Journal of Machine Learning Research 20 (2019) 1-38 Submitted 10/17; Published 2/19

# Delay and Cooperation in Nonstochastic Bandits

**Nicolò Cesa-Bianchi** NICOLO.CESA-BIANCHI@UNIMI.IT
*Department of Computer Science & DSRC*
*Università degli Studi di Milano*
*20133 Milano, Italy*

**Claudio Gentile** CLA.GENTILE@GMAIL.COM
*Google Research*
*New York, NY, USA*

**Yishay Mansour** MANSOUR@TAU.AC.IL
*Google Research and Tel-Aviv University*
*Tel-Aviv 6997801, Israel*

**Editor:** Network Analysis Special Issue

## Abstract

We study networks of communicating learning agents that cooperate to solve a common nonstochastic bandit problem. Agents use an underlying communication network to get messages about actions selected by other agents, and drop messages that took more than $d$ hops to arrive, where $d$ is a delay parameter. We introduce Exp3-Coop, a cooperative version of the Exp3 algorithm and prove that with $K$ actions and $N$ agents the average per-agent regret after $T$ rounds is at most of order $\sqrt{\big(d+1+\frac{K}{N}\alpha_{\leq d}\big)(T\ln K)}$, where $\alpha_{\leq d}$ is the independence number of the $d$-th power of the communication graph $G$. We then show that for any connected graph, for $d=\sqrt{K}$ the regret bound is $K^{1/4}\sqrt{T}$, strictly better than the minimax regret $\sqrt{KT}$ for noncooperating agents. More informed choices of $d$ lead to bounds which are arbitrarily close to the full information minimax regret $\sqrt{T\ln K}$ when $G$ is dense. When $G$ has sparse components, we show that a variant of Exp3-Coop, allowing agents to choose their parameters according to their centrality in $G$, strictly improves the regret. Finally, as a by-product of our analysis, we provide the first characterization of the minimax regret for bandit learning with delay.

**Keywords:** Multi-armed bandits, distributed learning, cooperative multi-agent systems, regret minimization, LOCAL communication

## 1. Introduction

Delayed feedback naturally arises in many sequential decision problems. For instance, a recommender system typically learns the utility of a recommendation by detecting the occurrence of certain events (e.g., a user conversion), which may happen with a variable delay after the recommendation was issued. Other examples are the communication delays experienced by interacting learning agents. Concretely, consider a network of geographically distributed ad servers using real-time bidding to sell their inventory. Each server sequentially learns how to set the auction parameters (e.g., reserve price) in order to maximize the network's overall revenue, and shares feedback information with other servers in order to speed up learning. However, the rate at which information is exchanged through the

communication network is slower than the typical rate at which ads are served. This causes each learner to acquire feedback information from other servers with a delay that depends on the network's structure.

Motivated by the ad network example, we consider networks of learning agents that cooperate to solve the same nonstochastic bandit problem, and study the impact of delay on the global performance of these agents. We introduce the Exp3-Coop algorithm, a distributed and cooperative version of the Exp3 algorithm of Auer et al. (2002b). Exp3-Coop works within a distributed and synced model where each agent runs an instance of the same bandit algorithm (Exp3). All bandit instances are initialized in the same way irrespective to the agent's location in the network (that is, agents have no preliminary knowledge of the network), and we assume the information about an agent's actions is propagated through the network with a unit delay for each crossed edge. In each round $t$, each agent selects an action and incurs the corresponding loss (which is the same for all agents that pick that action in round $t$). Besides observing the loss of the selected action, each agent obtains the information previously broadcast by other agents with a delay equal to the shortest-path distance between the agents. Namely, at time $t$ an agent learns what the agents at shortest-path distance $s$ did at time $t-s$ for each $s = 1, \dots, d$, where $d$ is a delay parameter. In this scenario, we aim at controlling the growth of the regret averaged over all agents (the so-called average welfare regret).

In the noncooperative case, when agents ignore the information received from other agents, the average welfare regret grows like $\sqrt{KT}$ (the minimax rate for standard bandit setting), where $K$ is the number of actions and $T$ is the time horizon. We show that, using cooperation, $N$ agents with communication graph $G$ can achieve an average welfare regret of order $\sqrt{\left(d+1+\frac{K}{N}\alpha_{\leq d}\right)(T\ln K)}$. Here $\alpha_{\leq d}$ denotes the independence number of the $d$-th power of $G$ (i.e., the graph $G$ augmented with all edges between any two pair of nodes at shortest-path distance less than or equal to $d$). When $d = \sqrt{K}$ this bound is at most $K^{1/4}\sqrt{T\ln K} + \sqrt{K}(\ln T)$ for any connected graph —see Remark 8 in Section 4.1— which is asymptotically better than $\sqrt{KT}$.

Networks of nonstochastic bandits were also investigated by Awerbuch and Kleinberg (2008) in a setting where the distribution over actions is shared among the agents without delay. Awerbuch and Kleinberg (2008) prove a bound on the average welfare regret of order $\sqrt{(1+\frac{K}{N})T}$ ignoring polylog factors.[1] We recover the same bound as a special case of our bound when $G$ is a clique and $d = 1$. In the clique case our bound is also similar to the bound $\sqrt{\frac{K}{N}(T\ln K)}$ achieved by Seldin et al. (2014) in a single-agent bandit setting where, at each time step, the agent can choose a subset of $N \leq K$ actions and observe their loss. In the case when $N = 1$ (single agent), our analysis can be applied to the nonstochastic bandit problem where the player observes the loss of each played action with a delay of $d$ steps. In this case we improve on the previous result of $\sqrt{(d+1)KT}$ by Neu et al. (2010, 2014), and give the first characterization (up to logarithmic factors) of the minimax regret, which is of order $\sqrt{(d+K)\,T}$.

1. The rate proven in (Awerbuch and Kleinberg, 2008, Theorem 2.1) has a worse dependence on $T$, but we believe this is due to the fact that their setting allows for dishonest agents and agent-specific loss vectors.

In principle, the problem of delays in online learning could be tackled by simple reductions. Yet, these reductions give rise to suboptimal results. In the single agent setting, where the delay is constant and equal to $d$, one can use the technique of Weinberger and Ordentlich (2002) and run $d+1$ instances of an online algorithm for the nondelayed case, where each instance is used every $d+1$ steps. This delivers a suboptimal regret bound of $\sqrt{(d+1)KT}$. In the case of multiple delays, like in our multi-agent setting, one can repeat the same action for $d+1$ steps while accumulating information from the other agents, and then perform an update on scaled-up losses. The resulting (suboptimal) bound on the average welfare regret would be of the form $\sqrt{(d+1)\big(1+\frac{K}{N}\alpha_{\leq d}\big)(T\ln K)}$.

Rather than using reductions, the analysis of Exp3-Coop rests on quantifying the performance of suitable importance weighted estimates. In fact, in the single-agent setting with delay parameter $d$, using Exp3-Coop reduces to running the standard Exp3 algorithm performing an update as soon a new loss becomes available. This implies that at any round $t > d$, Exp3 selects an action without knowing the losses incurred during the last $d$ rounds. The resulting regret is bounded by relating the standard analysis of Exp3 to a detailed quantification of the extent to which the distribution maintained by Exp3 can drift in $d$ steps.

In the multi-agent case, the importance weighted estimate of Exp3-Coop is designed in such a way that at each time $t > d$ the instance of the algorithm run by an agent $v$ updates all actions that were played at time $t-d$ by agent $v$ or by other agents not further away than $d$ from $v$. Compared to the single agent case, here each agent can exploit the information circulated by the other agents. However, in order to compute the importance weighted estimates used locally by each agent, the probabilities maintained by the agents must be propagated together with the observed losses. Here, further concerns may show up, like the amount of communication, and the location of each agent within the network. In particular, when $G$ has sparse components, we show that a variant of Exp3-Coop, allowing agents to choose their parameters according to their centrality within $G$, strictly improves on the regret of Exp3-Coop.

Finally, we propose a second variant of Exp3-Coop where each agent is able to use loss information as soon as it becomes available. This implies that the updates performed at time $t$ now involve losses with different delays. For this reason, the new variant of Exp3-Coop combines many loss estimators, each defined for a different level of delay, through a fixed distribution $D$. In the analysis we show how the introduction of combined loss estimators affects the average welfare regret.

A preliminary version of this work appeared as an extended abstract in (Cesa-Bianchi et al., 2016).

## 2. Additional Related Work

Many important ideas in delayed online learning, including the observation that the effect of delays can be limited by controlling the amount of change in the agent strategy, were introduced by Mesterharm (2005) —see also (Mesterharm, 2007, Chapter 8). A more recent investigation on delayed online learning is due to Neu et al. (2010, 2014), who analyzed exponential weights with delayed feedbacks. Furher progress is made by Joulani et al. (2013), who also study delays in the general partial monitoring setting. Additional works (Joulani

et al., 2016; Quanrud and Khashabi, 2015) prove regret bounds for the full-information case of the form $\sqrt{(D+T)\ln K}$, where $D$ is the total delay experienced over the $T$ rounds. In the stochastic case, bandit learning with delayed feedback was considered by Dudík et al. (2011); Joulani et al. (2013) and in a harder anonymized model by Pike-Burke et al. (2018).

To the best of our knowledge, the first paper about nonstochastic cooperative bandit networks is (Awerbuch and Kleinberg, 2008). More papers analyze the stochastic setting, and the closest one to our work is perhaps (Szorenyi et al., 2013). In that paper, delayed loss estimates in a network of cooperating stochastic bandits are analyzed using a dynamic P2P random networks as communication model. A more recent paper is (Landgren et al., 2015), where the communication network is a fixed graph and a cooperative version of the UCB algorithm is introduced which uses a distributed consensus algorithm to estimate the mean rewards of the arms. The main result is an individual (per-agent) regret bound that depends on the network structure without taking delays into account.

Another interesting paper about cooperating bandits in a stochastic setting is (Kar et al., 2011). Similar to our model, agents sit on the nodes of a communication network. However, only one designated agent observes the rewards of actions he selects, whereas the others remain in the dark. This designated agent broadcasts his sampled actions through the networks to the other agents, who must learn their policies relying only on this indirect feedback. The paper shows that in any connected network this information is sufficient to achieve asymptotically optimal regret. Cooperative bandits with asymmetric feedback are also studied by Barrett and Stone (2011). In their model, an agent must teach the reward distribution to another agent while keeping the discounted regret under control. Tekin and van der Schaar (2015) investigate a stochastic contextual bandit model where each agent can either privately select an action or have another agent select an action on his behalf. In a related paper, Tekin et al. (2014) look at a stochastic bandit model with combinatorial actions in a distributed recommender system setting, and study incentives among agents who can now recommend items taken from other agents' inventories. A more recent paper (Kolla et al., 2016) studies the performance of the UCB algorithm by Auer et al. (2002a) in a multi-agent setting, where at each time step each agent observes also the losses of actions chosen by any agent located in the same neighborhood of the communication network.

A further line of relevant work involves problems of decentralized bandit coordination. For example, Stranders et al. (2012) consider a bandit coordination problem where the reward function is global and can be represented as a factor graph in which each agent controls a subset of the variables.

A parallel thread of research concerns networks of bandits that compete for shared resources. A paradigmatic application domain is that of cognitive radio networks, in which a number of channels are shared among many users and any two or more users interfere whenever they simultaneously try to use the same channel. The resulting bandit problem is one of coordination in a competitive environment, because every time two or more agents select the same action at the same time step they both get a zero reward due to the interference —see (Rosenski et al., 2015) for recent work on stochastic competitive bandits and (Kleinberg et al., 2009) for a study of more general congestion games in a game-theoretic setting.

Finally, there exists an extensive literature on the adaptation of gradient descent and related algorithms to distributed computing settings, where asynchronous processors natu-

rally introduce delays —see, e.g., (Zinkevich et al., 2009; Agarwal and Duchi, 2011; Li et al., 2013; McMahan and Streeter, 2014; Quanrud and Khashabi, 2015; Liu et al., 2015; Duchi et al., 2015). However, none of these works considers bandit settings, which are an essential ingredient for our analysis.

## 3. Preliminaries

We now establish our notation, along with basic assumptions and preliminary facts related to our algorithms. Notation and setting here both refer to the single agent case. The cooperative setting with multiple agents (and notation thereof) will be introduced in Section 4.

Let $A = \{1, \dots, K\}$ be the action set. A learning agent runs an exponentially-weighted algorithm with weights $w_t(i)$, and learning rate $\eta > 0$. Initially, $w_1(i) = 1$ for all $i \in A$. At each time step $t = 1, 2, \dots$, the agent draws action $I_t$ with probability $\mathbb{P}(I_t = i) = p_t(i) = w_t(i)/W_t$, where $W_t = \sum_{j \in A} w_t(j)$. After observing the loss $\ell_t(I_t) \in [0, 1]$ associated with the chosen action $I_t$, and possibly some additional information, the agent computes, for each $i \in A$, nonnegative loss estimates $\widehat{\ell}_t(i)$, and performs the exponential update

$$w_{t+1}(i) = p_t(i) \exp\bigl(-\eta\, \widehat{\ell}_t(i)\bigr) \tag{1}$$

to these weights. The following two lemmas are general results that control the evolution of the probability distributions in the exponentially-weighted algorithm. As we said in the introduction, bounding the extent to which the distribution used by our algorithms can drift in $d$ steps is key to controlling regret in a delayed setting. The first result bounds the *additive* change in the probability of any action, and it holds no matter how $\widehat{\ell}_t(i)$ is defined.

**Lemma 1** *Under the update rule (1), for all $t \geq 1$ and for all $i \in A$,*

$$-\eta\, p_t(i)\widehat{\ell}_t(i) \leq p_{t+1}(i) - p_t(i) \leq \eta\, p_{t+1}(i) \sum_{j \in A} p_t(j)\widehat{\ell}_t(j)$$

*holds deterministically with respect to the agent's randomization.*

**Proof** Directly from the definition of the update (1), $w_{t+1}(i) \leq p_t(i)$ for all $i \in A$, so that $W_{t+1} \leq 1$, which in turn implies $w_{t+1}(i) \leq w_{t+1}(i)/W_{t+1} = p_{t+1}(i)$. Therefore

$$p_{t+1}(i) - p_t(i) \geq w_{t+1}(i) - p_t(i) = p_t(i)\left(e^{-\eta\, \widehat{\ell}_t(i)} - 1\right) \geq -\eta\, p_t(i)\widehat{\ell}_t(i)\ ,$$

the last inequality using $1 - e^{-x} \leq x$, for $x \geq 0$. Similarly,

$$\begin{aligned}
p_{t+1}(i) - p_t(i) &\leq p_{t+1}(i) - w_{t+1}(i) \\
&= p_{t+1}(i) - p_{t+1}(i) W_{t+1} \\
&= p_{t+1}(i) \sum_{j \in A} \bigl(p_t(j) - w_{t+1}(j)\bigr) \\
&= p_{t+1}(i) \sum_{j \in A} p_t(j) \left(1 - e^{-\eta\, \widehat{\ell}_t(j)}\right) \\
&\leq \eta\, p_{t+1}(i) \sum_{j \in A} p_t(j)\widehat{\ell}_t(j)\ ,
\end{aligned}$$

hence concluding the proof. ■

The second result delivers a *multiplicative* bound on the change in the probability of any action when the loss estimates $\widehat{\ell}_t(i)$ are of the following form:

$$\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_{t-d}(i)}{q_{t-d}(i)} B_{t-d}(i) & \text{if } t > d, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where $d \geq 0$ is a delay parameter, $B_{t-d}(i) \in \{0,1\}$, for $i \in A$, are indicator functions, and $q_{t-d}(i) \geq p_{t-d}(i)$ for all $i$ and $t > d$. In all later sections, $B_{t-d}(i)$ will be instantiated to the indicator function of the event that action $i$ has been played at time $t-d$ by some agent, and $q_{t-d}(i)$ will be the (conditional) probability of this event.

**Lemma 2** *Let $\widehat{\ell}_t(i)$ be of the form (2) for each $t \geq 1$ and $i \in A$. If $\eta \leq \frac{1}{Ke(d+1)}$ in the update rule (1), then*

$$p_{t+1}(i) \leq \left(1 + \frac{1}{d}\right) p_t(i)$$

*holds for all $t \geq 1$ and $i \in A$, deterministically with respect to the agent's randomization.*

**Proof** We proceed by induction over $t$. For all $t \leq d$, $\widehat{\ell}_t(\cdot) = 0$. Hence $p_t(\cdot) = 1/K$, and the lemma trivially holds. For $t > d$ we can write

$$\begin{aligned}
\sum_{i \in A} p_t(i)\widehat{\ell}_t(i) &= \sum_{i \in A} p_t(i) \frac{\ell_{t-d}(i)}{q_{t-d}(i)} B_{t-d}(i) \\
&\leq \sum_{i \in A} \frac{p_t(i)}{q_{t-d}(i)} && \text{(because } B_{t-d}(i)\ell_{t-d}(i) \leq 1\text{)} \\
&\leq \sum_{i \in A} \left(1 + \frac{1}{d}\right)^d \frac{p_{t-d}(i)}{q_{t-d}(i)} && \text{(by the inductive hypothesis)} \\
&\leq \left(1 + \frac{1}{d}\right)^d K && \text{(because } q_{t-d}(i) \geq p_{t-d}(i)\text{)} \\
&\leq Ke \ .
\end{aligned}$$

Hence, using Lemma 1,

$$p_{t+1}(i)\big(1 - \eta\, Ke\big) \leq p_{t+1}(i) \left(1 - \eta \sum_{j \in A} p_t(j)\widehat{\ell}_t(j)\right) \leq p_t(i)$$

which implies $p_{t+1}(i) \leq \big(1 + \frac{1}{d}\big)\, p_t(i)$ whenever $\eta \leq \frac{1}{Ke(d+1)}$ . ■

As we said in Section 1, the idea of controlling the drift of the probabilities in order to bound the effects of delayed feedback is not new. In particular, variants of Lemma 1 were already derived in the work of Neu et al. (2010, 2014). However, Lemma 2 appears to be new, and this is the key result to achieving our improvements.

## 4. The Cooperative Setting on a Communication Network

In our multi-agent bandit setting, there are $N$ agents sitting on the vertices of a connected and undirected communication graph $G = (V, E)$, with $V = \{1, \dots, N\}$. The agents cooperate to solve the same instance of a nonstochastic bandit problem while limiting the communication among them. Let $N_s(v)$ be the set of nodes $v' \in V$ whose shortest-path distance $\mathrm{dist}_G(v, v')$ from $v$ in $G$ is exactly $s$. At each time step $t = 1, 2, \dots$, each agent $v \in V$ draws an action $I_t(v)$ from the common action set $A$. Note that each action $i \in A$ delivers the same loss $\ell_t(i) \in [0, 1]$ to all agents $v$ such that $I_t(v) = i$. At the end of round $t$, each agent $v$ observes his own loss $\ell_t\big(I_t(v)\big)$, and sends to his neighbors in $G$ the message

$$m_t(v) = \Big\langle t, v, I_t(v), \ell_t\big(I_t(v)\big), \boldsymbol{p}_t(v) \Big\rangle$$

where $\boldsymbol{p}_t(v) = \big(p_t(1, v), \dots, p_t(K, v)\big)$ is the distribution of $I_t(v)$. Moreover, $v$ also receives from his neighbors a variable number of messages $m_{t-s}(v')$. Each message $m_{t-s}(v')$ that $v$ receives from a neighbor is used to update $\boldsymbol{p}_t(v)$ and then forwarded to the other neighbors only if $s < d$, otherwise it is dropped.[2] Here $d$ is the maximum delay, a parameter of the communication protocol. Therefore, at the end of round $t$, each agent $v$ receives one message $m_{t-s}(v')$ from each agent $v'$ such that $\mathrm{dist}_G(v, v') = s$, where $s \in \{1, \dots, d\}$. Graph $G$ can thus be seen as a synchronous multi-hop communication network where messages are broadcast, each hop causing a delay of one time step. Our learning protocol is summarized in Figure 1, while Figure 2 contains a pictorial example.

Our model is similar to the LOCAL communication model in distributed computing (Linial, 1992; Suomela, 2013), where the output of a node depends only on the inputs of other nodes in a constant-size neighborhood of it, and the goal is to derive algorithms whose running time is independent of the network size. (The main difference is that the task here has no completion time, however, also in our model influence on a node is only through a constant-size neighborhood of it.)

One aspect deserving attention is that, apart from the common delay parameter $d$, the agents need not share further information. In particular, the agents need not know neither the topology of the graph $G$ nor the total number of agents $N$. In Section 5, we show that our distributed algorithm can also be analyzed when each agent $v$ uses a personalized delay $d(v)$, thus doing away with the need of a common delay parameter, and guaranteeing a generally better performance.

Further graph notation is needed at this point. Given $G$ as above, let us denote by $G_{\le d}$ the graph $(V, E_{\le d})$ where $(u, v) \in E_{\le d}$ if and only if the shortest-path distance between agents $u$ and $v$ in $G$ is *at most* $d$ (hence $G_{\le 1} = G$). Graph $G_{\le d}$ is sometimes called the $d$-th power of $G$. We also use $G_0$ to denote the graph $(V, \emptyset)$. Recall that an independent set of $G$ is any subset $T \subseteq V$ such that no two $i, j \in T$ are connected by an edge in $E$. The largest size of an independent set is the *independence number* of $G$, denoted by $\alpha(G)$. Let $d_G$ be the *diameter* of $G$ (maximal length over all possible shortest paths between all pairs of nodes); then $G_{\le d_G}$ is a clique, and one can easily see that $N = \alpha(G_0) > \alpha(G) \ge$

2. Dropping messages older than $d$ rounds is clearly immaterial with respect to proving bandit regret bounds. We added this feature just to prove a point about the message complexity of the protocol. See Remark 10 in Section 5 for further discussion.

**The cooperative bandit protocol**
**Parameters:** Undirected communication graph $G = (V, E)$, hidden loss vectors $\boldsymbol{\ell}_t = \big(\ell_t(1), \ldots, \ell_t(K)\big) \in [0,1]^K$ for $t \geq 1$, delay $d$.
**For** $t = 1, 2, \ldots$

1. Each agent $v \in V$ plays action $I_t(v) \in A$ drawn according to distribution $\boldsymbol{p}_t(v)$;
2. Each agent $v \in V$ observes loss $\ell_t\big(I_t(v)\big)$, sends to his neighbors the message $m_t(v)$, and receives from his neighbors messages $m_{t-s}(v')$;
3. Each agent $v \in V$ drops any message $m_{t-s}(v')$ received from some neighbor such that $s \geq d$, and forwards to the other neighbors the remaining messages.

Figure 1: The cooperative bandit protocol where all agents share the same delay parameter $d$.

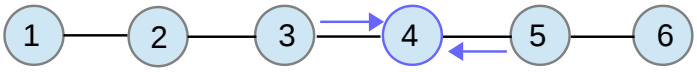


Figure 2: In this example, $G$ is a line graph with $N = 6$ agents, and delay $d = 2$. At the end of time step $t$, Agent 4 sends to his neighbors 3 and 5 message $m_t(4)$, receives from Agent 3 messages $m_{t-1}(3)$ and $m_{t-2}(2)$, and from Agent 5 messages $m_{t-1}(5)$ and $m_{t-2}(6)$. Finally, Agent 4 forwards to Agent 5 message $m_{t-1}(3)$ and forwards to Agent 3 message $m_{t-1}(5)$. Any message older than $t-1$ received by Agent 4 at the end of round $t$ will not be forwarded to his neighbors.

$\alpha(G_{\leq 2}) \geq \cdots \geq \alpha(G_{\leq d_G}) = 1$. We show in Section 4.1 that the collective performance of our algorithms depends on $\alpha(G_{\leq d})$. If the graph $G$ under consideration is directed (see Section 5), then $\alpha(G)$ is the independence number of the undirected graph obtained from $G$ by disregarding edge orientation.

The adversary generating losses is oblivious: loss vectors $\boldsymbol{\ell}_t = \big(\ell_t(1), \ldots, \ell_t(K)\big) \in [0,1]^K$ do not depend on the agents' internal randomization. The agents' goal is to control the *average welfare* regret $R_T^{\text{coop}}$, defined as

$$R_T^{\text{coop}} = \left( \frac{1}{N} \sum_{v \in V} \mathbb{E}\left[ \sum_{t=1}^{T} \ell_t\big(I_t(v)\big) \right] - \min_{i \in A} \sum_{t=1}^{T} \ell_t(i) \right) ,$$

the expectation being with respect to the internal randomization of each agent's algorithm. In the sequel, we write $\mathbb{E}_t[\cdot]$ to denote the expectation w.r.t. the product distribution $\prod_{v \in V} \boldsymbol{p}_t(v)$, conditioned on $I_1(v), \ldots, I_{t-1}(v)$, $v \in V$.

### 4.1. The Exp3-Coop algorithm

Our first algorithm, called Exp3-Coop (Cooperative Exp3) is described in Figure 3. The algorithm works in the learning protocol of Figure 1. Each agent $v \in V$ runs the exponentially-weighted algorithm (1), combined with a "delayed" importance-weighted loss estimate

**The Exp3-Coop Algorithm**
**Parameters:** Undirected communication graph $G = (V, E)$; delay $d$; learning rate $\eta$.
**Init:** Each agent $v \in V$ sets weights $w_1(i, v) = 1$ for all $i \in A$.
**For** $t = 1, 2, \ldots$

1. Each agent $v \in V$ plays action $I_t(v) \in A$ drawn according to distribution $\boldsymbol{p}_t(v) = (p_t(1, v), \ldots, p_t(K, v))$ , where

$$p_t(i, v) = \frac{w_t(i, v)}{W_t(v)},\ i = 1, \ldots, K, \qquad \text{and} \qquad W_t(v) = \sum_{j \in A} w_t(j, v)\ ;$$

2. Each agent $v \in V$ observes loss $\ell_t\big(I_t(v)\big)$ and exchanges messages with his neighbors (Steps 2 and 3 of the protocol in Figure 1);
3. Each agent $v \in V$ performs the update $w_{t+1}(i, v) = p_t(i, v)\, \exp\big(-\eta\, \widehat{\ell}_t(i, v)\big)$ for all $i \in A$, where

$$\widehat{\ell}_t(i, v) = \begin{cases} \dfrac{\ell_{t-d}(i)}{q_{d,t-d}(i, v)}\, B_{d,t-d}(i, v) & \text{if } t > d, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and $B_{d,t-d}(i, v) = \mathbb{I}\{\exists v' \in N_{\le d}(v) \,:\, I_{t-d}(v') = i\}$ with

$$q_{d,t-d}(i, v) = 1 - \prod_{v' \in N_{\le d}(v)} \big(1 - p_{t-d}(i, v')\big)\ .$$

Figure 3: The Exp3-Coop algorithm where all agents share the same delay parameter $d$.

$\widehat{\ell}_t(i, v)$ that incorporates the delayed information sent by the other agents. Specifically, denote by $N_{\le d}(v) = \bigcup_{s \le d} N_s(v)$ the set of nodes in $G$ whose shortest-path distance from $v$ is at most $d$, and note that, for all $v$, $\{v\} = N_{\le 0}(v) \subseteq N_{\le 1}(v) \subseteq N_{\le 2}(v) \subseteq \cdots$ . If any of the agents in $N_{\le d}(v)$ has played at time $t - d$ action $i$ (that is, $B_{d,t-d}(i, v) = 1$ in (3)), then the corresponding loss $\ell_{t-d}(i)$ is incorporated by $v$ into $\widehat{\ell}_t(i, v)$. The denominator $q_{d,t-d}(i, v)$ is simply, conditioned on the history, the probability of $B_{d,t-d}(i, v) = 1$, i.e., $q_{d,t-d}(i, v) = \mathbb{E}_t[B_{d,t-d}(i, v)]$. Observe that $\{v\} \subseteq N_{\le d}(v)$ for all $d \ge 0$ implies $q_{d,t-d}(i, v) \ge p_{t-d}(i, v)$, as required by (2) so as to make Lemma 2 hold. It is also worth mentioning that, despite this is not strictly needed by our learning protocol, each agent $v$ actually exploits the loss information gathered from playing action $I_t(v)$ only $d$ time steps later. An extension of Exp3-Coop, where each loss is exploited as soon as it is made available to an agent, is analyzed in Section 6. Section 7, instead, studies an important special case of this setting where there is a single bandit agent receiving delayed feedback.

By their very definition, the loss estimates $\widehat{\ell}_t(\cdot, \cdot)$ at time $t$ are determined by the realizations of $I_s(\cdot)$, for $s = 1, \ldots, t - d$. This implies that the numbers $p_t(\cdot, \cdot)$ defining $q_{d,t-d}(\cdot, \cdot)$, are determined by the realizations of $I_s(\cdot)$ for $s = 1, \ldots, t - d - 1$ (because the probabilities $\boldsymbol{p}_t(v)$ at time $t$ are determined by the loss estimates up to time $t - 1$, see (1)). We have, for

all $t > d$, $i \in A$, and $v \in V$,

$$\mathbb{E}_{t-d}\Big[\widehat{\ell}_t(i,v)\Big] = \ell_{t-d}(i) \ . \tag{4}$$

Further, because of what we just said about $p_t(\cdot,\cdot)$ and $q_{d,t-d}(\cdot,\cdot)$ being determined by $I_1(\cdot), \ldots, I_{t-d-1}(\cdot)$, we also have

$$\mathbb{E}_{t-d}\Big[p_t(i,v)\widehat{\ell}_t(i,v)\Big] = p_t(i,v)\ell_{t-d}(i) \ , \quad \mathbb{E}_{t-d}\Big[p_t(i,v)\widehat{\ell}_t(i,v)^2\Big] = p_t(i,v)\frac{\ell_{t-d}(i)^2}{q_{d,t-d}(i,v)} \ . \tag{5}$$

The next lemma relates the variance of the estimates (3) to the structure of the communication graph $G$. The lemma is stated for a generic undirected communication graph $G$, but our application of it actually involves graph $G_{\le d}$.

**Lemma 3** *Let $G = (V,E)$ be an undirected graph with independence number $\alpha(G)$. For each $v \in V$, let $N_{\le 1}(v)$ be the neighborhood of node $v$ (including $v$ itself), and $\boldsymbol{p}(v) = \big(p(1,v), \ldots, p(K,v)\big)$ be a probability distribution over $A = \{1, \ldots, K\}$. Then, for all $i \in A$,*

$$\sum_{v \in V} \frac{p(i,v)}{q(i,v)} \le \frac{1}{1-e^{-1}} \left( \alpha(G) + \sum_{v \in V} p(i,v) \right) \quad \textit{where} \quad q(i,v) = 1 - \prod_{v' \in N_{\le 1}(v)} \big(1 - p(i,v')\big) \ .$$

**Proof** Fix $i \in A$ and set for brevity $P(i,v) = \sum_{v' \in N_{\le 1}(v)} p(i,v')$. We can write

$$\sum_{v \in V} \frac{p(i,v)}{q(i,v)} = \underbrace{\sum_{v \in V \,:\, P(i,v) \ge 1} \frac{p(i,v)}{q(i,v)}}_{\text{(I)}} + \underbrace{\sum_{v \in V \,:\, P(i,v) < 1} \frac{p(i,v)}{q(i,v)}}_{\text{(II)}} \ ,$$

and proceed by upper bounding the two terms (I) and (II) separately. Let $r(v)$ be the cardinality of $N_{\le 1}(v)$. We have, for any given $v \in V$,

$$\min \left\{ q(i,v) \ : \sum_{v' \in N_{\le 1}(v)} p(i,v') \ge 1 \right\} = 1 - \left(1 - \frac{1}{r(v)}\right)^{r(v)} \ge 1 - e^{-1} \ .$$

The equality is due to the fact that the minimum is achieved when $p(i,v') = \frac{1}{r(v)}$ for all $v' \in N_{\le 1}(v)$, and the inequality comes from $r(v) \ge 1$ (for, $v \in N_{\le 1}(v)$). Hence

$$\text{(I)} \le \sum_{v \in V \,:\, P(i,v) \ge 1} \frac{p(i,v)}{1-e^{-1}} \le \sum_{v \in V} \frac{p(i,v)}{1-e^{-1}} \ .$$

As for (II), using the inequality $1 - x \le e^{-x}, x \in [0,1]$, with $x = p(i,v')$, we can write

$$q(i,v) \ge 1 - \exp\left( - \sum_{v' \in N_{\le 1}(v)} p(i,v') \right) = 1 - \exp\left(-P(i,v)\right) \ .$$

In turn, because $P(i,v) < 1$ in terms (II), we can use the inequality $1 - e^{-x} \geq (1 - e^{-1})\, x$, holding when $x \in [0,1]$, with $x = P(i,v)$, thereby concluding that $q(i,v) \geq (1-e^{-1})P(i,v)$. Thus

$$\text{(II)} \leq \sum_{v \in V\,:\,P(i,v)<1} \frac{p(i,v)}{(1-e^{-1})P(i,v)} \leq \frac{1}{1-e^{-1}} \sum_{v \in V} \frac{p(i,v)}{P(i,v)} \leq \frac{\alpha(G)}{1-e^{-1}}$$

where in the last step we used (Alon et al., 2017, Lemma 10). Note that despite the statement of this lemma refers to a directed graph and its maximum acyclic subgraph, in the special case of undirected graphs, the size of the maximum acyclic subgraph coincides with the independence number. Moreover, observe that $p(i,1), \dots, p(i,N) \geq 0$ need not sum to one in order for this lemma to hold. ■

The following theorem quantifies the behavior of Exp3-Coop in terms of a free parameter $\gamma$ in the learning rate, the tuning of which will be addressed in the subsequent Theorem 5.

**Theorem 4** *The regret of* Exp3-Coop *run over a network $G = (V,E)$ of $N$ agents, each using delay $d$ and learning rate $\eta = \gamma \big/ \big(Ke(d+1)\big)$, for $\gamma \in (0,1]$, satisfies*

$$R_T^{\mathrm{coop}} \leq 2d + \frac{Ke(d+1)\ln K}{\gamma} + \gamma \left( \frac{\alpha(G_{\leq d})}{2(1-e^{-1})(d+1)N} + \frac{3}{Ke} \right) T \ .$$

**Proof** The standard analysis of the exponentially-weighted algorithm with importance-sampling estimates —see, e.g., the proof of (Alon et al., 2017, Lemma 1)— gives for each agent $v$ and each action $k$ the deterministic bound

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v) \leq \sum_{t=1}^{T} \widehat{\ell}_t(k,v) + \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v)^2 \ . \tag{6}$$

We take expectations of the three (double) sums in (6) separately. As for the first sum, note that an iterative application of Lemma 1 gives, for $t > d$,

$$p_t(i,v) \geq p_{t-d}(i,v) - \eta \sum_{s=1}^{d} p_{t-s}(i,v) \widehat{\ell}_{t-s}(i,v) \ ,$$

so that, setting for brevity $A_t(i,v) = \sum_{s=1}^{d} p_{t-s}(i,v) \widehat{\ell}_{t-s}(i,v)$, we have

$$\begin{aligned} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v) &\geq \sum_{t=2d+1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v) \\ &\geq \sum_{t=2d+1}^{T} \sum_{i=1}^{K} p_{t-d}(i,v) \widehat{\ell}_t(i,v) - \eta \sum_{t=2d+1}^{T} \sum_{i=1}^{K} A_t(i,v)\, \widehat{\ell}_t(i,v) \ . \end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i,v)\widehat{\ell}_t(i,v)\right] &\geq \mathbb{E}\left[\sum_{t=2d+1}^{T}\sum_{i=1}^{K} p_{t-d}(i,v)\widehat{\ell}_t(i,v)\right] - \eta\,\mathbb{E}\left[\sum_{t=2d+1}^{T}\sum_{i=1}^{K} A_t(i,v)\,\widehat{\ell}_t(i,v)\right] \\
&= \mathbb{E}\left[\sum_{t=2d+1}^{T}\sum_{i=1}^{K} p_{t-d}(i,v)\,\mathbb{E}_{t-d}\left[\widehat{\ell}_t(i,v)\right]\right] - \eta\,\mathbb{E}\left[\sum_{t=2d+1}^{T}\sum_{i=1}^{K} A_t(i,v)\,\mathbb{E}_{t-d}\left[\widehat{\ell}_t(i,v)\right]\right] \\
&\quad \text{(since } p_t(i,v) \text{ is determined by } I_1(\cdot),\ldots,I_{t-d-1}(\cdot)\text{)} \\
&= \mathbb{E}\left[\sum_{t=2d+1}^{T}\sum_{i=1}^{K} p_{t-d}(i,v)\,\ell_{t-d}(i)\right] - \eta\,\mathbb{E}\left[\sum_{t=2d+1}^{T}\sum_{i=1}^{K} A_t(i,v)\,\ell_{t-d}(i)\right] \qquad \text{(using (4))} \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i,v)\,\ell_t(i)\right] - 2d - \eta\,T\,d\ .
\end{aligned}$$

The last step uses

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{K} A_t(i,v)\,\ell_{t-d}(i)\right] &\leq \mathbb{E}\left[\sum_{i=1}^{K} A_t(i,v)\right] = \mathbb{E}\left[\sum_{i=1}^{K}\sum_{s=1}^{d} p_{t-s}(i,v)\widehat{\ell}_{t-s}(i,v)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{s=1}^{d} p_{t-s}(i,v)\ell_{t-s-d}(i)\right] \leq \mathbb{E}\left[\sum_{i=1}^{K}\sum_{s=1}^{d} p_{t-s}(i,v)\right] = d
\end{aligned}$$

holding for $t \geq 2d+1$. Similarly, for the second sum in (6), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}_t(k,v)\right] = \sum_{t=d+1}^{T}\ell_{t-d}(k) \leq \sum_{t=1}^{T}\ell_t(k)\ .$$

Finally, for the third sum in (6), an iterative application of Lemma 2 yields, for $t > d$,

$$p_t(i,v) \leq \left(1+\frac{1}{d}\right)^d p_{t-d}(i,v) \leq e\,p_{t-d}(i,v)\ ,$$

so that we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i,v)\widehat{\ell}_t(i,v)^2\right] &= \mathbb{E}\left[\sum_{t=d+1}^{T}\sum_{i=1}^{K}\mathbb{E}_{t-d}\left[p_t(i,v)\widehat{\ell}_t(i,v)^2\right]\right] \\
&\leq \mathbb{E}\left[\sum_{t=d+1}^{T}\sum_{i=1}^{K}\frac{p_t(i,v)}{q_{d,t-d}(i,v)}\right] \qquad \text{(using (5) and } \ell_t(\cdot)\leq 1\text{)} \\
&\leq e\,\mathbb{E}\left[\sum_{t=d+1}^{T}\sum_{i=1}^{K}\frac{p_{t-d}(i,v)}{q_{d,t-d}(i,v)}\right]\ .
\end{aligned}$$

The last inequality comes from an iterative application of Lemma 2, and the observation that $\left(1+\frac{1}{d}\right)^d \leq e$.

Hence, summing over all agents $v$, dividing by $N$, and using Lemma 3 on $G_{\leq d}$ gives

$$\begin{aligned}\frac{1}{N}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{v\in V} p_t(i,v)\widehat{\ell}_t(i,v)^2\right] &\leq \frac{e}{N}\,\mathbb{E}\left[\sum_{t=d+1}^{T}\sum_{i=1}^{K}\sum_{v\in V}\frac{p_{t-d}(i,v)}{q_{d,t-d}(i,v)}\right]\\ &\leq \frac{e}{(1-e^{-1})\,N}\,\mathbb{E}\left[\sum_{t=d+1}^{T}\sum_{i=1}^{K}\left(\alpha(G_{\leq d})+\sum_{v\in V} p_{t-d}(i,v)\right)\right]\\ &\leq \frac{e}{1-e^{-1}}\,T\left(\frac{K}{N}\,\alpha(G_{\leq d})+1\right)\ .\end{aligned}$$

Finally, putting together as in (6), setting $\eta=\gamma/\big(Ke(d+1)\big)$, and overapproximating, we obtain the desired bound. ■

We might be tempted to optimize the bound of Theorem 4 for $\gamma$. However, this is not a legal learning rate setting in a distributed scenario, for the optimized value of $\gamma$ would depend on the global quantities $N$ and $\alpha(G_{\leq d})$. Thus, instead of this global tuning, we let each agent set its own learning rate $\gamma$ through a "doubling trick" played locally. The doubling trick[3] works as follows. For each $v\in V$, we let $\gamma_r(v)=Ke(d+1)\sqrt{(\ln K)/2^r}$ for each $r=r_0,r_0+1,\dots$, where $r_0=\big\lceil \log_2\ln K+2\log_2(Ke(d+1))\big\rceil$ is chosen in such a way that $\gamma_r(v)\leq 1$ for all $r\geq r_0$. Let $T_r$ be the random set of consecutive time steps where the same $\gamma_r(v)$ was used. Whenever the local algorithm at $v$ is running with $\gamma_r(v)$ and detects $\sum_{s\in T_r} Q_s(v)>2^r$, where $Q_s(v)$ is the quantity defined in (9) —see the proof of Theorem 5 in the appendix— then we restart this algorithm with $\gamma(v)=\gamma_{r+1}(v)$.

We have the following result (proof in the appendix).

**Theorem 5** *The regret of* Exp3-Coop *run over a network $G=(V,E)$ of $N$ agents, each using delay $d$, and an individual learning rate $\eta(v)=\gamma(v)/\big(Ke(d+1)\big)$, where $\gamma(v)\in(0,1]$ is adaptively selected by each agent through the above doubling trick, satisfies, when $T$ grows large,*[4]

$$R_T^{\mathrm{coop}}=\mathcal{O}\left(\sqrt{(\ln K)\left(d+1+\frac{K}{N}\,\alpha(G_{\leq d})\right)T}+d\,\log T\right)\ .$$

**Remark 6** *Theorem 5 shows a natural trade-off between delay and information. To make it clear, suppose $N\approx K$. In this case, the regret bound becomes of order $\sqrt{\big(d+\alpha(G_{\leq d})\big)T\ln K}+d\ln T$. Now, if $d$ is as big as the diameter $d_G$ of $G$, then $\alpha(G_{\leq d})=1$. This means that at every time step all $N\approx K$ agents observe (with some delay) the losses of each other's actions. This is very much reminiscent of a full information scenario, and in fact our bound becomes of order $\sqrt{(d_G+1)T\ln K}+d_G\ln T$, which is close to the full information minimax rate $\sqrt{(d+1)T\ln K}$ when feedback has a constant*

3. There has been some recent work on adaptive learning rate tuning applied to nonstochastic bandit algorithms (Kocák et al., 2014; Neu, 2015). One might wonder whether the same techniques may apply here as well. Unfortunately, the specific form of our update (1) makes this adaptation nontrivial, and this is why we resorted to a more traditional "doubling trick".

4. The big-oh notation here hides additive terms that are independent of $T$ and do depend polynomially on the other parameters.

*delay $d$ (Weinberger and Ordentlich, 2002). When $G$ is sparse (i.e., $d_G$ is likely to be large, say $d_G \approx N$), then agents have no advantage in taking $d = d_G$ since $d_G \approx N \approx K$. In this case, agents may even give up cooperation (choosing $d = 0$ in Figure 3), and fall back on the standard bandit bound $\sqrt{TK \ln K}$, which corresponds to running* Exp3-Coop *on the edgeless graph $G_0$. (No doubling trick is needed in this case, hence no extra $\log T$ term appears.)*

**Remark 7** *When $d = d_G$, each neighborhood $N_{\leq d}(v)$ used in the loss estimate (3) is equal to $V$, hence all agents receive the same feedback. Because they all start off from the same initial weights, the agents end up computing the same updates. This in turn implies that: (1) the individual regret incurred by each agent is the same as the average welfare regret $R_T^{\text{coop}}$; (2) the messages exchanged by the agents (see Figure 1) may be shortened by dropping the distribution part $\boldsymbol{p}_{t-s}(v')$.*

**Remark 8** *An interesting question is whether the agents can come up with a reasonable choice for the value of $d$ even when they lack any information whatsoever about the global structure of $G$. A partial answer to this question follows. It is easy to show that the choice $d = \sqrt{K}$ in Theorem 5 yields a bound on the average welfare regret of the form $K^{1/4}\sqrt{T \ln K} + \sqrt{K}(\ln T)$ for all $G$ (and irrespective to the value of $N = |V|$), provided $G$ is connected. This holds because, for any connected graph $G$, the independence number $\alpha(G_{\leq d})$ is always bounded by*[5] *$\lceil 2N/(d+2) \rceil$. To see why this latter statement is true, observe that the neighborhood $N_{\leq d/2}(v)$ of any node $v$ in $G_{\leq d/2}$ contains at least $d/2+1$ nodes (including $v$), and any pair of nodes $v', v'' \in N_{\leq d/2}(v)$ are adjacent in $G_{\leq d}$. Therefore, no independent set of $G_{\leq d}$ can have size bigger than $\lceil 2N/(d+2) \rceil$. A more detailed bound is contained, e.g., in (Firby and Haviland, 1997).*

**Remark 9** *The choice of $d$ minimizing the bound $d + \frac{K}{N}\,\alpha(G_{\leq d})$ requires prior knowledge of the independence numbers $\alpha(G_{\leq d})$ for different values of $d$. Computing these quantities is NP-hard in general, even when $G$ is fully known.*

## 5. Extension I: Cooperation with Individual Parameters

In this section, we analyze a modification of Exp3-Coop that allows each agent $v$ in the network to use a delay parameter $d(v)$ different from that of the other agents. We then show how such individual delays may improve the average welfare regret of the agents. In the previous setting, where all agents use the same delay parameter $d$, messages have an implicit time-to-live equal to $d$. In this setting, however, agents may not have a detailed knowledge of the delay parameters used by the other agents. For this reason we allow an agent $v$ to generate messages with a time-to-live $ttl(v)$ possibly different from the delay parameter $d(v)$. Note that the role of the two parameters $d(v)$ and $ttl(v)$ is inherently different. Whereas $d(v)$ rules the extent to which $v$ uses the messages received from the other agents, $ttl(v)$ limits the number of times a message from $v$ is forwarded to the other agents, thereby limiting the message complexity of the algorithm. In order to accomodate this additional parameter, we are required to modify the cooperative bandit protocol of

5. Because it holds for a worst-case (connected) $G$, this upper bound on $\alpha(G_{\leq d})$ can be made tighter when specific graph topologies are considered.

Figure 1. As in Section 4, we have an undirected communication network $G = (V, E)$ over the agents. However, in this new protocol the message that at the end of round $t$ each agent $v$ sends to his neighbors in $G$ has the format

$$m_t(v) = \Big\langle t, v, ttl(v), I_t(v), \ell_t\big(I_t(v)\big), \boldsymbol{p}_t(v) \Big\rangle$$

where $ttl(v)$ is the time-to-live parameter of agent $v$. Each message $m_{t-s}(v')$, which $v$ receives from a neighbor, first has its time-to-leave decremented by one. If the resulting value is positive, the message is forwarded to the other neighbors, otherwise it is dropped. Moreover, $v$ uses this message to update $\boldsymbol{p}_t(v)$ only if $s \leq d(v)$. Hence, at time $t$ an agent $v$ uses the message sent at time $t - s$ by $v'$ if and only if $\mathrm{dist}_G(v', v) = s$ with $s \leq \min\{d(v), ttl(v')\}$, where $\mathrm{dist}_G(v, v')$ is the shortest-path distance from $v'$ to $v$ in $G$.

Based on the collection $\mathcal{P} = \{d(v), ttl(v)\}_{v \in V}$ of individual parameters, we define the *directed* graph $G_{\mathcal{P}} = (V, E_{\mathcal{P}})$ as follows: arc $(v', v) \in E_{\mathcal{P}}$ if and only if $\mathrm{dist}_G(v, v') \leq \min\{d(v), ttl(v')\}$. The in-neighborhood $N^-_{\mathcal{P}}(v)$ of $v$ thus contains the set of all $v' \in V$ whose distance from $v$ is not larger than $\min\{d(v), ttl(v')\}$. Notice that, with this definition, $v \in N^-_{\mathcal{P}}(v)$, so that $(V, E_{\mathcal{P}})$ includes all self-loops $(v, v)$. Figure 4(a) illustrates these concepts through a simple pictorial example.

**Remark 10** *It is important to remark that the communication structure encoded by $\mathcal{P}$ is an exogenous parameter of the regret minimization problem, and so our algorithms cannot trade it off against regret. In addition to that, the parameterization $\mathcal{P} = \{d(v), ttl(v)\}_{v \in V}$ defines a simple and static communication graph which makes it relatively easy to express regret as a function of the amount of available communication. This would not be possible if we had each individual node $v$ decide whether to forward a message based, say, on its own local delay parameter $d(v)$. To see why, consider the situation where nodes $v$ and $v'$ are along the route of a message that is reaching $v$ before $v'$. The decision of $v$ to drop the message may clash with the willingness of $v'$ to receive it, and this may clearly happen when $d(v) < d(v')$. The structure of the communication graph resulting from this individual behavior of the nodes would be rather complicated. On the contrary, the time-to-live-based parametrization, which is commonly used in communication networks to control communication complexity, does not have this issue.*

Figure 5 contains our algorithm (called Exp3-Coop2) for this setting. Exp3-Coop2 is a strict generalization of Exp3-Coop, and so is its analysis. The main difference between the two algorithms is that Exp3-Coop2 deals with directed graphs. This fact prevents us from using the same techniques of Section 4.1 in order to control the regret. Intuitively, adding orientations to the edges reduces the information available to the agents and thus increases the variance of their loss estimates. Thus, in order to control this variance, we need a lower bound[6] on the probabilities $p_t(i, v)$. From Figure 5, one can easily see that

$$1 = \sum_{i \in A} \frac{w_t(i, v)}{W_t(v)} \leq \widetilde{P}_t(v) \leq \sum_{i \in A} \left( \frac{w_t(i, v)}{W_t(v)} + \frac{\delta}{K} \right) = 1 + \delta \tag{7}$$

6. We find it convenient to derive this lower bound without mixing with the uniform distribution over $A$ —see, e.g., (Auer et al., 2002b)— but in a slightly different manner. This facilitates our delayed feedback analysis.

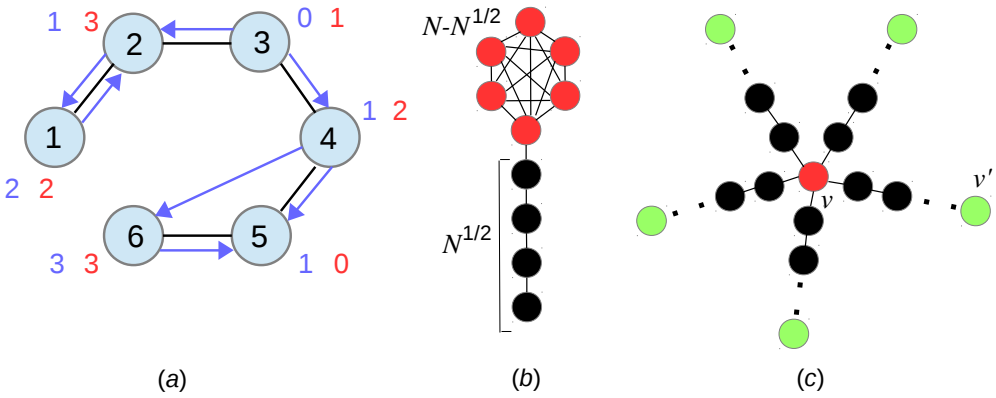


Figure 4: **(a)** In this example, the communication network $G$ is an undirected line graph with $N = 6$ agents, whose edges are depicted in black. Close to each node $v = 1, \ldots, 6$ is the individual delay $d(v)$ (in blue), and the individual time-to-leave $ttl(v)$ (in red). The arcs (a.k.a., directed edges) of the induced directed graph $G_{\mathcal{P}}$ are also depicted in blue. Self-loops are not depicted. For instance, we have $N^-_{\mathcal{P}}(5) = \{4, 5, 6\}$ and $N^-_{\mathcal{P}}(3) = \{3\}$. **(b)** A communication network having a dense (red nodes) and a sparse (black nodes) region. The black region has $N^{1/2}$ agents, the red one has $N - N^{1/2}$ agents. **(c)** A star graph with long rays. The center $v$ (in red) sets a small $d(v)$ and a large $ttl(v)$. The peripheral nodes $v'$ (in green) set a large $d(v')$ and a small $ttl(v')$.

implying the lower bound $p_t(i, v) \geq \frac{\delta}{K(1+\delta)}$, holding for all $i$, $t$, and $v$.

The following theorem (proof in the appendix) is the main result of this section.

**Theorem 11** *The regret of* Exp3-Coop2 *run over a network $G = (V, E)$ of $N$ agents, each agent $v$ using individual delay $d(v)$, individual time-to-leave $ttl(v)$, exploration parameter $\delta = 1/T$, and learning rate $\eta$ such that $\eta \to 0$ as $T \to \infty$ satisfies, when $T$ grows large,*

$$R_T^{\text{coop}} = \mathcal{O}\left(\frac{\ln K}{\eta} + \eta\Big(\bar{d}_V + \frac{K}{N}\,\alpha\left(G_{\mathcal{P}}\right)\,\ln(TNK)\Big)T\right) \quad , \quad \text{where} \quad \bar{d}_V = \frac{1}{N}\sum_{v \in V} d(v) \ .$$

Using a doubling trick in much the same way we used it to prove Theorem 5, we can state the following result (proof in the appendix).

**Corollary 12** *The regret of* Exp3-Coop2 *run over a network $G = (V, E)$ of $N$ agents, each agent $v$ using individual delay $d(v)$, individual time-to-leave $ttl(v)$, exploration parameter $\delta = 1/T$, and individual learning rate $\eta(v)$ adaptively selected by each agent through a doubling trick, satisfies, when $T$ grows large*

$$R_T^{\text{coop}} = \mathcal{O}\left(\sqrt{(\ln K)\left(\bar{d}_V + 1 + \frac{K}{N}\,\alpha(G_{\mathcal{P}})\ln(TNK)\right)T} + \bar{d}_V\left(\ln T + \ln\ln(TNK)\right)\right) \ .$$

To illustrate the advantage of having individual delays as opposed to sharing the same delay value, it suffices to consider a communication network including regions of different density.

**The Exp3-Coop2 Algorithm**
**Parameters:** Undirected graph $G = (V, E)$; learning rate $\eta$; exploration parameter $\delta > 0$.
**Init:** Each $v \in V$ sets weights $w_1(i, v) = 1$, for all $i \in A$, delay $d(v)$, and time-to-live $ttl(v)$.
**For** $t = 1, 2, \ldots$

1. Each agent $v \in V$ plays action $I_t(v) \in A$ drawn according to distribution $\boldsymbol{p}_t(v) = (p_t(1, v), \ldots, p_t(K, v))$ , where

$$p_t(i, v) = \frac{\widetilde{p}_t(i, v)}{\widetilde{P}_t(v)}, \qquad \widetilde{P}_t(v) = \sum_{j \in A} \widetilde{p}_t(j, v),$$

and

$$\widetilde{p}_t(i, v) = \max\left\{\frac{w_t(i, v)}{W_t(v)}, \frac{\delta}{K}\right\}, \qquad W_t(v) = \sum_{j \in A} w_t(j, v)\,;$$

2. Each agent $v \in V$ observes loss $\ell_t\big(I_t(v)\big)$ and exchanges messages with his neighbors (see main text for an explanation);
3. Each agent $v \in V$ performs the update $w_{t+1}(i, v) = p_t(i, v)\, \exp\big(-\eta\, \widehat{\ell}_t(i, v)\big)$ for all $i \in A$, where

$$\widehat{\ell}_t(i, v) = \begin{cases} \dfrac{\ell_{t-d(v)}(i)}{q_{\mathcal{P}, t-d(v)}(i, v)}\, B_{\mathcal{P}, t-d(v)}(i, v) & \text{if } t > d(v), \\ 0 & \text{otherwise,} \end{cases}$$

and $B_{\mathcal{P}, t-d(v)}(i, v) = \mathbb{I}\{\exists v' \in N^-_{\mathcal{P}}(v) \,:\, I_{t-d(v)}(v') = i\}$, with

$$q_{\mathcal{P}, t-d(v)}(i, v) = 1 - \prod_{v' \in N^-_{\mathcal{P}}(v)} \big(1 - p_{t-d(v)}(i, v')\big)\ .$$

Figure 5: The Exp3-Coop2 algorithm with individual delay and time-to-live parameters.

Concretely, consider the graph in Figure 4(b) with a large densely connected region (red agents) and a small sparsely connected region black agents). In this example, the black agents prefer a large value of their individual delay so as to receive more information from nearby agents, but this comes at the price of a larger bias for their estimators $\widehat{\ell}_t(i, v)$. On the contrary, information from nearby agents is readily available to the red agents, so that they do not gain any regret improvement from a large delay parameter. A similar argument applies here to the individual time-to-live values: red agents $v$ will set a small $ttl(v)$ to reduce communication. Black agents $v'$ may decide to set $ttl(v')$ depending on their intention to reach the red nodes. But because the red agents have set a small $d(v)$, any effort made by $v'$ trying to reach them would be a communication waste. Hence, it is reasonable for a black

agent $v'$ to set a moderately large value for $ttl(v')$, but perhaps not so large as to reach the red agents. One can read this off the bounds in both Theorem 11 and Corollary 12, as explained next. Suppose for simplicity that $K \approx N$ so that, disregarding log factors, these bounds depend on parameters $\mathcal{P}$ only through the quantity $H = \bar{d}_V + \alpha(G_{\mathcal{P}})$. Now, in the case of a common delay parameter $d$ (Section 4.1), it is not hard to see that the best setting for $d$ in order to minimize $H$ is of the form $d = N^{1/4}$, resulting in $H = \Theta(N^{1/4})$. On the other hand, the best setting for the individual delays is $d(v) = 1$ when $v$ is red, and $d(v) = \sqrt{N}$ when $v$ is black, resulting in $H = \Theta(1)$.

The time-to-live parameters $ttl(v)$ affect the regret bound only through $\alpha(G_{\mathcal{P}})$, but they clearly play the additional role of bounding the message complexity of the algorithm. In our example of Figure 4(b), we essentially have $d(v) \approx ttl(v)$ for all $v$. A typical scenario where agents may have $d(v) \neq ttl(v)$ is illustrated in Figure 4(c). In this case, we have star-like graph where a central agent is connected through long rays to all others agents. The center $v$ prefers to set a small $d(v)$, since it has a large degree, but also a large $ttl(v)$ in order to reach the green peripheral nodes. The green nodes $v'$ are reasonably doing the opposite: a large $d(v')$ in order to gather information from other nodes, but also a smaller time-to-live than the center, for the information transmitted by $v'$ is comparatively less valuable to the whole network than the one transmitted by the center.

Agents can set their individual parameters in a topology-dependent manner using any algorithm for assessing the centrality of nodes in a distributed fashion —e.g., (Wehmuth and Ziviani, 2013), and references therein. This can be done at the beginning in a number of rounds which only depends on the network topology (but not on $T$). Hence, this initial phase would affect the regret bound only by an additive constant.

## 6. Extension II: Cooperation with Mixed Delays

The two algorithms we designed so far do not use the loss information in the most effective way, as they both postpone the update step by $d$ (Figure 3) or $d(v)$ (Figure 5) time steps. The advantage of postponing updates is that loss estimates are simple to design, because the updates at time $t$ all involve losses generated at the same time $t - d$ or $t - d(v)$. In this section, instead, we study generalized versions of Exp3-coop and Exp3-coop2 where all losses $\ell_{t-s}(i)$ sent from agents at distance $s$ to any given agent $v$ are used by $v$ at time $t$; that is, they are used as soon as they become available to $v$. Unlike before, the updates performed at time $t$ now involve losses with different delays, and for this reason these generalized algorithms combine many loss estimators, each defined for a different level of delay, through a fixed distribution $D$ (e.g., a distribution emphasizing recent losses). As we show, in the resulting regret bounds both delays and independence numbers end up correspondingly mixed according to the distribution $D$. In the rest of this section, we present Exp3-Coop-Mix, which generalizes Exp3-Coop to mixed delayed estimators. The learning protocol remains the same (Figure 1). A similar extension exists for Exp3-Coop2, where instead of having individual delays $d(v)$, we have individual distributions over delay values. We decided to omit this further extension from the paper because it does not add any extra value to the overall discussion.

Exp3-Coop-Mix is described in Figure 6, where we simply replaced estimator (3) by its mixed version (8). The next theorem (whose proof is in the appendix) and the subsequent

**The Exp3-Coop-Mix Algorithm**
**Parameters:** Undirected communication graph $G = (V, E)$; maximal delay $d$; delay distribution $D$ over $\{0, 1, \ldots, d-1\}$; learning rate $\eta > 0$.
**Init:** Each agent $v \in V$ sets weights $w_1(i, v) = 1$ for all $i \in A$.
**For** $t = 1, 2, \ldots$

1. Each agent $v \in V$ plays action $I_t(v) \in A$ drawn according to distribution $\boldsymbol{p}_t(v) = (p_t(1, v), \ldots, p_t(K, v))$, where

$$p_t(i, v) = \frac{w_t(i, v)}{W_t(v)},\ i = 1, \ldots, K \qquad \text{and} \qquad W_t(v) = \sum_{j \in A} w_t(j, v)\ ;$$

2. Each agent $v \in V$ observes loss $\ell_t\big(I_t(v)\big)$ and exchanges messages with his neighbors (Steps 2 and 3 of the protocol in Figure 1);
3. Each agent $v \in V$ performs the update $w_{t+1}(i, v) = p_t(i, v)\, \exp\big(-\eta\, \widehat{\ell}_t(i, v)\big)$ for all $i \in A$, where

$$\widehat{\ell}_t(i, v) = \begin{cases} \displaystyle\sum_{s=0}^{d-1} D(s)\, \frac{\ell_{t-s}(i)}{q_{s,t-s}(i, v)}\, B_{s,t-s}(i, v) & \text{if } t > d, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

and $B_{s,t-s}(i, v) = \mathbb{I}\{\exists v' \in N_{\le s}(v)\ :\ I_{t-s}(v') = i\}$ with

$$q_{s,t-s}(i, v) = 1 - \prod_{v' \in N_{\le s}(v)} \big(1 - p_{t-s}(i, v')\big)\ .$$

Figure 6: The Exp3-Coop-Mix algorithm where all agents share the same delay distribution $D$.

discussion investigate the impact of using a distribution over delay levels on the average welfare regret bound.

**Theorem 13** *The regret of* Exp3-Coop-Mix *run over a network $G = (V, E)$ of $N$ agents, each using the same delay distribution $D$, and learning rate $\eta \le 1/(Ke(d+1))$, satisfies*

$$R_T^{\text{coop}} \le 3d + \frac{\ln K}{\eta} + \frac{e^2 \eta\, T}{1 - e^{-1}} \left(1 + \mu_D + \frac{2K}{N}\, \bar{\alpha}_D\right),$$

*where $\bar{\alpha}_D = \sum_{s=0}^{d-1} D(s) \alpha(G_{\le s})$ is the expected independence number of $G$'s power as determined by mixture $D$, and $\mu_D$ is the expectation of distribution $D$.*

Through a doubling trick played locally by each agent, which is utterly similar to the one described before the statement of Theorem 5, we can state (proof in the appendix) the following corollary.

**Corollary 14** *The regret of* Exp3-Coop-Mix *run over a network $G = (V, E)$ of $N$ agents, each using the same delay distribution $D$, and individual learning rate $\eta(v)$ adaptively selected by each agent through a doubling trick, satisfies, when $T$ grows large*

$$R_T^{\text{coop}} = \mathcal{O}\left(\sqrt{(\ln K)\left(\mu_D + 1 + \frac{K}{N}\,\bar{\alpha}_D)\right)T} + \mu_D\,\log T\right)\ .$$

## 7. Delayed Losses (for a Single Agent)

Exp3-Coop can be specialized to the setting where a single agent is facing a bandit problem in which the loss of the chosen action is observed with a fixed delay $d$. In this setting, at the end of each round $t$ the agent incurs loss $\ell_t(I_t)$ and observes $\ell_{t-d}(I_{t-d})$, if $t > d$, and nothing otherwise. The regret is defined in the usual way,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(I_t)\right] - \min_{i=1,\dots,K}\sum_{t=1}^{T}\ell_t(i)\ .$$

This problem was studied by Weinberger and Ordentlich (2002) in the full information case, for which they proved that $\sqrt{(d+1)T\ln K}$ is the optimal order for the minimax regret. The result was extended to the bandit case by Neu et al. (2010, 2014) —see also Joulani et al. (2013)— whose techniques can be used to obtain a regret bound of order $\sqrt{(d+1)KT}$. Yet, no matching lower bound was available for the bandit case.

As a matter of fact, the upper bound $\sqrt{(d+1)KT}$ for the bandit case is easily obtained: just run in parallel $d+1$ instances of the minimax optimal bandit algorithm for the standard (no delay) setting, achieving $R_T \le \sqrt{KT}$ (ignoring constant factors). At each time step $t = (d+1)r + s$ (for $r = 0, 1, \dots$ and $s = 0, \dots, d$), use instance $s+1$ for the current play. Hence, the no-delay bound applies to every instance and, assuming $d+1$ divides $T$, we immediately obtain $R_T \le \sum_{s=1}^{d+1}\sqrt{K\frac{T}{d+1}} \le \sqrt{(d+1)KT}$ , again, ignoring constant factors.

Next, we show that the machinery we developed in Section 4.1 delivers an improved upper bound on the regret for the bandit problem with delayed losses, and then we complement this result by providing a lower bound matching the upper bound up to log factors, thereby characterizing (up to log factors) the minimax regret for this problem.

**Corollary 15** *In the nonstochastic bandit setting with $K \ge 2$ actions and delay $d \ge 0$, where at the end of each round $t$ the predictor has access to the losses $\ell_1(I_1), \dots, \ell_s(I_s) \in [0,1]^K$ for $s = \max\{1, t-d\}$, the minimax regret is of order $\sqrt{(K+d)T}$ , ignoring logarithmic factors.*

**Proof** In order to prove the upper bound, we use the exponentially-weighted algorithm with estimate (3) specialized to the case of one agent only, namely $B_{t-d}(i) = \mathbb{I}\{I_{t-d} = i\}$ and $q_{d,t-d}(i) = p_{t-d}(i)$. Notice that this amounts to running the standard Exp3 algorithm performing an update as soon a new loss becomes available. In this case, because $N = \alpha(G_{\le d}) = 1$, the bound of Theorem 4, with a suitable choice of $\gamma$ (which depends on $T$, $K$, and $d$) reduces to

$$R_T = \mathcal{O}\left(d + \sqrt{(K+d)\,T\ln K}\right)\ .$$

We now prove a lower bound matching our upper bound up to logarithmic factors. The proof hinges on combining the known lower bound $\Omega\big(\sqrt{KT}\big)$ for bandits without delay of Auer et al. (2002b) with the following argument by Weinberger and Ordentlich (2002) that provides a lower bound for the full information case with delay. The proof of the latter bound is by contradiction: we show that a low-regret full information algorithm for delay $d > 0$ can be used to design a low-regret full information algorithm for the $d = 0$ (no delay) setting. We then apply the known lower bound for the minimax regret in the no-delay setting to derive a lower bound for the setting with delay.

Fix $d > 0$ and let $\mathcal{A}$ be a predictor for the full-information online prediction problem with delay $d$. Let $\boldsymbol{p}_t$ be the probability distribution used by $\mathcal{A}$ at time $t$. We now apply algorithm $\mathcal{A}$ to design a new algorithm $\mathcal{A}'$ for a full information online prediction problem with arbitrary loss vectors $\boldsymbol{\ell}'_1, \ldots, \boldsymbol{\ell}'_B \in [0,1]^K$ and no delay. More specifically, we create a sequence $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T \in [0,1]^K$ of loss vectors such that $T = (d+1)B$ and $\boldsymbol{\ell}_t = \boldsymbol{\ell}'_b$ where $b = \big\lceil t/(d+1) \big\rceil$. At each time $b = 1, \ldots, B$ algorithm $\mathcal{A}'$ uses the distribution

$$\boldsymbol{p}'_b = \frac{1}{d+1} \sum_{s=1}^{d+1} \boldsymbol{p}_{(d+1)(b-1)+s}$$

where $\boldsymbol{p}_t = \big(\frac{1}{K}, \ldots, \frac{1}{K}\big)$ for all $t \le 1$. Note that $\boldsymbol{p}'_b$ is defined using $\boldsymbol{p}_{(d+1)(b-1)+1}, \cdots, \boldsymbol{p}_{(d+1)b}$. These are in turn defined using the same loss vectors $\boldsymbol{\ell}'_1, \ldots, \boldsymbol{\ell}'_{b-1}$ since, by definition, each $\boldsymbol{p}_{t+1}$ uses $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{t-d}$, and $\big\lceil (t-d)/(d+1) \big\rceil = b-1$ for all $t = (d+1)(b-1), \ldots, (d+1)b-1$. So $\mathcal{A}'$ is a legitimate full-information online algorithm for the problem $\boldsymbol{\ell}'_1, \ldots, \boldsymbol{\ell}'_B$ with no delay. As a consequence,

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{i=1}^{K} \ell_t(i) p_t(i) &= \sum_{b=1}^{B} \sum_{s=1}^{d+1} \sum_{i=1}^{K} \ell'_b(i) p_{(d+1)(b-1)+s}(i) \\
&= (d+1) \sum_{b=1}^{B} \sum_{i=1}^{K} \frac{1}{d+1} \sum_{s=1}^{d+1} \ell'_b(i) p_{(d+1)(b-1)+s}(i) \\
&= (d+1) \sum_{b=1}^{B} \sum_{i=1}^{K} \ell'_b(i) p'_b(i) \ .
\end{aligned}$$

Moreover,

$$\min_{k \in A} \sum_{t=1}^{T} \ell_t(k) = (d+1) \min_{k \in A} \sum_{b=1}^{B} \ell'_b(k) \ .$$

Since we know that for any predictor $\mathcal{A}'$ there exists a loss sequence $\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2, \ldots$ such that the regret of $\mathcal{A}'$ is at least $\big(1 - o(1)\big)\sqrt{(T/2)\ln K}$, where $o(1) \to 0$ for $K, B \to \infty$, we have that the regret of $\mathcal{A}$ is at least

$$(d+1) R_{T/(d+1)}(\mathcal{A}') = \big(1 - o(1)\big)(d+1)\sqrt{\frac{T}{2(d+1)} \ln K} = \big(1 - o(1)\big)\sqrt{(d+1)\frac{T}{2} \ln K} \ ,$$

where $R_{T/(d+1)}(\mathcal{A}')$ is the regret of $\mathcal{A}'$ over $T/(d+1)$ time steps. The proof is completed by observing that that the regret of any predictor in the bandit setting with delay $d$ cannot be

smaller than the regret of the predictor in the bandit setting with no delay or smaller than the regret of the predictor in the full information setting with delay $d$. Hence, the minimax regret in the bandit setting with delay $d$ must be at least of order

$$\max\left\{\sqrt{KT}, \sqrt{(d+1)T\ln K}\right\} = \Omega\left(\sqrt{(K+d)\,T}\right) \ ,$$

ignoring log factors. This concludes the proof. ■

Recent results by Joulani et al. (2016)— see also (Quanrud and Khashabi, 2015)— consider a full information setting with variable delays $d_t > 0$. At the end of each round $t$, the agent simultaneously observes all loss vectors $\boldsymbol{\ell}_s$ such that $s + d_s = t$. For this setting, they prove a regret bound of order $\sqrt{(D+T)\ln K}$, where $D$ is the sum of the delays $d_t$. The proof is based on a generic algorithm (SOLID) which simply feeds the loss vectors, as soon as they become available, to any deterministic predictor BASE designed to operate in a setting without delays. At the beginning of each round $t$, SOLID predicts using the current distribution $\boldsymbol{p}_t$ of BASE. A general argument shows that the regret of SOLID is equal to the regret of BASE on the sequence of loss vectors permuted according to the delays, plus a term that accounts for the drift of the distributions $\boldsymbol{p}_t$. For the bandit setting, we conjecture an upper bound on the regret of order $\sqrt{(D+KT)\ln K}$. However, we have not been able to prove this result via a direct application of our techniques.

## 8. Conclusions and Open Questions

We have investigated a cooperative and nonstochastic bandit scenario where cooperation comes at the price of delayed information. We have proven average welfare regret bounds that exhibit a natural tradeoff between amount cooperation and delay, the tradeoff being ruled by the underlying communication network topology. As a by-product of our analysis, we have also provided the first characterization to date of the regret of learning with (constant) delayed feedback in an adversarial bandit setting. There are a number of possible extensions of this work:

1. Our analysis only delivers average welfare regret bounds. It would be interesting to show simultaneous regret bounds that hold for each agent individually. We conjecture that the individual regret bound of an agent $v$ is of the form

$$\sqrt{(\ln K)\left(d + \frac{K}{|N_{\le d}(v)|}\right) T}$$

   where $|N_{\le d}(v)|$ is the degree of $v$ in $G_{\le d}$ (plus one). Such bound would in fact imply, e.g., the one in Theorem 5. A possible line of attack to solve this problem could be the use of graph sparsity along the lines of (Pan et al., 2015; Duchi et al., 2013; Mania et al., 2015; McMahan and Streeter, 2014).

2. It would be nice to characherize the average welfare regret by complementing our upper bounds with suitable *lower* bounds. For example, is the upper bound of Theorem 5 optimal in the communication model considered here?

3. It is natural to think of ways to adaptively tune our algorithms so as to automatically determine the best local parameters, e.g., the delay parameter $d$. For instance, disregarding message complexity, is there a way for each agent to adaptively tune $d$ locally so to minimize the bound in Theorem 5?

4. Our messages $m_t(v)$ contain both action/loss information and distribution information. Is it possible to drop the distribution information and still achieve average welfare regret bounds similar to those in Theorems 4, 11, and 13?

5. Besides settling the conjecture advanced at the end of Section 7, we generally think that the study of learning on a communication network with time-varying delays, and its impact on the regret rates, is a topic worth of attention.

## Acknowledgments

Parts of this work were done while NCB and CG were visiting YM at Tel-Aviv University. NCB gratefully acknowledges partial support from EU through the COST Action CA16228 (GAMENET). YM was supported in part by a grant from the Israel Science Foundation (ISF), and the Israeli Centers of Research Excellence (I-CORE) program (Center No. 4/11).

## Proofs from Section 4.1

**Proof of Theorem 5.** We start off from first part of the proof of Theorem 4 which, after rearranging terms, gives the following bound for each agent $v$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i,v)\ell_t(i)\right] - \sum_{t=1}^{T}\ell_t(k) \\
&\quad\le 2d + \mathbb{E}\left[\frac{\ln K}{\eta(v)} + \eta(v)\, d^2 + \eta(v)\sum_{t=d+1}^{T}\left(d + \frac{e}{2}\sum_{i=1}^{K}\frac{p_{t-d}(i,v)}{q_{d,t-d}(i,v)}\right)\right] \\
&\quad\le 3d + \mathbb{E}\left[\frac{Ke(d+1)\ln K}{\gamma(v)} + \frac{\gamma(v)}{Ke(d+1)}\sum_{t=1}^{T}\underbrace{\left(\mathbb{I}\{t>d\}\, d + \frac{e}{2}\left(\sum_{i=1}^{K}\frac{p_{t-d}(i,v)}{q_{d,t-d}(i,v)}\right)\mathbb{I}\{t>d\}\right)}_{Q_t(v)}\right] .
\end{aligned} \tag{9}$$

Note that the optimal tuning of $\gamma(v)$ depends on the random quantity

$$\overline{Q}_T(v) = \sum_{t=1}^{T} Q_t(v) \ .$$

We now apply the doubling trick to each instance of Exp3-Coop. Recall that, for each $v \in V$, we let $\gamma_r(v) = Ke(d+1)\sqrt{(\ln K)/2^r}$ for each $r = r_0, r_0+1, \ldots$, where $r_0 = \left\lceil \log_2 \ln K + 2\log_2(Ke(d+1))\right\rceil$ is chosen in a way that $\gamma_r(v) \le 1$ for all $r \ge r_0$. Let $T_r$ be the random set of consecutive time steps where the same $\gamma_r(v)$ was used. Whenever the algorithm is running with $\gamma_r(v)$ and detects $\sum_{s\in T_r} Q_s(v) > 2^r$, then we restart the algorithm with $\gamma(v) = \gamma_{r+1}(v)$. The largest $r = r(v)$ we need is $\left\lceil \log_2 \overline{Q}_T(v)\right\rceil$ and

$$\sum_{r=r_0}^{\left\lceil \log_2 \overline{Q}_T(v)\right\rceil} 2^{r/2} < 5\sqrt{\overline{Q}_T(v)} \ .$$

Because of (9), the regret agent $v$ suffers when using $\gamma_r(v)$ within $T_r$ is at most $3d + 2\sqrt{(\ln K)2^r}$. Now, since we pay at most regret $d$ at each restart, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i} p_t(i,v)\ell_t(i)\right] - \sum_{t=1}^{T}\ell_t(k) &\le 3d + 4Ke(d+1)\ln K \\
&\quad + \mathbb{E}\left[10\sqrt{(\ln K)\overline{Q}_T(v)} + 3d\left\lceil \log_2 \overline{Q}_T(v)\right\rceil\right] \ .
\end{aligned}$$

The term $3d + 4Ke(d+1)\ln K$ bounds the regret when the algorithm is never restarted implying that only $\gamma_{r_0}(v)$ is used.

Taking averages with respect to $v$, using Jensen's inequality multiple times, and applying the deterministic bound

$$\frac{1}{N}\sum_{v\in V}\overline{Q}_T(v) \le \left(d + \frac{e}{2(1-e^{-1})}\,\frac{K\left(\alpha(G_{\le d})+1\right)}{N}\right) T$$

derived with the aid of Lemma 3 at the end of the proof of Theorem 4, gives

$$\begin{aligned} R_T^{\mathrm{coop}} &\le 3d + 4Ke(d+1)\ln K \\ &\quad + 10\sqrt{(\ln K)\mathbb{E}\left[\frac{1}{N}\sum_{v\in V}\overline{Q}_T(v)\right]} + 3d\log_2\left(\mathbb{E}\left[\frac{1}{N}\sum_{v\in V}\overline{Q}_T(v)\right]\right) \\ &\le 10\sqrt{(\ln K)\left(d + \frac{e}{2(1-e^{-1})}\,\frac{K\left(\alpha(G_{\le d})+1\right)}{N}\right)T} + 3d\log_2 T + C\,, \end{aligned}$$

where $C$ is independent of $T$ and depends polynomially on the other parameters. Hence, as $T$ grows large,

$$R_T^{\mathrm{coop}} = \mathcal{O}\left(\sqrt{(\ln K)\left(d+1+\frac{K}{N}\,\alpha(G_{\le d})\right)T} + d\,\log T\right)$$

as claimed. ■

## Proofs from Section 5

We first need to adapt the preliminary Lemmas 1 and 2 to the new update rule of Exp3-Coop2 contained in Figure 5.

**Lemma 16** *Under the update rule contained in Figure 5, for all $t \ge 1$, for all $i \in A$, and for all $v \in V$*

$$\begin{aligned} -p_t(i,v)\left(\eta\widehat{\ell}_t(i,v)+\delta\right) &\le p_{t+1}(i,v) - p_t(i,v) \\ &\le p_{t+1}(i,v)\sum_{j=1}^{K} p_t(j,v)\left(1 - \mathbb{I}\{\widetilde{p}_{t+1}(i,v) > \delta/K\}\big(1-\eta\,\widehat{\ell}_t(i,v)\big)\right) \end{aligned}$$

*holds deterministically with respect to the agents' randomization.*

**Proof** For the lower bound, we have

$$p_{t+1}(i,v) - p_t(i,v) = \frac{\widetilde{p}_{t+1}(i,v)}{\widetilde{P}_{t+1}(v)} - p_t(i,v) \ge \frac{w_{t+1}(i,v)}{W_{t+1}(v)\,\widetilde{P}_{t+1}(v)} - p_t(i,v)\ .$$

Since $W_{t+1}(v) = \sum_{i\in A} p_t(i,v)e^{-\eta\widehat{\ell}_t(i,v)} \le \sum_{i\in A} p_t(i,v) = 1$, and $\widetilde{P}_{t+1}(v) \le 1+\delta$ by (7), we can write

$$\begin{aligned} p_{t+1}(i,v) - p_t(i,v) &\ge \frac{w_{t+1}(i,v)}{1+\delta} - p_t(i,v) \\ &= p_t(i,v)\left(\frac{e^{-\eta\widehat{\ell}_t(i,v)}}{1+\delta} - 1\right) \\ &\ge p_t(i,v)\left(\frac{1-\eta\widehat{\ell}_t(i,v)}{1+\delta} - 1\right) \qquad (\text{using } e^{-x} \ge 1-x) \\ &\ge p_t(i,v)\left(-\delta - \eta\widehat{\ell}_t(i,v)\right) \end{aligned}$$

as claimed. As for the upper bound, we first claim that

$$\frac{w_{t+1}(i,v)}{W_{t+1}(v)} \geq p_{t+1}(i,v)\mathbb{I}\{\widetilde{p}_{t+1}(i,v) > \delta/K\}\,. \tag{10}$$

To prove (10), we recall that $\widetilde{p}_{t+1}(i,v) = \max\left\{\frac{w_{t+1}(i,v)}{W_{t+1}(v)}, \frac{\delta}{K}\right\}$. Then we distinguish two cases:

1. If $\frac{w_{t+1}(i,v)}{W_{t+1}(v)} \leq \frac{\delta}{K}$, then $\widetilde{p}_{t+1}(i,v) = \delta/K$, and $w_{t+1}(i,v)/W_{t+1}(v) > 0$ by definition, hence (10) holds;

2. If $\frac{w_{t+1}(i,v)}{W_{t+1}(v)} > \frac{\delta}{K}$ then $\widetilde{p}_{t+1}(i,v) = \frac{w_{t+1}(i,v)}{W_{t+1}(v)}$, so that $p_{t+1}(i,v) \leq p_{t+1}(i,v)\,\widetilde{P}_{t+1}(v) = \widetilde{p}_{t+1}(i,v)$ and (10) again holds.

Then, setting for brevity $C = \mathbb{I}\{\widetilde{p}_{t+1}(i,v) > \delta/K\}$, we can write

$$\begin{aligned}
p_{t+1}(i,v) - p_t(i,v) &\leq p_{t+1}(i,v) - w_{t+1}(i,v) && \text{(from the update (1))}\\
&\leq p_{t+1}(i,v) - W_{t+1}(v)p_{t+1}(i,v)\,C && \text{(using (10))}\\
&= p_{t+1}(i,v)\big(1 - W_{t+1}(v)\,C\big)\\
&= p_{t+1}(i,v)\left(\sum_{j\in A}\big(p_t(j,v) - C\,w_{t+1}(j,v)\big)\right)\\
&= p_{t+1}(i,v)\sum_{j\in A} p_t(j,v)\left(1 - C\,e^{-\eta\,\widehat{\ell}_t(j,v)}\right)\\
&\leq p_{t+1}(i,v)\sum_{j\in A} p_t(j,v)\left(1 - C(1 - \eta\,\widehat{\ell}_t(j,v))\right)
\end{aligned}$$

where in the last step we again used $e^{-x} \geq 1 - x$. This concludes the proof. ■

**Lemma 17** *Under the update rule contained in Figure 5, if $\delta \leq 1/d(v)$ and $\eta \leq \frac{1}{Ke(d(v)+1)}$, then*

$$p_{t+1}(i,v) \leq \left(1 + \frac{1}{d(v)}\right) p_t(i,v) \tag{11}$$

*holds for all $t \geq 1$ and $i \in A$, deterministically with respect to the agents' randomization.*

**Proof** If $\widetilde{p}_{t+1}(i,v) = \delta/K$ then, from (7), we have $\delta/K = p_{t+1}(i,v)\widetilde{P}_{t+1}(v) \geq p_{t+1}(i,v)$, and $p_t(i,v) \geq \frac{\delta}{K(1+\delta)}$. Hence, $\frac{p_{t+1}(i,v)}{p_t(i,v)} \leq \frac{\delta/K}{\delta/(K(1+\delta))} = 1 + \delta$, so the claim follows from $\delta \leq \frac{1}{d(v)}$. On the other hand, if $\widetilde{p}_{t+1}(i,v) > \delta/K$, then the proof is exactly the same as the proof of Lemma 2, for the second inequality in the statement of Lemma 16 turns out to be exactly the same as the corresponding inequality in the statement in Lemma 1. ■

Next, we generalize Lemma 3 to the case of directed graphs. This is where we need a lower bound on the probabilities $p_t(i,v)$. If $G = (V,E)$ is a directed graph, then for each $v \in V$ let $N^-_{\leq 1}(v)$ be the in-neighborhood of node $v$ (i.e., the set of $v' \in V$ such that arc $(v',v) \in E$), including $v$ itself.

**Lemma 18** *Let $G = (V, E)$ be a directed graph with independence number $\alpha(G)$. Let $\boldsymbol{p}(v) = \big(p(1,v), \dots, p(K,v)\big)$ be a probability distribution over $A = \{1, \dots, K\}$ such that $p(i,v) \geq \frac{\delta}{K(1+\delta)}$. Then, for all $i \in A$,*

$$\sum_{v \in V} \frac{p(i,v)}{q(i,v)} \leq \frac{1}{1-e^{-1}} \left( 6\,\alpha(G) \ln\left(1 + \frac{N^2 K(1+\delta)}{\delta}\right) + \sum_{v \in V} p(i,v) \right) ,$$

*where $q(i,v) = 1 - \prod_{v' \in N^-_{\leq 1}(v)} \big(1 - p(i,v')\big)$.*

**Proof** We follow the notation and the proof of Lemma 3, where it is shown that

$$\sum_{v \in V} \frac{p(i,v)}{q(i,v)} \leq \frac{1}{1-e^{-1}} \sum_{v \in V} \left( \frac{p(i,v)}{P(i,v)} + p(i,v) \right) .$$

In order to bound from above the sum $\sum_{v \in V} \frac{p(i,v)}{P(i,v)}$, we combine (Alon et al., 2017, Lemma 14 and 16) and derive the upper bound

$$\sum_{v \in V} \frac{p(i,v)}{P(i,v)} \leq 6\,\alpha(G) \ln\left(1 + \frac{N^2 K(1+\delta)}{\delta}\right)$$

holding when $p(i,v) \geq \frac{\delta}{K(1+\delta)}$. Again, the probabilities $p(i,1), \dots, p(i,N) \geq 0$ need not sum to one in order for this lemma to apply. ■

With the above three lemmas handy, we are ready to prove Theorem 11.

**Proof of Theorem 11.** This proof is similar to the proof of Theorem 4, hence we only emphasize the differences between the two.

From the update rule in Figure 5, we have, for each $v \in V$,

$$\begin{aligned}
W_{T+1}(v) &= \sum_{i=1}^{K} \frac{\widetilde{p}_T(i)}{\widetilde{P}_T(v)} e^{-\eta \widehat{\ell}_T(i,v)} \\
&\geq \sum_{i=1}^{K} \frac{w_T(i,v)}{W_T(v)\widetilde{P}_T(v)} e^{-\eta \widehat{\ell}_T(i,v)} && \text{(since } \widetilde{p}_T(i) \geq w_T(i,v)/W_T(v)\text{)} \\
&= \sum_{i=1}^{K} \frac{\widetilde{p}_{T-1}(i,v) e^{-\eta \widehat{\ell}_{T-1}(i,v)} e^{-\eta \widehat{\ell}_T(i,v)}}{W_T(v)\widetilde{P}_{T-1}(v)\widetilde{P}_T(v)} \\
&\;\;\vdots \\
&\geq \sum_{i=1}^{K} \frac{w_1(i,v)\, e^{-\eta \sum_{t=1}^{T} \widehat{\ell}_t(i,v)}}{W_1(v) \cdots W_T(v) \widetilde{P}_1(v) \cdots \widetilde{P}_T(v)} .
\end{aligned}$$

Now, because $w_1(i,v) = 1$, $W_1(v) = K$, and $\widetilde{P}_t(v) \leq 1 + \delta$ for all $t$, see (7), the above chain of inequalities implies that, for any fixed action $k \in A$,

$$(1+\delta)^T K \left( \prod_{t=1}^{T} W_{t+1}(v) \right) \geq e^{-\eta \sum_{t=1}^{T} \widehat{\ell}_t(k,v)} . \tag{12}$$

As usual, the quantity $W_{t+1}(v)$ can be upper bounded as

$$\begin{aligned} W_{t+1}(v) &= \sum_{i=1}^{K} p_t(i,v) e^{-\eta \widehat{\ell}_t(i,v)} \\ &\le \sum_{i=1}^{K} p_t(i,v) \left(1 - \eta \widehat{\ell}_t(i,v) + \frac{\eta^2}{2} \widehat{\ell}_t(i,v)^2\right) \\ &\quad \text{(from } e^{-x} \le 1 - x + x^2/2 \text{ for all } x \ge 0) \\ &= 1 - \eta \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v) + \frac{\eta^2}{2} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v)^2 \ . \end{aligned}$$

Substituting into (12) and taking logs of both sides gives

$$\begin{aligned} &T \ln(1+\delta) + \ln K + \sum_{t=1}^{T} \ln \left(1 - \eta \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v) + \frac{\eta^2}{2} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v)^2\right) \\ &\qquad \ge -\eta \sum_{i=1}^{K} \widehat{\ell}_t(k,v) \, . \end{aligned}$$

Finally, using $\ln(1+x) \le x$, dividing by $\eta$, using $\delta = 1/T$, and rearranging yields

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v) \le \frac{1 + \ln K}{\eta} + \sum_{t=1}^{T} \widehat{\ell}_t(k,v) + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v)^2 \ , \qquad (13)$$

hence arriving at the counterpart of (6). From this point on, we proceed as in the proof of Theorem 4 by taking expectation on the three (double) sums in (13). Note that we do still have, for all $v \in V$, $t > d(v)$, and $i \in A$,

$$\begin{aligned} \mathbb{E}_{t-d(v)} \left[\widehat{\ell}_t(i,v)\right] &= \ell_{t-d(v)}(i) \\ \mathbb{E}_{t-d(v)} \left[p_t(i,v) \widehat{\ell}_t(i,v)\right] &= p_t(i,v) \ell_{t-d(v)}(i) \\ \mathbb{E}_{t-d(v)} \left[p_t(i,v) \widehat{\ell}_t(i,v)^2\right] &= p_t(i,v) \frac{\ell_{t-d(v)}(i)^2}{q_{\mathcal{P},t-d(v)}(i,v)} \ . \end{aligned}$$

We can write

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v)\right] &\ge \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \, \ell_t(i)\right] - 2d(v) - (\eta + \delta) \, T \, d(v) \\ \mathbb{E}\left[\sum_{t=1}^{T} \widehat{\ell}_t(k,v)\right] &\le \sum_{t=1}^{T} \ell_t(k) \end{aligned}$$

and, as in the proof of Theorem 4,

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i,v) \widehat{\ell}_t(i,v)^2\right] \le e \, \mathbb{E}\left[\sum_{t=d(v)+1}^{T} \sum_{i=1}^{K} \frac{p_{t-d(v)}(i,v)}{q_{\mathcal{P},t-d(v)}(i,v)}\right] \ .$$

Summing over all agents $v$, dividing by $N$, and applying Lemma 18 to the directed graph $G_{\mathcal{P}}$, the last inequality gives

$$\frac{1}{N}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{v\in V} p_t(i,v)\widehat{\ell}_t(i,v)^2\right] \leq \frac{e}{1-e^{-1}}\,T\left(\frac{6K}{N}\,\alpha\left(G_{\mathcal{P}}\right)\,\ln\left(1+2TN^2K\right)+1\right)\ .$$

Combining as in (13), recalling that $\delta = 1/T$, and setting for brevity $\bar{d}_V = \frac{1}{N}\sum_{v\in V} d(v)$, we have thus obtained that the average welfare regret of Exp3-Coop2 satisfies

$$\begin{aligned} R_T^{\mathrm{coop}} &\leq 3\bar{d}_V + \eta\,T\,\bar{d}_V + \frac{1+\ln K}{\eta} + \frac{e\eta}{2(1-e^{-1})}\,T\left(\frac{6K}{N}\,\alpha\left(G_{\mathcal{P}}\right)\,\ln\left(1+2TN^2K\right)+1\right) \\ &= \mathcal{O}\left(\eta\,T\,\bar{d}_V + \frac{\ln K}{\eta} + \frac{\eta\,TK}{N}\,\alpha\left(G_{\mathcal{P}}\right)\,\ln\left(TNK\right)\right) \end{aligned}$$

as $T$ grows large. This concludes the proof. ■

## Proofs from Section 6

The following are further generalizations of Lemmas 1 and 2 that apply to mixed delay estimators.

**Lemma 19** *Let $\widehat{\ell}_t(i,v)$ be of the form (8) for each $t \geq 1$, $i \in A$, and $v \in V$. If $\eta \leq \frac{1}{Ke(d+1)}$ in the update rule $w_{t+1}(i,v) = p_t(i,v)\,\exp\bigl(-\eta\,\widehat{\ell}_t(i,v)\bigr)$ then*

$$p_{t+1}(i,v) \leq \left(1+\frac{1}{d}\right) p_t(i,v)$$

*holds for all $t \geq 1$, $i \in A$, and $v \in V$ deterministically with respect to the agent's randomization.*

**Proof** The proof follows a similar inductive argument as in the proof of Lemma 2. We can write, for $t > d$,

$$\begin{aligned} \sum_{i\in A} p_t(i,v)\widehat{\ell}_t(i,v) &= \sum_{i\in A} p_t(i,v)\sum_{s=0}^{d-1} D(s)\,\frac{\ell_{t-s}(i,v)}{q_{s,t-s}(i,v)}B_{s,t-s}(i,v) \\ &\leq \sum_{i\in A} p_t(i,v)\sum_{s=0}^{d-1} D(s)\,\frac{1}{q_{s,t-s}(i,v)} \\ &\leq \sum_{i\in A}\sum_{s=0}^{d-1}\left(1+\frac{1}{d}\right)^s D(s)\,\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)} \\ &\leq \sum_{i\in A}\sum_{s=0}^{d-1}\left(1+\frac{1}{d}\right)^s D(s) \\ &\leq Ke\ . \end{aligned}$$

Combining with Lemma 1 as in the proof of Lemma 2 concludes the proof. ■

**Lemma 20** *Under the same assumptions and notation as in Lemma 19, we have that*

$$p_{t+1}(i,v) - p_t(i,v) \geq -e\eta \sum_{s=0}^{d-1} D(s)\, \frac{B_{s,t-s}(i,v)\, p_{t-s}(i,v)}{q_{s,t-s}(i,v)}$$

*holds for all $t \geq 1$, $i \in A$, and $v \in V$ deterministically with respect to the agent's randomization.*

**Proof** When $t \leq d$ the statement is trivially verified, hence we continue by assuming $t > d$. From Lemma 1 we have $p_{t+1}(i,v) - p_t(i,v) \geq -\eta\, p_t(i,v)\widehat{\ell}_t(i,v)$. Then, proceeding as in the proof of Lemma 19, we have

$$\begin{aligned}
p_t(i,v)\widehat{\ell}_t(i,v) &= p_t(i,v) \sum_{s=0}^{d-1} D(s)\, \frac{B_{s,t-s}(i,v)\, \ell_{t-s}(i)}{q_{s,t-s}(i,v)} \\
&\leq \sum_{s=0}^{d-1} D(s)\, \frac{B_{s,t-s}(i,v)\, p_t(i,v)}{q_{s,t-s}(i,v)} \\
&\leq \sum_{s=0}^{d-1} D(s)\, \left(1+\frac{1}{d}\right)^s \frac{B_{s,t-s}(i,v)\, p_{t-s}(i,v)}{q_{s,t-s}(i,v)} \\
&\leq e \sum_{s=0}^{d-1} D(s)\, \frac{B_{s,t-s}(i,v)\, p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\,.
\end{aligned}$$

Putting together proves the claim. ■

**Proof of Theorem 13.** As in the proof of Theorem 4, we start off from (6). Using the definition of $\widehat{\ell}_t(i,v)$ in (8) this can be written as

$$\begin{aligned}
\sum_{t=d+1}^{T} \sum_{i=1}^{K} \sum_{s=0}^{d-1} D(s) \frac{B_{s,t-s}(i,v)\, p_t(i,v)\, \ell_{t-s}(i)}{q_{s,t-s}(i,v)} &\leq \sum_{t=d+1}^{T} \sum_{s=0}^{d-1} D(s) \frac{B_{s,t-s}(k,v)\, \ell_{t-s}(k)}{q_{s,t-s}(k,v)} \\
&\quad + \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=d+1}^{T} \sum_{i=1}^{K} p_t(i,v) \left( \sum_{s=0}^{d-1} D(s) \frac{B_{s,t-s}(i,v)\, \ell_{t-s}(i)}{q_{s,t-s}(i,v)} \right)^2 . \quad (14)
\end{aligned}$$

We now bound the first and the third sum in (14) separately. As for the first sum, a repeated application of Lemma 20 for $t \geq 2d+1$ and $s = 1, \ldots, d$ leads to

$$p_t(i,v) \geq p_{t-s}(i,v) - e\eta \sum_{h=1}^{s} \sum_{r=0}^{d-1} D(r) \frac{B_{r,t-h-r}(i,v)\, p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)}\,,$$

so that

$$\begin{aligned}\sum_{t=d+1}^{T}\sum_{i=1}^{K}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,p_t(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)} &\geq \sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,p_t(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}\\ &\geq \sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)} - e\eta\sum_{t=2d+1}^{T}\sum_{i=1}^{K} S_t(i,v)\,,\end{aligned}$$

where the slack $S_t(i,v)$ satisfies

$$S_t(i,v) = \sum_{s=1}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}\sum_{h=1}^{s}\sum_{r=0}^{d-1} D(r)\frac{B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)}\,. \tag{15}$$

As for the third sum, Jensen's inequality and $\ell_{i,t-s}\in[0,1]$ jointly give

$$\left(\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}\right)^2 \leq \sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)}{q_{s,t-s}(i,v)^2}\,.$$

Moreover, recalling that $s\in\{1,\dots,d\}$, a repeated application of Lemma 19 yields

$$p_t(i,v) \leq \left(1+\frac{1}{d}\right)^s p_{t-s}(i,v) \leq \left(1+\frac{1}{d}\right)^d p_{t-s}(i,v) \leq e\,p_{t-s}(i,v)\;,$$

so that

$$\sum_{t=d+1}^{T}\sum_{i=1}^{K} p_t(i,v)\left(\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}\right)^2 \leq e\sum_{t=d+1}^{T}\sum_{i=1}^{K}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)}{q_{s,t-s}(i,v)^2}\;.$$

Putting together and summing over $v\in V$ we obtain

$$\begin{aligned}\underbrace{\sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{v\in V}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}}_{\text{(I)}} &\leq \underbrace{\sum_{t=d+1}^{T}\sum_{v\in V}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(k,v)\,\ell_{t-s}(k)}{q_{s,t-s}(k,v)}}_{\text{(II)}}\\ &+\frac{N\,\ln K}{\eta} + e\eta\underbrace{\sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{v\in V} S_t(i,v)}_{\text{(III)}} + \frac{e\eta}{2}\underbrace{\sum_{t=d+1}^{T}\sum_{i=1}^{K}\sum_{v\in V}\sum_{s=0}^{d-1} D(s)\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)}{q_{s,t-s}(i,v)^2}}_{\text{(IV)}}\;.\end{aligned}$$

Similar to before, we have, for any $i\in A$, $v\in V$, $t=1,2,\dots$, and $s=0,\dots,d-1$,

$$\begin{aligned}\mathbb{E}_{t-s}\left[\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}\right] &= \frac{p_{t-s}(i,v)\,\ell_{t-s}(i)}{q_{s,t-s}(i,v)}\,\mathbb{E}_{t-s}\big[B_{s,t-s}(i,v)\big] = p_{t-s}(i,v)\,\ell_{t-s}(i)\\ \mathbb{E}_{t-s}\left[\frac{B_{s,t-s}(k,v)\,\ell_{t-s}(k)}{q_{s,t-s}(k,v)}\right] &= \frac{\ell_{t-s}(k)}{q_{s,t-s}(k,v)}\,\mathbb{E}_{t-s}\big[B_{s,t-s}(k,v)\big] = \ell_{t-s}(k)\\ \mathbb{E}_{t-s}\left[\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)}{\big(q_{s,t-s}(i,v)\big)^2}\right] &= \frac{p_{t-s}(i,v)}{\big(q_{s,t-s}(i,v)\big)^2}\,\mathbb{E}_{t-s}\big[B_{s,t-s}(i,v)\big] = \frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\,.\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\left[(\mathrm{I})\right] &= \sum_{i=1}^{K}\sum_{v\in V}\sum_{t=2d+1}^{T}\sum_{s=0}^{d-1} D(s)p_{t-s}(i,v)\,\ell_{t-s}(i) \\
&\geq \sum_{i=1}^{K}\sum_{v\in V}\sum_{t=2d+1}^{T-d+1} p_t(i,v)\,\ell_t(i) \\
&\geq \sum_{i=1}^{K}\sum_{v\in V}\left(\sum_{t=1}^{T} p_t(i,v)\,\ell_t(i) - \sum_{t=1}^{2d} p_t(i,v) - \sum_{t=T-d+2}^{T} p_t(i,v)\right) \\
&\geq \sum_{v\in V}\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i,v)\,\ell_t(i) - 3dN
\end{aligned}$$

and

$$\mathbb{E}\left[(\mathrm{II})\right] = \sum_{v\in V}\sum_{t=2d+1}^{T}\sum_{s=0}^{d-1} D(s)\ell_{t-s}(k) \leq \sum_{v\in V}\sum_{t=1}^{T}\ell_t(k)\,.$$

Moreover, from Lemma 3,

$$\begin{aligned}
\mathbb{E}\left[(\mathrm{IV})\right] &= \sum_{i=1}^{K}\sum_{t=d+1}^{T}\sum_{s=0}^{d-1} D(s)\sum_{v\in V}\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)} \\
&\leq \sum_{i=1}^{K}\sum_{t=d+1}^{T}\sum_{s=0}^{d-1} D(s)\left(\frac{1}{1-e^{-1}}\left(\alpha(G_{\leq s}) + \sum_{v\in V} p_{t-s}(i,v)\right)\right) \\
&= \frac{K}{1-e^{-1}}\sum_{t=d+1}^{T}\sum_{s=0}^{d-1} D(s)\alpha(G_{\leq s}) + \frac{1}{1-e^{-1}}\sum_{t=d+1}^{T}\sum_{s=0}^{d-1} D(s)\sum_{v\in V}\sum_{i=1}^{K} p_{t-s}(i,v) \\
&= \frac{K}{1-e^{-1}}\,(T-d)\,\bar{\alpha}_D + \frac{1}{1-e^{-1}}\,(T\,N - d\,N) \\
&< \frac{T}{1-e^{-1}}\,\left(K\bar{\alpha}_D + N\right)\,,
\end{aligned}$$

where $\bar{\alpha}_D = \sum_{s=0}^{d-1} D(s)\alpha(G_{\leq s})$.

Finally, we are left with upper bounding $\mathbb{E}[(\mathrm{III})]$. Consider each slack term $S_t(i,v)$ in (15) We can then rewrite $S_t(i,v)$ as

$$S_t(i,v) = \sum_{s=1}^{d-1}\sum_{h=1}^{s}\sum_{r=0}^{d-1} D(s)D(r)\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\,.$$

Now, we consider three cases, depending on the value of the indices $s, h$, and $r$ in the triple sum.

**Case 1:** $t-s > t-h-r$ (i.e., $s < h+r$). We have

$$\begin{aligned}
&\mathbb{E}\left[\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\right] \\
&= \mathbb{E}\left[\mathbb{E}_{t-s}\left[\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\right]\right] \\
&= \mathbb{E}\left[\frac{\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\mathbb{E}_{t-s}\left[B_{s,t-s}(i,v)\right]\right] \\
&= \mathbb{E}\left[\frac{\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)}\right] \\
&\le \mathbb{E}\left[\frac{p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)}\,E_{t-h-r}[B_{r,t-h-r}(i,v)]\right] \\
&= \mathbb{E}\left[p_{t-h-r}(i,v)\right]\ .
\end{aligned}$$

**Case 2:** $t-s < t-h-r$ (i.e., $s > h+r$). We can write

$$\begin{aligned}
&\mathbb{E}\left[\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\right] \\
&= \mathbb{E}\left[\mathbb{E}_{t-h-r}\left[\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\right]\right] \\
&= \mathbb{E}\left[\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\mathbb{E}_{t-h-r}\left[B_{r,t-h-r}(i,v)\right]\right] \\
&\le \mathbb{E}\left[\frac{B_{s,t-s}(i,v)\,p_{t-h-r}(i,v)}{q_{s,t-s}(i,v)}\right] \\
&\le \left(1+\frac{1}{d}\right)^{s-h-r}\mathbb{E}\left[\frac{B_{s,t-s}(i,v)\,p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\right] \qquad \text{(repeatedly using Lemma 19)} \\
&\le e\,\mathbb{E}\left[\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\,\mathbb{E}_{t-s}\left[B_{s,t-s}(i,v)\right]\right] \\
&= e\,\mathbb{E}\left[p_{t-s}(i,v)\right]\ .
\end{aligned}$$

**Case 3:** $t-s = t-h-r$ (i.e., $s = h+r$). We have $p_{t-h-r}(i,v) = p_{t-s}(i,v)$. Moreover, since $h \ge 1$ we have $s > r$, so that $B_{s,t-s}(i,v)\,B_{r,t-h-r}(i,v) = B_{r,t-h-r}(i,v)$. We can write

$$\begin{aligned}
&\mathbb{E}\left[\frac{B_{s,t-s}(i,v)\,\ell_{t-s}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\right] \\
&= \mathbb{E}\left[\mathbb{E}_{t-h-r}\left[\frac{\ell_{t-h-r}(i)\,B_{r,t-h-r}(i,v)\,p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\right]\right] \\
&\le \mathbb{E}\left[\frac{p_{t-h-r}(i,v)}{q_{r,t-h-r}(i,v)\,q_{s,t-s}(i,v)}\,\mathbb{E}_{t-h-r}\left[B_{r,t-h-r}(i,v)\right]\right] \\
&= \mathbb{E}\left[\frac{p_{t-h-r}(i,v)}{q_{s,t-s}(i,v)}\right] \\
&= \mathbb{E}\left[\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\right]\ .
\end{aligned}$$

Therefore, piecing together and overapproximating yields

$$\mathbb{E}\left[S_t(i,v)\right] \le e\,\mathbb{E}\Bigg[\sum_{s,h,r\,:\,s<h+r} D(s)D(r)\,p_{t-h-r}(i,v) + \sum_{s,h,r\,:\,s>h+r} D(s)D(r)\,p_{t-s}(i,v) + \sum_{s,h,r\,:\,s=h+r} D(s)D(r)\,\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\Bigg]\,.$$

This allows us to write

$$\begin{aligned}
\mathbb{E}[(\mathrm{III})] &\le e\sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{v\in V}\mathbb{E}\Bigg[\sum_{s,h,r\,:\,s<h+r} D(s)D(r)\,p_{t-h-r}(i,v) \\
&\qquad + \sum_{s,h,r\,:\,s>h+r} D(s)D(r)\,p_{t-s}(i,v) + \sum_{s,h,r\,:\,s=h+r} D(s)D(r)\,\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\Bigg] \\
&= e\sum_{t=2d+1}^{T}\sum_{v\in V}\sum_{s,h,r\,:\,s<h+r} D(s)D(r) + e\sum_{t=2d+1}^{T}\sum_{v\in V}\sum_{s,h,r\,:\,s>h+r} D(s)D(r) \\
&\qquad + e\sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{s,h,r\,:\,s=h+r} D(s)D(r)\,\mathbb{E}\left[\sum_{v\in V}\frac{p_{t-s}(i,v)}{q_{s,t-s}(i,v)}\right] \\
&\le e\,T\,N\sum_{s,h,r\,:\,s\neq h+r} D(s)D(r) \\
&\qquad + \frac{e}{1-e^{-1}}\sum_{t=2d+1}^{T}\sum_{i=1}^{K}\sum_{s,h,r\,:\,s=h+r} D(s)D(r)\left(\alpha(G_{\le s}) + \sum_{v\in V} p_{t-s}(i,v)\right) \\
&= e\,T\,N\sum_{s,h,r\,:\,s\neq h+r} D(s)D(r) \\
&\qquad + \frac{T\,e}{1-e^{-1}}\left(K\sum_{s,h,r\,:\,s=h+r} D(s)D(r)\alpha(G_{\le s}) + N\sum_{s,h,r\,:\,s=h+r} D(s)D(r)\right) \\
&< \frac{T\,e}{1-e^{-1}}\left(K\sum_{s,h,r\,:\,s=h+r} D(s)D(r)\alpha(G_{\le s}) + N\sum_{s,h,r} D(s)D(r)\right)
\end{aligned}$$

Now,

$$\sum_{s,h,r} D(s)D(r) = \sum_{s=1}^{d-1} D(s)\sum_{h=1}^{s}\sum_{r=0}^{d-1} D(r) = \sum_{s=1}^{d-1} D(s)\,s = \mu_D$$

and

$$\sum_{s,h,r\,:\,s=h+r} D(s)D(r)\alpha(G_{\le s}) = \sum_{s=1}^{d-1} D(s)\alpha(G_{\le s})\sum_{h=1}^{s} D(s-h) \le \sum_{s=1}^{d-1} D(s)\alpha(G_{\le s}) = \bar{\alpha}_D$$

so that

$$\mathbb{E}[(\mathrm{III})] \leq \frac{T\,e}{1-e^{-1}} \left(N\mu_D + K\bar{\alpha}_D\right).$$

Putting pieces together, dividing by $N$, and further overapproximating gives

$$R_T^{\mathrm{coop}} \leq 3d + \frac{\ln K}{\eta} + \frac{e^2\eta\,T}{1-e^{-1}} \left(1 + \mu_D + \frac{2K}{N}\,\bar{\alpha}_D\right),$$

thereby concluding the proof. ■

## References


Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 873–881. Curran Associates, Inc., 2011.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal of Computing*, 46(6):1785–1826, 2017.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Baruch Awerbuch and Robert Kleinberg. Competitive collaborative learning. *Journal of Computer and System Sciences*, 74(8):1271–1288, 2008.

Samuel Barrett and Peter Stone. Ad hoc teamwork modeled with multi-armed bandits: An extension to discounted infinite rewards. In *Proceedings of 2011 AAMAS Workshop on Adaptive and Learning Agents*, pages 9–14, 2011.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622, 2016.

John Duchi, Michael I Jordan, and Brendan McMahan. Estimation, optimization, and parallelism when data is sparse. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2832–2840. Curran Associates, Inc., 2013.

John C Duchi, Sorathan Chaturapruek, and Christopher Ré. Asynchronous stochastic convex optimization. *arXiv preprint arXiv:1508.00882*, 2015.

Miroslav Dudík, Daniel J. Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 169–178, 2011.

P. Firby and J. Haviland. Independence and average distance in graphs. *Discrete Applied Mathematics*, 75:27–37, 1997.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1453–1461, 2013.

Pooria Joulani, András György, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 1744–1750, 2016.

Soummya Kar, H Vincent Poor, and Shuguang Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, pages 1771–1778. IEEE, 2011.

Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 533–542. ACM, 2009.

Tomáš Kocák, Gergely Neu, Michal Valko, and Remi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems 27*, pages 613–621. 2014.

Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative learning of stochastic bandits over a social network. In *54th Annual Allerton Conference on Communication, Control, and Computing*, pages 1228–1235. IEEE, 2016.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. *arXiv preprint arXiv:1512.06888*, 2015.

Mu Li, David G. Andersen, and Alexander Smola. Distributed delayed proximal gradient methods. In *NIPS Workshop on Optimization for Machine Learning*, 2013.

Nathan Linial. Locality in distributed graph algorithms. *SIAM J. Comput.*, 21(1):193–201, 1992.

Ji Liu, Stephen J Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *The Journal of Machine Learning Research*, 16(1):285–322, 2015.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *arXiv preprint arXiv:1507.06970*, 2015.

Brendan McMahan and Matthew Streeter. Delay-tolerant algorithms for asynchronous distributed online learning. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2915–2923. Curran Associates, Inc., 2014.

Chris Mesterharm. On-line learning with delayed label feedback. In *Algorithmic Learning Theory*, pages 399–413. Springer, 2005.

Chris Mesterharm. *Improving Online Learning*. PhD thesis, Department of Computer Science, Rutgers University, 2007.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23*, pages 1804–1812. Curran Associates, Inc., 2010.

Gergely Neu, Andras Gyorgy, Csaba Szepesvari, and Andras Antos. Online markov decision processes under bandit feedback. *Automatic Control, IEEE Transactions on*, 59(3):676–691, 2014.

Xinghao Pan, Dimitris Papailiopoulos, Samet Oymak, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Parallel correlation clustering on big graphs. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 82–90. Curran Associates, Inc., 2015.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4102–4110, 2018.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1270–1278. Curran Associates, Inc., 2015.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits – a musical chairs approach. *CoRR*, abs/1512.02866, 2015.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of The 31st International Conference on Machine Learning*, pages 280–287, 2014.

Ruben Stranders, Long Tran-Thanh, Francesco M Delle Fave, Alex Rogers, and Nicholas R Jennings. Dcops and bandits: Exploration and exploitation in decentralised coordination. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 289–296. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

Jukka Suomela. Survey of local algorithms. *ACM Computing Surveys*, 45(2):24, 2013.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedüs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *30th International Conference on Machine Learning (ICML 2013)*, volume 28, pages 19–27. ACM Press, 2013.

Cem Tekin and Mihaela van der Schaar. Distributed online learning via cooperative contextual bandits. *IEEE Transactions on Signal Processing*, 63(14):3700–3714, 2015.

Cem Tekin, Simpson Z. Zhang, and Mihaela van der Schaar. Distributed online learning in social recommender systems. *J. Sel. Topics Signal Processing*, 8(4):638–652, 2014.

Klaus Wehmuth and Artur Ziviani. Daccer: Distributed assessment of the closeness centrality ranking in complex networks. *Computer Networks*, 57(13):2536–2548, 2013.

Marcelo J Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

Martin Zinkevich, John Langford, and Alex J. Smola. Slow learners are fast. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2331–2339. Curran Associates, Inc., 2009.