Appendices of Distributed Multi-Agent Online Learning Based on Global Feedback

Jie Xu*, Cem Tekin, Simpson Zhang and Mihaela van der Schaar

APPENDIX A: PROOF OF THEOREM 1

In order to prove $\min_{\pi\in\Pi^c} R(T;\pi) = \min_{\pi\in\Pi^d} R(T;\pi)$, we will show that *(i)* there exists an optimal algorithm π^* that is deterministic and *(ii)* any deterministic algorithm can be implemented in the distributed scenario if agent's have identical observations of the overall reward realization. If (i) and (ii) are true, then the optimal algorithm π^* belongs to π^d and therefore $\min_{\pi\in\Pi^c} R(T;\pi) = \min_{\pi\in\Pi^d} R(T;\pi)$. We prove the claims in the following.

(i) Suppose at time $t \le T$, the reward history is $\mathcal{H}_t$. Let $\text{Reward}(T;\mathcal{H}_t)$ be the expected sum of rewards from time t to T given the history $\mathcal{H}_t$. For an optimal algorithm π^*, the following must be satisfied

$$\text{Reward}(T;\mathcal{H}_t) = \max_{\pi^*(\mathcal{H}_t)} \mathbb{E}[r(\pi^*(\mathcal{H}_t)) + \text{Reward}(T;\{\mathcal{H}_t, r(\pi^*(H_t))\})] \quad (1)$$

If $\pi^*(\mathcal{H}_t)$ is a mixed strategy, then it implies that there exist at least two pure joint arms $\boldsymbol{a}'$ and $\boldsymbol{a}''$ such that the algorithm is indifferent between these two joint arms in terms of maximizing the expected sum reward, i.e.

$$\text{Reward}(T;\mathcal{H}_t) = \text{Reward}(T;\mathcal{H}_t,\boldsymbol{a}') = \text{Reward}(T;\mathcal{H}_t,\boldsymbol{a}'') \quad (2)$$

Therefore, by setting $\pi^*(\mathcal{H}_t)$ to be any one of the pure joint arms, the algorithm does not lose any expected reward. Since this argument holds for any time t and any history $\mathcal{H}_t$, there must exist a deterministic algorithm that is optimal.

(ii) If an algorithm is deterministic and the overall reward realization can be perfectly observed by all agents, then the reward history is public and identical for all agents. Each agent's algorithm depends only

* J. Xu, T. Cem and M. van der Schaar are with the Dept. of Electrical Engineering (EE), University of California Los Angeles (UCLA). S. Zhang is with the Dept. of Economics, University of California Los Angeles (UCLA).

on its observed history, and so each agent can correctly infer from the public reward history the arms to be selected by all other agents. This implies that there is no need for message exchange among agents at run-time, because each agent will know the exact arm chosen by any other agent at every time t. Hence, any deterministic algorithm can be implemented in a distributed setting.

APPENDIX B: PROOF OF PROPOSITION 1

We prove this by constructing an example showing that the regret achieved by UCB1 is linear. Consider a network with 2 agents and each agent having two arms to select from. The expected overall reward of the 4 joint arms is listed in Table 1.

Table 1. Reward matrix for an illustrative system with two agents.

		Agent 1	
		Arm 1	Arm 2
Agent 2	Arm 1	10	0
	Arm 2	0	8

To simplify the analysis, we will assume that the realization of the overall reward at each time is exactly the expected overall reward at all times. However, agents may observe different noisy versions of the realization. We consider a special case where agent 1 perfectly observes the reward realization without any error in all slots, while agent 2 observes the reward realization with errors in the first 4 slots but without errors in the remaining slots. The error is drawn uniformly from the space $\{-2,0,2\}$ and independent across time. In the UCB1 algorithm, agents start by selecting each of the 4 joint arms once and updating their reward estimates. Consider an error sequence $\{$-2,0,0,2$\}$ in the first 4 slots. Given this error sequence, at the beginning of slot 5, the reward estimates are (the superscript indicates the agent): for agent 1, $\overline{r}^1(\{1,1\})=10$, $\overline{r}^1(\{1,2\})=0$, $\overline{r}^1(\{2,1\})=0$, $\overline{r}^1(\{2,2\})=8$; for agent 2, $\overline{r}^2(\{1,1\})=8$, $\overline{r}^2(\{1,2\})=0$, $\overline{r}^2(\{2,1\})=0$, $\overline{r}^2(\{2,2\})=10$. According to UCB1, both agents calculate the indices for all joint arms which are: for agent 1, $g^1(\{1,1\})=10+\sqrt{\frac{2\ln 5}{1}}$, $g^1(\{1,2\})=0+\sqrt{\frac{2\ln 5}{1}}$, $g^1(\{2,1\})=0+\sqrt{\frac{2\ln 5}{1}}$, $g^1(\{2,2\})=8+\sqrt{\frac{2\ln 5}{1}}$; for agent 2, $g^2(\{1,1\})=8+\sqrt{\frac{2\ln 5}{1}}$, $g^2(\{1,2\})=0+\sqrt{\frac{2\ln 5}{1}}$, $g^2(\{2,1\})=0+\sqrt{\frac{2\ln 5}{1}}$, $g^2(\{2,2\})=10+\sqrt{\frac{2\ln 5}{1}}$. Therefore, agent 1 selects arm 1 and believes that agent 2 will also select arm 1 while agent 2 selects arm 2 and believes that agent 1 will also select arm 2. Since the actual selected joint arm is $\{1,2\}$, the realized reward in slot 5 is

0. At this point, agent 1 updates the reward estimate for the joint arm $\{1,1\}$ and agent 2 updates the reward estimate for the joint arm $\{2,2\}$. Hence, $\overline{r}^1(\{1,1\})=5$ and $\overline{r}^2(\{2,2\})=5$. We note that the reward estimates and indices are still "symmetric" in the sense $\overline{r}^1(\{1,1\})=\overline{r}^2(\{2,2\})$, $\overline{r}^1(\{2,2\})=\overline{r}^2(\{1,1\})$, $g^1(\{1,1\})=g^2(\{2,2\})$ and $g^1(\{2,2\})=g^2(\{1,1\})$. It can be easily shown that such "symmetry" will persist for all remaining periods. Suppose at some time agent 1 believes that it needs to select $\{1,2\}$ (or $\{2,1\}$), then agent 2 will also believe that it needs to select arm $\{1,2\}$ (or $\{2,1\}$). Both agents will update the rewards of $\{1,2\}$ (or $\{2,1\}$) correctly. However, if agent 1 believes it needs to select arm $\{1,1\}$ (or $\{2,2\}$), then agent 2 will believe it needs to select arm $\{2,2\}$ (or $\{1,1\}$). Both agents will update the rewards wrongly but in the same way. Hence, in all time slots after the first 4 slots, the realized rewards are 0. Since the error sequence $\{-2,0,0,2\}$ occurs with a positive probability ($\frac{1}{3^4}$), there is a constant gap from the optimal reward. Hence, the regret bound is linear by running such UCB1 algorithm when observing the reward realization is subject to private errors.

APPENDIX C: PROOF OF THEOREM 2

It is clear that $\overline{r}^n(l)=\overline{r}^n(\boldsymbol{a}^l),\ \forall n$ where a^l is the joint arm selected in the l^{th} relative index in the exploration phase. Since each joint arm is selected once in each exploration phase, it is equivalent to the case where agents maintain the reward estimates for all joint arms. Moveover, in the exploitation slot, agents select the component arms of the joint arm that maximizes the reward estimate.

First we prove that after P exploration phases, for each agent n, the probability that a non-optimal joint arm $\boldsymbol{a}\neq\boldsymbol{a}^*$ is selected in an exploitation slot is at most $e^{-\frac{P}{2}\left(\frac{\Delta_a}{D}\right)^2}$ where $\Delta_a=\mu(\boldsymbol{a}^*)-\mu(\boldsymbol{a})$. A non-optimal joint arm $\boldsymbol{a}\neq\boldsymbol{a}^*$ is selected by agent n in an exploitation slot only if $\overline{r}^n(\boldsymbol{a})\geq\overline{r}^n(\boldsymbol{a}^*)$.

Since

$$P(\overline{r}^n(\boldsymbol{a})<\overline{r}^n(\boldsymbol{a}^*))>P(\overline{r}^n(\boldsymbol{a}^*)>\mu(\boldsymbol{a}^*)-0.5\Delta_a)\times P(\overline{r}^n(\boldsymbol{a})<\mu(\boldsymbol{a})+0.5\Delta_a) \quad (3)$$

we have,

$$\begin{aligned}&P(\overline{r}^n(\boldsymbol{a})\geq\overline{r}^n(\boldsymbol{a}^*))=1-P(\overline{r}^n(\boldsymbol{a})<\overline{r}^n(\boldsymbol{a}^*))\\&<1-P(\overline{r}^n(\boldsymbol{a}^*)>\mu(\boldsymbol{a}^*)-0.5\Delta_a)\times P(\overline{r}^n(\boldsymbol{a})<\mu(\boldsymbol{a})+0.5\Delta_a) \quad (4)\\&<P(\overline{r}^n(\boldsymbol{a}^*)\leq\mu(\boldsymbol{a}^*)-0.5\Delta_a)+P(\overline{r}^n(\boldsymbol{a})\geq\mu(\boldsymbol{a})+0.5\Delta_a)\end{aligned}$$

Since the reward is bounded by D and the reward estimate is obtained using P realizations, by Hoeffding's inequality,

$$P(\overline{r}^n(\boldsymbol{a}^*) \le \mu(\boldsymbol{a}^*) - 0.5\Delta_{\boldsymbol{a}}) = P(\overline{r}^n(\boldsymbol{a}) \ge \mu(\boldsymbol{a}) + 0.5\Delta_{\boldsymbol{a}}) \le e^{-\frac{P}{2}\left(\frac{\Delta_{\boldsymbol{a}}}{D}\right)^2} \le e^{-\frac{P}{2}\left(\frac{\Delta^{min}}{D}\right)^2} \quad (5)$$

Therefore, $P(\overline{r}^n(\boldsymbol{a}) \ge \overline{r}^n(\boldsymbol{a}^*)) \le 2e^{-\frac{P}{2}\left(\frac{\Delta^{min}}{D}\right)^2}$. Since there are L_1 sub-optimal joint arms, the probability that agent n selects any one of the sub-optimal joint arm is less than $2L_1 e^{-\frac{P}{2}\left(\frac{\Delta^{min}}{D}\right)^2}$. Since there are N agents, the probability that there is at least one agent that selects the joint arm is less than $2L_1 N e^{-\frac{P}{2}\left(\frac{\Delta^{min}}{D}\right)^2}$.

Now we bound the regret of the DisCo algorithm. The regret consists of two parts $R(T) = R_1(T) + R_2(T)$ where $R_1(T)$ is the regret incurred in the exploration phases and $R_2(T)$ is the regret incurred in the exploitation phases up to slot T.

We bound $R_1(T)$ first. Since the algorithm starts with the exploration phase, it ensures that, at any time t, at most $\lceil \zeta(T) \rceil$ exploration phases have been gone through. In each exploration phase, the maximum regret is achieved when the agents select the worst joint arm in every slot and hence, the regret in one exploration phase is bounded by $L_1\Delta^{max}$. Therefore, $R_1(T)$ is at most

$$R_1(T) < \lceil \zeta(T) \rceil L_1\Delta^{max} \le (\zeta(T)+1)L_1\Delta^{max} < AL_1\Delta^{max}\ln T + L_1\Delta^{max} \quad (6)$$

Next we bound $R_2(T)$. We know that, at any time slot $t < T$ when it is an exploitation slot, the probability that a non-optimal joint arm is selected is at most $2L_1 N e^{-\frac{\zeta(t)}{2}\left(\frac{\Delta^{min}}{D}\right)^2}$ since the algorithm ensures that at any exploitation slot at least $\lceil \zeta(t) \rceil$ exploration phases have been gone through. Therefore, the expected regret in any exploitation slot by selecting a non-optimal joint arm is at most

$$2L_1 N e^{-\frac{\zeta(t)}{2}\left(\frac{\Delta^{min}}{D}\right)^2}\Delta^{max} \le 2L_1 N e^{-\frac{A\ln t}{2}\left(\frac{\Delta^{min}}{D}\right)^2}\Delta^{max} \quad (7)$$

Therefore, the expected regret $R_2(T)$ incurred in the exploitation phase is bounded by

$$R_2(T) \le \sum_t^{\infty} 2NL_1\Delta^{max} t^{-\frac{A}{2}\left(\frac{\Delta^{min}}{D}\right)^2} \quad (8)$$

If we let $A > 2\left(\frac{D}{\Delta^{min}}\right)^2$, then $\sum_t^{\infty} t^{-\frac{A}{2}\left(\frac{\Delta^{min}}{D}\right)^2}$ is finite. Combining the bounds on $R_1(T)$ and $R_2(T)$ we get the result.

APPENDIX D: PROOF OF THEOREM 3

First we prove that after P exploration phases, the probability that a non-optimal arm $a_n \neq a_n^*$ is selected by agent n is at most $2e^{-\frac{P}{2}\left(\frac{\Delta_n^{min}}{D}\right)^2}$. A non-optimal arm $a_n \neq a_n^*$ is selected only if $\overline{r}^n(a_n) > \overline{r}^n(a_n^*)$. The proposed algorithm ensures that in the n^{th} exploration subphase, agents $i \neq n$ are selecting the same arms while agent n is learning each of its own arms. Let $\boldsymbol{a}_{-n}(p)$ be the set of arms selected by other agents in the n^{th} subphase of the $p^{th}(p \leq P)$ exploration phase. Then the expectation of the reward estimate $\overline{r}^n(a_n)$ is

$$\mathbb{E}[\overline{r}^n(a_n)] = \frac{1}{P}\sum_{p=1}^{P} \mu(a_n, \boldsymbol{a}_{-n}(p)) \tag{9}$$

Since $\mu(a_n^*, \boldsymbol{a}_{-n}) - \mu(a_n, \boldsymbol{a}_{-n}) \geq \Delta_n^{min}, \forall \boldsymbol{a}_{-n}$, we also have

$$\mathbb{E}[\overline{r}^n(a_n^*)] - \mathbb{E}[\overline{r}^n(a_n)] \geq \Delta_n^{min} \tag{10}$$

Because the realized reward is bounded, according to Hoeffding's inequality, we have

$$P(\overline{r}^n(a_n) > \overline{r}^n(a_n^*)) \leq 2e^{-\frac{P}{2}\left(\frac{\Delta_n^{min}}{D}\right)^2} \tag{11}$$

Therefore, the probability that agent n selects a suboptimal arm is at most $2K_n e^{-\frac{P}{2}\left(\frac{\Delta_n^{min}}{D}\right)^2}$

Now, we prove the regret bound of the learning algorithm which consists of two parts $R(T) = R_1(T) + R_2(T)$ where $R_1(T)$ is the regret incurred in the exploration phases and $R_2(T)$ is the regret incurred in the exploitation phases up to slot T.

First we bound $R_1(T)$. In each exploration phase, the maximum regret is achieved when the agents select the worst joint arm in every slot and hence, the regret in one exploration phase is bounded by $L_2\Delta^{max}$. Since the algorithm ensures that at any time T, at most $\lceil\zeta(T)\rceil$ exploration phases have been gone through, $R_1(T)$ is at most

$$R_1(T) < \lceil\zeta(T)\rceil L_2\Delta^{max} \leq (\zeta(T)+1)L_2\Delta^{max} = AL_2\Delta^{max}\ln T + L_2\Delta^{max} \tag{12}$$

Next we bound $R_2(T)$. At any time $t < T$ when it is an exploitation period, the expected regret by choosing a non-optimal arm is at most

$$2\sum_{n=1}^{N} K_n e^{-\frac{\zeta(T)}{2}\left(\frac{\Delta_n^{min}}{D}\right)^2}\Delta^{max} \leq 2L_2 e^{-\frac{A\ln T}{2}\left(\frac{\Delta_{FI}^{min}}{D}\right)^2}\Delta^{max} \tag{13}$$

Hence, the expected regret in the exploitation slots up to time T is at most

$$R_2(T) = \sum_{t=1}^{T}\sum_{n=1}^{N} 2K_n \Delta^{max} t^{-\frac{A}{2}\left(\frac{\Delta_{FI}^{min}}{D}\right)^2} < 2L_2\Delta^{max}\sum_{t=1}^{\infty} t^{-\frac{A}{2}\left(\frac{\Delta_{FI}^{min}}{D}\right)^2} \quad (14)$$

Because $A \geq 2\left(\frac{D}{\Delta_{FI}^{min}}\right)^2$, $\sum_{t=1}^{\infty} t^{-\frac{A}{2}\left(\frac{\Delta_{FI}^{min}}{D}\right)^2}$ is finite. Combining the bounds on $R_1(T)$ and $R_2(T)$ we get the result.

APPENDIX E: PROOF OF THEOREM 4

Using the similar techniques in the proofs of Theorem 1 and Theorem 2, we can show that after P exploration phases, the probability that a non-optimal group-joint arm is selected by at least one agent in group g_m is at most $2N_m S_m e^{-\frac{P}{2}\left(\frac{\Delta_m^{min}}{D}\right)^2}$. With this, we prove the regret bound of the DisCo-GO algorithm. The regret consists of two parts $R(T) = R_1(T) + R_2(T)$ where $R_1(T)$ is the regret incurred in the exploration phases and $R_2(T)$ is the regret incurred in the exploitation phases up to slot T.

First we bound $R_1(T)$. In each exploration phase, the maximum regret is achieved when the agents select the worst joint arm in every slot and hence, the regret in one exploration phase is bounded by $L_3\Delta^{max}$. Since the algorithm ensures that at any time T, at most $\lceil\zeta(T)\rceil$ exploration phases have been gone through, $R_1(T)$ is at most

$$R_1(T) < \lceil\zeta(T)\rceil L_3\Delta^{max} \leq AL_3\Delta^{max}\ln T + L_3\Delta^{max} \quad (15)$$

Next we bound $R_2(T)$. At any time $t < T$ when it is an exploitation slot, the expected regret by choosing a non-optimal arm $\boldsymbol{a}_m \neq \boldsymbol{a}_m^*$ for agent group m is at most

$$2N_m S_m \Delta^{max} e^{-\frac{\zeta(T)}{2}\left(\frac{\Delta_m^{min}}{D}\right)^2} = 2N_m S_m \Delta^{max} t^{-\frac{A}{2}\left(\frac{\Delta_{PI}^{min}}{D}\right)^2} \quad (16)$$

Hence, the expected regret in the exploitation slots up to time T is at most

$$R_2(T) = \sum_{t=1}^{T}\sum_{m=1}^{M} 2N_m \prod_{n\in g_m} K_n \Delta^{max} t^{-\frac{A}{2}\left(\frac{\Delta_{PI}^{min}}{D}\right)^2} < 2\sum_{m=1}^{M} N_m \prod_{n\in g_m} K_n \Delta^{max}\sum_{t=1}^{\infty} t^{-\frac{A}{2}\left(\frac{\Delta_{PI}^{min}}{D}\right)^2} \quad (17)$$

Because $A \geq 2\left(\frac{D}{\Delta_{PI}^{min}}\right)^2$, $\sum_{t=1}^{\infty} t^{-\frac{A}{2}\left(\frac{\Delta_{PI}^{min}}{D}\right)^2}$ is finite. Combining the bounds on $R_1(T)$ and $R_2(T)$ we get the result.

APPENDIX F: PROOF OF PROPOSITION 2

We only prove for the basic DisCo algorithm. Recall that the regret consists of two parts $R(T) = R_1(T) + R_2(T)$ where $R_1(T)$ is the regret incurred in the exploration phase and $R_2(T)$ is the regret incurred in the exploitation phase. In the case with missing feedbacks, the expected length of each exploration phase is L_1 / p. Therefore $R_1(T)$ can be bounded by

$$R_1(T) < \left\lceil \zeta(T) \right\rceil L_1 / p\Delta^{max} \le A / pL_1\Delta^{max} \ln T + 1 / pL_1\Delta^{max} \tag{18}$$

$R_2(T)$ is bounded in the same way as that in the case without missing feedbacks. Hence, we get the desired result.

APPENDIX G: PROOF OF PROPOSITION 3

We only prove for the basic DisCo algorithm. Recall that the regret consists of two parts $R(T) = R_1(T) + R_2(T)$ where $R_1(T)$ is the regret incurred in the exploration phase and $R_2(T)$ is the regret incurred in the exploitation phase. In the case with delays, $R_1(T)$ can be bounded by

$$\begin{aligned} R_1(T) &< \left\lceil \zeta(T) \right\rceil L_1\Delta^{max} \le (\zeta(T)+1)(\prod_{n=1}^{N} K_n + L_{max})\Delta^{max} \\ &< (\zeta(T)+1)(\prod_{n=1}^{N} K_n + L_{max})\Delta^{max} + (\zeta(T)+1)L_{max}\Delta^{max} \end{aligned} \tag{19}$$

$R_2(T)$ is bounded in the same way as that in the case without delays. Hence, we get the desired result.

APPENDIX H: PROOF OF PROPOSITION 4

We only prove for the first case. At any exploitation slot t, the reward estimates use A reward realizations in the most recent Γ slots. Due to the slowly varying expected rewards, $P(\overline{r}_t(\boldsymbol{a}) \ge \overline{r}_t(\boldsymbol{a}^*(t))) < 2e^{-\frac{A}{2}\left(\frac{\underline{\Delta}^{min} - 2c_\Gamma}{D}\right)^2}$. Therefore, the expected reward loss in any exploitation slot t is bounded by $2NL_1 e^{-\frac{A}{2}\left(\frac{\underline{\Delta}^{min} - 2c_\Gamma}{D}\right)^2} \overline{\Delta}^{max}$. Because in any time window Γ, there are at most A exploration phases and hence, at most AL_1 exploration slots. Because the regret in an exploration slot is at most $\overline{\Delta}^{max}$, hence the time average regret is at most

$$\frac{R(T)}{T} \le \frac{AL_1 + 2(\Gamma - AL_1)NL_1 e^{-\frac{A}{2}\left(\frac{\underline{\Delta}^{min} - 2c_\Gamma}{D}\right)^2}}{\Gamma} \overline{\Delta}^{max} \tag{20}$$