Gradient-free Online Learning in Games with Delayed Rewards

Amélie Héliou [1] **Panayotis Mertikopoulos** [2 1] **Zhengyuan Zhou** [3]

Abstract

Motivated by applications to online advertising and recommender systems, we consider a game-theoretic model with delayed rewards and asynchronous, payoff-based feedback. In contrast to previous work on delayed multi-armed bandits, we focus on multi-player games with continuous action spaces, and we examine the long-run behavior of strategic agents that follow a no-regret learning policy (but are otherwise oblivious to the game being played, the objectives of their opponents, etc.). To account for the lack of a consistent stream of information (for instance, rewards can arrive out of order, with an a priori unbounded delay, etc.), we introduce a gradient-free learning policy where payoff information is placed in a priority queue as it arrives. In this general context, we derive new bounds for the agents' regret; furthermore, under a standard diagonal concavity assumption, we show that the induced sequence of play converges to Nash equilibrium (NE) with probability 1, even if the delay between choosing an action and receiving the corresponding reward is unbounded.

1. Introduction

A major challenge in the application of learning theory to online advertising and recommender systems is that there is often a significant delay between action and reaction: for instance, a click on an ad can be observed within seconds of the ad being displayed, but the corresponding sale can take hours or days to occur – if it occurs at all. Putting aside all questions of causality and "what if" reasoning (e.g., the attribution of the sale to a given click), this delay has an adverse effect on all levels of the characterization between marketing actions and a user's decisions.

[1]Criteo AI Lab [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France [3]Stern School of Business, NYU, and IBM Research. Correspondence to: Panayotis Mertikopoulos <panayotis.mertikopoulos@imag.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Similar issues also arise in operations research, online machine learning, and other fields where online decision-making is the norm; as an example, we mention here the case of traffic allocation and online path planning, signal coveriance optimization in signal processing, etc. In view of all this, a key question that arises is *a*) to quantify the impact of a delayed reward / feedback structure on multi-agent learning; and *b*) to design policies that exploit obsolete information in a way as to minimize said impact.

Context. In this paper, we examine the above questions in the general framework of online learning in games with continuous action spaces. In more detail, we focus on recurrent decision processes that unfold as follows: At each stage $t = 1, 2, \dots$, the decision-maker (or player) selects an action X_t from a set of possible actions $\mathcal{X}$. This action subsequently triggers a reward $u_t(X_t)$ based on some (a priori unknown) payoff function $u_t \colon \mathcal{X} \to \mathbb{R}$. However, in contrast to the standard online optimization setting, this reward is only received by the player d_t stages later, i.e., at round $t + d_t$. As a result, the player may receive no information at round t, or they may receive older, obsolete information from some previous round $s < t$.

This very broad framework places no assumptions on the governing dynamics between actions and rewards, the payoff-generating process, or the delays encountered by the player. As such, the most common performance measure for a realized sequence of actions is the player's *regret*, i.e., the difference between the player's cumulative payoff over a given horizon and that of the best fixed action in hindsight. Thus, in the absence of more refined knowledge about the environments, the most sensible choice would be to deploy a policy which, at the very least, leads to *no regret*.

A specific instance of this "agnostic" framework – and one that has attracted considerable interest in the literature – is when the rewards of a given player are determined by the player's interactions with other players, even though the dynamics of these interactions can be unknown to the decision-making players beforehand. For instance, when placing a bid for reserving ad space, the ultimate payoff of a bidder will be determined by the bids of all other participating players and the rules of the underlying auction. The exact details of the auction (e.g., its reserve price) may be unknown to the bidders, and the bidders may not know

anything about whom they are bidding against, but their rewards are still determined by a fixed mechanism – that of an N-*player game*.

With all this in mind, our paper focuses on the following questions that arise naturally in this context: *Is there a policy leading to no regret in online optimization problems with delayed, payoff-based feedback?* And, assuming all players subscribe to such a policy, *does the induced sequence of play converge to a stable, equilibrium state?*

Our contributions. Our first contribution is to design a policy for online learning in this setting, which we call *gradient-free online learning with delayed feedback* (GOLD). The backbone of this policy is the online gradient descent (OGD) algorithm of Zinkevich (2003), but with two important modifications designed to address the challenges of the current setting. The first modification is the inclusion of a zeroth-order gradient estimator based on the simultaneous perturbation stochastic approximation (SPSA) mechanism of Spall (1997) and Flaxman et al. (2005). By virtue of this stochastic approximation mechanism, the player can estimate – albeit in a biased way – the gradient of their payoff function by receiving the reward of a nearby, perturbed action. The second element of GOLD is the design of a novel information pooling strategy that records information in a priority queue as they arrive, and subsequently dequeues them following a first-in, first-out (FIFO) scheme. The main challenge that occurs here is that the stream of information received by an agent may be highly unbalanced, e.g., consisting of intermittent batches of obsolete information followed by periods of feedback silence. This suggests that an agent should exercise a certain "economy of actions" and refrain from burning through batches of received information too quickly; the proposed pooling policy achieves precisely this by dequeueing at most one bit of feedback, even if more is available at any given stage.

From a theoretical viewpoint, the principal difficulty that arises is how to fuse these two components and control the errors that accrue over time from the use of obsolete – and biased – gradient estimates. This requires a delicate shadowing analysis and a careful tweaking of the method's parameters – specifically, its step-size sequence and the query radius of the SPSA estimator. In so doing, our first theoretical result is that GOLD guarantees no regret, even if the delays encountered by the agent are unbounded. Specifically, if the reward of the t-th round is received up to $o(t^\alpha)$ rounds later, then the GOLD algorithm enjoys a regret bound of the form $\mathcal{O}(T^{3/4} + T^{2/3+\alpha/3})$. In particular, this means that GOLD guarantees no regret even under *unbounded* delays that might grow over time at a sublinear rate.

Our third contribution is to derive the game-theoretic implications of concurrently running GOLD in a multi-agent setting. A priori, the link between no regret and Nash equilibrium (as opposed to coarse correlated equilibrium) is quite weak. Nevertheless, if the game in question satisfies a standard monotonicity condition due to Rosen (1965), we show that the sequence of actions generated by the GOLD policy converges to Nash equilibrium with probability 1. To the best of our knowledge, this is the first Nash equilibrium convergence result for game-theoretic learning with delayed, payoff-based feedback.

Related work. The no-regret properties of OGD in settings with delayed feedback was recently considered by Quanrud & Khashabi (2015) who proposed a natural extension of OGD where the player performs a *batched* gradient update the moment gradients are received. Doing so, Quanrud & Khashabi (2015) showed that if the total delay over a horizon T is $D_T = \sum_{t=1}^{T} d_t$, OGD enjoys a regret bound of the form $\mathcal{O}(\sqrt{T + D_T})$. This bound echoes a string of results obtained in the multi-armed bandit (MAB) literature under different assumptions: for instance, Joulani et al. (2013) and Vernade et al. (2017) assume that the origin of the information is known; Quanrud & Khashabi (2015) and Pike-Burke et al. (2018) do not make this assumption and instead consider an "anonymized" feedback enviroment; etc.

When the action space is finite, online learning with delayed feedback has also been explored in the context of adversarial MABs. In this context, Thune et al. (2019) bound the regret in this case with the cumulative delay, which, in our notation, would be $\mathcal{O}(T^{1+\alpha})$. Taking into account the non-square-root scaling of the regret due to the lack of gradient observations, this would conceivably lead to a bound similar to that of Theorem 1 for a MAB setting. Related papers which provide adaptive tuning to the unknown sum of delays are the works of Joulani et al. (2016), Zimmert & Seldin (2020), while Bistritz et al. (2019) and (Zhou et al., 2019) provide further results in adversarial and linear contextual bandits respectively. However, the algorithms used in these works have little to do with OGD.

Likewise, no-regret learning in bandit convex optimization has a long history dating back at least to Kleinberg (2004) and Flaxman et al. (2005). The standard OGD policy with SPSA gradient estimates achieves an $\mathcal{O}(T^{3/4})$ regret bound, and the $T^{3/4}$ term in our bound is indeed related to this estimate. Using sophisticated kernel estimation techniques, Bubeck & Eldan (2016, 2017) decreased this bound to $\mathcal{O}(T^{1/2})$, suggesting an interesting interplay with our work. However, very little is known when the learner has to cope *simultaneously* with delayed and payoff-based feedback.

In the MAB setting, the work of Joulani et al. (2013) provides an answer for mixed-strategy learning over finite-action spaces, but the online convex optimization case is

completely different. In particular, a major difficulty that arises is that the batch update approach of Quanrud & Khashabi (2015) cannot be easily applied with stochastic estimates of the received gradient information (or when attempting to infer such information from realized payoffs). This issue was highlighted in the work of Zhou et al. (2017a) who employed a batching strategy similar to that of Quanrud & Khashabi (2015) in a game-theoretic context with *perfect* gradient information. Because of this, online learning in the presence of delayed reward/feedback structures requires new tools and techniques.

On the game theory side, Krichene et al. (2015) and Balandat et al. (2016) studied the Nash equilibrium convergence properties of no-regret learning in specific classes of continuous games (zero-sum and potential games). The work of Mertikopoulos & Zhou (2019) and its follow-ups (Lin et al., 2020, Mertikopoulos et al., 2019, Zhou et al., 2017b, 2018, 2020) provided an extension to the class of monotone games with varying degrees of generality; however, all these works rely on the availability of gradients in the learning process. In sharp contrast to this, Bervoets et al. (2018) recently considered payoff-based learning in games with one-dimensional action sets, and they established convergence to Nash equilibrium under a synchronous, two-point, "sample-then-play" bandit strategy. More recently, Bravo et al. (2018) showed that no-regret learning with payoff-based feedback converges to Nash equilibrium in strongly monotone games, but it is assumed that actions are synchronized across players and rewards are assumed to arrive instantaneously. A model of learning with delays was provided by Zhou et al. (2017a) but their analysis and learning strategy only applies to perfect gradient information: the case of noisy – or, worse, *payoff-based* – delayed feedback was stated in that paper as a challenging open issue. Our paper settles this open question in the affirmative.

2. The model

2.1. The general framework

The general online optimization framework that we consider can be represented as the following sequence of events (presented for the moment from the viewpoint of a single, focal agent):

- At each stage $t = 1, 2, \dots$, of the process, the agent picks an *action* X_t from a compact convex subset $\mathcal{X}$ of a n-dimensional real space $\mathbb{R}^n$.
- The choice of action generates a *reward* $\hat{u}_t = u_t(X_t)$ based on a concave function $u_t \colon \mathcal{X} \to \mathbb{R}$ (assumed unknown to the player at stage t).
- Simultaneously, X_t triggers a *delay* $d_t \geq 0$ which determines the round $t + d_t$ at which the generated reward $\hat{u}_t$ will be received.
- The agent receives the rewards from all previous rounds $\mathcal{R}_t = \{s : s + d_s = t\}$, and the process repeats.

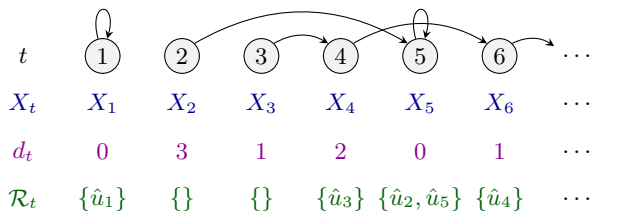

Figure 1: Schematic illustration of the delayed feedback framework considered in the paper. Arrows illustrate the round to which the payoff is deferred.

The above model has been stated in an abstract way that focuses on a single agent so as to provide the basis for the analysis to come. The setting where there are no assumptions on the process generating the agent's payoff functions will be referred to as the *unilateral* setting; by contrast, in the multi-agent, *game-theoretic* setting, the payoff functions of the focal agent will be determined by the stream of actions of the other players (see below for the details). In the latter case, all variables other than the running counter t will be indexed by i to indicate their dependence on the i-th player; for example, the action space of the i-th player will be written $\mathcal{X}^i$, the corresponding action chosen by at stage t will be denoted X_t^i, etc. For concreteness, we provide a diagrammatic illustration in Fig. 1 above.

In both cases, our blanket assumptions for the stream of payoffs and the delays encountered by the players will be as follows:

Assumption 1. For each $t = 1, 2, \dots$, u_t is concave in x, V_*-Lipschitz continuous, and β-Lipschitz smooth. Specifically, the gradient $V_t(x) \equiv \nabla_x u_t(x)$ of u_t is bounded by V_* and satisfies $\|V_t(\tilde{x}) - V_t(x)\| \leq \beta \|\tilde{x} - x\|$ for all $x, \tilde{x} \in \mathcal{X}$.

Assumption 2. The delays d_t grow asymptotically as $d_t = o(t^\alpha)$ for some $\alpha < 1$.

Regarding the delay assumption above, large-scale analytic studies have shown that long delays *are* observed in practice: in a study by Chapelle (2014) with data from the real-time bidding company Criteo, it was found that more than 10% of the conversions were more than two weeks old. Moreover, the conclusion of the same study was that the distribution of delays in online advertising can be fitted reasonably well by long-tailed distributions, especially when conditioning on context and feature variables available to the advertiser, thus justifying the assumption of a possibly unbounded delay between choosing an action and receiving a reward. We also note here that we are making no further assumptions on the

way the sequence of delays is generated: conceivably, delays could even be determined adversarially, as in Quanrud & Khashabi (2015).

2.2. Multi-agent considerations

For the multi-agent case, suppose there is a finite set of players $\mathcal{N} = \{1, \dots, N\}$, each with their own action space $\mathcal{X}^i \subseteq \mathbb{R}^{n^i}$ (always assumed convex and compact). In this case, it will be convenient to encode the players' joint action profile $x = (x^i)_{i\in\mathcal{N}} \in \mathcal{X} \equiv \prod_{i\in\mathcal{N}} \mathcal{X}^i$ by means of the shorthand $(x^i; x^{-i}) \equiv (x^1, \dots, x^i, \dots, x^N)$ which highlights the action $x^i \in \mathcal{X}^i$ of the i-th player against the action profile $x^{-i} \in \mathcal{X}^{-i} \equiv \prod_{j\neq i} \mathcal{X}^j$ of i's opponents. The payoff to each player $i \in \mathcal{N}$ for a given action profile $x \in \mathcal{X}$ will then be determined by an associated *payoff* (or *utility*) *function* $u^i \colon \mathcal{X} \to \mathbb{R}$, assumed here and throughout to be concave in the action variable x^i of the player in question. We will refer to the tuple $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ as an *N-player continuous game* (Debreu, 1952, Fudenberg & Tirole, 1991, Rosen, 1965).

In this context, if $X_t = (X_t^1, \dots, X_t^N) \in \mathcal{X}$ is a sequence of joint actions, the payoff function encountered by the i-th player at stage t will be given by

$$u_t^i(x^i) \equiv u^i(x^i; X_t^{-i}) \quad \text{for all } x^i \in \mathcal{X}^i, \tag{1}$$

leading to the gradient expression

$$V_t^i(x^i) \equiv \nabla_{x^i} u_t^i(x^i; X_t^{-i}) = V^i(x^i; X_t^{-i}) \tag{2}$$

where

$$V^i(x) = \nabla_{x^i} u^i(x^i; x^{-i}). \tag{3}$$

denotes the individual payoff gradient of the i-th player at the action profile $x \in \mathcal{X}$. In the rest of our paper, we will assume that u^i is Lipschitz continuous and Lipschitz smooth, so Assumption 1 is satisfied by default in this case.

2.3. Regret and equilibrium

With these preliminaries at hand, our principal performance indicators will be the minimization of *regret* and the notion of a *Nash equilibrium*. Starting with the former, the regret of an agent in the unilateral setting is defined over a horizon of T stages as

$$\mathrm{Reg}(T) = \max_{x\in\mathcal{X}} \sum_{t=1}^{T} [u_t(x) - u_t(X_t)]. \tag{4}$$

and, in the presence of randomness, we similarly introduce the agent's mean (or pseudo-) regret as

$$\overline{\mathrm{Reg}}(T) = \max_{x\in\mathcal{X}} \mathbb{E}[\sum_{t=1}^{T} [u_t(x) - u_t(X_t)]]. \tag{5}$$

Accordingly, we will say that a sequence of actions $X_t \in \mathcal{X}$, $t = 1, 2, \dots$, leads to *no regret* if $\mathrm{Reg}(T) = o(T)$.

On the other hand, the notion of a *Nash equilibrium* (NE) is a purely game-theoretic concept which characterizes those action profiles that are resilient to unilateral deviations. In more detail, we say that $x_* \in \mathcal{X}$ is a Nash equilibrium of $\mathcal{G}$ when

$$u^i(x_*) \geq u^i(x^i; x_*^{-i}) \tag{NE}$$

for all $x^i \in \mathcal{X}^i$ and all $i \in \mathcal{N}$. In full generality, the relation between Nash equilibria and regret minimization is feeble at best: if all players play a Nash equilibrium for all $t = 1, 2, \dots$, they will trivially have no regret; the converse however fails by a longshot, see e.g., Viossat & Zapechelnyuk (2013) and references therein.[1]

In the game-theoretic literature, existence and uniqueness of equilibrium points has been mainly studied under a condition known as *diagonal strict concavity* (DSC) (Rosen, 1965), which we define here as:

$$\sum_{i\in\mathcal{N}} \lambda^i \langle V^i(\tilde{x}) - V^x(x), \tilde{x}^i - x^i \rangle < 0 \tag{DSC}$$

for some $\lambda^i > 0$ and all $x, \tilde{x} \in \mathcal{X}$ with $\tilde{x} \neq x$.

In optimization, this condition is known as *monotonicity* (Bauschke & Combettes, 2017), so we will interchange the terms "diagonally strictly concave" and "*monotone*"for games that satisfy (DSC). Under (DSC), Rosen (1965) showed the existence of a unique Nash equilibrium; this is of particular importance to online advertising because of the following auction mechanism that can be seen as a monotone game:

Example 2.1 (Kelly auctions). Consider a provider with a splittable commodity (such as advertising time or website traffic to which a given banner will be displayed). Any fraction of this commodity can be auctioned off to a set of N bidders (players) who can place monetary bids $x^i \geq 0$ up to each player's total budget b^i to acquire it. Once all players have placed their respective bids, the commodity is split among the bidders proportionally to each player's bid; specifically, the i-th player gets a fraction $\rho^i = x^i \big/ (c + \sum_{j\in\mathcal{N}} x_j)$ of the auctioned commodity (where $c \geq 0$ is an "entry barrier" for bidding on the resource). A simple model for the utility of player i is then given by the *Kelly auction mechanism* (Kelly et al., 1998):

$$u^i(x^i; x^{-i}) = g^i \rho^i - x^i, \tag{6}$$

where g^i represents the marginal gain of player i from a unit of the commodity. Using standard arguments, it is easy

[1] Specifically, Viossat & Zapechelnyuk (2013) show that ther are games whose set of coarse correlated equilibria contain strategies that assign positive probability *only* to strictly dominated strategies.

to show that the resulting game satisfies (DSC), see e.g., Goodman (1980).

Other example of games satisfying (DSC) are (strictly) convex-concave zero-sum games (Juditsky et al., 2011), routing games (Nisan et al., 2007), Cournot oligopolies (Mertikopoulos & Zhou, 2019), power control (Mertikopoulos et al., 2017, Scutari et al., 2010), etc. For an extensive discussion of monotonicity in game theory, see Facchinei & Kanzow (2007), Laraki et al. (2019), Pang et al. (2010), Sandholm (2015) and references therein. In the rest of our paper, we will assume that all games under consideration satisfy (DSC).

3. The GOLD algorithm

We are now in a position to state the proposed *gradient-free online learning with delayed feedback* (GOLD) method. As the name suggests, the method concurrently addresses the two aspects of the online learning framework presented in the previous section, namely the delays encountered and the lack of gradient information. We describe each component in detail below, and we provide a pseudocode implementation of the method as Algorithm 1 above; for convenience and notational clarity, we take the viewpoint of a focal agent throughout, and we do not carry the player index i.

3.1. Delays

To describe the way that the proposed method tackles delays, it is convenient to decouple the two issues mentioned above and instead assume that, at time t, along with the generated rewards $\hat{u}_s$ for $s \in \mathcal{R}_t = \{s : s + d_s = t\}$, the agent also receives *perfect gradient information* for the corresponding rounds, i.e., gets to observe $V_s(X_s)$ for $s \in \mathcal{R}_t$. We stress here that this assumption is *only* made to illustrate the way that the algorithm is handling delays, and will be dropped in the sequel.

With this in mind, the first thing to note is that the set of information received at a given round might be empty, i.e., we could have $\mathcal{R}_t = \varnothing$ for some t. To address this sporadic shortage of information, we introduce a *pooling strategy*, not unlike the one considered by Joulani et al. (2013) in the context of multi-armed bandit problems. Specifically, we assume that, as information is received over time, the agent adds it to an *information pool* $\mathcal{P}_t$, and then uses the oldest information available in the pool (where "oldest" refers to the time at which the information was generated).

Specifically, starting at $t = 0$ with an empty pool $\mathcal{P}_0 = \varnothing$ (since there is no information at the beginning of the game), the agent's information pool is updated following the recursive rule

$$\mathcal{P}_t = \mathcal{P}_{t-1} \cup \mathcal{R}_t \setminus \{q_t\} \tag{7}$$

where

$$q_t = \min(\mathcal{P}_{t-1} \cup \mathcal{R}_t) \tag{8}$$

denotes the oldest round from which the agent has unused information at round t. Heuristically, this scheme can be seen as a priority queue in which data $V_s(X_s)$, $s \in \mathcal{R}_t$, arrives at time t and is assigned priority s (i.e., the round from which the data originated); subsequently, gradient data is dequeued one at a time, in ascending priority order (i.e., oldest information is utilized first).

In view of the above, if we let $\hat{V}_t = V_{q_t}(X_{q_t})$ denote the gradient information dequeued at round t, we will use the basic gradient update

$$X_{t+1} = \Pi(X_t + \gamma_t \hat{V}_t), \tag{9}$$

where $\gamma_t > 0$ is a variable step-size sequence (discussed extensively in the sequel), and $\Pi(y) = \arg\min_{x \in \mathcal{X}} \|x - y\|$ denotes the Euclidean projection to the agent's action space $\mathcal{X}$. Of course, an important issue that arises in the update step (7) is that, despite the parsimonious use of gradient information, it may well happen that the agent's information pool $\mathcal{P}_t$ is empty at time t (e.g., if at time $t = 1$, we have $d_1 > 0$). In this case, following the standard convention $\inf \varnothing = \infty$, we set $q_t = \infty$ (since it is impossible to ever have information about the stage $t = \infty$), and, by convention, we also set $V_\infty = 0$. Under this convention, (9) can be written in more explicit form as

$$\begin{aligned} X_{t+1} &= \Pi(X_t + \gamma_t \mathbb{1}_{\mathcal{P}_t \neq \varnothing} \hat{V}_t) \\ &= \begin{cases} X_t & \text{if } \mathcal{P}_t = \varnothing, \\ \Pi(X_t + \gamma_t \hat{V}_t) & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

In this way, the gradient update (9) can be seen as a delayed variant of Zinkevich's online gradient descent policy; however, in contrast to "batching-type" policies (Quanrud & Khashabi, 2015, Zhou et al., 2017a), there is no gradient aggregation: received gradients are introduced in the algorithm one at a time, oldest information first.

3.2. Payoff-based gradient estimation

We now proceed to describe the process with which the agent infers gradient information from the received rewards. To that end, following Spall (1997) and Flaxman et al. (2005), we will use a one-point, simultaneous perturbation stochastic approximation (SPSA) approach that was also recently employed by Bravo et al. (2018) for game-theoretic learning with bandit feedback (but no delays or asynchronicities). In our delayed reward setting (and always from the viewpoint of a single, focal agent), this process can be described as follows:

1. Pick a pivot state X_t to estimate its payoff gradient.

Algorithm 1: gradient-free online learning with delayed feedback (GOLD) [focal player view]

Require: step-size $\gamma_t > 0$, sampling radius $\delta_t > 0$, safety set $\mathbb{B}_r(p) \subseteq \mathcal{X}$

choose $X_1 \in \mathcal{X}$; set $\mathcal{P}_0 \leftarrow \varnothing$, $\hat{u}_\infty = 0$, $Z_\infty = 0$ # initialization
for $t = 1, 2, \dots$ **do**
  draw Z_t uniformly from $\mathbb{S}^n$ # perturbation direction
  set $W_t \leftarrow Z_t - (X_t - p)/r$ # feasibility adjustment
  play $\hat{X}_t \leftarrow X_t + \delta_t W_t$ # player chooses action
  generate payoff $\hat{u}_t = u(\hat{X}_t)$ # associated payoff
  trigger delay d_t # delay for payoff
  collect rewards $\mathcal{R}_t = \{s : s + d_s = t\}$ # receive past payoffs
  update pool $\mathcal{P}_t \leftarrow \mathcal{P}_{t-1} \cup \mathcal{R}_t$ # enqueue received info
  take $q_t = \min \mathcal{P}_t$; set $\mathcal{P}_t \leftarrow \mathcal{P}_t \setminus \{q_t\}$ # dequeue oldest info
  set $\hat{V}_t \leftarrow (n/\delta_{q_t}) \hat{u}_{q_t} Z_{q_t}$ # estimate gradient
  update $X_{t+1} \leftarrow \Pi(X_t + \gamma_t \hat{V}_t)$ # update pivot
end for

2. Pick a sampling radius $\delta_t > 0$ (detailed below) and draw a random sampling direction Z_t from the unit sphere $\mathbb{S}^n$.

3. Introduce an adjustment W_t to Z_t to ensure feasibility of the sampled action

$$\hat{X}_t = X_t + \delta_t W_t \tag{11}$$

4. Generate the reward $\hat{u}_t = u_t(\hat{X}_t)$ and estimate the gradient of u_t at X_t as

$$\hat{\nabla}_t = \frac{n}{\delta_t} \hat{u}_t Z_t \tag{12}$$

More precisely, the feasibility adjustment mentioned above is a skewing operation of the form

$$W_t = Z_t - r^{-1}(X_t - p) \tag{13}$$

where $p \in \mathcal{X}$ and $r > 0$ are such that the radius-r ball $\mathbb{B}_r(p)$ has $\mathbb{B}_r(p) \subseteq \mathcal{X}$, ensuring in this way that $\hat{X}_t \in \mathcal{X}$ whenever $X_t \in \mathcal{X}$; for more details, see Bubeck & Cesa-Bianchi (2012).

3.3. Learning with delayed, payoff-based feedback

Of course, the main problem in the SPSA estimator (12) lies in the fact that, in a delayed reward structure, the payoff generated at time t would only be observed at stage $t + d_t$. With this in mind, we make the following bare-bones assumptions:

- Expectations are taken relative to the inherent randomness in the sampling direction Z_t.
- The agent retains in memory the chosen sampling direction Z_s for all $s \leq t$ that have not yet been utilized, i.e., for all $s \in \mathcal{U}_t \equiv \{1, \dots, t\} \setminus \{q_\ell : \ell = 1, \dots, t\}$.[2]

[2] In the appendix, we show that $|\mathcal{U}_t| \leq \max_{1 \leq s \leq t} d_s$, so this requirement is fairly mild (linear) relative to the delays, especially when the delay distribution is exponential – e.g., as in the online advertising study of Chapelle (2014).

In this way, to combine the two frameworks described above (delays *and* bandit feedback), we will employ the gradient estimator

$$\hat{V}_t = \mathbb{1}_{\mathcal{P}_t \neq \varnothing} \hat{\nabla}_{q_t} = \frac{n}{\delta_{q_t}} \hat{u}_{q_t} Z_{q_t} \tag{14}$$

with the convention $\hat{u}_\infty = 0$, $Z_\infty = 0$ if $q_t = \infty$ – i.e., if the player's information pool $\mathcal{P}_t$ is empty at stage t. Thus, putting everything together, we obtain the *gradient-free online learning with delayed feedback* (GOLD) policy:

$$\begin{aligned} \hat{X}_t &= X_t + \delta_t W_t \\ X_{t+1} &= \Pi(X_t + \gamma_t \hat{V}_t) \end{aligned} \tag{GOLD}$$

with W_t and $\hat{V}_t$ given by Eqs. (13) and (14) respectively (for a pseudocode implementation of the policy, see Algorithm 1). We will examine the learning properties of this policy in the next section.

4. Analysis and guarantees

4.1. Statement and discussion of main results

We are now in a position to state and prove our main results for the GOLD algorithm under Assumptions 1 and 2. We begin with the algorithm's regret guarantees in the unilateral setting:

Theorem 1. *Suppose that an agent is running* (GOLD) *with step-size and sampling radius sequences of the form* $\gamma_t = \gamma/t^c$ *and* $\delta_t = \delta/t^b$ *for some* $\gamma, \delta > 0$ *and* $b = \min\{1/4, 1/3 - \alpha/3\}$, $c = \max\{3/4, 2/3 + \alpha/3\}$. *Then, the agent enjoys the mean regret bound*

$$\overline{\mathrm{Reg}}(T) = \tilde{\mathcal{O}}\big(T^{3/4} + T^{2/3+\alpha/3}\big). \tag{15}$$

Remark. In the above, $\tilde{\mathcal{O}}(\cdot)$ stands for "$\mathcal{O}(\cdot)$ up to logarithmic factors". The actual multiplicative constants that are hidden in the Landau "big oh" notation have a complicated

dependence on the diameter of $\mathcal{X}$, the dimension of the ambient space, the range of the players' utility functions; we provide more details on this in the paper's appendix.

For the game-theoretic setting, we will focus on games satisfying Rosen's diagonal strict concavity condition (e.g., as the Kelly auction example described in Section 2). In this general context, we have:

Theorem 2. *Let $\mathcal{G}$ be a continuous game satisfying* (DSC)*, and suppose that each agent follows* (GOLD) *with step-size and sampling radius sequences $\gamma_t = \gamma/t^c$ and $\delta_t = \delta/t^b$ for some $\gamma, \delta > 0$ and b, c satisfying the conditions:*

$$2c - b > 1 + \alpha, \tag{16a}$$
$$b + c > 1, \tag{16b}$$
$$2c - 2b > 1. \tag{16c}$$

Then, with probability 1, the sequence of play $\hat{X}_t$ induced by (GOLD) *converges to the game's* (*necessarily*) *unique Nash equilibrium.*

The above results are our main guarantees for (GOLD) so, before discussing their proof, some remarks are in order. The first concerns the tuning of the algorithm's hyperparameters, i.e., the exponents b and c. Even though the conditions stated in Theorem 1 may appear overly precise, we should note that agents have considerably more leeway at their disposal. Specifically, as part of the proof, we show that any choice of the exponents b and c satisfying (16) also leads to no regret – albeit possibly at a worse rate. This is particularly important for the interplay between no regret and convergence to Nash equilibrium because it shows that the two guarantees are fairly well aligned as long as (GOLD) is the class of no-regret policies under consideration.

We should also note here that the $T^{3/4}$ term is the standad regret bound that one obtains in the bandit online convex optimization framework. On the other hand, the term $T^{2/3+\alpha/3}$ describes the advent of the delays which, combined with the bias of the SPSA gradient estimator, contribute a significant amount of regret over time (recall in particular that d_t is a priori unbounded). This is of particular importance to applications to online advertising where delays can often become arbitrarily large.

For concreteness, we also plot in Fig. 2 the region of allowed step-size and sampling radius exponents. This plot reveals the interesting property that, if the feedback delays do not grow too large over time – specifically, if $d_t = o(t^{1/4})$ – then they have no impact on the allowable choices of b and c. This is also reflected in the regret bound (15) where, for $\alpha = 1/4$, the regret-specific term becomes $T^{3/4}$ as well; in particular, in the *constant regret* case $d_t = \mathcal{O}(1)$, the delays are invisible in (15). These considerations illustrate the impact of each source of feedback scarcity (bandit vs. delays)

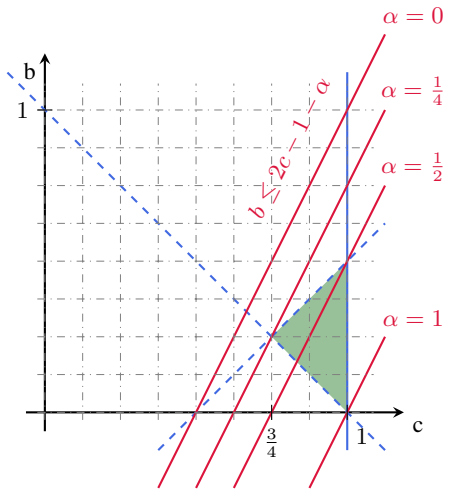

Figure 2: The allowable region (green shaded areas) of possible values of the sampling radius and step-size exponents b and c for various values of the groth exponent α of the encountered delays. The dashed blue lines corresponding to the last two terms in (16) indicate hard boundaries leading to logarithmic terms in the regret instead of constants.

on the performance of (GOLD) and provides a clear insight on the different mechanisms affecting the algorithm's regret and convergence guarantees.

4.2. Analysis and sketch of proof

The rest of this section is devoted to a high-level sketch of the proof of Theorems 1 and 2. We begin by decomposing $\hat{V}_t$ as a noisy estimate of $V(X_{q_t})$ into the following elements:

$$\hat{V}_t = V(X_{q_t}) + U_{q_t+1} + b_{q_t}. \tag{17}$$

The various terms in (17) above are defined as follows:

1. First, we set

$$U_{q_t+1} = \hat{V}_t - \mathbb{E}[\hat{V}_t \,|\, \mathcal{F}_t] \tag{18}$$

where the filtration $\mathcal{F}_t$ contains all the random variables that have been realized at the beginning of the t-th iteration of the algorithm; more precisely, we let

$$\mathcal{F}_t = \sigma(\varnothing, X_1, \dots, \hat{u}_{q_t-1}, Z_{t-1}, X_t) \tag{19}$$

with the convention $\hat{u}_\infty = 0$, $Z_\infty = 0$ if $q_t = \infty$. We note for posterity that U_{q_t} is a martingale difference sequence relative to $\mathcal{F}_t$, i.e., $\mathbb{E}[U_{q_t+1} \,|\, \mathcal{F}_t] = 0$.

2. Second, we let

$$b_{q_t} = \mathbb{E}[\hat{V}_t \,|\, \mathcal{F}_t] - V(X_{q_t}) \tag{20}$$

denote the systematic error of the estimator $\hat{V}_t$ relative to the gradient of the dequeued state X_{q_t} (i.e., the

error remaining after any zero-sum component has been averaged out). In contrast to U, this term is not zero-mean; instead, as we discuss in the appendix, the SPSA gradient estimation process that we employ induces a bias of order $\|b_{q_t}\| = \mathcal{O}(\delta_{q_t})$. This bias term grows smaller with t but its variance increases, leading to a bias-variance trade-off in our setting.

With all this in hand, the workhorse of our calculations is the distance of the sequence X_t to a given "benchmark" action $p \in \mathcal{X}$ (the best fixed action in hindsight, or the game's equilibrium, depending on the context). Specifically, letting

$$D_t = \frac{1}{2}\|X_t - p\|^2 \tag{21}$$

we have the following template inequality:

Lemma 1. *If (GOLD) is run with assumptions as above, then, for all $p \in \mathcal{X}$, we have*

$$\begin{aligned} D_{t+1} \leq D_t &+ \gamma_t \langle V(X_{q_t}), X_t - p\rangle && \text{(22a)}\\ &+ \gamma_t \langle U_{q_t+1}, X_t - p\rangle && \text{(22b)}\\ &+ \gamma_t \langle b_{q_t}, X_t - p\rangle && \text{(22c)}\\ &+ \tfrac{1}{2}\gamma_t^2 \|\hat{V}_t\|^2 && \text{(22d)} \end{aligned}$$

This lemma follows from the decomposition (17), the non-expansivity of the projection mapping, and the regularity assumption (1) which allows us to control the terms (22a) and (22c) above; to streamline our discussion, we defer the details to the paper's supplement. Moving forward, with this estimate at our disposal, the analysis branches for Theorems 1 and 2 as indicated below.

Regret analysis. To bound the agent's regret, we need to isolate the scalar product in (22a) and telescope through $t = 1, 2, \dots, T$ after dividing by the step-size γ_t. Deferring the ensuing lengthy calculations to the appendix, we ultimately obtain a bound of the form

$$\overline{\mathrm{Reg}}(T) = \mathcal{O}\left(\frac{1}{\gamma_T}\sum_{t=1}^{T}\left(\gamma_t \sum_{s=q_t}^{t-1}\frac{\gamma_s}{\delta_s} + \gamma_t\delta_{q_t} + \frac{\gamma_t^2}{\delta_{q_t}^2}\right)\right) \tag{23}$$

As a result, to proceed, we need to provide a specific bound for each of the above summands. The difficulty here is the mixing of different quantities at different time-stamps, e.g., as in the product term $\gamma_t\delta_{q_t}$. Bounding these terms requires a delicate analysis of the delay terms in order to estimate the maximum distance between t and q_t. We will return to this point below; for now, with some hindsight, we only stress that the terms in (23) correspond on a one-to-one basis with the conditions (16) for the parameters of (GOLD).

Game-theoretic analysis. The game-theoretic analysis is significantly more involved and relies on a two-pronged approach:

1. We first employ a version of the Robbins–Siegmund theorem to show that the random variable $D_t = (1/2)\|X_t - x_*\|^2$ converges pointwise as $t \to \infty$ to a random variable D_∞ that is bounded in expectation (here x_* denotes the game's unique equilibrium).

2. Subsequently, we use a series of probabilistic arguments (more precisely, a law of large numbers for martingale difference sequences and Doob's submartingale convergence theorem) to show that (GOLD) admits a (possibly random) subsequence X_{t_s} converging to x_*.

Once these two distinct elements have been obtained, we can readily deduce that $X_t \to x_*$ with probability 1 as $t \to \infty$. Hence, given that $\|X_t - \hat{X}_t\| = \mathcal{O}(\delta_t)$ and $\lim_t \delta_t = 0$, our claim would follow.

However, applying the probabilistic arguments outlined above requires in turn a series of summability conditions. Referring to the paper's supplement for the details, these requirements boil down to showing that the sequences

$$A_t = \gamma_t \sum_{s=q_t}^{t-1}\frac{\gamma_s}{\delta_s}, \quad B_t = \gamma_t\delta_{q_t}, \quad \text{and} \quad C_t = \frac{\gamma_t^2}{\delta_{q_t}^2}, \tag{24}$$

are all summable. Importantly, each of these three sums has a clear and concise interpretation in our learning context:

1. The first term (A_t) is the cumulative error induced by using outdated information.

2. The second term (B_t) is the error propagated from the bias of the SPSA estimator.

3. Finally, the third term (C_t) corresponds to the variance (or, rather, the mean square) of the SPSA estimator.

As a result, as long as these terms are all summable, their impact on the learning process should be relatively small (if not outright negligible).

Comparing the above term-by-term to (23) is where the game-theoretic analysis rejoins the regret analysis. As we said above, this requires a careful treatment of the delay process, which we outline below.

Delay analysis. A key difficulty in bounding the sums in (23) is that the first term (A_t in (24) is a sum of $t - q_t$ terms, so it can grow quite rapidly in principle. However, our pooling strategy guarantees that $t - q_t$ cannot grow faster than the delay (which is sublinear by assumption). This observation (detailed in the supplement) guarantees the convergence of the sum. A further hidden feature of (22) is in the noise term U_t: in the case of batching or reweighted strategies (e.g., as in Zhou et al., 2017a), this term incorporates a sum of terms arriving from different stages of the process, making

it very difficult (if not impossible) to control. By contrast, the pooling strategy that defines the GOLD policy allows us to treat this as an additional "noise" variable; we achieve this by carefully choosing the step-size and sampling radius parameters based on the following lemma:

Lemma 2. *Suppose that* (GOLD) *is run with step-size and sampling radius parameters of the form* $\gamma_t \propto \gamma/t^c$ *and* $\delta_t \propto \delta/t^b$*, with* $b, c > 0$*. Then:*

1. *If* $2c - b \geq 1 + \alpha$*, then* $\sum_{t=1}^{T} A_t = \mathcal{O}(\log T)$*; in addition, if the inequality is strict,* A_t *is summable.*
2. *If* $c + b \geq 1$*, then* $\sum_{t=1}^{T} B_t = \mathcal{O}(\log T)$*; in addition, if the inequality is strict,* B_t *is summable.*
3. *If* $2c - 2b \geq 1$*, then* $\sum_{t=1}^{T} C_t = \mathcal{O}(\log T)$*; in addition, if the inequality is strict,* C_t *is summable.*

Proving this lemma requires a series of intermediate results that we defer to the paper's supplement.

5. Concluding remarks

Our aim in this paper was to examine the properties of bandit online learning in games with continuous action spaces and a delayed reward structure (with a priori unbounded delays). The proposed GOLD policy is the first in the literature to simultaneously achieve no regret and convergence to Nash equilibrium with delayed rewards *and* bandit feedback. From a regret perspective, it matches the standard $\mathcal{O}(T^{3/4})$ bound of Flaxman et al. (2005) if the delay process is tame (specifically, if d_t grows no faster than $o(t^{1/4})$); in addition, from game-theoretic standpoint, it converges to equilibrium with probability 1 in all games satisfying Rosen's DSC condition.

One important direction for future research concerns the case of anonymous – i.e., not time-stamped – rewards. This complicates the matters considerably because it is no longer possible to match a received reward to an action; as a result, the GOLD policy would have to be redesigned from the ground up in this context. Another important avenue is that the kernel-based estimation techniques of Bubeck & Eldan (2016, 2017) achieve a faster $\mathcal{O}(T^{1/2})$ regret minimization rate with bandit feedback; whether this is still achievable with a delayed reward structure, and whether this can also lead to fast(er) convergence to Nash equilibrium is another direction for future research.

Acknowledgments

This research was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET). P. Mertikopoulos is also grateful for financial support by the French National Research Agency (ANR) under grant no. ANR–16–CE33–0004–01 (ORACLESS), and the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02) and the LabEx PERSYVAL (ANR-11-LABX-0025-01. Zhengyuan Zhou is grateful for the IBM Goldstine fellowship.

References

Balandat, M., Krichene, W., Tomlin, C., and Bayen, A. Minimizing regret on reflexive Banach spaces and Nash equilibria in continuous zero-sum games. In *NIPS '16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

Bervoets, S., Bravo, M., and Faure, M. Learning with minimal information in continuous games. https://arxiv.org/abs/1806.11506, 2018.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pp. 11345–11354, 2019.

Bravo, M., Leslie, D. S., and Mertikopoulos, P. Bandit learning in concave N-person games. In *NIPS '18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

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

Bubeck, S. and Eldan, R. Multi-scale exploration of convex functions and bandit convex optimization. In *COLT '16: Proceedings of the 29th Annual Conference on Learning Theory*, 2016.

Bubeck, S. and Eldan, R. Kernel-based methods for bandit convex optimization. In *STOC '17: Proceedings of the 49th annual ACM SIGACT symposium on the Theory of Computing*, 2017.

Chapelle, O. Modeling delayed feedback in display advertising. In *ACMSIGKDD '14 Proceedings of the 20th international conference on Knowledge discovery and data mining*, 2014.

Debreu, G. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences of the USA*, 38(10): 886–893, October 1952.

Facchinei, F. and Kanzow, C. Generalized Nash equilibrium problems. *4OR*, 5(3):173–210, September 2007.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA '05: Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Fudenberg, D. and Tirole, J. *Game Theory*. The MIT Press, 1991.

Goodman, J. C. Note on existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 48(1): 251, 1980.

Joulani, P., Gyorgy, A., and Szepesvári, C. Online learning under delayed feedback. In *ICML '13: Proceedings of the 30th International Conference on Machine Learning*, 2013.

Joulani, P., György, A., and Szepesvári, C. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *AAAI '16: Proceedings of the 30th Conference on Artificial Intelligence*, 2016.

Juditsky, A., Nemirovski, A. S., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Kelly, F. P., Maulloo, A. K., and Tan, D. K. H. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252, March 1998.

Kleinberg, R. D. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS' 04: Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2004.

Krichene, S., Krichene, W., Dong, R., and Bayen, A. Convergence of heterogeneous distributed learning in stochastic routing games. In *Allerton '15: Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing*, 2015.

Laraki, R., Renault, J., and Sorin, S. *Mathematical Foundations of Game Theory*. Universitext. Springer, 2019.

Lin, T., Zhou, Z., Mertikopoulos, P., and Jordan, M. I. Finite-time last-iterate convergence for multi-agent learning in games. *arXiv preprint arXiv:2002.09806*, 2020.

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

Mertikopoulos, P., Belmega, E. V., Negrel, R., and Sanguinetti, L. Distributed stochastic optimization via matrix exponential learning. *IEEE Trans. Signal Process.*, 65(9):2277–2290, May 2017.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. 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.

Nisan, N., Roughgarden, T., Tardos, É., and Vazirani, V. V. (eds.). *Algorithmic Game Theory*. Cambridge University Press, 2007.

Pang, J.-S., Scutari, G., Palomar, D. P., and Facchinei, F. Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach. *IEEE Trans. Signal Process.*, 58(6):3251–3271, June 2010.

Pike-Burke, C., Shipra, A., Szepesvári, C., and Grunewalder, S. Bandits with delayed, aggregated anonymous feedback. In *ICML '18: Proceedings of the 35th International Conference on Machine Learning*, 2018.

Quanrud, K. and Khashabi, D. Online learning with adversarial delays. In *NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015.

Rosen, J. B. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33(3):520–534, 1965.

Sandholm, W. H. Population games and deterministic evolutionary dynamics. In Young, H. P. and Zamir, S. (eds.), *Handbook of Game Theory IV*, pp. 703–778. Elsevier, 2015.

Scutari, G., Facchinei, F., Palomar, D. P., and Pang, J.-S. Convex optimization, game theory, and variational inequality theory in multiuser communication systems. *IEEE Signal Process. Mag.*, 27(3):35–49, May 2010.

Spall, J. C. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112, 1997.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. Nonstochastic multiarmed bandits with unrestricted delays. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Vernade, C., Cappé, O., and Perchet, V. Stochastic banit models for delayed conversions. In *UAI' 17: Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence*, 2017.

Viossat, Y. and Zapechelnyuk, A. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, March 2013.

Zhou, Z., Mertikopoulos, P., Bambos, N., Glynn, P. W., and Tomlin, C. Countering feedback delays in multi-agent learning. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017a.

Zhou, Z., Mertikopoulos, P., Moustakas, A. L., Bambos, N., and Glynn, P. W. Mirror descent learning in continuous games. In *CDC '17: Proceedings of the 56th IEEE Annual Conference on Decision and Control*, 2017b.

Zhou, Z., Mertikopoulos, P., Athey, S., Bambos, N., Glynn, P. W., and Ye, Y. Learning in games with lossy feedback. In *NIPS '18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pp. 5198–5209, 2019.

Zhou, Z., Mertikopoulos, P., Moustakas, A. L., Bambos, N., and Glynn, P. W. Robust power management via learning and game design. Operations Research, 2020.

Zimmert, J. and Seldin, Y. An optimal algorithm for adversarial bandits with arbitrary delays. In *AISTATS '20: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

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