The convergence rate of regularized learning in games: From bandits and uncertainty to optimism and beyond

Angeliki Giannou
Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
giannouangeliki@gmail.com

Emmanouil V. Vlatakis-Gkaragkounis
Department of Computer Science
Columbia University
New York, NY 10025
emvlatakis@cs.columbia.edu

Panayotis Mertikopoulos
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG &
Criteo AI Lab
panayotis.mertikopoulos@imag.fr

Abstract

In this paper, we examine the convergence rate of a wide range of regularized methods for learning in games. To that end, we propose a unified algorithmic template that we call *"follow the generalized leader"* (FTGL), and which includes as special cases the canonical "follow the regularized leader" algorithm, its optimistic variants, extra-gradient schemes, and many others. The proposed framework is also sufficiently flexible to account for several different feedback models – from full information to bandit feedback. In this general setting, we show that FTGL algorithms converge locally to strict Nash equilibria at a rate which *does not depend* on the level of uncertainty faced by the players, but only on the geometry of the regularizer near the equilibrium. In particular, we show that algorithms based on entropic regularization – like the exponential weights algorithm – enjoy a linear convergence rate, while Euclidean projection methods converge to equilibrium in a *finite* number of iterations, even with bandit feedback.

1 Introduction

In the presence of uncertainty, the players of a game may not have full knowledge of its structure, "or the ability and inclination to go through any complex reasoning process to calculate an equilibrium. But the participants are still supposed to adapt by accumulating empirical information on the relative advantages of the various pure strategies at their disposal". This aphorism – originally due to Nash [36, p. 21] – constitutes the driving principle of game-theoretic learning, and highlights one of the field's most central questions: *Does learning with empirical observations lead to a Nash equilibrium? And, if so, at what rate?*

These questions have been at the forefront of game-theoretic research ever since the early days of the field, and they have recently received renewed attention via their connection to multi-agent reinforcement learning [45], generative adversarial networks [18], auctions [46], and many other applications where online decision-making plays a major role. Still, any attempt to provide a positive answer to these questions must wrestle with a major roadblock: the well-known impossibility result of Hart and Mas-Colell [20] shows that there are no uncoupled dynamics that converge to Nash equilibrium in *all* games, thus shattering any hope of obtaining a universal convergence result.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), .

In view of the above, contemporary research on game-theoretic learning has focused on relaxing the feedback requirements of the players' learning processes, and understanding the stability – and *instability* – properties of different kinds of equilibria under popular learning algorithms. One property that stands out in this regard is the so-called "folk theorem" of evolutionary game theory [21], which can be stated as follows: Under the replicator dynamics – the continuous-time limit of the multiplicative / exponential weights (EW) algorithm [2, 31, 47] – *a Nash equilibrium is stable and attracting if and only if it is strict* (i.e., if and only if each player has a unique best response).

The replicator dynamics are the most widely studied model for evolution in population games, so the above equivalence essentially delineates what is and what isn't achievable in an evolutionary setting. In the context of online learning (our paper's main focus), a similar equivalence was obtained only recently [11, 15, 32], but it extends to the entire family of *"follow the regularized leader"* (FTRL) dynamics [43, 44], in both continuous [11, 32] and discrete time [15]. In particular, [15] studied the convergence of discrete-time FTRL models in the presence of uncertainty, and proved a high-probability, stochastic version of this equivalence that holds for several different types of feedback (full information, bandit, etc.). Thus, coupled with the prominence of FTRL in online and game-theoretic learning, strict Nash equilibria emerge as the only stable limit points of regularized learning under uncertainty.

Our contributions. One important limitation of the above results is that they are qualitative in nature. Indeed, even though asymptotic stability guarantees that a learning process converges locally to a strict equilibrium, it provides no information about the *speed* of this convergence. In particular, especially for discrete-time models of regularized learning, asymptotic stability does not provide any guidance on how to tune the algorithm's hyperparameters (learning rate, mixing, etc.), and/or what to expect in terms of the number of iterations required to reach a neighborhood of a Nash equilibrium.

Our paper aims to provide quantitative answers to these questions for a wide array of regularized learning methods in the presence of uncertainty and limited information. To do so, we first introduce a flexible algorithmic framework – dubbed *"follow the generalized leader"* (FTGL) – that incorporates a broad spectrum of action choice mechanisms and feedback models. In more detail (and in analogy to FTRL), the FTGL template maintains a cumulative estimate for the payoff of each action available to the learner, and then selects a mixed strategy via a suitable "regularized" choice map. Specifically:

1. In terms of regularization, the FTGL template includes as special cases the standard logit choice and Euclidean projection methods (as well as all other standard regularizers used in practice).
2. In terms of the information used to update the "aggregate score" of each pure strategy, FTGL includes "vanilla" FTRL, its optimistic variants [10, 40–42], extra-gradient and mirror-prox methods [25, 27, 37], with either full, oracle-based, or bandit feedback.

In this general context, our main result may be summarized as follows. First, we introduce a "rate function" ϕ that depends *only* on the regularizer defining the learning process, and which captures the sensitivity of the induced choice map to external stimuli: for example, $\phi(x) = \exp(x)$ for entropic / logit choice models, whereas $\phi(x) = [x]_+$ for methods run with Euclidean projections. We then show that, with probability at least $1 - \delta$, the algorithm's local rate of convergence to a strict equilibrium x^* is of the form $\|X_n - x^*\| \leq \phi(d - c\sum_{s=1}^{n} \gamma_s)$, where γ_n is the method's learning rate and c, d are constants with $c > 0$.

This result shows that the convergence speed of FTGL methods depends *only* on the choice of regularizer and learning rate: for example, EW methods run with a constant step size converge to an equilibrium at an exponential rate, whereas Euclidean regularization attains convergence in a *finite* number of iterations. From a regret-theoretic point of view, this is somewhat surprising because the regret guarantees of entropic FTRL (the EW algorithm) are far superior to those of FTRL with Euclidean regularization [5, 43].

Equally surprising is the fact that the type of feedback employed *does not affect the method's rate of convergence:* ceteris paribus, the base sequence of states generated by an FTGL method attains the *same* rate of convergence to strict Nash equilibria, whether run with full, partial, or bandit / payoff-based feedback. This comes into stark contrast with the corresponding rates of regret minimization, which depend crucially on the type of feedback received [6, 29]; in a certain, precise sense, this robustness in the face of uncertainty shows that regret minimization and convergence to Nash equilibrium are fundamdentally different questions.

Related work. The convergence speed of methods based on the FTRL template – "vanilla", optimistic, or otherwise – have been studied extensively in the context of monotone games and variational inequalities; for a (highly incomplete) list of recent references, see [9, 10, 16, 17, 22, 24, 30, 33–35] and references therein. In this branch of the literature, there are two distinct threads: results concerning the convergence of the "time-average" of the process [16, 25, 35, 37], and those focusing on the algorithm's "last-iterate" [9, 10, 17, 22, 24, 30]. In the latter case (which is the one closest to our setting), the fastest achievable speed of convergence is exponential when the method is run with a finetuned constant step-size, perfect payoff gradient observations, and the operator defining the problem is strongly monotone and Lipschitz smooth. When run with stochastic gradients, the corresponding min-max optimal rate is $\mathcal{O}(1/T)$ under the same assumptions (zeroth-order rates are usually much worse). The apparent gulf between the rates of convergence obtained for monotone games and those obtained herein have to do with two crucial factors: first, we are studying *finite games*, which are typically not monotone; second, we are examining the algorithm's rate of convergence to *strict* equilibria, which are corner points of the problem's domain. This means that the geometry of the problem around a strict equilibrium is fundamentally sharper than the geometry around a solution of a generic monotone variational inequality, a fact which in turn explains the qualitatively different nature of the rates we obtain.

In the context of finite games, there have been several works examining the speed of convergence to the game's set of coarse correlated equilibria (CCE) by leveraging the algorithm's regret minimization properties, cf. [3, 4, 12, 13, 38, 46] and references therein. However, in addition to examining the algorithm's empirical average – as opposed to the induced day-to-day sequence of play – these results focus almost exclusively on CCE, which means that it is not possible to draw any conclusions about convergence to the game's Nash set – qualitatively or quantitatively. To the best of our knowledge, the closest work to our own in the literature is the paper of Cohen et al. [8] who showed that the EXP3 algorithm with explicit exploration converges at a sub-geometric rate in potential games; our analysis allows for a wider range of learning rates, so we are able to obtain faster convergence rates than Cohen et al. [8]. We are not aware of any other comparable results in the literature.

2 Preliminaries

Finite games. Throughout this work we consider N-players finite games in normal form. Formally, each *player*, indexed by $i \in \mathcal{N} = \{1, \dots, N\}$, has a finite set of *pure strategies* $\alpha_i \in \mathcal{A}_i = \{1, \dots, A_i\}$, and a *payoff function* $u_i \colon \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} \coloneqq \prod_i \mathcal{A}_i$ is the space of all pure strategy profiles. For concision, we will denote such a game as a tuple $\Gamma = \Gamma(\mathcal{N}, \mathcal{A}, u)$.

During play, players can also play *mixed strategies*, i.e., probability distributions $x_i \in \mathcal{X}_i \coloneqq \Delta(\mathcal{A}_i)$ over their pure strategies. In this case, we will write $x_{i\alpha_i}$ for the probability that player $i \in \mathcal{N}$ selects $\alpha_i \in \mathcal{A}_i$ under x_i, $x = (x_1, \dots, x_N)$ for the players' *mixed strategy profile*, and $\mathcal{X} \coloneqq \prod_i \mathcal{X}_i$ for the set thereof. Finally, when focusing on the mixed strategy of a particular player $i \in \mathcal{N}$, we will use the shorthand $(x_i; x_{-i}) \coloneqq (x_1, \dots, x_i, \dots, x_N)$ – and, similarly, $(\alpha_i; \alpha_{-i})$ for pure strategies.

Now, the expected payoff of player i in a mixed strategy profile $x \in \mathcal{X}$ is 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 can be seen as a vector field $v_i \colon \mathcal{X} \to \mathcal{Y}_i = \mathbb{R}^{\mathcal{A}_i}$ which can be written in components as

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

Again, we will write $v(x) = (v_1(x), \dots, v_N(x))$ for the ensemble of payoff vectors of all players and $\mathcal{Y} = \prod_i \mathcal{Y}_i$ for the space of payoff vectors 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 write $v_i(\alpha) = (u_i(\alpha_i; \alpha_{-i}))_{\alpha_i \in \mathcal{A}_i}$. for the corresponding *pure payoff vector*.

Nash equilibrium. The most widely used solution concept in game theory is that of a Nash equilibrium i.e., a (possibly) mixed strategy profile $x^* \in \mathcal{X}$ that discourages unilateral deviations;

formally, $x^* \in \mathcal{X}$ is said to be 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 $\operatorname{supp}(x^*_i) = \{\alpha_i \in \mathcal{A}_i : x^*_{i\alpha_i} > 0\}$. In turn, the size of the support of x^* leads to the following dichotomy: x^* is called *pure* if $\operatorname{supp}(x^*_i) \equiv \prod_{i\in N} \operatorname{supp}(x^*_i)$ is a singleton and *mixed* otherwise.

Finally, we will also say that a Nash equilibrium x^* is *strict* if (NE) holds as a *strict* inequality whenever $x_i \neq x^*_i$; obviously, strict equilibria are also pure, but the converse need not hold. Strict Nash equilibria play a key role in game theory because any unilateral deviation incurs a strict loss to the deviating player; put differently, if x^* is strict, every player has a unique best response. In addition, they are the only equilibria that remain invariant under small generic perturbations of the game [14]; these robustness properties of strict equilibria will play a key role in the sequel.

3 Regularized learning

Throughout our paper, we will focus on a wide family of learning schemes that unfold as follows: At each stage $n = 1, 2, \dots$, every player maintains a "score vector" $Y_{i,n} \in \mathcal{Y}_i$ whose components indicate the player's propensity to play a given pure strategy. More formally, this is captured by a player-specific *"regularized choice" map* $Q_i : \mathcal{Y}_i \to \mathcal{X}_i$ which outputs the player's mixed strategy $X_{i,n} = Q_i(Y_{i,n})$ as a function of $Y_{i,n}$ (see below for a detailed definition). Then, after selecting their actions and collecting their rewards, players also receive – or otherwise construct – an estimate $V_{i,n}$ of their mixed payoff vectors, which is used to increment their score variables and continue playing.

Formally, this learning process, which we call *"follow the generalized leader"* (FTGL), can be described via the round-by-round recursive rule

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

where $\gamma_n > 0$ is a "learning rate" parameter such that $\sum_n \gamma_n = \infty$. The terminology FTGL alludes to the widely known "follow the regularized leader" algorithm, which is, historically speaking, the parent-scheme of FTGL. The link to regularization is provided by the method's choice map, which we detail below; the assumptions for the signal sequence $V_{i,n}$ are provided right after.

3.1. The choice map. The guiding principle behind the definition of the players' choice maps $Q_i : \mathcal{Y}_i \to \mathcal{X}_i$, $i \in \mathcal{N}$, as follows: Because the players' score variables $Y_{i,n}$ are assumed to represent an estimate of each strategy's cumulative payoff over time, Q_i is defined as a "regularized" version of the best-response correspondence $y_i \mapsto \arg\max_{x_i \in \mathcal{X}_i} \langle y_i, x_i \rangle$.[1] On that account, we will consider *regularized best responses* of the general form

$$Q_i(y_i) = \arg\max_{x_i \in \mathcal{X}_i} \{\langle y_i, x_i \rangle - h_i(x_i)\} \tag{3}$$

where $h_i : \mathcal{X}_i \to \mathbb{R}$ denotes the i-th player's *regularization function*.

For concreteness, we will focus on a class of decomposable regularizers of the form $h_i(x_i) = \sum_{\alpha_i \in \mathcal{A}_i} \theta_i(x_i)$ where the so-called "kernel function" $\theta_i : [0,1] \to \mathbb{R}$ is assumed continuous on $[0,1]$, twice differentiable on $(0,1]$, and strongly convex, i.e., $\inf_{(0,1]} \theta_i'' > 0$. Of course, different regularizers give rise to different instances of (FTGL); two of the most widely used and cited examples are as follows:

Example 3.1 (Entropic regularization and multiplicative/exponential weights)**.** Perhaps the most common representative of regularization functions is given by the entropic kernel $\theta(x) = x \log x$ i.e., $h(x_i) = \sum_{\alpha_i \in \mathcal{A}_i} x_{i\alpha_i} \log x_{i\alpha_i}$. This choice of regularizer is well-known to provide the *logit choice map* $\Lambda_i(y_i) = (\exp(y_{i\alpha_i}))_{\alpha_i \in \mathcal{A}_i} \big/ \sum_{\alpha_i \in \mathcal{A}_i} \exp(y_{i\alpha_i})$. The resulting algorithm is known in the literature as the multiplicative/exponential weights algorithm [1, 2, 31, 43, 47].

Example 3.2 (Euclidean projection)**.** Another popular regularizer is the quadratic penalty $h(x) = \sum_\alpha {x_\alpha}^2/2$, which yields the *payoff projection* map $\Pi(y) = \arg\min_{x \in \Delta} \|y - x\|^2$, cf. [28, 48].

[1] In this context, regularization can be seen as a means to reinforce exploration so as to avoid committing prematurely to a given strategy.

Remark 3.1. Examples 3.1 and 3.2 are archetypes of a fundamental dichotomy between regularization functions: in the former case, we have $\theta'(0) = -\infty$, so h becomes *steep* at the boundary of the player's strategy space; in the later case, θ is differentiable at 0, so h is non-steep. We will see that this steep/non-steep dichotomy has a crucial impact on the method's rate of convergence.

3.2. The feedback model. As we mentioned in the beginning of the section, the "payoff signal" V_n contains information about the structure of the algorithm as well as the setting under consideration. Thus to account for as broad a range of algorithms as possible, we will assume that the players' signal sequence is of the general form

$$V_n = v(X_n) + Z_n \tag{4}$$

for some abstract error process $Z_n = (Z_{i,n})_{i\in\mathcal{N}}$. Tp be clear though, we should stress that *we do not assume* that V_n is directly correlated to – or obtained by – the chosen mixed strategy X_n; this will be made clear in the range of models we present below.

To distinguish between random (zero-mean) and systematic (non-zero-mean) errors, we will further decompose Z_n as $Z_n = U_n + b_n$, where

$$b_n = \mathbb{E}[Z_n \,|\, \mathcal{F}_n] \quad \text{and} \quad \mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0 \tag{5}$$

with $\mathcal{F}_n$ denoting the history of X_n up to stage n (inclusive). Notice that, since the feedback signal is generated only *after* the player chooses a strategy, V_n is not $\mathcal{F}_n$-measurable in general. On this account, we will make the following blanket assumptions for the input signal sequence V_n:

1. *Vanishing bias:* b_n converges uniformly to 0 as $n \to \infty$. (A1)
2. *Bounded variance:* $\mathbb{E}[\|U_n\|_*^q \,|\, \mathcal{F}_n] \leq \sigma_n^q$ for some $q > 2$. (A2)

In the above σ_n is assumed to be a deterministic, stage-specific, and possibly increasing bound on the variance of the noise component U_n; our precise assumptions for its growth (relative to b_n or otherwise) will be detailed later in this section.

Specific models. So far, the formulation of (FTGL) has been kept intentionally abstract and devoid of any modeling assumptions for how the players' payoff signals are generated or estimated. To illustrate the width and breadth of (FTGL), we present of series of specific models below, including the popular FTRL and optimistic FTRL methods:

Model 1 (FTRL with oracle-based feedback)**.** Assume that each player chooses an action based on a given mixed strategy, and once every player has chosen an action, an oracle reveals to each player their corresponding pure payoff vector. Formally, at each round $n = 1, 2, \dots$, each player chooses a pure strategy $\alpha_{i,n} \in \mathcal{A}_i$ based on a mixed strategy $X_{i,n} \in \mathcal{X}_i$ and subsequently observes the payoff vector

$$V_{i,n} = v_i(\alpha_n) = (u_i(\alpha_i; \alpha_{-i,n}))_{\alpha_i\in\mathcal{A}_i}. \tag{6}$$

Thus, in this case, (FTGL) boils down to the standard *"follow the regularized leader"* (FTRL) algorithm of [43, 44]. As for our basic feedback assumptions, we readily see that $b_{i,n} = 0$ and $U_{i,n} = v_i(\alpha_n) - v_i(X_n)$; hence:

- (A1) is trivially satisfied since $b_{i,n} = 0$.
- (A2) is again satisfied because $\|U_{i,n}\|_* = \|v_i(\alpha_n) - v_i(X_n)\|_* \leq 2\max_{\alpha\in\mathcal{A}}\|v_i(\alpha)\|_*$, so U_n has uniformly bounded moments for all $q \in [1, \infty]$. §

Model 2 (FTRL with bandit feedback)**.** If the players only observe their realized rewards, they have to *construct* a model for V_n based on incomplete information. This is the standard setting for multi-armed bandits [5, 6, 29], so it is often referred to as the "bandit feedback" model. In this case, the players' action selection process is as in Model 1 above, but the feedback signal sequence V_n is now reconstructed by means of the importance-weighted estimator

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

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 now recalibrated by an *explicit exploration* parameter $\varepsilon_n \to 0$ whose role is to stabilize the learning process. The idea behind this adjustment 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.

Feedback	FTRL	OptFTRL	EG / MP
Full information	$b_n = 0$ $M_n = 0$	$\|b_n\|_* = \mathcal{O}(\gamma_n)$ $M_n = 0$	$\|b_n\|_* = \mathcal{O}(\gamma_n)$ $M_n = 0$
Oracle-based	$b_n = 0$ $M_n = \mathcal{O}(1)$	$\|b_n\|_* = \mathcal{O}(\gamma_n)$ $M_n = \mathcal{O}(1)$	$\|b_n\|_* = \mathcal{O}(\gamma_n)$ $M_n = \mathcal{O}(1)$
Bandit (payoff-based)	$\|b_n\|_* = \mathcal{O}(\varepsilon_n)$ $M_n = \Theta(1/\varepsilon_n)$	$\|b_n\|_* = \mathcal{O}(\varepsilon_n)$ $M_n = \Theta(1/\varepsilon_n)$	$\|b_n\|_* = \mathcal{O}(\varepsilon_n)$ $M_n = \Theta(1/\varepsilon_n)$

Table 1: Recasting different online learning algorithms within the general template of (FTGL).

The importance-weighted estimator (IWE) estimator may be seen as a special case of the model (4) with $U_{i,n} = V_{i,n} - v_i(\hat{X}_n)$ and $b_{i,n} = v_i(\hat{X}_n) - v_i(X_n)$. Both assumptions (A1),(A2) are again satisfied; indeed:

- For (A1): A standard calculation performed in Appendix D reveals that $\|b_{i,n}\|_* = O(\varepsilon_n)$. Thus our assumption is satisfied since $\varepsilon_n \to 0$.
- For (A2): Again a standard calculation presented in Appendix D reveals that $\|V_{i,n} - v_i(\hat{X}_n)\|_* = O(1/\varepsilon_n)$ and thus the noise has bounded moments, $\sigma_n = \Theta(1/\varepsilon_n)$ for all $q \in [1, \infty]$.

Model 3 (OptFTRL with oracle-based feedback)**.** Following Rakhlin and Sridharan [42], the so-called "optimistic" variant of FTRL is given by the recursive formula:

$$\tilde{Y}_{i,n} = Y_{i,n} + \gamma_n V_{i,n-1} \qquad \tilde{X}_{i,n} = Q_i(\tilde{Y}_{i,n}) \qquad Y_{i,n+1} = Y_{i,n} + \gamma_n V_{i,n} \tag{OptFTRL}$$

In the above the payoff signal $V_{i,n}$, which depends on the state $\tilde{X}_n$, is generated as follows: at each round $n = 1, 2, \dots$, every player $i \in \mathcal{N}$ picks an action $\alpha_{i,n} \in \mathcal{A}_i$ based on $\tilde{X}_{i,n} \in \mathcal{X}_i$ and observes the pure payoff vector $v_i(\alpha_n) \equiv (u_i(\alpha_i; \alpha_{-i,n}))_{\alpha_i \in \mathcal{A}_i}$. At first glance, it seems difficult to reconcile the above update structure with that of (FTGL); however, it is indeed possible to integrate (OptFTRL) within (FTGL) by considering the auxiliary states $X_n = Q(Y_n)$ (which are never played and are only used here for the analysis).

Indeed, each player's input signal is a special case of (4) with payoff feedback $V_{i,n} = v_i(\alpha_n)$, zero-mean noise $U_{i,n} = v_i(\alpha_n) - v_i(\tilde{X}_n)$ and bias $b_{i,n} = v_i(\tilde{X}_n) - v_i(X_n)$ that satisfy both the assumptions (A1),(A2). In more detail, we have:

- For (A1): $\|b_{i,n}\|_* = \|v_i(\tilde{X}_n) - v_i(X_n)\|_* = O(\gamma_n)$, which goes uniformly to 0 whenever $\gamma_n \to 0$.
- For (A2): $\|U_{i,n}\|_* = \|v_i(\alpha_n) - v_i(\tilde{X}_n)\|_* \leq 2 \max_{\alpha \in \mathcal{A}} \|v_i(\alpha)\|_*$ and thus the noise has bounded moments for all $q \in [1, \infty]$.

Remark 3.2. Based on the structure of the aforementioned algorithms, it is easy to check that we capture *a-fortiori* the model of a full-information payoff signal; for a more complete account of the different algorithms and feedback models see Table 1.

4 Analysis & Results

We are now in a position to state our main convergence results for (FTGL). We begin with a precise statement and discussion in Section 4.1; subsequently, we present the main proof techniques in Section 4.2.

4.1. Statement and discussion of our main results. Our analysis will focus exclusively on strict Nash equilibria. As we discussed in the introduction, the reason for this is that only strict Nash equilibria can be asymptotically stable under FTRL [11, 15], so they are the only reasonable candidates to consider when examining the rate of convergence of a regularized learning algorithm.[2]

[2] As a sidenote, we should remark here that FTGL also contains the optimistic FTRL algorithm, which *does* converge to mixed Nash equilibria in bilinear zero-sum games with *perfect, deterministic* feedback [16, 27, 34]. At first glance, this would seem to contradict the results of [11, 15], but one needs to bear in mind that the convergence of (OptFTRL) to mixed equilibria only occurs in settings with *perfect information* (i.e., $V_n = v(X_n)$ for all $n = 1, 2, \dots$). In the presence of uncertainty, this convergence is destroyed [7, 23], so there is no contradiction with the results of [15]. Because we are primarily interested in learning with limited information and/or under uncertainty, we will not treat this somewhat fragile case.

To proceed, we will need one technical assumption linking the learning rate of (FTGL) and the bias/variance parameters of the driving feedback sequence V_n. This is as follows:

$$\text{The sequence } \delta_n := \frac{\sum_{k=1}^{n} \gamma_k^{1+\frac{q}{2}} \sigma_k^q}{\left[\sum_{k=1}^{n} \gamma_k\right]^{1+\beta q/2}} \text{ is summable for some } \beta < 1. \tag{A3}$$

Assumption (A3) imposes a growth condition on the method's learning rate relative to the bias and variance parameters of the input signal sequence V_n, but it is otherwise a technical prerequisite for the analysis to come. What is more important for our purposes is that the concrete models that we discussed in the previous section satisfy it for a wide range of the player-chosen parameters γ_n (and ε_n in the case of bandit-based algorithms); to streamline our presentation, we postpone a more precise discussion of this issue until after the statement of our main results.

The last element that we need to introduce concerns the players' choice of regularizer: clearly, since propensities are transformed to strategies via each player's individual choice map $Q_i : \mathcal{Y}_i \to \mathcal{X}_i$, it stands to reason that the underlying regularization function h plays a major role in the method's rate of convergence. Indeed, given an update of the form $Y_{n+1} \leftarrow Y_n + \gamma_n V_n$, the method's strategy variable will move forward as $X_{n+1} \leftarrow X_n + \gamma_n JQ(V)V_n + \mathcal{O}(\gamma_n^2)$, where JQ denotes the Jacobian matrix of Q. Thus, to leading order in γ_n, the update $X_{n+1} \leftarrow X_n$ is dominated by the first derivatives of Q.

By a relatively straightforward application of the Legendre identity from convex analysis ($Q = (\partial h)^{-1}$ in our context; see below for a precise statement), this rate of change is related to the inverse mapping of the derivative each θ_i (the kernel of the underlying regularizer). Motivated by this observation, we introduce below the algorithm's so-called *rate function*:

$$\phi_i(t) = \begin{cases} (\theta_i')^{-1}(t) & \text{if } t > \theta_i'(0^+), \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The definition of the rate function ϕ captures the sensitivity of the choice map Q in a very precise way: If the score difference corresponding to two pure strategies $\alpha, \beta \in \mathcal{A}_i$ grows as $y_\beta - y_\alpha = t$ for some $t > 0$, then the probability of playing the strategy with the lesser score must be less than the probabiity of playing the strategy with the higher score. The precise amount of this disparity of course depends on the player's choice function Q and ϕ acts as a "transfer" function in this regard. Specifically, as we show in detail later, we have $x_\alpha = \phi(-\Theta(t))$, i.e., ϕ determines the rate at which x_α vanishes. For different regularizers we present the corresponding rates in Table 2.

With all this in hand, our main result can be stated as follows:

Theorem 1. *Let x^* be a strict Nash equilibrium of Γ, and fix some confidence level $\delta > 0$. If Assumptions (A1)–(A3) hold, there exists an unbounded open set of initial conditions $\mathcal{W}_{\text{init}} \subseteq \mathcal{Y}$ and constants d_i, c_i' with $c_i' > 0$ such that, if $Y_1 \in \mathcal{W}_{\text{init}}$, we have:*

1. *X_n converges to x^* with probability at least $1 - \delta$.*
2. *Conditioned on the above, the rate of convergence for each player $i \in \mathcal{N}$ is given by*

$$\|X_{i,n} - x_i^*\|_1 \leq 2 \sum_{\alpha_i \in \mathcal{A}_i \setminus \operatorname{supp}(x_i^*)} \phi_i\Big(d_i - c_i' \sum_{k=1}^{n} \gamma_k\Big). \tag{8}$$

Armed with this general result, we readily obtain the following immediate consequences thereof:

Corollary 1. *If the regularizer employed is non-steep (i.e., θ_i is differentiable at 0), X_n converges to x^* in a finite number of iterations.*

Corollary 2. *Suppose that FTRL is run with oracle-based feedback as per Model 1 and a learning rate of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$. Then the conclusion of Theorem 1 holds.*

Corollary 3. *Suppose that FTRL is run with bandit feedback as per Model 2, a learning rate of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$ and a mixing parameter $\varepsilon_n \propto 1/n^r$, $r \in (0, 1/2)$. Then the conclusion of Theorem 1 holds.*

Remark 4.1. We stress out here that for all the bandit-feedback derived results, the convergence rates refer to $X_{i,n}$ instead of the explicit exporation term $\hat{X}_{i,n}$ whose rate is always dominated by the mixing parameter ε_n.

Corollary 4. *Suppose that Optimistic FTRL is run with oracle-based feedback as per Model 3 and a learning rate of the form $\gamma_n \propto 1/n^p$, $p \in (0, 1]$. Then the conclusion of Theorem 1 holds.*

ALGORITHM	KERNEL $\theta(\mathbf{x})$	RATE $\phi(-\mathbf{y})$
Multiplicative Weight Updates	$x \log x$	$\exp(-y)$
Projection Gradient Descent	$x^2/2$	$-y$
Inverse Updates	$-\log x$	$-1/y$
q-Replicator$_{q>0}$	$[q(1-q)]^{-1}(x - x^q)$	$[q^{-1} + (1-q)y]^{1/q-1}$

Table 2: Regularizers & correspinding rates.

More generally, we show in the supplement that the conclusion of Theorem 1 holds for all algorithms and feedback models presented in Table 1: in all cases therein, players can employ step-size policies of the form $\gamma_n \propto 1/n^p$, $p \in [0,1]$, and a mixing parameter $\varepsilon_n \propto 1/n^r$ with $r \in (0, 1/2)$ for the bandit models. The only case that does not follow as an immediate corollary of Theorem 1 is the case of constant step-sizes for Optimistic FTRL and EG/MP; however, a slightly more refined argument (that we present in the Appendix C) shows that constant step-sizes are also covered by the convergence rate guarantee (8) of Theorem 1.

4.2. Sketch of proof and main techniques. At a high level, the main idea of the proof of Theorem 1 relies on a tandem application of martingale limit theory and convex analysis in order to exploit the specific structure of (FTGL). While martingale limit theory emerges naturally to control the components of the noise, a delicate analysis of the contribution of h_i in the solution of the convex constrained optimization problem, that $x = Q_i(y)$ induces, is necessary to derive the aforementioned generic rates. Below we provide a sketch of the main steps in this analysis

Step 1. Our starting point is to explore the geometric properties that are induced by the existence of a strict Nash equilibrium. Indeed, the fact that (NE) holds as a strict inequality for each pure strategy against the equilibrium's strategy, ensures convergence properties for strict Nash equilibria. More precisely, an immediate implication of (NE) is that there exist neighborhood $\mathcal{U}$ of x^* and constants $c_1, \dots, c_N$ such that

$$v_{i\alpha_i^*}(x) - v_{i\alpha_i}(x) \geq c_i \text{ for all } x \in \mathcal{U} \text{ and } \alpha_i \neq \alpha_i^*, \alpha_i \in \mathcal{A}_i, i \in \mathcal{N} \tag{9}$$

In other words, in the neighborhood $\mathcal{U}$ the payoff of the equilibrium's strategy strictly dominates all other strategies' payoffs for each player. However, since the linchpin of (FTGL) is the interplay between $\mathcal{X}$ and $\mathcal{Y}$, for the purpose of our analysis, we need to investigate the variational structure of $\mathcal{U}$ in both spaces.

Informal Lemma 1. *There exists a neighborhood $\mathcal{U}$, constants $c_1, \dots, c_N$ and $M_1, \dots, M_N$ for which (9) holds such that $\prod_{i\in\mathcal{N}} Q_i(\mathcal{W}_{M_i}) \subseteq \mathcal{U}$, where $\mathcal{W}_{M_i}$ are open sets of the form* [3]

$$\mathcal{W}_{M_i} = \{Y_i : Y_{i\alpha_i^*} - Y_{i\alpha_i} > M_i \text{ for all } \alpha_i \neq \alpha_i^*, \alpha_i \in \mathcal{A}_i\} \text{ for } M_i > 0, i \in \mathcal{N} \tag{10}$$

Step 2. We now focus on one player $i \in \mathcal{N}$ and drop the index i altogether. First we prove that there exists an open set of initializations $\mathcal{W}_{\text{init}}$ of (FTGL), for which with arbitrary high probability the dual variable $(Y_k)_{k\in\mathbb{N}}$ never exits $\mathcal{W}_M$ and thus its image remains in the desired neighborhood $\mathcal{U}$. We start by writing the score differences between each pure strategy $\alpha \in \mathcal{A}$ and $\alpha^* \in \operatorname{supp}(x^*)$

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} = Y_{\alpha,1} - Y_{\alpha^*,1} + \sum_{k=1}^{n} \gamma_k(\text{drift}_k + \text{noise}_k + \text{bias}_k) \tag{11}$$

where $\text{drift}_k = v_\alpha(X_k) - v_{\alpha^*}(X_k)$, $\text{noise}_k = U_{\alpha,k} - U_{\alpha^*,k}$, $\text{bias}_k = b_{\alpha,k} - b_{\alpha^*,k}$. We will prove by induction our forward-invariant statement; let $Y_k \in \mathcal{W}_M$ and thus $X_k \in \mathcal{U}$ for all $k = 1, \dots, n$ then

- By (9) we have $\sum_{k=1}^n \gamma_k \text{drift}_k \leq -c \sum_{k=1}^n \gamma_k$ for all $k = 1, \dots, n$.
- By the triangle inequality and (A1), the term $\sum_{k=1}^n \gamma_k \text{bias}_k$ is dominated by the term $\sum_{k=1}^n \gamma_k \text{drift}_k$ for all $n = 1, 2, \dots$.
- Subsequently, by leveraging the machinery of martingale's maximal inequalities and assumption (A2), which we present in Appendix A and using learning rates that respect (A3), we prove that with probability at least $1-\delta$, for any fixed confidence level δ, $\sum_{k=1}^n \gamma_k \text{noise}_k$, which is a martingale, is also dominated by the term $\sum_{k=1}^n \gamma_k \text{drift}_k$ for all $n = 1, 2, \dots$
- We now define the open set of initial conditions $\mathcal{W}_{\text{init}}$, which is of the form described in (10), with constant M_{init}. By choosing[4] $M_{\text{init}} \geq M + \sum_{k=1}^n \gamma_k(\text{noise}_k + \text{bias}_k) - (c - c') \sum_{k=1}^n \gamma_k$, for any $c' < c$ and any $n \geq 1$, since $Y_1 \in \mathcal{W}_{\text{init}}$ we have that $Y_{\alpha,n+1} - Y_{\alpha^*,n+1} \leq -M$ for all $n \geq 1$.

[3] It is worth mentioning that the images of these open sets belong to neighborhoods of x^*, which are nested as M_i increases.
[4] such a M_{init} exists since both the bias and the noise terms are dominated by the term $-(c - c') \sum_{k=1}^n \gamma_k$.

By substituting the inequality for M_{init} in (11) we get $Y_{\alpha,n+1} - Y_{\alpha^*,n+1} \leq -M - c' \sum_{k=1}^{n} \gamma_k$ and convergence occurs as an immediate consequence; Indeed $X_{\alpha^*,n} \to 1$, since whenever $Y_\alpha - Y_{\alpha^*} \to -\infty$, it holds that each $\alpha \in \mathcal{A} \setminus \text{supp}(x^*)$ becomes extinct i.e., $X_\alpha \to 0$.

Step 3. We now proceed to the delineation of the rates of convergence. Using the KKT conditions (Lemma B.1) combined with Eq. (11),Eq. (9) and the fact that $Y_1 \in \mathcal{W}_{\text{init}}$ we have

$$\theta'(X_{\alpha,n+1}) - \theta'(X_{\alpha^*,n+1}) = Y_{\alpha,n+1} - Y_{\alpha^*,n+1} \leq -M_{\text{init}} - c \sum_{k=1}^{n} \gamma_k + \sum_{k=1}^{n} \gamma_k (\text{noise}_k + \text{bias}_k)$$

Recall that θ is strong convex, or equivalently θ' is strictly increasing; by rearranging and substituting to the above inequality we get

$$\theta'(X_{\alpha,n+1}) \leq \theta'(X_{\alpha^*,n+1}) - M - c' \sum_{k=1}^{n} \gamma_k \leq d - c' \sum_{k=1}^{n} \gamma_k \tag{12}$$

where $d = -M + \theta'(1)$ and $\alpha \in \mathcal{A}$, $\alpha \neq \alpha^*$. By aggregating over all $\alpha \in \mathcal{A}, \alpha \neq \alpha^*$

$$\|x^* - X_{n+1}\|_1 = 2(1 - X_{\alpha^*,n+1}) \leq 2 \sum_{\alpha \in \mathcal{A} \neq \alpha^*} \phi(d - c' \sum_{k=1}^{n} \gamma_k) \tag{13}$$

which indicates the rate of convergence and completes our proof.

Remark 4.2. *The bounds we provide are indeed sharp. To see this, consider a single-player game with two actions "0" and "1", and payoffs $u(0) = 0$, $u(1) = 1$. Then, if e.g., FTRL is run with "full information" feedback, the probability that the player plays "1" at time t is exactly equal to*

$$X_t = 1 - \phi(c - \sum_{s=1}^{t} \gamma_s u(1)) = 1 - \phi(c - \sum_{s=1}^{t} \gamma_s)$$

where ϕ is the rate function of Eq. (7) and c is an initialization constant. This simple derivation shows that MWU converges to the game's (strict) equilibrium at a rate of exactly $\exp(-\Theta(\sum_{s=1}^{t} \gamma_s))$, whereas Euclidean methods achieve an equilibrium after a finite number of iterations – in particular, as soon as $\sum_{s=1}^{t} \gamma_s$ exceeds c. It thus follows that the rates provided by Theorem 1 are, in general, unimprovable.

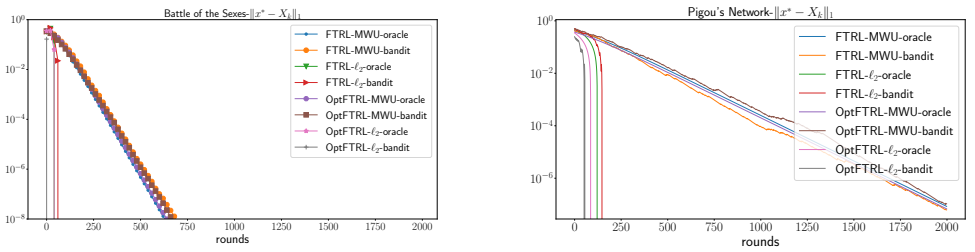
5 Numerical experiments

In this section we perform a series of numerical experiments to validate our theoretical findings. Specifically we are interested in verifying both the correctness in the computation of the rates via ϕ_i for different regularizers and at the same time the fact that convergence speed is invariant to different feedback models and algorithmic variants of (FTGL).

To do this, we start by examining variations of (FTGL) in the archetypal game of *Battle of the Sexes*, a popular two-player benchmark of the coordination games, which however involves elements of conflict as well. This game exhibits two strict Nash equilibria and one mixed equilibrium (for the exact definition, see Appendix E). We then seek to experimentally study the performance of (FTGL) while the number of the players scales up. To do this we use the atomic version of classical *Pigou's congestion game* [39], where N units of traffic (e.g., rush-hour drivers) leave from O (origin) to D (destination) at the same time and each driver has the same dominant pure strategy/path for this trip. Accordingly to Table 2 the decay rate for the entropic regularization is exponential while for the case of euclidean is linear, which indeed yield linear and constant-time convergence as Fig. 1 illustrates.

We defer a detailed exposition of various configurations with different step-sizes, alternative discretization methods like MirrorProx and ExtraGradient and feedback models with the presence (or not) of extra heavy-tailed/uniform/gaussian noise again to the paper's supplement.

It is worth mentioning that the sharpness of the provided rates of Theorem 1 can clearly be observed in the list of the extensive numerical experiments we present in Fig. 1 and Appendix E. In particular, the faster convergence rate of Euclidean algorithms is somewhat surprising since a regret-based viewpoint would suggest the use of entropic regularization (which, ceteris paribus, has much better regret guarantees) as optimal in this regard. Interestingly, however our analysis shows that a Euclidean regularizer is much more suitable for achieving convergence to equilibrium in a game-theoretic setting. It is for this reason that we insisted on the comparison between entropic and Euclidean regularization in the simulations. (*The Pigou network example of Fig. 1b is especially telling in this regard.*)



(a) Battle of the Sexes

(b) Pigou Network

Figure 1: For the *Battle of the Sexes* experiment, we initialize uniformly randomly our executions from $Y_{init} \in [-1,1] \times [-1,1]$ and examine the instantiations of Model 1-3 using constant-step size and exploration rate $\varepsilon_n \propto 1/\sqrt[3]{n}$. For the *Pigou's* game, our setup includes two alternative disjoint paths for $N = 1000$ drivers. The first path has linear latency $\ell_1(x) = x/N$ while the second one has constant unit congestion, $\ell_2(x) = 1$, where x denotes the population of the drivers that uses the corresponding path.

6 Concluding remarks

A key take-away from this study is that the questions of regret minimization and convergence to Nash equilibrium are fundamentally different. In particular, much of the conventional wisdom that has been accrued for regret-minimization strategies (such as which regularizer to use, with what learning rate, etc.) ceases to apply when the figure of merit is convergence to an equilibrium. Because the only states that are stable under leader-following policies are the game's strict Nash equilibria, the agents can be significantly more firm and confident in their choices, without compromising their final limit state; as a result, this extra degree of "confidence" allows for rates of convergence that are well beyond the operational envelope of regret minimization problems. We believe that this polar shift in perspective constitutes an important - and under-explored - issue in game-theoretic learning, and charting out its ramifications for multi-agent learning is a particularly fruitful direction for future research.

Acknowledgments and Disclosure of Funding

This research was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET) and the Onassis Foundation under Scholarship ID: F ZN 010-1/2017-2018. P. Mertikopoulos is also grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grant ALIAS (ANR-19-CE48-0018-01). E.V. Vlatakis-Gkaragkounis is grateful to be supported by NSF grants CCF-1703925, CCF1763970, CCF-1814873, CCF-1563155, and by the Simons Collaboration on Algorithms and Geometry.

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] Avrim Blum and Yishay Mansour. Learning, regret minimization, and equilibria. In Noam Nisan, Tim Roughgarden, Éva Tardos, and V. V. Vazirani, editors, *Algorithmic Game Theory*, chapter 4. Cambridge University Press, 2007.

[4] Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: on convergence to Nash equilibria of regret-minimizing in routing games. In *PODC '06: Proceedings of the 25th annual ACM SIGACT-SIGOPS symposium on Principles of Distributed Computing*, pages 45–52, 2006.

[5] 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.

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

[7] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[8] 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.

[9] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS '19: Proceedings of the 10th Conference on Innovations in Theoretical Computer Science*, 2019.

[10] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[11] 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.

[12] Dean Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21 (1):40–55, October 1997.

[13] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

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

[15] Angeliki Giannou, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT '21: Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

[16] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[17] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

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

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

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

[22] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019.

[23] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[24] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *COLT '21: Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

[25] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[26] David Kelsey and Sara Le Roux. An experimental study on the effect of ambiguity in a coordination game. *Theory and Decision*, 79(4):667–688, 2015.

[27] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12: 747–756, 1976.

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

[29] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

[30] Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael I. Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *ICML '20: Proceedings of the 37th International Conference on Machine Learning*, 2020.

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

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

[33] 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.

[34] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[35] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. https://arxiv.org/pdf/1906.01115.pdf, 2019.

[36] John F. Nash. *Non-cooperative games*. PhD thesis, Princeton University, 1950.

[37] Arkadi Semen Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15 (1):229–251, 2004.

[38] Noam Nisan, Tim Roughgarden, Éva Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[39] Arthur Cecil Pigou. *The Economics of Welfare*. Macmillan, London, UK, 1920.

[40] Leonid Denisovich Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[41] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

[42] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013.

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

[44] 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.

[45] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[46] 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.

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

[48] 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.

A Martingale limit theory

Our analysis leverages tools from martingale limit theory. Below we present the two main theorems that we utilize in the main body of our proofs.

- (Doob's inequality), also known as *Kolmogorov's submartingale inequality* gives a bound on the probability that a stochastic process exceeds any given value over a given interval of time.
- (Burkholder's inequality), also known the *Burkholder-Davis-Gundy inequality* is a remarkable result relating the maximum of a local martingale with its quadratic variation.

Theorem A.1 (Doob's inequality)**.** *Let S_n be a martingale with respect to the filtration $\mathcal{F}_n$, then for each $\varepsilon > 0$ and $q \geq 1$,*

$$\mathbb{P}(\sup_{1 \leq k \leq n} |S_k| \geq \varepsilon) \leq \frac{\mathbb{E}\,|S_n|^q}{\varepsilon^q} \qquad \text{(Doob's inequality)}$$

Theorem A.2 (Burkholder's inequality)**.** *Let S_n be a martingale with respect to the filtration $\mathcal{F}_n$ and $X_n = S_n - S_{n-1}$. Then for all $1 < q < \infty$, there exists constant C_q depending only on q such that*

$$\mathbb{E}\,|S_n|^q \leq C_q\, \mathbb{E}\left|\sum_{k=1}^{n} X_k^2\right|^{q/2} \qquad \text{(Burkholder's inequality)}$$

Proofs for these two theorems can be found in [19].

B A dichotomy between the regularizers

Our main result (Theorem 1) provides a mechanism to compute the convergence rate to a strict Nash Equilibrium universally for all smooth convex regularizers $h_i(x) = \sum_{\alpha_i \in \mathcal{A}_i} \theta_i(x_{\alpha_i})$. An important implication of our main theorem (Corollary 1) is that for the case of non-steep kernels (i.e., θ_i is differentiable at 0), X_n converges to x^* in a finite number of iterations. Below we give some intuition for the interested reader about the differences between the *steep* and *non-steep* case.

Steep vs non-steep. In this section we elaborate in detail the dichotomy among the different regularizers mentioned in Sections 3.1 and 4. As we established in Section 3.1, different players may apply different regularizers h_i in their choice maps $Q_i(y_i)$. Depending on the regularizer chosen, the behavior of (FTGL) could vary significantly. To investigate more this diversity, we start by describing formally the strategy-choice step $x_i = Q_i(y_i)$ as a convex constrainted minimization problem.

$$Q_i(y_i) = -\arg\min_{x_i \in \mathcal{X}_i} \{h_i(x_i) - \langle x_i, y_i \rangle\}\,. \tag{B.1}$$

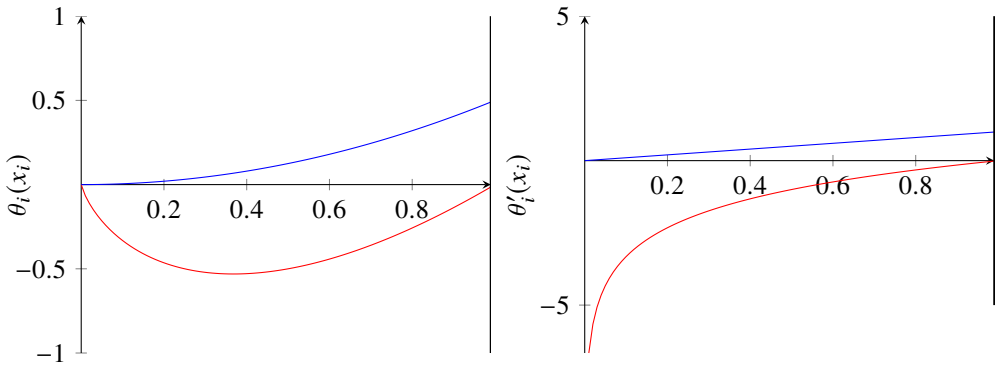
Following also the folklore convention from convex analysis, we express h as an extended-real valued function $h : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ with value ∞ outside of the simplex $\mathcal{X}$. Additionally, the subdifferential of h at $x \in \mathcal{V}$ is defined as:

$$\partial h(x) = \{y \in \mathcal{V}^* : h(x') \geq h(x) + \langle y, x' - x \rangle \ \forall x' \in \mathcal{V}\} \tag{B.2}$$

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 more compactly $\mathrm{ri}(\mathcal{X}) \subseteq \mathrm{dom}\,\partial h \equiv \{x \in \mathcal{X} : \partial h(x) \neq \emptyset\} \subseteq \mathrm{dom}\, h \subseteq \mathcal{X}$. Notice that when the gradient of h exists, then its subgradient always contains it. Leveraging the property that local and global minima coincides in the case of convex function, Fermat's rule provides a simple characterization of the minimizers of a function as the zeros of its subdifferential:

Fact (Fermat's Rule)**.** *For a proper convex function f,* $\mathrm{argmin} f \equiv \mathrm{zer}\partial f = \{x \in \mathcal{X} \mid 0 \in \partial f(x)\}$

With these in mind, we present a typical separation between the different regularizers,, focusing on the more simple case of decomposable ones $h(x) = \sum_{\alpha \in \mathcal{A}} \theta_\alpha(x)$. On the one hand, *steep* regularizers have differentiable kernels on $(0, 1]$ and become infinitely steep as x approaches the boundary or $\theta'(0) = -\infty$. On the other hand, for the *non-steep* case the kernel is differentiable in all of $[0, 1]$. As a result of Fermat's Rule, when a steep regularizer is employed the points of the boundary are



$$\begin{bmatrix} \text{steep} & h_1(x) = \sum_{\alpha\in\mathcal{A}} \theta_i(x_\alpha) = \sum_{\alpha\in\mathcal{A}} x_\alpha \log(x_\alpha) \\ \text{non-steep} & h_2(x) = \sum_{\alpha\in\mathcal{A}} \theta_i(x_\alpha) = \sum_{\alpha\in\mathcal{A}} \frac{1}{2} x_\alpha^2 \end{bmatrix}$$

Figure 2: Steep vs. non-steep regularizers (note in particular the singular behavior of the gradient at the boundary in the case of steep regularizers).

infeasible not only as initial conditions but also as part of the sequence of play, while non-steep ones allow completely the sequence of play to transfer between the different faces of the simplex. The qualititative difference in behavior between these cases is illustrated in Fig. 2 (which shows the very different behavior of the derivates of h near the boundary of the state space).

Having discussed the connection between the choice map and the properties of the regularizer, the following lemma quantifies the gulf between the steep and non-steep case and provides the relation between mixed strategies and score vectors and the mirror map (3) that defines the dynamics (FTGL). More precisely, we focus on the perspective of an arbitrary player, say i, and for ease of notation we write Q, x and y instead of Q_i, x_i and y_i respectively.

Lemma B.1. *$x = Q(y)$ if and only if there exist $\mu \in \mathbb{R}$ and $\nu_\alpha \in \mathbb{R}_+$ such that, for all $\alpha \in \mathcal{A}$, we have: a) $y_\alpha = \frac{\partial h}{\partial x_\alpha} + \mu - \nu_\alpha$; and b) $x_\alpha \nu_\alpha = 0$ In particular, if h is steep, we have $\nu_\alpha = 0$ for all $\alpha \in \mathcal{A}$.*

Proof. Recall that

$$\begin{aligned} Q(y) &= \arg\max_{x\in\mathcal{K}} \{\langle y|x\rangle - h(x)\} \\ &= \arg\max \left\{ \sum_{\alpha\in\mathcal{A}} y_\alpha x_\alpha - h(x) : \sum_{\alpha\in\mathcal{A}} x_\alpha = 1 \text{ and } \forall \alpha \in \mathcal{A} : x_\alpha \geq 0 \right\} \end{aligned}$$

The result follows by applying the Karash-Kuhn Tucker (KKT) conditions to this optimization problem and noting that, since the constraints are affine, the KKT conditions are sufficient for optimality. Our Langragian is

$$\mathcal{L}(x, \mu, \nu) = (\sum_{\alpha\in\mathcal{A}} y_\alpha x_\alpha - h(x)) - \mu(\sum_{\alpha\in\mathcal{A}} x_\alpha - 1) + \sum_{\alpha\in\mathcal{A}} \nu_\alpha x_\alpha$$

where the set of constraints (i) of the statement of the lemma are the stationarity constraints, which in our case are $\nabla\mathcal{L}(x, \mu, \nu) = 0 \Leftrightarrow \nabla(\sum_{\alpha\in\mathcal{A}} y_\alpha x_\alpha - h(x)) = \mu\nabla(\sum_{\alpha\in\mathcal{A}} x_\alpha - 1) - \sum_{\alpha\in\mathcal{A}} \nu_\alpha \nabla x_\alpha$, while the set of constraints (ii) of the statement of the lemmas are the complementary slackness constraints. Note that complementary slackness implies that whenever $\nu_\alpha > 0$ whenever $\alpha \notin \text{supp}(x)$. Finally, if h is steep, we have $|\partial_\alpha h(x)| \to \infty$ as $x \to \text{bd}(\mathcal{X})$, which implies that the KKT conditions admit a solution with $\nu_\alpha = 0$. ∎

C Proof of Main Theorem

Our first lemma shows a property of strict Nash equilibria. More precisely, we prove the existence of a neighborhood $\mathcal{U}$ in which each player's payoff corresponding to the strategy of the equilibrium outweighs the payoff of any other pure strategy.

Lemma C.1. *Let $x^* = (\alpha_1^*, \dots, \alpha_N^*) \in \mathcal{A}$ be a strict Nash equilibrium. Then there exists a neighborhood $\mathcal{U}$ of x^* and constants c_i such that for each player $i \in \mathcal{N}$:*

$$v_{i\alpha_i^*}(x) - v_{i\alpha_i}(x) \geq c_i \text{ for all } x \in \mathcal{U} \text{ and } \alpha_i \neq \alpha_i^*,\ \alpha_i \in \mathcal{A}_i. \tag{C.1}$$

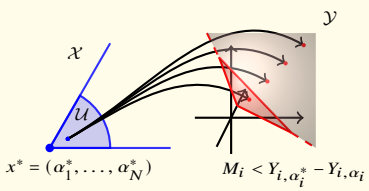
Proof. Our claim is a consequence of the definition of strict Nash equilibria. Specifically, from (NE) for each player $i \in \mathcal{N}$ we have that

$$v_{i\alpha_i^*}(x^*) > v_{i\alpha_i}(x^*) \text{ for all } \alpha_i \in \mathcal{A}_i, \alpha_i \neq \alpha_i^* \tag{C.2}$$

By continuity there exists a neighborhood $\mathcal{U} \subseteq \mathcal{X}$ and $c_i > 0$ for each player $i \in \mathcal{N}$ such that

$$v_{i\alpha_i^*}(x) - v_{i\alpha_i}(x) \geq c_i \text{ for all } x \in \mathcal{U} \tag{C.3}$$

■



The following lemma plays a central role in the proof of our main theorem (Theorem 1). More precisely, Lemma C.2 provides a detailed analysis of the topology of a neighborhood $\mathcal{U}$ where variational inequality (C.1) holds both in primal space $\mathcal{X}$ and dual space $\mathcal{Y}$. In order to achieve that we introduce the notion of "(α_i^*, M_i)-score-dominant" open set for a player $i \in \mathcal{N}$, which we denote $\mathcal{W}_i(M_i)$.

Definition (*Score-Dominant Collection*)**.** Let $x^* = (\alpha_1^*, \dots, \alpha_N^*) \in \mathcal{A}$ be a strict Nash equilibrium of a finite game Γ. Then a collection is said to be "$(\alpha_i^*, M_i)_{i\in\mathcal{N}}$-score-dominant" if there exist positive constants $M_i > 0$ corresponding open sets $\mathcal{W}_i(M_i)$ of the form

$$\mathcal{W}_i(M_i) = \{Y_i : Y_{i\alpha_i^*} - Y_{i\alpha_i} > M_i \text{ for all } \alpha_i \neq \alpha_i^*, \alpha_i \in \mathcal{A}_i\} \text{ for each player } i \in \mathcal{N} \tag{C.4}$$

Lemma C.2. *Let $x^* = (\alpha_1^*, \dots, \alpha_N^*) \in \mathcal{A}$ be a strict Nash equilibrium. Then for every $\varepsilon \in (0, 1)$, there exist constants $M_{i,\varepsilon}$ and the corresponding score-dominant open sets for each player $i \in \mathcal{N}$ such that: $\prod_{i\in\mathcal{N}} Q_i(\mathcal{W}_i(M_{i,\varepsilon})) \subseteq \mathcal{U}_\varepsilon$, where $\mathcal{U}_\varepsilon = \{x \in \mathcal{X} : x_{i\alpha_i^*} > 1 - \varepsilon$ for every player $i \in \mathcal{N}\}$*

Proof. For an arbitrary player $i \in \mathcal{N}$ let $\mathcal{W}_i(M_{i,\varepsilon})$ be a *score-dominant* open set. We will show that any $M_{i,\varepsilon} > \theta_i'(1) - \theta_i'(\frac{\varepsilon}{|\mathcal{A}_i|}) > 0$ satisfies the desideratum. Indeed, again by using Lemma B.1 for a $Y_i \in \mathcal{W}_i(M_{i,\varepsilon})$ with $x_i = Q_i(Y_i)$ we have that

$$Y_{i\alpha_i^*} - Y_{i\alpha_i} > M_{i,\varepsilon} \tag{C.5}$$

$$\theta'(x_{i\alpha_i^*}) - \theta_i'(x_{i\alpha_i}) - (v_{\alpha_i^*} - v_{\alpha_i}) > M_{i,\varepsilon}. \tag{C.6}$$

with $v_{\alpha_i} \geq 0$ and $x_{i\alpha_i} = 0$ whenever $x_{i\alpha_i} > 0$. Notice that since $M_{i,\varepsilon} > 0$ and θ_i' is strictly increasing, it holds that $x_{i\alpha_i} < x_{i\alpha_i^*}$. Indeed, assume by contradiction that $x_{i\alpha_i} \geq x_{i\alpha_i^*}$ for some α_i, then we examine two different cases:

(*i*) If $x_{i\alpha_i^*} = 0$, then $x_{i\alpha_i} \geq x_{i\alpha_i^*}$ for all $\alpha_i \in \mathcal{A}_i$ with $x_{i\alpha_i} > 0$ for at least one $\alpha_i \in \mathcal{A}_i, \alpha_i \neq \alpha_i^*$ which is a contradiction to (C.6).

(*ii*) if $x_{i\alpha_i^*} > 0$, then (C.6) implies that $M_{i,\varepsilon} \leq \theta'(x_{i\alpha_i^*}) - \theta_i'(x_{i\alpha_i}) < 0$ which is again a contradiction.

Therefore $v_{\alpha_i^*} = 0$ and (C.6) can be rewritten for all $\alpha_i \neq \alpha_i^*$ with $x_{i\alpha_i} > 0$ as

$$\theta_i'(x_{i\alpha_i}) < -M_{i,\varepsilon} + \theta'(x_{i\alpha_i^*}) < -M_{i,\varepsilon} + \theta'(1) < \theta_i'(\tfrac{\varepsilon}{|\mathcal{A}_i|}) \tag{C.7}$$

where last inequality holds by the choice of $M_{i,\varepsilon} > \theta_i'(1) - \theta_i'(\frac{\varepsilon}{|\mathcal{A}_i|}) > 0$. Again, since θ' is strictly increasing, this implies that for all $\alpha_i \neq \alpha_i^*$ either $x_{i\alpha_i} = 0$ or $0 < x_{i\alpha_i} < \frac{\varepsilon}{|\mathcal{A}_i|}$. By union bound, this implies that $x_{i\alpha_i^*} > 1 - \varepsilon$ and equivalently that $x \in \mathcal{U}_\varepsilon$. ■

Remark C.1. It is easy to check that as M_i' increases the score-dominant sets and their corresponding images are nested. Indeed if $M' \geq M_\varepsilon \Rightarrow \mathcal{W}(M) \subseteq \mathcal{W}(M') \Rightarrow Q(\mathcal{W}(M)) \subseteq Q(\mathcal{W}(M'))$, since $Y_{i\alpha_i^*} - Y_{i\alpha_i} > M > M_\varepsilon$ for all $\alpha_i \neq \alpha_i^*, \alpha_i \in \mathcal{A}_i$.

Remark C.2. Notice that since the above analysis is for each strategy $\alpha_i \in \mathcal{A}_i$ of player i, it implies that not only the images $Q_i(\mathcal{W}_{M_i})$ are nested, but also that if $x_i = Q_i(Y_i), Y_i \in \mathcal{W}_{M_i}$ all $x_{i\alpha_i} \to 0$ for $\alpha_i \neq \alpha_i^*$ as $M_i \to \infty$.

Theorem 1. *Let x^* be a strict Nash equilibrium of Γ, and fix some confidence level $\delta > 0$. If Assumptions (A1)–(A3) hold, there exists an unbounded open set of initial conditions $\mathcal{W}_{\text{init}} \subseteq \mathcal{Y}$ and constants d_i, c_i' with $c_i' > 0$ such that, if $Y_1 \in \mathcal{W}_{\text{init}}$, we have:*

1. *X_n converges to x^* with probability at least $1 - \delta$.*
2. *Conditioned on the above, the rate of convergence for each player $i \in \mathcal{N}$ is given by*

$$\|X_{i,n} - x_i^*\|_1 \leq 2 \sum\nolimits_{\alpha_i \in \mathcal{A}_i \setminus \text{supp}(x_i^*)} \phi_i\Big(d_i - c_i' \sum\nolimits_{k=1}^n \gamma_k\Big). \tag{8}$$

Remark C.3. The probability guarantee is over only the potential randomness that the payoff oracle. i.e., when players have access to a perfect payoff oracle; the results hold with probability 1.

Proof. Fix a confidence level δ and the parameters of the algorithm respecting (A1)–(A3). We will prove that there exists a "score-dominant" open set of initial conditions $\mathcal{W}_{\text{init}}$

$$\mathcal{W}_{\text{init}} \equiv \{Y : M_{\text{init}} < Y_{\alpha^*} - Y_\alpha \text{ for all } \alpha \neq \alpha^*, \alpha \in \mathcal{A}\} \subseteq \mathcal{Y} \text{ for some } M_{\text{init}} > 0$$

such that whenever $Y_1 \in \mathcal{W}_{\text{init}}$ then with probability at least $1 - \delta$ the sequence of play generated by (FTGL) converges to x^* with rate given by the function ϕ_i

$$\phi_i(t) = \begin{cases} (\theta_i')^{-1}(t) & \text{if } t > \theta_i'(0^+), \\ 0 & \text{otherwise.} \end{cases} \tag{C.8}$$

which depends on the choice of the kernel θ_i of each player and the payoff matrix of the game.

For convenience of notation we focus on an arbitrary player in the proof, without loss of generality let it be the i-th one, and we completely drop the index i. Since the equilibrium is strict by Lemmas C.1 and C.2 there exist a neighborhood $\mathcal{U}_{\text{strict}}$, $c_{\text{strict}} > 0$ and $M_{\text{strict}} > 0$ such that

$$v_{\alpha^*}(x) - v_\alpha(x) \geq c_{\text{strict}} \quad \text{for all } \alpha \neq \alpha^*, \alpha \in \mathcal{A} \text{ and } x \in \mathcal{U}_{\text{strict}} \tag{C.9}$$

$$Y_\alpha^* - Y_\alpha > M_{\text{strict}} \quad \text{for all } \alpha \neq \alpha^*, \alpha \in \mathcal{A} \text{ and } x = Q(Y) \in \mathcal{U}_{\text{strict}} \tag{C.10}$$

We start by proving the following claim:

Claim 1. *Let $\mathcal{W}(M)$ be a "score-dominant" open set for the strict Nash equilibrium x^* . Then there exists $M_{\text{init}} > 0$ such that if $Y_1 \in \mathcal{W}(M_{\text{init}}) = \mathcal{W}_{\text{init}}$ then with probability at least $1 - \delta$ the sequence of play $(Y_n)_{n\in\mathbb{N}}$ stays in $\mathcal{W}(M_{\text{strict}})$.*

Proof of Claim. By definition of (FTGL) for the score differences we have

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} = Y_{\alpha,1} - Y_{\alpha^*,1} + \sum_{k=1}^n \gamma_k(\text{drift}_k + \text{noise}_k + \text{bias}_k) \tag{C.11}$$

where $\text{drift}_k = v_\alpha(X_k) - v_{\alpha^*}(X_k)$, $\text{noise}_k = U_{\alpha,k} - U_{\alpha^*,k}$, $\text{bias}_k = b_{\alpha,k} - b_{\alpha^*,k}$. Notice that

- (*Bias*) By (A1): $\sum_{k=1}^n \gamma_k \text{bias}_k \leq 2 \sum_{k=1}^n \gamma_k \|b_k\|_* = o(\sum_{k=1}^n \gamma_k)$ (C.12)
- (*Payoff*) By Lemma C.1: $\sum_{k=1}^n \gamma_k \text{drift}_k \leq -c \sum_{k=1}^n \gamma_k$ (C.13)
- (*Zero-mean Noise*) For the remaining term, $R_n = \sum_{k=1}^n \gamma_k \text{noise}_k$, firstly notice that it is trivially a martingale. We will prove that with probability at least $1 - \delta$ this martingale is bounded above by a term ξ_n which is dominated by the term $\sum_{k=1}^n \gamma_k$. Consider the event $D_{n,\xi_n} = \{\sup_{1\leq k\leq n} |R_k| \geq \xi_n\}$; we will show that the union of these events $\mathcal{E} = \bigcup_{n=1}^\infty D_{n,\xi_n}$ occurs with probability at most δ when $\xi_n = \xi(\sum_{k=1}^n \gamma_k)^a$ with $a < 1$. Using Theorem A.1 and Theorem A.2 we have

$$\mathbb{P}(D_{n,\xi_n}) \leq \frac{\mathbb{E}[|R_n|^q]}{{\xi_n}^q} \leq \frac{c_q \, \mathbb{E}[(\sum_{k=1}^n \gamma_k^2 \|U_k\|_*^2)^{q/2}]}{\xi_n^q} \tag{C.14}$$

Fact (Generalized Hölder's Inequality)**.** *We will now consider a variation of the Hölder'sinequality*

$$\left(\sum_{k=1}^{n} a_k b_k\right)^r \leq \left(\sum_{k=1}^{n} a_k^{\frac{\mu r}{r-1}}\right)^{r-1} \sum_{k=1}^{n} a_k^{(1-\mu)r} b_k^r \text{ for all } r > 1, \mu \in (0,1) \tag{GH}$$

Applying (GH) for $a_k = \gamma_k^2$, $b_k = \|U_k\|_*^2$, $r = q/2$ and $\mu = (r-1)/2r = (q-2)/2q$, we get

$$\mathbb{P}(D_{n,\xi_n}) \leq \frac{c_q (\sum_{k=1}^n \gamma_k)^{\frac{q-2}{2}} \sum_{k=1}^n \gamma_k^{1+q/2} \mathbb{E}[\|U_k\|_*^q]}{\xi_n^q} \tag{C.15}$$

$$\leq \frac{c_q (\sum_{k=1}^n \gamma_k)^{\frac{q-2}{2}} \sum_{k=1}^n \gamma_k^{1+q/2} \mathbb{E}[\mathbb{E}[\|U_k\|_*^q \mid \mathcal{F}_k]]}{\xi_n^q} \tag{C.16}$$

$$\leq \frac{c_q (\sum_{k=1}^n \gamma_k)^{\frac{q-2}{2}} \sum_{k=1}^n \gamma_k^{1+q/2} \sigma_k^q}{\xi_n^q} \tag{C.17}$$

Recall that $\xi_n = \xi\left(\sum_{k=1}^n \gamma_k\right)^a$ with $a < 1$ and let us denote $\delta_n = \frac{c_q}{\xi^q} \frac{\sum_{k=1}^n \gamma_k^{1+\frac{q}{2}} \sigma_k^q}{\left[\sum_{k=1}^n \gamma_k\right]^{1+(2a-1)q/2}}$ or equivalently $\delta_n = \frac{c_q}{\xi^q} \frac{\sum_{k=1}^n \gamma_k^{1+\frac{q}{2}} \sigma_k^q}{\left[\sum_{k=1}^n \gamma_k\right]^{1+\beta q/2}}$ for some $\beta < 1$. By assumption (A3), δ_n is summable and by controlling the parameter ξ we can ensure that

$$\sum_{n=1}^{\infty} \delta_n = \delta \tag{C.18}$$

Applying union bound to all the events D_{n,ξ_n} we have that with probability at least $1-\delta$ it is

$$\sum_{k=1}^{n} \gamma_k \text{noise}_k \leq \xi_n \text{ for all } n = 1, 2, \ldots \tag{C.19}$$

For the rest of the proof we condition to the event $\mathcal{E}^c$. Let us define a constant M_{init}, such that $M_{\text{init}} \geq \max\{M_{\text{strict}}, M_{\text{strict}} + \sup_{n\geq 1}\{\sum_{k=1}^n \gamma_k(\text{noise}_k + \text{bias}_k) - (c-c')\sum_{k=1}^n \gamma_k\}$, for any arbitray choice of $0 < c' < c_{\text{strict}}$ [5] . Let us recall the definition of a "score-dominant" open set

$$\mathcal{W}(M) = \{Y : Y_\alpha^* - Y_\alpha > M \text{ for all } \alpha \neq \alpha^*, \alpha \in \mathcal{A}\}.$$

We will prove by strong induction that $Y_n \in \mathcal{W}(M_{\text{strict}})$, for all $n \geq 1$.

- For the base of the induction, we have that $Y_1 \in \mathcal{W}(M_{\text{init}})$ and by the choice of M_{strict}, trivially we get that $Y_1 \in \mathcal{W}(M_{\text{strict}})$.
- For the inductive step, let us assume that $Y_k \in \mathcal{W}(M_{\text{strict}})$ for all $k = 1, 2, \ldots, n$, we will show below that $Y_{n+1} \in \mathcal{W}(M_{\text{strict}})$.

Combining (C.12),(C.13),(C.19) for the terms $\sum_{k=1}^n \gamma_k \text{drift}_k$, $\sum_{k=1}^n \gamma_k \text{noise}_k$, $\sum_{k=1}^n \gamma_k \text{bias}_k$ the claim's assumption $Y_1 \in \mathcal{W}(M_{\text{strict}})$ and the choice of M_{init}, (C.11) can be bounded as

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} = Y_{\alpha,1} - Y_{\alpha^*,1} + \sum_{k=1}^{n} \gamma_k(\text{drift}_k + \text{noise}_k + \text{bias}_k) \tag{C.20}$$

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} \leq Y_{\alpha,1} - Y_{\alpha^*,1} - c_{\text{strict}} \sum_{k=1}^{n} \gamma_k + \xi_n + 2\sum_{k=1}^{n} \gamma_k \|b_k\|_* \tag{C.21}$$

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} \leq -M_{\text{init}} - (c_{\text{strict}} - c') \sum_{k=1}^{n} \gamma_k + \xi_n + 2\sum_{k=1}^{n} \gamma_k \|b_k\|_* - c' \sum_{k=1}^{n} \gamma_k \tag{C.22}$$

$$Y_{\alpha,n+1} - Y_{\alpha^*,n+1} \leq -M_{\text{strict}} - c' \sum_{k=1}^{n} \gamma_k \leq -M_{\text{strict}} \tag{C.23}$$

and thus $Y_{n+1} \in \mathcal{W}(M_{\text{strict}})$. ∎

[5] such a $M_{\text{init}} > 0$ exists since both the bias and the noise terms are dominated by the term the terms $2\sum_{k=1}^n \gamma_k \|b_k\|_*, \xi_n$ and consequently by $-(c-c')\sum_{k=1}^n \gamma_k$.

The above claim immediately implies that $X_n \in \mathcal{U}$ for all $n = 1, 2, \ldots$. We will now prove that the sequence of play converges to x^*.

Proof of Convergence. Let's assume that ad absordum that there exists at least one strategy $\alpha \neq \alpha^*, \alpha \in \mathcal{A}$ such that $\limsup_{n\to\infty} X_{\alpha,n} \geq \varepsilon > 0$. for all sufficiently large n. Recall also that for $X \in \mathcal{U}_{\text{strict}}$, it holds that $X_{\alpha^*} > 0$ by construction in Lemma C.2.

Then by Lemma B.1 we have

$$Y_\alpha = \theta'(X_\alpha) + \mu - v_\alpha \tag{C.24}$$

where $\mu \in \mathbb{R}$ and $v_\alpha \geq 0$ while $v_\alpha = 0$ whenever $X_\alpha > 0$. Leveraging that *i)* the sequence of play is contained in $\mathcal{U}$, *ii)* by descending to a subsequence if necessary $X_{\alpha,m_i} > 0$ and *iii)* recall (C.23) for the subsequence we have

$$Y_{\alpha,m_{i+1}} - Y_{\alpha^*,m_{i+1}} = \theta'(X_{\alpha,m_{i+1}}) - \theta'(X_{\alpha^*,m_{i+1}}) \leq -M_{\text{strict}} - c' \sum_{k=1}^{m_i} \gamma_k \tag{C.25}$$

However, the RHS of the above inequality goes to $-\infty$ as $m_i \to \infty$, while the LHS of the above inequality is bounded by the constant $\theta'(\varepsilon) - \theta'(1)$ since θ' is strictly increasing, which is a contradiction[6]. ∎

Proof of Rate. We now proceed to the delineation of the exact rates achieved. Consider the function

$$\phi(t) = \begin{cases} (\theta')^{-1}(t) & \text{if } t > \theta'(0^+), \\ 0 & \text{otherwise.} \end{cases} \tag{C.26}$$

where $(\theta')^{-1}(z)$ is the inverse function of the kernel θ'[7]. Focusing on (C.25) we can derive that

$$\theta'(X_{\alpha,n+1}) \leq -M_{\text{strict}} + \theta'(X_{\alpha^*,n+1}) - c' \sum_{k=1}^{n} \gamma_k \tag{C.27}$$

$$\leq -M_{\text{strict}} + \theta'(1) - c' \sum_{k=1}^{n} \gamma_k \tag{C.28}$$

for all $\alpha \in \mathcal{A}_i$ and $n = 1, 2, \ldots$. As a result

$$X_{\alpha,n+1} \leq \phi(-M_{\text{strict}} + \theta'(1) - c' \sum_{k=1}^{n} \gamma_k) \tag{C.29}$$

Aggregating over all strategies $\alpha \in \mathcal{A}$, $\alpha \neq \alpha^*$ we have

$$\|x^* - X_{n+1}\|_1 = 2(1 - X_{\alpha^*,n+1}) \tag{C.30}$$

$$\leq \sum_{\alpha\in\mathcal{A}\neq\alpha^*} \phi(-M_{\text{strict}} + \theta'(1) - c' \sum_{k=1}^{n} \gamma_k) \tag{C.31}$$

$$\leq \sum_{\alpha\in\mathcal{A}\neq\alpha^*} \phi(d - c' \sum_{k=1}^{n} \gamma_k) \tag{C.32}$$

where $d = -M_{\text{strict}} + \theta'(1)$. ∎

∎

[6] The aforementioned by contradiction argument also provides a short proof of Remark C.2.

[7] θ' is strictly increasing and so does its inverse.

Corollary 1. *If the regularizer employed is non-steep (i.e., θ_i is differentiable at 0), X_n converges to x^* in a finite number of iterations.*

Proof. Additionally, in the case of non-steep regularizers we can prove that convergence occurs in finite time. More precisely, focusing on (C.28) and bearing in mind that $X_{\alpha,n+1} \geq 0$ for all $n = 1, 2, \ldots$ we have

$$\theta'(0) \leq \theta'(X_{\alpha,n+1}) \leq -M_{\text{strict}} + \theta'(1) - c' \sum_{k=1}^{n} \gamma_k \tag{C.33}$$

At the same time for finite n it holds

$$\sum_{k=1}^{n} \gamma_k \geq (-M_{\text{strict}} + \theta'(1) - \theta'(0))/c' \tag{C.34}$$

since $\theta'(0)$ is finite for non-steep regularizers. Rearranging the above inequality we have

$$-M_{\text{strict}} + \theta'(1) - c' \sum_{k=1}^{n} \gamma_k \leq \theta'(0) \tag{C.35}$$

which inevitably implies that $X_{\alpha,n+1} = 0$. ∎

D Models

We start by presenting the well-known algorithms *Follow the Regularized Leader* (FTRL), *Optimistic Follow the Regularized Leader* (OptFTRL) and *Mirror Prox* (MP), as special cases of our general algorithmic framework.

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

$$\tilde{Y}_{i,n} = Y_{i,n} + \gamma_n V_{i,n-1} \qquad \tilde{X}_{i,n} = Q_i(\tilde{Y}_{i,n}) \qquad Y_{i,n+1} = Y_{i,n} + \gamma_n V_{i,n} \tag{OptFTRL}$$

Remark D.1. (OptFTRL) requires two initializations and then at each stage the previous payoff signal is stored and is utilized to calculate the auxiliary cumulative payoff $\tilde{Y}_{i,n}$.

$$\begin{aligned} Y_{i,n+1/2} &= Y_{i,n} + \gamma_n V_{i,n} & Y_{i,n+1} &= Y_{i,n} + \gamma_n V_{i,n+1/2} \\ X_{i,n+1/2} &= Q_i(Y_{i,n+1/2}) & X_{i,n+1} &= Q_i(Y_{i,n+1}) \end{aligned} \tag{MirrorProx}$$

Remark D.2. (MirrorProx) requires only one initialization, but at each stage the algorithm generates two different states and correspondingly two payoff signals are needed.

For both the algorithms (OptFTRL),(MirrorProx) we can prove that for the cases of full information, oracle based feedback and noisy payoff feedback, the implicit bias for modeling their intermediate steps is $\|b_{i,n}\|_* = \mathcal{O}(\gamma_n)$. The bias is the same in all of the three cases and thus we only present the case of full information.

Proof. **Full information:**

- (OptFTRL): $V_{i,n} = v_i(X_n) + (v_i(\tilde{X}_n) - v_i(X_n))$. Thus

$$\begin{aligned} \|b_{i,n}\|_* &= \|v_i(\tilde{X}_n) - v_i(X_n)\|_* \leq C\|\tilde{X}_n - X_n\| && \text{(D.1)} \\ &= C\|Q_i(\tilde{Y}_n) - Q_i(Y_n)\| \leq C'\|\tilde{Y}_n - Y_n\|_* && \text{(D.2)} \\ &= \mathcal{O}(\gamma_n) && \text{(D.3)} \end{aligned}$$

- (MirrorProx): $V_{i,n} = v_i(X_n) + (v_i(X_{n+1/2}) - v_i(X_n))$. The proof is similar to the above and $\|b_{i,n}\|_* = \mathcal{O}(\gamma_n)$.

■

Below, we explain how the proof of Theorem 1 can be oriented to the specific structure of both (OptFTRL) and (MirrorProx), in order to achieve all the permitted step-sizes. We will not make an exact proof but we will thoroughly describe how the proof of Theorem 1 should be altered for the case of full information; the reader can follow similar steps for the case of oracle based feedback.

- *Optimistic Follow the Regularized Leader*
(OptFTRL) has an extra auxiliary cumulative payoff $\tilde{Y}_n$. We will first prove that if the two initializations of (OptFTRL) are appropriate then Theorem 1 holds without introducing any bias term.
Step 1: Notice that for the score differences of the auxiliary cumulative payoffs we have

$$\tilde{Y}_{\alpha,n+1} - \tilde{Y}_{\alpha^*,n+1} = Y_{\alpha,n} - Y_{\alpha^*,n} + \gamma_n\big(v_\alpha(\tilde{X}_{n-1}) - v_{\alpha^*}(\tilde{X}_{n-1})\big) \tag{D.4}$$

By substituting all the Y_n terms we have

$$\tilde{Y}_{\alpha,n+1} - \tilde{Y}_{\alpha^*,n+1} = Y_{\alpha,1} - Y_{\alpha^*,1} + \sum_{k=1}^{n-1} \gamma_k\big(v_\alpha(\tilde{X}_k) - v_{\alpha^*}(\tilde{X}_k)\big) + \gamma_n\big(v_\alpha(\tilde{X}_{n-1}) - v_{\alpha^*}(\tilde{X}_{n-1})\big) \tag{D.5}$$

Step 2: Assume that $\tilde{Y}_k \in \mathcal{W}_M$ as described in Theorem 1 and thus $\tilde{X}_k \in \mathcal{U}$ for all $k = 1, \dots, n$. We will prove by induction that $\tilde{Y}_{n+1} \in \mathcal{W}_M$. Notice that since $\tilde{X}_k \in \mathcal{U}$ it holds that

$$v_\alpha(\tilde{X}_k) - v_{\alpha^*}(\tilde{X}_k) \leq -c \text{ for all } k = 1, \dots, n \tag{D.6}$$

Step 3: From Eq. (D.5) we have

$$\tilde{Y}_{\alpha,n+1} - \tilde{Y}_{\alpha^*,n+1} \leq Y_{\alpha,1} - Y_{\alpha^*,1} - c\sum_{k=1}^{n} \gamma_k \tag{D.7}$$

By choosing $M_{\text{init}} > M$ our claim follows. We stress here that we have implicitly assumed that for the second initialization of (OptFTRL) it holds $\tilde{Y}_1 \in \mathcal{W}$.
Step 4: The rest of the proof holds as the one in Theorem 1, as all of the states $\tilde{X}_n$ remain in the desired neighborhood $\mathcal{U}$ in which the variational inequality holds.

- *Mirror Prox*
This algorithm, as we have already mentioned, calculates two different cumulative payoffs and primal states at each round.
Step 1: We will first prove by induction that that the cumulatve payoffs $Y_{n+1/2} \in \mathcal{W}_M$ for all $n = 1, 2, \dots$. Assume that $Y_{k+1/2} \in \mathcal{W}_M$ and thus $X_{k+1/2} \in \mathcal{U}$ for all $k = 1, \dots, n$ then for the score differences we have

$$Y_{\alpha,n+1/2} - Y_{\alpha^*,n+1/2} = Y_{\alpha,n} - Y_{\alpha^*,n} + \gamma_n(v_\alpha(X_n) - v_{\alpha^*}(X_n)) \tag{D.8}$$

$$= Y_{\alpha,1} - Y_{\alpha^*,1} + \sum_{k=1}^{n-1} \gamma_k\big(v_\alpha(X_{k-1/2}) - v_{\alpha^*}(X_{k-1/2})\big) \tag{D.9}$$

$$+ \gamma_n(v_\alpha(X_n) - v_{\alpha^*}(X_n)) \tag{D.10}$$

$$\leq Y_{\alpha,1} - Y_{\alpha^*,1} - c\sum_{k=1}^{n-1} \gamma_k + \gamma_n \max_{\alpha\in\mathcal{A}} \|v(\alpha)\|_* \tag{D.11}$$

Step 2: Choose $M_{\text{init}} > M + \gamma_n \max_{\alpha\in\mathcal{A}}\{\|v(\alpha)\|_*\}$ which is feasible for step-size of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$ and our claim follows.
Step 3: Continue with the proof as presented in Theorem 1.

Below we prove some properties concerning the case of payoff oracle/bandit feedback.

Proposition D.1. *In the bandit case, let $\tilde{X}_n$ be the state such that $\hat{X}_{i,n}$ is the mixed strategy of the i^{th} player at round n i.e., $\hat{X}_{i,n} = (1-\varepsilon_n)\tilde{X}_{i,n} + \varepsilon_n/|\mathcal{A}_i|$, based on which the pure strategy $\alpha_{i,n}$ is selected. Then the following properties hold*

1. $\mathbb{E}[U_{i,n} \,|\, \mathcal{F}_n] = 0$.
2. $\|U_{i,n}\|_* = \mathcal{O}(1/\varepsilon_n)$.
3. $\|b_{i,n}\|_* = \mathcal{O}(\varepsilon_n)$.

Remark D.3. In the case of (MirrorProx) $\tilde{X}_{i,n}$ is the state $X_{i,n-1/2}$.

Proof. The payoff signal which is estimated through the (IWE) can be rewritten as $V_{i,n} = v_i(X_n) + U_{i,n} + b_{i,n}$, where $U_{i,n} = V_{i,n} - v_i(\hat{X}_n)$ and $b_{i,n} = v_i(\hat{X}_n) - v_i(X_n)$.

1. Let $\mathcal{A}_i = \{\alpha_1, \dots, \alpha_{|\mathcal{A}_i|}\}$ be the pure strategies of player $i \in \mathcal{N}$; then
$$\mathbb{E}[V_{i,n}] = \sum_{\alpha_{-i} \in \mathcal{A}_{-i}} (u_i(\alpha_1; \alpha_{-i}), \dots, u_i(\alpha_{|\mathcal{A}_i|}))\hat{X}_{-i,n} = v_i(\hat{X}_n) \tag{D.12}$$
where with $\hat{X}_{-i,n}$ we symbolize the joint probability distribution for all players except for the i^{th} player.
2. We move on to the second part of this proposition.
$$\begin{aligned} \|U_{i,n}\|_* &= \|V_{i,n} - v_i(\hat{X}_n)\|_* && \text{(D.13)}\\ &\le \|V_{i,n}\|_* + \|v_i(\hat{X}_n)\|_* && \text{(D.14)}\\ &\le \max_{\alpha \in \mathcal{A}} |u_i(\alpha)||\mathcal{A}_i|/\varepsilon_n + \max_{\alpha \in \mathcal{A}} |u_i(\alpha)| && \text{(D.15)}\\ &= \mathcal{O}(1/\varepsilon_n) && \text{(D.16)} \end{aligned}$$
3. Finally for the norm of the bias term, let again $\mathcal{A}_i = \{\alpha_1, \dots, \alpha_{|\mathcal{A}_i|}\}$ be the pure strategies of player $i \in \mathcal{N}$; then
$$\begin{aligned} \|b_{i,n}\|_* &= \|v_i(\hat{X}_n) - v_i(X_n)\|_* && \text{(D.17)}\\ &= \|(u_i(\alpha_1; \hat{X}_{-i,n}) - u_i(\alpha_1; X_{-i,n}), \dots, u_i(\alpha_{|\mathcal{A}_i|}; \hat{X}_{-i;n}) - u_i(\alpha_{|\mathcal{A}_i|}; X_{-i;n}))\|_* && \text{(D.18)} \end{aligned}$$
It is sufficient to examine one of the elements of the vector $b_{i,n}$:
$$\begin{aligned} &|u_i(\alpha_1; \hat{X}_{-i,n}) - u_i(\alpha_1; X_{-i,n})| && \text{(D.19)}\\ &= |\sum_{\alpha_2 \in \mathcal{A}_2} \cdots \sum_{\alpha_N \in \mathcal{A}_N} (\hat{X}_{2\alpha_2,n} \dots \hat{X}_{N\alpha_N,n} - X_{2\alpha_2,n} \dots X_{N\alpha_N,n}) u_i(\alpha_1, \dots, \alpha_N)| && \text{(D.20)}\\ &\le \sum_{\alpha_2 \in \mathcal{A}_2} \cdots \sum_{\alpha_N \in \mathcal{A}_N} |\hat{X}_{2\alpha_2,n} \dots \hat{X}_{N\alpha_N,n} - X_{2\alpha_2,n} \dots X_{N\alpha_N,n}||u_i(\alpha_1, \dots, \alpha_N)| && \text{(D.21)}\\ &= \mathcal{O}(\varepsilon_n) && \text{(D.22)} \end{aligned}$$

■

In this section we provide different algorithms and feedback models which connect to our general algorithm (FTGL) and model described in Section 3.2. We first present a useful proposition in order to calculate the permitted parameters of the algorithm in order for assumption A3 to be satisfied.

Proposition D.2. *1. For all step sizes of the form $\gamma_n = \gamma/n^p$, with $p < 1$ and noise bounds $\sigma_n = \sigma n^r$ assumption A3 is satisfied if*
$$\frac{2}{q} - p + 2r < \beta(1-p) \text{ for some } \beta < 1 \tag{D.23}$$
Furthermore, it holds that
$$1/q + r < 1/2 \tag{D.24}$$
2. For all step-sizes of the form $\gamma_n = \gamma/n$ and $\sigma_n = \sigma n^r$, assumption A3 holds as long as
$$1/q + r < 1/2 \tag{D.25}$$

Proof. 1. Since $\gamma_n = \gamma/n^p$ and $\sigma_n = \sigma n^r$, assumption A3 is restated as

$$\delta_n = \frac{\sum_{k=1}^n \gamma_k^{1+q/2}\sigma_k^q}{[\sum_{k=1}^n \gamma_k]^{1+\beta q/2}} \tag{D.26}$$

$$= C_q(\sum_{k=1}^n 1/k^p)^{-1-\beta q/2}\sum_{k=1}^n 1/k^{p(1+\frac{q}{2})}k^{rq} \tag{D.27}$$

$$\leq C'_q n^{(1-p)(-1-\frac{\beta q}{2})}n^{1-p(1+\frac{q}{2})+rq} \tag{D.28}$$

$$\leq C'_q n^{-1-\frac{\beta q}{2}+p+\frac{p\beta q}{2}+1-p-\frac{pq}{2}+rq} \tag{D.29}$$

$$\leq C'_q n^{-\frac{\beta q}{2}+\frac{p\beta q}{2}-\frac{pq}{2}+rq} \tag{D.30}$$

Thus δ_n is summable if the exponent of n is less than -1:

$$-\frac{\beta q}{2}+\frac{p\beta q}{2}-\frac{pq}{2}+rq < -1 \tag{D.31}$$

$$\frac{2}{q}-p+2r < \beta(1-p) \tag{D.32}$$

The second expression of the proposition can be derived if we only keep the variable a in the RHS of the above inequality

$$\frac{2}{q}-p+2r < \beta(1-p) \tag{D.33}$$

$$(\frac{2}{q}-p+2r)/(1-p) < \beta < 1 \tag{D.34}$$

$$\frac{2}{q}-p+2r < 1-p \tag{D.35}$$

$$1/q+r < 1/2 \tag{D.36}$$

2. Let $\gamma_n = \gamma/n$ and $\sigma_n = \sigma n^r$, then for assumption A3 we have

$$\delta_n = \frac{\sum_{k=1}^n \gamma_k^{1+q/2}\sigma_k^q}{[\sum_{k=1}^n \gamma_k]^{1+\beta q/2}} \tag{D.37}$$

$$= C_q\frac{\sum_{k=1}^n \frac{1}{k^{1+q/2}}k^{rq}}{[\sum_{k=1}^n \frac{1}{k}]^{1+\beta q/2}} \tag{D.38}$$

$$\leq C'_q(\log(n+1))^{-1-\beta q/2}n^{1-1-q/2+rq} \tag{D.39}$$

$$\leq C'_q(\log(n+1))^{-1-\beta q/2}n^{-q/2+rq} \tag{D.40}$$

Since the sum $\sum_{n=1}^\infty 1/(\log^{1+\varepsilon}(n)n^{1+\varepsilon'})$ is finite for all $\varepsilon, \varepsilon' > 0$; assumption A3 is satisfied as long as

$$-q/2+rq < -1 \Rightarrow 1/q+r < 1/2 \tag{D.41}$$

■

Model D.1 ((FTRL) & Full information)**.** In this case players have access to their full payoff vector $v(X_n)$ for each round $n = 1, 2, \ldots$ and thus $V_{i,n} = v_i(X_n)$ for all $i \in \mathcal{N}$. All of the assumptions A1-A3 are satisfied; indeed

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Trivially satisfied since $U_{i,n} = 0$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$. §

Model D.2 ((FTRL) & Noisy payoff feedback)**.** In this setting at each round $n = 1, 2, \ldots$ players have access to a perturbed version of their full payoff vector $v(X_n)$ with a zero-mean noise U_n. Two examples of such noises that we consider in the experimental section are a zero-mean guassian noise and a uniform noise at $[-1.1]$. Both these noises satisfy (A2) with deterministic constant bounds for all $q \in [1, \infty]$. Thus

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Satisfied for all $q \in [1, \infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$. §

Model D.3 ((FTRL) & Oracle-based feedback)**.** Assume that each player chooses an action based on a given mixed strategy, and once every player has chosen an action, an oracle reveals to each player their corresponding pure payoff vector. Formally, at each round $n = 1, 2, \ldots$, each player chooses a pure strategy $\alpha_{i,n} \in \mathcal{A}_i$ based on a mixed strategy $X_{i,n} \in \mathcal{X}_i$ and subsequently observes the payoff vector

$$V_{i,n} = v_i(\alpha_n) = (u_i(\alpha_i; \alpha_{-i,n}))_{\alpha_i \in \mathcal{A}_i}. \tag{D.42}$$

Regarding our basic assumptions, we readily have $b_{i,n} = 0$ and $U_{i,n} = v_i(\alpha_n) - v_i(X_n)$; hence:

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Satisfied because $\|U_{i,n}\|_* = \|v_i(\alpha_n) - v_i(X_n)\|_* \leq 2 \max_{\alpha \in \mathcal{A}} \|v_i(\alpha)\|_*$, so U_n has uniformly bounded moments for all $q \in [1, \infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$. §

Model D.4 ((FTRL) & Payoff-based feedback)**.** If the players only observe their realized rewards, they have to *construct* a model for V_n based on incomplete information. This is the standard setting for multi-armed bandits, so it is often referred to as the "bandit feedback" model. In this case, the players' action selection process is as in Model D.3 above, but the feedback signal sequence V_n is now reconstructed by means of the importance-weighted estimator

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

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 now recalibrated by an *explicit exploration* parameter $\varepsilon_n \to 0$ whose role is to stabilize the learning process. The idea behind this adjustment 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.

The IWE estimator may be seen as a special case of the model (4) with $U_{i,n} = V_{i,n} - v_i(\hat{X}_n)$ and $b_{i,n} = v_i(\hat{X}_n) - v_i(X_n)$. All of the assumptions (A1)-(A3) are again satisfied; indeed:

- (A1): From Proposition D.1 $\|b_{i,n}\|_* = O(\varepsilon_n)$. Thus our assumption is satisfied since $\varepsilon_n \to 0$.
- (A2): Again from Proposition D.1 $\|V_{i,n} - v_i(\hat{X}_n)\|_* = O(1/\varepsilon_n)$ and thus the noise has bounded moments, $\sigma_n = \Theta(1/\varepsilon_n)$ for all $q \in [1, \infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$ and $\varepsilon_n \propto 1/n^r$, $r \in (0, 1/2)$.

§

Model D.5 ((OptFTRL) & Full information)**.** In this case the full payoff vector of each player is $V_{i,n} = v_i(\tilde{X}_n)$ for all $i \in \mathcal{N}$. As we proved above the state $\tilde{X}_n$ can be treated separately and thus

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Trivially satisfied since $U_{i,n} = 0$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$. §

Model D.6 ((OptFTRL) & Noisy payoff feedback)**.** Again in this setting at each round $n = 1, 2, \ldots$ players have access to a perturbed version of their full payoff vector $v(\tilde{X}_n)$ with a zero-mean noise U_n. Two examples of such noises that we consider in the experimental section are a zero-mean guassian noise and a uniform noise at $[-1.1]$. Both these noises satisfy (A2) with deterministic constant bounds for all $q \in [1, \infty]$. Thus

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Satisfied for all $q \in [1, \infty]$.
- (A3): From Proposition D.2 and our specific analysis for (OptFTRL) is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$. §

Model D.7 ((OptFTRL) & Oracle-based feedback)**.** In this case the payoff signal $V_{i,n}$, which depends on the state $\tilde{X}_n$, is generated as follows: at each round $n = 1, 2, \ldots$, every player $i \in \mathcal{N}$ picks an action $\alpha_{i,n} \in \mathcal{A}_i$ based on $\tilde{X}_{i,n} \in \mathcal{X}_i$ and observes the pure payoff vector $v_i(\alpha_n) \equiv (u_i(\alpha_i; \alpha_{-i,n}))_{\alpha_i \in \mathcal{A}_i}$.

Each player's input signal is a special case of (4) with payoff feedback $V_{i,n} = v_i(\alpha_n)$, zero-mean noise $U_{i,n} = v_i(\alpha_n) - v_i(\tilde{X}_n)$ and bias $b_{i,n} = 0$ that satisfy all of the assumptions A1 - A3. In more detail, we have:

- (A1): trivially satisfied since $b_{i,n} = 0$.
- (A2): $\|U_{i,n}\|_* = \|v_i(\alpha_n) - v_i(\tilde{X}_n)\|_* \leq 2\max_{\alpha\in\mathcal{A}}\|v_i(\alpha)\|_*$ and thus the noise has bounded moments for all $q \in [1,\infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0,1]$. §

Model D.8 ((OptFTRL) & Payoff-based feedback)**.** As we mentioned in Model D.4, in this case players only observe their realized rewards; thus they have to *construct* a model for V_n based on incomplete information. The players' action selection process is as in Model D.7 above, but the feedback signal sequence V_n is now reconstructed by means of the importance-weighted estimator

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

where $\hat{X}_{i,n} = (1-\varepsilon_n)\tilde{X}_{i,n} + \varepsilon_n/|\mathcal{A}_i|$ is the mixed strategy of the i-th player at stage n. Compared to $\tilde{X}_{i,n}$ the player's actual sampling strategy is now recalibrated by an *explicit exploration* parameter $\varepsilon_n \to 0$.

This type of feedback may be seen as a special case of the model (4) with $U_{i,n} = V_{i,n} - v_i(\hat{X}_n)$ and $b_{i,n} = v_i(\hat{X}_n) - v_i(X_n)$. All of the assumptions (A1)-(A3) are again satisfied; indeed:

- (A1): From Proposition D.1 $\|b_{i,n}\|_* = O(\varepsilon_n)$. Thus our assumption is satisfied since $\varepsilon_n \to 0$.
- (A2): Again from Proposition D.1 $\|V_{i,n} - v_i(\hat{X}_n)\|_* = O(1/\varepsilon_n)$ and thus the noise has bounded moments, $\sigma_n = \Theta(1/\varepsilon_n)$ for all $q \in [1,\infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0,1]$ and $\varepsilon_n \propto 1/n^r$, $r \in (0,1/2)$. §

Model D.9 ((MirrorProx) & Full information)**.** In this case players have access to their full payoff vector $v(X_n)$ for each round $n = 1,2,\ldots$; for the algorithm (MirrorProx) we observe two payoff vectors at each round and as stated in the specific analysis above, for each one of $v_i(X_{n+1/2})$ and $v_i(X_n)$, we have

- Assumption A1: Trivially satisfied since $b_{i,n} = 0$.
- (A2): Trivially satisfied since $U_{i,n} = 0$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0,1]$. §

Model D.10 ((MirrorProx) & Noisy payoff feedback)**.** As before at each round $n = 1,2,\ldots$ players have access to a perturbed version of their full payoff vector $v(X_n)$ with a zero-mean noise U_n. Two examples of such noises that we consider in the experimental section are a zero-mean guassian noise and a uniform noise at $[-1.1]$. Both these noises satisfy (A2) with deterministic constant bounds for all $q \in [1,\infty]$. Thus

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Satisfied for all $q \in [1,\infty]$.
- (A3): From Proposition D.2 and our specific analysis for (MirrorProx) is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0,1]$. §

We simply mention here that in the exact same way all of the assumptions (A1)-(A3) are satisfied for the second "intermediate" state of (MirrorProx).

Model D.11 ((MirrorProx) & Oracle-based feedback)**.** In this case, at each round n each player $i \in \mathcal{N}$ chooses two pure strategies $\alpha_{i,n}$ and $\alpha_{i,n+1/2}$ successively based on the mixed strategies $X_{i,n}$, $X_{i,n+1/2}$ equivalently. Thus, the first payoff signal is $V_{i,n} = v_i(\alpha_n)$ with $b_{i,n} = 0$ and $U_{i,n} = v_i(\alpha_n) - v_i(X_n)$. Hence:

- (A1): Trivially satisfied since $b_{i,n} = 0$.
- (A2): Satisfied because $\|U_{i,n}\|_* = \|v_i(\alpha_n) - v_i(X_n)\|_* \leq 2\max_{\alpha\in\mathcal{A}}\|v_i(\alpha)\|_*$, so U_n has uniformly bounded moments for all $q \in [1,\infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0,1]$, by also taking into account our specific analysis for (MirrorProx) presented above. §

The second payoff signal is $V_{i,n+1/2} = v_i(\alpha_{n+1/2})$ with $b_{i,n+1/2} = 0$ and $U_{i,n+1/2} = v_i(\alpha_{n+1/2}) - v_i(X_{n+1/2})$

- (A1): Trivially satisfied since $b_{i,n+1/2} = 0$.
- (A2): Satisfied because $\|U_{i,n+1/2}\|_* = \|v_i(\alpha_{n+1/2}) - v_i(X_{n+1/2})\|_* \leq 2\max_{\alpha\in\mathcal{A}}\|v_i(\alpha)\|_*$, so U_n has uniformly bounded moments for all $q \in [1, \infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$, by also taking into account our specific analysis for (MirrorProx) presented above. §

Model D.12 ((MirrorProx) & Payoff-based feedback)**.** In this case, as we have already mentioned, players only observe their realized rewards and the feedback signal sequence V_n is now reconstructed by means of the importance-weighted estimator

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

where $\hat{X}_{i,n} = (1 - \varepsilon_n)X_{i,n+1/2} + \varepsilon_n/|\mathcal{A}_i|$ is the mixed strategy of the i-th player at stage n, with $\varepsilon_n \to 0$.

The IWE estimator may be seen as a special case of the model (4) with $U_{i,n} = V_{i,n} - v_i(\hat{X}_n)$ and $b_{i,n} = v_i(\hat{X}_n) - v_i(X_n)$. All of the assumptions (A1)-(A3) are again satisfied; indeed:

- (A1): From Proposition D.1 $\|b_{i,n}\|_* = O(\varepsilon_n)$. Thus our assumption is satisfied since $\varepsilon_n \to 0$.
- (A2): Again from Proposition D.1 $\|V_{i,n} - v_i(\hat{X}_n)\|_* = O(1/\varepsilon_n)$ and thus the noise has bounded moments, $\sigma_n = \Theta(1/\varepsilon_n)$ for all $q \in [1, \infty]$.
- (A3): From Proposition D.2 is satisfied for all the step-sizes of the form $\gamma_n \propto 1/n^p$, $p \in [0, 1]$ and $\varepsilon_n \propto 1/n^r$, $r \in (0, 1/2)$.

E Experiments

We start this section by explaining in detail the two main games that our experiments are conducted.

E.1. Games.

1. In the archetypal game of *Battle of the Sexes*, a couple argues over which music to listen over the weekend. Both know that they want to spend the weekend together, but they cannot agree over what to do. The partner (A) prefers to audit a *Rock* band concert, whereas the partner (B) prefers a *Pop* music show. This is a classical example of a coordination game, analysed in game theory for its applications in many fields, such as business management or military operations. For the interested reader, check [26]. Since the couple wants to spend time together, if they go separate ways, they will receive no utility (set of payoffs will be $0, 0$). If they go either to a *Rock* or a *Pop* musical, both will receive some utility from the fact that they're together, but one of them will actually enjoy the activity. The description of this game in strategic form is therefore as follows:

Battle of Sexes

	Rock	*Pop*
Rock	$(2, 1)$	$(0, 0)$
Pop	$(0, 0)$	$(1, 2)$

Figure 3: Equilibrium Structure: This game has two strict Nash equilibria, one where both go to the *Rock* concert, and another where both go to the *Pop* concert. There is also a mixed Nash equilibrium, where the players go to their preferred event more often than the other. For the described payoffs, each player attends their preferred event with probability $3/5$.

2. In the selfish routing game of *Pigou's Congestion Network*, we consider the simple network shown in Fig. 4. Two disjoint edges/paths connect a source vertex O to a destination vertex D. Each edge is labeled with a cost function, which describes the cost (e.g., travel time) incurred by users of the edge, as a function of the amount of traffic routed on the edge. In the atomic version of the game the population of the drivers that uses a specific edge is an integer $x \in \{0, \cdots, N\}$. The upper edge has the constant latency function $\ell_1(x) = 1$, and thus it represents a route that is relatively long but immune to congestion. In the linear latency setting, the cost of the lower edge, which is governed by the function $\ell_2(x) = x/N$, increases as the edge gets more congested. In particular, the lower edge is cheaper than the upper edge if and only if less than N drivers uses it.

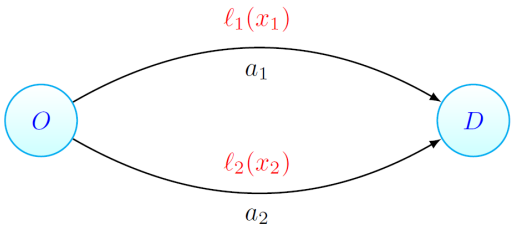
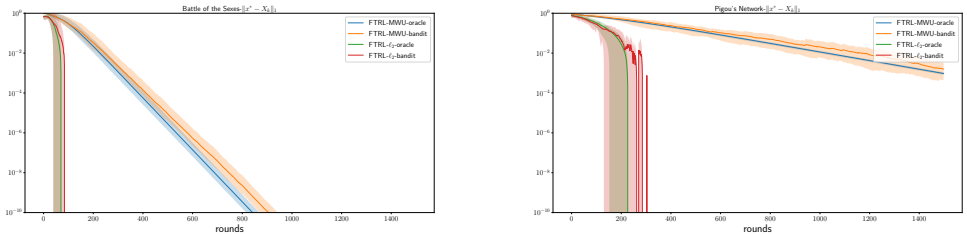


Figure 4: Pigou's Network

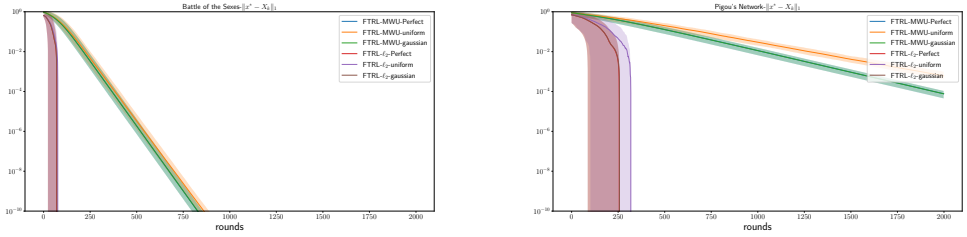
E.2. Experimental setup and methodology. Below, we will present separately the three archetypal instantiations of (FTGL) that we discussed in Appendix D, namely (FTRL),(OptFTRL) and (MirrorProx). All algorithms were run on *a*) a game of the Battle of the Sexes; and *b*) Pigou's linear version with $N = 1000$ atomic drivers. For each algorithm and each model we will present the performance of two well-studied regularizers: • entropic : $\theta_\alpha(x) = x_\alpha \log x_\alpha$ • euclidean : $\theta_\alpha(x) = x_\alpha^2/2$.

We will group our models with the following way: The first collection of figures for each algorithmic subsection will include the {oracle-based,payoff based/bandit} feedback model for the two aforementioned games for constant step-size and inverse-polynomial $\gamma_n \propto 1/n^{1/2}$. The latter one will present the {perfect,uniform-noise,gaussian-noise} feedback. Finally, the shaded areas around the curves represent the error bars in the execution for different random initializations.



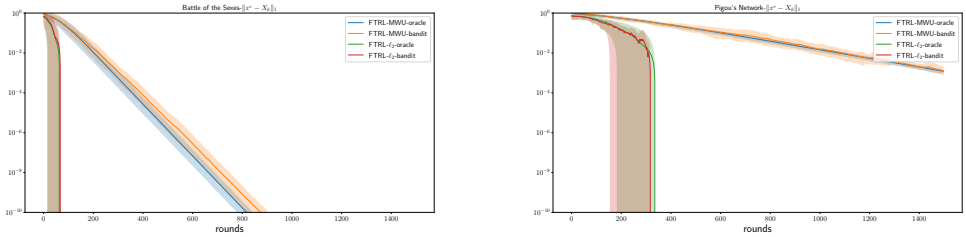
(a) Battle of the Sexes **(b)** Pigou Network

Figure 5: FTRL: oracle-based, bandit; $\gamma_n = 0.05$



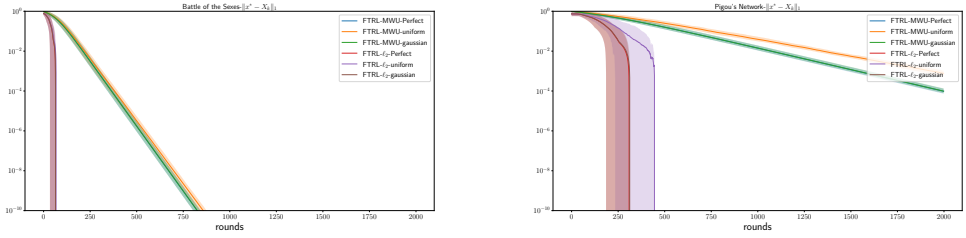
(a) Battle of the Sexes **(b)** Pigou Network

Figure 6: FTRL: uniform, gaussian; $\gamma_n = 0.05$.



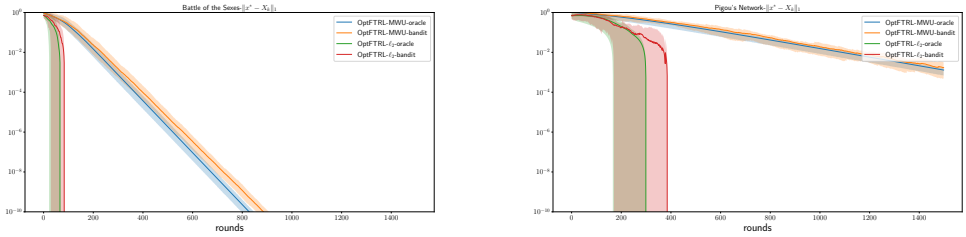
(a) Battle of the Sexes **(b)** Pigou Network

Figure 7: FTRL oracle, bandit; $\gamma_n \propto 1/n^{1/2}$



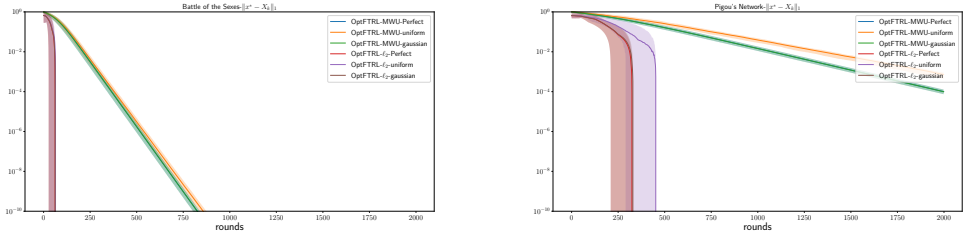
(a) Battle of the Sexes **(b)** Pigou Network

Figure 8: FTRL: uniform, gaussian; $\gamma_n \propto 1/n^{1/2}$



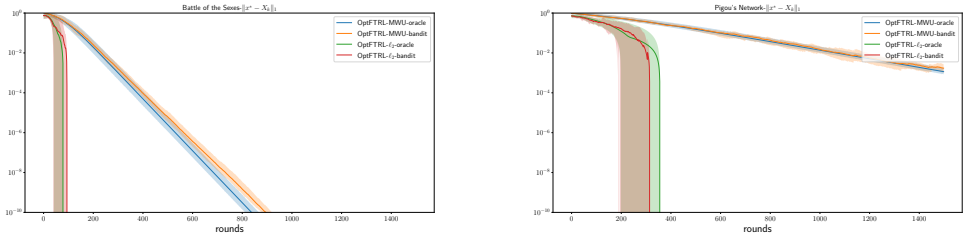
(a) Battle of the Sexes **(b)** Pigou Network

Figure 9: OptFTRL: oracle-based, bandit; $\gamma_n = 0.05$



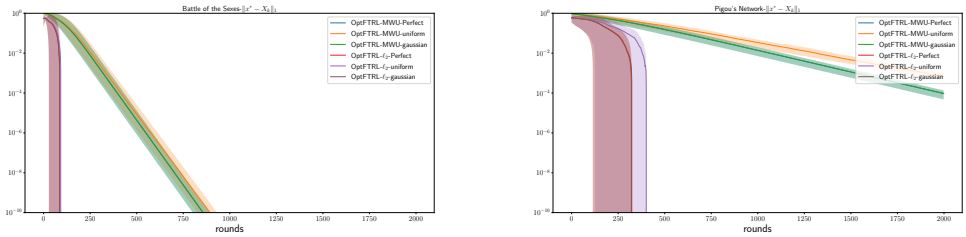
(a) Battle of the Sexes **(b)** Pigou Network

Figure 10: OptFTRL: uniform, gaussian; $\gamma_n = 0.05$



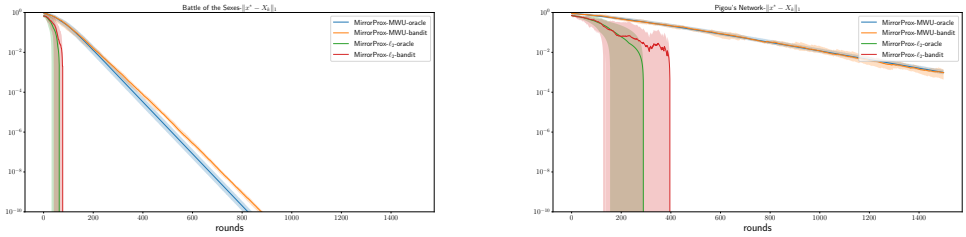
(a) Battle of the Sexes **(b)** Pigou Network

Figure 11: OptFTRL: oracle-based, bandit; $\gamma_n \propto 1/n^{1/2}$



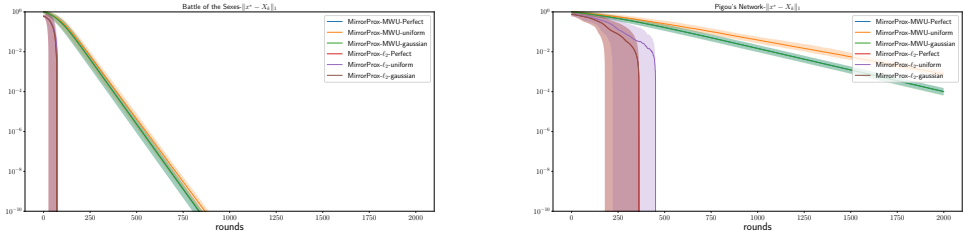
(a) Battle of the Sexes **(b)** Pigou Network

Figure 12: OptFTRL: uniform, gaussian; $\gamma_n \propto 1/n^{1/2}$



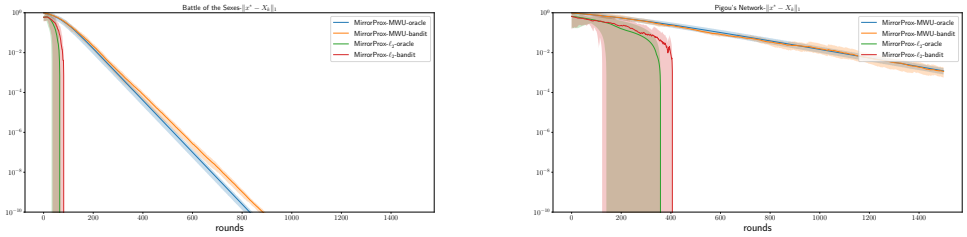
(a) Battle of the Sexes **(b)** Pigou Network

Figure 13: MP: oracle-based, bandit;$\gamma_n = 0.05$



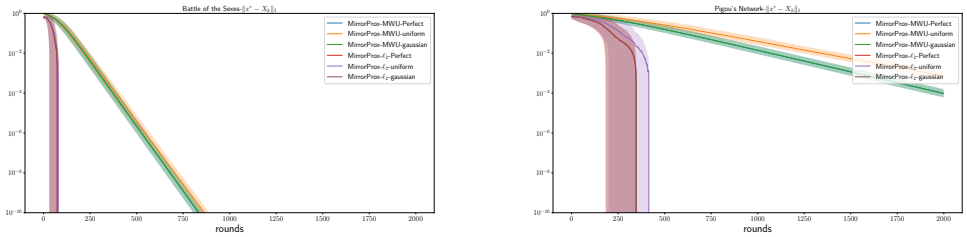
(a) Battle of the Sexes **(b)** Pigou Network

Figure 14: MP: uniform, gaussian; $\gamma_n = 0.05$



(a) Battle of the Sexes **(b)** Pigou Network

Figure 15: mirror-prox (MP): oracle-based, bandit; $\gamma_n \propto 1/n^{1/2}$



(a) Battle of the Sexes **(b)** Pigou Network

Figure 16: MP: uniform, gaussian; $\gamma_n \propto 1/n^{1/2}$