{(A.28)} \\
&\leq \gamma_n^2 \, \mathbb{E}[\|Z_n\|_*^2 \,|\, \mathcal{F}_n] \|\mathcal{X}\|^2 && \text{(A.29)} \\
&\leq \gamma_n^2 M_n^2 \|\mathcal{X}\|^2 && \text{(A.30)}
\end{aligned}$$

since

$$\begin{aligned}
\mathbb{E}[\|Z_n\|_*^2 \,|\, \mathcal{F}_n] &= \mathbb{E}[\|\hat{v}_n - \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n]\|_*^2 \,|\, \mathcal{F}_n] && \text{(A.31)} \\
&= \mathbb{E}[\|\hat{v}_n\|_*^2 - 2\langle \hat{v}_n, \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] \rangle + \|\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n]\|_*^2 \,|\, \mathcal{F}_n] && \text{(A.32)} \\
&= \mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] - \|\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n]\|_*^2 && \text{(A.33)} \\
&\leq \mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] \leq M_n^2 && \text{(A.34)}
\end{aligned}$$

where M_n^2 is the upper bound of $\mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n]$ described in Section 3.1.
Obviously, $\sum_{n=0}^{\infty} \tau_n^{-2} \, \mathbb{E}[|\psi_n|^2 \,|\, \mathcal{F}_n] < \infty$ and so by the strong law of large number for martingales (Theorem 3) yields that almost surely

$$\lim_{n \to \infty} \tau_n^{-1} S_n = 0 \tag{A.35}$$

- Let $W_n = \sum_{k=0}^{n} \gamma_k \langle b_k, X_k - x^* \rangle$ then by Cauchy-Schwarz inequality

$$\begin{aligned}
|\tau_n^{-1} W_n| \leq |\tau_n^{-1} \sum_{k=0}^{n} \gamma_k \langle b_k, X_k - x^* \rangle| &\leq \tau_n^{-1} \sum_{k=0}^{n} \gamma_k |\langle b_k, X_k - x^* \rangle| \\
&\leq \tau_n^{-1} \sum_{k=0}^{n} \gamma_k \|b_k\|_* \|\mathcal{X}\|
\end{aligned} \tag{A.36}$$

Let $J_n = \sum_{k=0}^{n} \gamma_k \|b_k\|_* \|\mathcal{X}\|$. Notice that $W_n \leq J_n$ and that from Fact 1 J_n is a submartingale with

$$\mathbb{E}[J_n] = \|\mathcal{X}\| \sum_{k=0}^{n} \gamma_k \, \mathbb{E}[\|b_k\|_*] \leq \|\mathcal{X}\| \sum_{k=0}^{n} \gamma_k \, \mathbb{E}[\mathbb{E}[\|b_k\|_* \,|\, \mathcal{F}_k]] \leq \|\mathcal{X}\| \sum_{k=0}^{n} \gamma_k B_k < \infty \tag{A.37}$$

where B_n is the upper bound of $\mathbb{E}[\|b_n\|_* \,|\, \mathcal{F}_n]$. Thus, J_n is a L_1 bounded submartingale and by Doob's convergence theorem (Theorem 4) almost surely

$$\lim_{n \to \infty} \tau_n^{-1} J_n = 0 \tag{A.38}$$

As a result, $\tau_n^{-1} W_n \to 0$.

- Finally, we will examine the term $\sum_{k=0}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle$. Recall that we had assumed that $X_n \in \mathcal{R} \setminus U$ for all $n \geq 0$, while variational stability holds in $\mathcal{R}$, so by continuity there exists $c > 0$, such that for all $n \geq 0$

$$\langle v(X_n), X_n - x^* \rangle \leq -c \tag{A.39}$$

We return to (A.24) and we equivalently we have that

$$\begin{aligned} F_h(x^*, Y_{n+1}) &\leq F_h(x^*, Y_0) + \sum_{k=0}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle + \tau_n(\tau_n^{-1} W_n + \tau_n^{-1} R_n + \tau_{-1}^n S_n) \\ &\leq F_h(x^*, Y_0) - c\tau_n + \tau_n(\tau_n^{-1} W_n + \tau_n^{-1} R_n + \tau_n^{-1} S_n) \end{aligned} \tag{A.40}$$

Thus, $F_h(x^*, Y_{n+1}) \sim -c\sum_{k=0}^{\infty} \gamma_k \to -\infty$.

By Proposition A.4 we conclude to a contradiction. This implies that some instance of the sequence of play is included to every neighborhood U of x^* and thus there exists subsequence X_{n_k} of X_n that almost surely converges to x^*. □

Theorem 7 (Restatement of Theorem 1)**.** *Let x^* be a strict Nash equilibrium. If* (FTRL) *is run with payoff feedback that satisfies* (A1)-(A2)*, then x^* is stochastically asymptotically stable.*

Proof. Fix a confidence level δ and let $U_\varepsilon = \{x : D_h(x^*, x) < \varepsilon\}$ and $U_\varepsilon^* = \{y \in \mathcal{Y} : F_h(x^*, y) < \varepsilon\}$.

- By Proposition A.2 for all $x \in U_\varepsilon$ it holds that $\|x - x^*\|^2 < 2\varepsilon/K$.
- By Proposition A.4 for all $x = Q(y)$, $y \in U_\varepsilon^*$ it holds that $\|x - x^*\|^2 < 2\varepsilon/K$.
- Notice that from Proposition A.4 $Q(U_\varepsilon^*) \subseteq U_\varepsilon$ and $Q^{-1}(U_\varepsilon) = U_\varepsilon^*$.

Thus we conclude that whenever $y \in U_\varepsilon^*$, $x = Q(y) \in U_\varepsilon$. Finally, by (Reciprocity) U_ε is a neighborhood of x^* . Since x^* is a strict Nash equilibrium, pick ε sufficiently small such that (VS) holds for all $x \in U_{4\varepsilon}$.

(Stability).
Assume now that $Y_0 \in U_\varepsilon^*$ and thus $F_h(x^*, Y_0) < \varepsilon \leq 4\varepsilon$. We will prove by induction that $Y_n \in U_{4\varepsilon}^*$ for all $n \geq 1$ with probability at least $1 - \delta$. Suppose that $F_h(x^*, Y_k) < 4\varepsilon$ for all $1 \leq k \leq n$ and we will prove that $Y_{n+1} \in U_{4\varepsilon}^*$ and consequently $X_{n+1} \in U_{4\varepsilon}$.

From Proposition A.4 we have

$$F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_n) + \gamma_n \langle \hat{v}_n, X_n - x^* \rangle + \frac{1}{2K} \gamma_n^2 \|\hat{v}_n\|_*^2 \tag{A.41}$$

For the payoff feedback, it holds $\hat{v}_n = v(X_n) + Z_n + b_n$. Then by telescoping the above inequality and substituting we get

$$\begin{aligned} F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_0) &+ \sum_{k=0}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle + \sum_{k=0}^{n} \gamma_k \langle Z_k, X_k - x^* \rangle \\ &+ \sum_{k=0}^{n} \gamma_k \langle b_k, X_k - x^* \rangle + \frac{1}{2K} \sum_{k=0}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2 \end{aligned} \tag{A.42}$$

We will study each term of the inequality separately.

- Let $R_n = \frac{1}{2K} \sum_{k=0}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2$ and $F_{n,\varepsilon} = \left\{\sup_{0 \leq k \leq n} R_k \geq \varepsilon\right\}$. As we discussed in Lemma A.1, R_n is a submartingale with $\mathbb{E}[R_n] \leq \sum_{k=0}^{n} \gamma_k^2 M_k^2$. Doob's maximal inequality (Theorem 5) yields

$$\mathbb{P}(F_{n,\varepsilon}) \leq \frac{\mathbb{E}[R_n]}{\varepsilon} \leq \frac{\sum_{k=0}^{n} \gamma_k^2 M_k^2}{2K\varepsilon} \tag{A.43}$$

By demanding $\sum_{k=0}^{\infty}\gamma_k^2 M_k^2 \leq 2K\varepsilon\delta/3$ the event $F_\varepsilon = \bigcup_{n=0}^{\infty} F_{\varepsilon,n}$ will occur with probability at most $\delta/3$.

- Let $S_n = \sum_{k=0}^{n}\gamma_k\langle Z_k, X_k - x^*\rangle$ and $E_{n,\varepsilon} = \{\sup_{0\leq k\leq n} S_k \geq \varepsilon\}$. Since S_n is a martingale, as we discussed in Lemma A.1, Doob's maximal inequality (Theorem 6) yields

$$\mathbb{P}(E_{n,\varepsilon}) \leq \frac{\mathbb{E}[{S_n}^2]}{\varepsilon^2} \leq \frac{\|\mathcal{X}\|^2\sum_{k=0}^{n}\gamma_k^2 M_k^2}{\varepsilon^2} \tag{A.44}$$

In order to calculate the above upper bound, we define $\psi_k = \langle Z_k, X_k - x^*\rangle$. Notice that $S_n^2 = \sum_{k=0}^{n}|\psi_k|^2 + 2\sum_{k<\ell}^{n}\psi_k\psi_\ell$. Indeed it holds that

$$\mathbb{E}[|\psi_k|^2] \leq \mathbb{E}[\mathbb{E}[\|Z_k\|_*^2\|X_k - x^*\|^2 \,|\, \mathcal{F}_k]] \tag{A.45}$$
$$\leq \mathbb{E}[\mathbb{E}[\|Z_k\|_*^2 \,|\, \mathcal{F}_k]]\|\mathcal{X}\|^2 \tag{A.46}$$

where,

$$\mathbb{E}[\|Z_k\|_*^2 \,|\, \mathcal{F}_k] = \mathbb{E}[\|\hat{v}_k - \mathbb{E}[\hat{v}_k \,|\, \mathcal{F}_k]\|_*^2 \,|\, \mathcal{F}_k] \tag{A.47}$$
$$= \mathbb{E}[\|\hat{v}_k\|_*^2 - 2\langle\hat{v}_k, \mathbb{E}[\hat{v}_k \,|\, \mathcal{F}_k]\rangle + \|\mathbb{E}[\hat{v}_k \,|\, \mathcal{F}_k]\|_*^2 \,|\, \mathcal{F}_k] \tag{A.48}$$
$$= \mathbb{E}[\|\hat{v}_k\|_*^2 \,|\, \mathcal{F}_k] - \|\mathbb{E}[\hat{v}_k \,|\, \mathcal{F}_k]\|_*^2 \leq M_k^2 \tag{A.49}$$
$$\leq \mathbb{E}[\|\hat{v}_k\|_*^2 \,|\, \mathcal{F}_k] \leq M_k^2 \tag{A.50}$$

Furthermore, for all $k \neq \ell$ it holds that $\mathbb{E}[\psi_k\psi_\ell] = \mathbb{E}[\mathbb{E}[\psi_k\psi_\ell \,|\, \mathcal{F}_{k\vee\ell}]] = 0$.
Thus, by demanding $\sum_{k=0}^{\infty}\gamma_k^2 M_k^2 \leq \dfrac{\varepsilon^2\delta}{3\|\mathcal{X}\|^2}$ we ensure that the event $E_\varepsilon = \bigcup_{n=0}^{\infty} E_{\varepsilon,n}$ will occur with probability at most $\delta/3$.

- Let $W_n = \sum_{k=0}^{n}\gamma_k\langle b_k, X_k - x^*\rangle$, $J_n = \sum_{k=0}^{n}\gamma_k\|b_k\|_*\|\mathcal{X}\|$ as we discussed in Lemma A.1

$$W_n \leq J_n \tag{A.51}$$

where J_n is a submartingale with $\mathbb{E}[J_n] \leq \|\mathcal{X}\|\sum_{k=0}^{n}\gamma_k B_k$. Similarly to the previous steps let $D_{\varepsilon,n} = \{\sup_{0\leq k\leq n} J_k \geq \varepsilon\}$, then Doob's maximal inequality (Theorem 5) yields

$$\mathbb{P}(D_{\varepsilon,n}) \leq \frac{\mathbb{E}[J_n]}{\varepsilon} \leq \frac{\|\mathcal{X}\|\sum_{k=0}^{n}\gamma_k B_k}{\varepsilon} \tag{A.52}$$

By demanding $\sum_{k=0}^{\infty}\gamma_k B_k \leq \dfrac{\varepsilon\delta}{3\|\mathcal{X}\|}$ then the event $D_\varepsilon = \cup_{n=0}^{\infty} D_{\varepsilon,n}$ will happen with probability at most $\delta/3$, which implies that with probability at most $\delta/3$ W_n will exceed ε for all $n \geq 0$.

- Furthermore, if X_k belongs to a neighborhood in which (VS) holds for all $0 \leq k \leq n$, we have

$$\langle v(X_k), X_k - x^*\rangle \leq 0 \text{ for all } n \geq 0 \tag{A.53}$$

By demanding the parameters of the algorithm to satisfy:

$$\sum_{k=0}^{\infty}\gamma_k^2 M_k^2 \leq \min\left\{\frac{\varepsilon^2\delta}{3\|\mathcal{X}\|^2}, \frac{2K\varepsilon\delta}{3}\right\} \quad \& \quad \sum_{k=0}^{\infty}\gamma_k B_k \leq \frac{\varepsilon\delta}{3\|\mathcal{X}\|\|\mathcal{Y}\|_*}$$

If all of $\bar{E}_\varepsilon, \bar{F}_\varepsilon, \bar{D}_\varepsilon$ hold, this happens with probability $\mathbb{P}(\bar{E}_\varepsilon \bigcap \bar{F}_\varepsilon \bigcap \bar{D}_\varepsilon) \geq 1-\delta$ and from (A.42) we have $F_h(x^*, Y_{n+1}) < 4\varepsilon$. This immediately yields that $Y_{n+1} \in U_{4\varepsilon}^*$ and consequently as we explained in the begin of the proof $X_{n+1} \in U_{4\varepsilon}$, in which variational stability holds, with probability at least $1-\delta$.

(Convergence).
By Lemma A.1 there exists a subsequence X_{n_k} that converges to x^*. By (Reciprocity) we have that $\liminf_{n\to\infty} F_h(x^*, Y_n) = 0$. In order to complete the proof, it is sufficient to prove that the limit of $F_h(x^*, Y_n)$ exists. Notice that since the sequence of play remains in $U_{4\varepsilon}$ variational stability holds and thus $\langle v(X_n), X_n - x^*\rangle \leq 0$. Again using Proposition A.4 we have:

$$F_h(x^*, Y_{n+1}) \leq F_h(x^*, Y_n) + \gamma_n\langle \hat{v}_n, X_n - x^*\rangle + \frac{1}{2K}\gamma_n^2\|\hat{v}_n\|_*^2 \tag{A.54}$$

$$\mathbb{E}[F_h(x^*, Y_{n+1}) \,|\, \mathcal{F}_n] \leq F_h(x^*, Y_n) + \gamma_n\,\mathbb{E}[\langle b_n, X_n - x^*\rangle \,|\, \mathcal{F}_n] + \frac{1}{2K}\gamma_n^2\,\mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] \tag{A.55}$$

$$\leq F_h(x^*, Y_n) + \gamma_n\,\mathbb{E}[\langle b_n, X_n - x^*\rangle \,|\, \mathcal{F}_n] + \frac{1}{2K}\gamma_n^2 M_n^2 \tag{A.56}$$

Notice that since from Proposition A.4 $F_h(x^*, Y) \geq 0$, if we apply absolute values in the above inequality we have

$$\mathbb{E}[F_h(x^*, Y_{n+1}) \,|\, \mathcal{F}_n] = |\mathbb{E}[F_h(x^*, Y_{n+1}) \,|\, \mathcal{F}_n]| \tag{A.57}$$

$$\leq |F_h(x^*, Y_n)| + \gamma_n\,\mathbb{E}[|\langle b_n, X_n - x^*\rangle| \,|\, \mathcal{F}_n] + \frac{1}{2K}\gamma_n^2 M_n^2 \tag{A.58}$$

$$\leq F_h(x^*, Y_n) + \gamma_n\,\mathbb{E}[\|b_n\|_* \,|\, \mathcal{F}_n]\|\mathcal{X}\| + \frac{1}{2K}\gamma_n^2 M_n^2 \tag{A.59}$$

$$\leq F_h(x^*, Y_n) + \gamma_n B_n\|\mathcal{X}\| + \frac{1}{2K}\gamma_n^2 M_n^2 \tag{A.60}$$

Let

$$R_n = F_h(x^*, Y_n) + \|\mathcal{X}\|\sum_{k=n}^{\infty}\gamma_k B_k + \frac{1}{2K}\sum_{k=n}^{\infty}\gamma_k^2 M_k^2 \tag{A.61}$$

Then

$$\mathbb{E}[R_{n+1} \,|\, \mathcal{F}_n] \leq \mathbb{E}[F_h(x^*, Y_{n+1}) \,|\, \mathcal{F}_n] + \sum_{k=n+1}^{\infty}\gamma_k B_k\|\mathcal{X}\| + \frac{1}{2K}\sum_{k=n+1}^{\infty}\gamma_k^2 M_k^2 \tag{A.62}$$

$$\leq F_h(x^*, Y_n) + \sum_{k=n}^{\infty}\gamma_k B_k\|\mathcal{X}\| + \frac{1}{2K}\sum_{k=n}^{\infty}\gamma_k^2 M_k^2 \tag{A.63}$$

$$= R_n \tag{A.64}$$

Therefore R_n is a supermartingale and it is also L_1 bounded (each one of the terms is bounded) and so from Doob's convergence theorem (Theorem 4) R_n converges to a finite random variable and so does $F_h(x^*, Y_n)$. Inevitably, $\liminf_{n\to\infty} F_h(x^*, Y_n) = \lim_{n\to\infty} F_h(x^*, Y_n) = 0$ and by Proposition A.4, $Q(Y_n) = X_n \to x^*$.

The above analysis shows that whenever $Y_0 \in U_\varepsilon^*$ and thus $X_0 \in U_\varepsilon \cap \operatorname{im} Q$, $X_n \in U_{4\varepsilon} \cap \operatorname{im} Q$ and converges to x^* with arbitrary high probability. Hence, x^* is stochastically asymptotically stable.

□

Appendix B. Proof of instability of mixed Nash equilibria

B.1. Proofs of assumptions for Model 2, Model 1

Below we provide a proof for our claim in Model 2 that $b_n = \mathcal{O}(\varepsilon_n)$, $M_n^2 = \mathcal{O}(1/\varepsilon_n)$. Focusing on one player $i \in \mathcal{N}$, notice that

$$\mathbb{E}[\hat{v}_{i,n} \,|\, \mathcal{F}_n] = \sum_{\alpha_{-i} \in \mathcal{A}_{-i}} \hat{X}_{-i,n}(u_i(\alpha_{i,1}; \alpha_{-i}), \dots, u_i(\alpha_{i,|\mathcal{A}_i|}; \alpha_{-i})) = v_i(\hat{X}_n) \tag{B.1}$$

Having this in mind $\hat{v}_{i,n}$ can be viewed as

$$\hat{v}_{i,n} = v_i(X_n) + Z_{i,n} + b_{i,n} \tag{B.2}$$

where $Z_{i,n} = \hat{v}_{i,n} - \mathbb{E}[\hat{v}_{i,n} \,|\, \mathcal{F}_n] = \hat{v}_{i,n} - v_i(\hat{X}_n)$ and $b_{i,n} = v_i(\hat{X}_n) - v_i(X_n)$. Thus, since $v_i(x)$ is multi-linear in x and $\hat{X}_{i,n} = (1 - \varepsilon_n)X_{i,n} + \varepsilon_n/|\mathcal{A}_i|$ it follows that $b_n = \mathcal{O}(\varepsilon_n)$. Finally, similarly to (B.1) we can conclude that $M_n^2 = \mathcal{O}(1/\varepsilon_n)$.

We continue by proving that assumption (A3) is indeed satisfied for both Model 1, Model 2. This is due to the genericity of the game. Actually in the following lemma and corollaries we show that there exist player $i \in \mathcal{N}$, strategies $a, b \in \operatorname{supp}(x_i^*)$ and pure strategy profile $\alpha_{-i} \in \operatorname{supp}(x_{-i}^*)$, where x^* is a mixed Nash equilibrium such that $|u_i(a; \alpha_{-i}) - u_i(b; \alpha_{-i})| \geq \beta$ for some $\beta > 0$. In order to acquire the exact statement of (A3), we have to take into account the round in which the game is evolved. Let $n > 0$ be this round, then when examining the stochastic asymptotic stability of a mixed Nash equilibrium x^*, the sequence of play is contained in a neighborhood of x^* and thus all of the strategies belonging to the support of x^* have strictly positive probability to be chosen, verifying the statement of (A3).

Lemma B.1. *If the game is generic and has a mixed Nash equilibrium x^*, then there exist player $i \in \mathcal{N}$, pure strategies $a, b \in \operatorname{supp}(x_i^*)$ ($a \neq b$) and pure strategy profile $\alpha_{-i} \in \operatorname{supp}(x_{-i}^*)$ such that $u_i(a; \alpha_{-i}) \neq u_i(b; \alpha_{-i})$.*

Proof. Assume that for all players $i \in \mathcal{N}$, pure strategy profiles $\alpha_{-i} \in \operatorname{supp}(x_{-i}^*)$ and pure strategies $a, b \in \operatorname{supp}(x_i^*)$ it is

$$u_i(a; \alpha_{-i}) = u_i(b; \alpha_{-i}) \tag{B.3}$$

Then for each player i, this implies that all of the payoffs corresponding to pure strategy profiles, which consists of the support of the equilibrium, are equal. Then each pure strategy profile $(\alpha_i; \alpha_{-i}) \in \operatorname{supp}(x^*)$ is a pure Nash equilibrium, which is a contradiction to the genericity of the game. □

Immediate implications of Lemma B.1 are:

Corollary 3. *There exists player $i \in \mathcal{N}$ and pure strategy profile $(\alpha_i; \alpha_{-i}) \in \mathrm{supp}(x^*)$, such that $u_i(\alpha_i; \alpha_{-i}) \neq 0$.*

Corollary 4. *There exist $\beta' > 0$, player i, strategies $a, b \in \mathrm{supp}(x_i^*)$ and pure strategy profile $\alpha_{-i} \in \mathrm{supp}(x_{-i}^*)$ such that $|u_i(a; \alpha_{-i}) - u_i(b; \alpha_{-i})| \geq \beta'$. There also exist $\beta'' > 0$ and $(\alpha_i; \alpha_{-i}) \in \mathrm{supp}(x^*)$ such that $|u_i(\alpha_i; \alpha_{-i})| \geq \beta''$.*

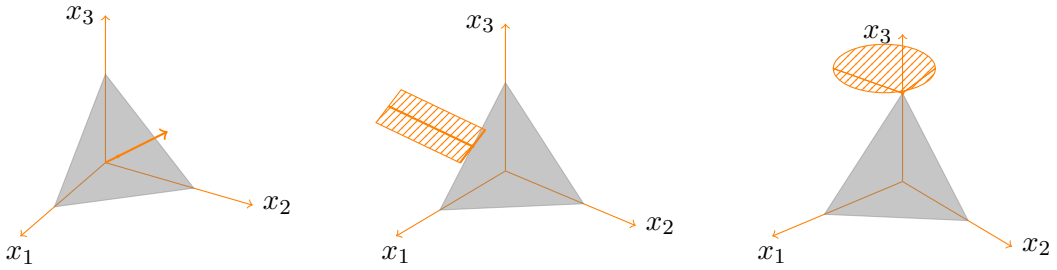

Figure 3: The polar cone corresponding to different points of the simplex. For an interior point this is a line perpendicular to the simplex. For a point of the boundary, it is a plane perpendicular to the simplex tangential to the point of the boundary. For an edge the polar cone corresponds to a cone.

B.2. Deferred proof of Theorem 2

Before moving on our proof we first provide some intuition derived from the notion of the polar cone (Appendix A.3). Looking at Fig. 3, the polar cone corresponding to *fully mixed* or *mixed* Nash equilibria has a key difference with the one corresponding to *strict* Nash equilibria. The latter, in contrast to the former, is fully dimensional. Thus intuitively, considering a sufficiently small neighborhood of a *mixed* Nash equilibrium, the slightest perturbation in the dual space of the payoffs, will lead to instability of the system. Our result is based on this intuition; we prove by contradiction that there exists a sufficiently small neighborhood of a *mixed* Nash equilibrium, from which the sequence of play will escape with strictly positive probability. The decomposability assumption of the regularizers ensures that the proof holds also for steep regularizers (See Appendix A.2).

Below, leveraging the definition of the polar cone in simplex, we prove a useful property for the difference of the aggregated payoffs of FTRL for a sequence of play that shares common pure strategies.

Lemma B.2. *Let $X_i = Q(Y_i) \in \mathcal{X}_i$ be a mixed strategy profile and $a, b \in \mathrm{supp}(X_i)$ be two pure strategies, for some player $i \in \mathcal{N}$. Then it holds:*

$$\langle Y_i, e_a - e_b \rangle = \langle \nabla h_i(X_i), e_a - e_b \rangle$$

Additionally, if (FTRL) is run then for a sequence of play $X_{i,n_1}, \ldots, X_{i,n_2}$ that maintains in its support both pure strategies $a, b \in \mathcal{A}_i$ it holds

$$\langle Y_{i,k_1} - Y_{i,k_2}, e_a - e_b \rangle = \langle \nabla h_i(X_{i,k_1}) - \nabla h_i(X_{i,k_2}), e_a - e_b \rangle \; \forall k_1, k_2 \in \{n_1, \ldots, n_2\}$$

Proof. From Proposition A.3, Y_i can be analyzed as $Y_i = \nabla h_i(X_i) + G$, $G \in \mathrm{PC}(X_i)$. Notice that $\nabla h_i(X_i) = (\theta_i(X_{i,\alpha_1}), \dots, \theta_i(X_{i,\alpha_{|\mathcal{A}_i|}}))$. Since X_i assigns positive probability to both a, b, by definition of the polar cone it is $G_a = G_b$. Thus,

$$\langle Y_i, e_a - e_b\rangle = G_a + \theta_i'(X_{i,a}) - G_b - \theta_i'(X_{i,b}) \tag{B.4}$$
$$= \langle \nabla h_i(X_i), e_a - e_b\rangle \tag{B.5}$$

For the second part, by applying (B.5) for both cases of Y_{i,k_1}, Y_{i,k_2} we have:

$$\langle Y_{i,k_1}, e_a - e_b\rangle = \langle \nabla h_i(X_{i,k_1}), e_a - e_b\rangle \tag{B.6}$$
$$\langle Y_{i,k_2}, e_a - e_b\rangle = \langle \nabla h_i(X_{i,k_2}), e_a - e_b\rangle \tag{B.7}$$

From the subtraction of the above equations, we derive the desideratum:

$$\langle Y_{i,k_1} - Y_{i,k_2}, e_a - e_b\rangle = \langle \nabla h_i(X_{i,k_1}) - \nabla h_i(X_{i,k_2}), e_a - e_b\rangle \tag{B.8}$$

□

Theorem 8 (Restatement of Theorem 2)**.** *Let x^* be a mixed Nash equilibrium. If* (FTRL) *is run with any feedback model that satisfies* (A3)*, then x^* cannot be stochastically asymptotically stable for any choice of step-schedules.*

Proof. We start by determining all the parameters of the algorithm (FTRL) and we assume ad absurdum that x^* is a mixed Nash equilibrium, which is stochastically asymptotically stable. Then for all neighborhoods U of x^* and $\delta > 0$, there exists some neighborhood U_0 such that whenever $X_0 \in U_0$, it holds that $X_n \in U$ for all $n \geq 0$ with probability at least $1 - \delta$. This equivalently implies that for all $\varepsilon, \delta > 0$ if $X_0 \in U_0$, $\|X_n - x^*\| < \varepsilon$ for all $n \geq 0$, with probability at least $1 - \delta$. We leave ε to be chosen at the end of our analysis, but we will consider it to be fixed.
For each player $i \in \mathcal{N}$ and round n if $X_{i,n}, X_{i,n+1}$ are two consecutive instances of the sequence of play; then $\|X_{i,n} - x_i^*\| < \varepsilon$, $\|X_{i,n+1} - x_i^*\| < \varepsilon$ and by the triangle inequality

$$\|X_{i,n+1} - X_{i,n}\| < 2\varepsilon \tag{B.9}$$

We fix a round n and focus on player $i \in \mathcal{N}$ who has the property of (A3); Since for two pure strategies $a, b \in \mathrm{supp}(x_i^*)$ of player $i \in \mathcal{N}$, holds that $\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta \,|\, \mathcal{F}_n) > 0$ for all $n \geq 0$, there exists for each round $n \geq 0$, $\pi_n > 0$ such that $\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta \,|\, \mathcal{F}_n) = \pi_n$. Choose δ such that $\delta < \pi_n$ and consequently

$$1 - \delta > 1 - \pi_n \tag{B.10}$$

This is possible, since π_n is strictly positive and δ can be chosen arbitrarily small.
Consider now the projection of the aggregate payoffs $Y_{i,n}, Y_{i,n+1}$ in the difference of the directions of these two strategies. From Lemma B.2 we have

$$\langle Y_{i,n+1} - Y_{i,n}, e_a - e_b\rangle = \langle \nabla h_i(X_{i,n+1}) - \nabla h_i(X_{i,n}), e_a - e_b\rangle \tag{B.11}$$

However, by definition of (FTRL) $Y_{i,n+1} - Y_{i,n} = \gamma_n \hat{v}_{i,n}$ and by taking into consideration that the regularizers used are decomposable, we get

$$(\theta_i'(X_{ia,n+1}) - \theta_i'(X_{ib,n+1}) - (\theta_i'(X_{ia,n}) - \theta_i'(X_{ib,n}))) = \gamma_n \langle \hat{v}_{i,n}, e_a - e_b\rangle \tag{B.12}$$

By rearranging we have

$$\left(\theta_i'(X_{ia,n+1}) - \theta_i'(X_{ia,n})\right) - \left(\theta_i'(X_{ib,n+1}) - \theta_i'(X_{ib,n})\right) = \gamma_n(\hat{v}_{ia,n} - \hat{v}_{ib,n}) \tag{B.13}$$

As a consequence of θ_i being continuously differentiable in all of $(0,1]$, θ_i' is continuous in $[L(\varepsilon),1]$, where $L(\varepsilon)$ is the lower bound of X_{ia}, X_{ib} whenever $\|X_i - x_i^*\| < \varepsilon$. $L(\varepsilon)$ can be guaranteed to be positive for a sufficiently small $\varepsilon < \varepsilon'$, which ensures that all the points of the neighborhood contain the support of the equilibrium for player i. Therefore, from extreme value theorem in θ_i', there exist finite C_a, C_b corresponding to a, b equivalently, such that

$$\left|\theta_i'(X_{ia,n+1}) - \theta_i'(X_{ia,n})\right| \leq C_a|X_{ia,n+1} - X_{ia,n}| < 2 \cdot C_a \cdot \varepsilon \tag{B.14}$$

$$\left|\theta_i'(X_{ib,n+1}) - \theta_i'(X_{ib,n})\right| \leq C_b|X_{ib,n+1} - X_{ib,n}| < 2 \cdot C_b \cdot \varepsilon \tag{B.15}$$

By applying the triangle inequality in (B.13) and using (B.14),(B.15) we get

$$\gamma_n|\hat{v}_{ia,n} - \hat{v}_{ib,n}| < (2 \cdot C_a + 2 \cdot C_b) \cdot \varepsilon \tag{B.16}$$

Equivalently,

$$|\hat{v}_{ia,n} - \hat{v}_{ib,n}| < \frac{2 \cdot C_a + 2 \cdot C_b}{\gamma_n} \cdot \varepsilon \tag{B.17}$$

The above inequality holds with probability $1-\delta$. Thus, if the sequence of play X_n is contained to an $\varepsilon-$neighborhood of x^* i.e., $\|X_n - x^*\| < \varepsilon$ for all $n \geq 0$, then the difference of the feedback, for some player $i \in \mathcal{N}$, to two strategies of the equilibrium is $O(\varepsilon/\gamma_n)$ with probability at least $1-\delta$. We now fix ε to be

$$\varepsilon < \min\left\{\varepsilon', \frac{\gamma_n}{2 \cdot C_a + 2 \cdot C_b}\beta\right\} \tag{B.18}$$

and consequently

$$\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| < \beta \,|\, \mathcal{F}_n) \geq 1 - \delta \tag{B.19}$$

However, from assumption (A3), it holds that

$$\mathbb{P}(|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta) \geq \pi_n \tag{B.20}$$

Combining (B.19),(B.20) we conclude

$$1 = \mathbb{P}\left[\{|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta\} \cup \{|\hat{v}_{ia,n} - \hat{v}_{ib,n}| < \beta\}\right] \tag{B.21}$$

$$= \mathbb{P}\left[|\hat{v}_{ia,n} - \hat{v}_{ib,n}| \geq \beta\right] + \mathbb{P}\left[|\hat{v}_{ia,n} - \hat{v}_{ib,n}| < \beta\right] \tag{B.22}$$

$$\geq \pi_n + 1 - \delta \tag{B.23}$$

$$> 1 \tag{B.24}$$

which is a contradiction.

Thus, a mixed Nash equilibrium cannot be stochastically asymptotically stable, under (FTRL) for types of payoff feedback described in Section 3.1. Notice that this analysis holds even for the first round. Once the parameters of the algorithm have been determined, asymptotic instability can be derived in whichever finite round. □