Multi-agent Multi-armed Bandit Regret Complexity and Optimality

Anonymous submission
Anonymous affiliation

A Discussions on the linear regret lower bound

As part of the contributions of this paper, it is shown that the regret lower bound is of order $\Omega(T)$ when the graph is disconnected. However, we would like to include that it is possible to get sublinear regret when the graph is disconnected by adding some assumptions on the problem settings as follows. The assumptions could be regarding both graphs and rewards which determine the problem complexity. Uniformly strongly connected graphs, give $O(\log T)$ [Zhu and Liu, 2023]. Also, with random graphs, e.g. the E-R model where each graph observation can be disconnected, the regret is $O(\log T)$ and $O(\sqrt{T})$ [Xu and Klabjan, 2023a]. For other disconnected graphs, sublinear regret is ensured if each client shares the same mean reward values (homogeneous) and plays with their own MAB optimally. More broadly, having heterogeneity within each connected component while ensuring homogeneity across these components is adequate. This paper shows that linear regret results from the difference in optimal arms. If monotonicity of rewards over arms or the choice of the optimal arm is the same across the connected components, then designing optimal methods within the connected components ensures sublinear regret.

B Proof of Results in Section 4

B.1 Proof of Theorem 2

Proof. On a complete graph, each client can observe the rewards of all arms at M clients, where the number of observations is thereby upper bounded by KM. Henceforth, we consider Theorem 4 in [Shamir, 2014] to obtain

$$R_T^F \geq \sqrt{\frac{KT}{1+KM}} = \Omega(\sqrt{T}).$$

This completes the first part of the statement.

For the instance-dependent regret lower bounds, we assume that the number of arms is 2 and the rewards of arms satisfies the assumptions in [Goldenshluger and Zeevi, 2013]. Then based on the result established by specifying a contextual linear bandit with $\alpha = 1$ as in [Goldenshluger and Zeevi, 2013], which reads as Theorem 2, we obtain

$$R_T^F \geq \Omega(\log T).$$

We add that the lower bound result for the bandit setting holds for the full-information setting by noting the analysis essentially uses the observations that are given by the full information setting.

This concludes the instance-dependent lower bound in the full information setting and thereby completes the proof.

□

B.2 Proof of Theorem 3

Proof. The instance-dependent regret bound presents nontrivial challenges to the analysis. We start with complete graphs. We specify $K = 2$ and assume $\mu_1 > \mu_2$ without loss of generality. Consider the centralized problem which has times when the clients pull the same arm (agreement) and times when the clients pull distinct arms (disagreement). We denote the number of time steps of agreement and disagreement as T_a and T_d, respectively. We observe that $T_a + T_d = T$. For T_d, there exist clients pulling the worse arm, which implies that for any policy $\pi \in \Pi_B$

$$\begin{aligned} R_T^\pi &= \frac{1}{M}\sum_m \sum_{t\in T_d} (\mu_1 - \mu_{a_t^m}) + \frac{1}{M}\sum_m \sum_{t\in T_a} (\mu_1 - \mu_{a_t^m}) \\ &= \sum_{t\in T_d} \Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a} (\mu_1 - \mu_{a_t^m}) \\ &= T_d\Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a} (\mu_1 - \mu_{a_t^m}). \end{aligned} \tag{1}$$

Note that when $T_d = \Omega(\log T)$, we immediately derive that $E[R_T^B] \geq \Omega(\log T)$, which concludes the proof.

From now on, we assume $T_d = o(\log T)$, which implies that $T_a = T - o(\log T)$ and $\frac{T_a}{T} \to 1$ as T goes to ∞. We denote the value $t_0 = \log T$ and divide the time horizon into $\bigcup_{j=0}^{t_0} [2^j, 2^{j+1} - 1]$. It is clear that 1) the number of intervals is $\log T$ and 2) the length of the j^{th} interval is 2^{j-1}. Let $t_d = \max\{t | t \in T_d\} + 1$. Since $T_d = o(\log T)$, we have $|[t_d, T]| \geq 2^{\frac{1}{2}\log T}$ for all large enough T.

Meanwhile, we observe that for T_a, it is equivalent to a single-agent multi-objective bandit problem [Xu and Klabjan, 2023b]

since the global reward of a single arm i is given as a reward vector $(r_i^{m,t})_{m=1}^M$ and is revealed to all the clients at each time step.

Note that $\frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1 - \mu_{a_t^m}^m) = \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1 - \mu_{a_t}^m) = \sum_{t\in T_a}(\mu_1 - \mu_{a_t})$ where the first equality is by the definition of T_a and the second equality uses the definition of μ_1 and μ_{a_t}. We denote $T_a^d = T_a \cap [t_d, T] = [t_d, T]$.

At the same time, the Pareto pseudo regret reads $R_{T_a^d,M} = Dist(\sum_{t\in T_a^d}(\mu_{a_t}^m)_m, O)$ where $Dist(\cdot)$ is the distance measure between a reward vector and the Pareto optimal set O as introduced in [Xu and Klabjan, 2023b], and satisfies that $R_{T_a^d,M} \geq \Omega(\log T_a^d)$ for any policy $\{a_t\}$ based on Theorem 6 in [Xu and Klabjan, 2023b].

By specifying the rewards homogeneous, i.e. $\mu_{a_t}^1 = \mu_{a_t}^2 = \ldots = \mu_{a_t}^M$ and following a similar analysis as on Theorem 6 in [Xu and Klabjan, 2023b], we obtain $R_{T_a^d,M} = Dist(\sum_{t\in T_a^d}(\mu_{a_t}^m)_m, O) = \sum_{t\in T_a^d}(\mu_1 - \mu_{a_t})$ which yields

$$\sum_{t\in T_a}(\mu_1 - \mu_{a_t}) \geq \sum_{t\in T_a^d}(\mu_1 - \mu_{a_t})$$
$$\geq \Omega(\log T_a^d) = \Omega(\log(2^{\frac{1}{2}\log T})) = \Omega(\log T). \quad (2)$$

To put everything together, we have that for any policy $\pi \in \Pi_B$ $R_T^\pi \geq T_d\Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1 - \mu_{a_t^m})) \geq \Omega(\log T)$ where the second inequality holds by (2).

Subsequently, we obtain $\min_{\pi\in\Pi_B} R_T^\pi \geq T_d\Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1 - \mu_{a_t^m})) \geq \Omega(\log T)$, which concludes the analysis of complete graphs.

The remaining cases follow from the monotonicity of the regret in the graph complexity as follows. We first consider the full-information setting. For any $0 < c \leq 1$, we denote $\sigma_c^t = \sigma(\{\{I_j^s\}_{j\in\mathcal{N}_m^c(s)}\}_{s\leq t})$. We observe that $\sigma_1^t = \sigma(\{I_1^s, \ldots, I_M^s\}_{s\leq t})$. We have $\sigma_c^t \subset \sigma_1^t$. We define policy set Π_c as $\{f_t\}$ where the domain of f_t is on σ_c^{t-1}.

For any policy $\pi \in \Pi_c$, i.e. $\pi = \{h_t\}_{t=1}^T$, we have that it only leverages the neighborhood information σ_c^{t-1} to determine a decision rule at each time step. Since $\sigma_c^{t-1} \subset \sigma_1^{t-1}$, σ_1^{t-1} also has the neighborhood information that h_t requires. This leads to $\pi \in \Pi_1$, and subsequently yields $\Pi_c \subset \Pi_1$. We hence obtain that in the full-information setting $\min_{\pi\in\Pi_1} R_T^\pi \leq \min_{\pi\in\Pi_c} R_T^\pi$.

By the above discussion on c and the statement for complete graphs, or equivalently, with respect to Π_1, we obtain $\Omega(\log T) \leq \min_{\pi\in\Pi_1} R_T^\pi$, in the instance-dependent sense and subsequently $\Omega(\log T) \leq \min_{\pi\in\Pi_c} R_T^\pi$.

By Theorem 1, we have $R_T^B \geq \Omega(\log T)$. This completes the E-R case. All remaining cases follow the same logic.

□

B.3 Proof of Theorem 4

Proof. Consider a disconnected graph G with a clique connected component C_G including clients $c_1, \ldots, c_Q$ without loss of generality. Since G is disconnected, for any other node $m \notin V(C_G)$, there is no path between m and any node in C_G.

Let $\Delta > 0$. For client $m \notin C_G$, the reward distributions read as $(\frac{M-1}{M-Q}\Delta, 0, \ldots, 0)$, which indicates that the optimal arm is arm 1. For client $m \in C_G$, however, the reward distribution reads as $(0, \frac{2}{Q}\Delta, 0, \ldots, \ldots, 0)$, implying that arm 2 is the optimal arm. It is straight-forward that the global mean reward value of arm 1 is $\frac{(M-1)}{M}\Delta$ that is larger than that of arm 2 which is $\frac{2\Delta}{M}$. The subsequent sub-optimality gap is $\Delta_2 = \frac{M-3}{M}\Delta$. Any no-regret (consistent as proposed in [Lattimore and Szepesvári, 2020]) algorithms π at client $j \in C_G$, where the regret with respect to the available information is defined on the rewards of client $j \in C_G$, leads to $E[n_{j,2}(T)] = O(T)$. However, in this situation, the global regret satisfies

$$\begin{aligned} E[R_T^\pi] &= \frac{1}{M}\sum_m \sum_{t=1}^T (E[\mu_1 - \mu_{a_t^m}]) \\ &\geq \frac{1}{M}\sum_{t=1}^T (E[\mu_1 - \mu_{a_t^j}]) \\ &\geq \frac{1}{M} E[n_{j,2}(T)] \cdot \Delta_1 \\ &= \frac{1}{M}\cdot\frac{M-3}{M}\Delta\cdot\Omega(T) = \Omega(T) \end{aligned}$$

where the first inequality is by only considering client j and the second inequality uses the fact that arm 2 is not a global optimal arm.

This completes the proof of the linear regret in the case when clients perform local consistent learning on disconnected graphs. □

B.4 Proof of Theorem 5

Proof. Again, we consider a disconnected graph G with a clique C_G including clients $c_1, \ldots, c_Q$ without loss of generality.

We assume there are two arms labeled as arm 1 and 2 and consider the instance at clients as follows by referencing [Alon *et al.*, 2015]. Let random variable X follow a uniform distribution in $\{0, 1\}$ and be fixed once determined, and for any time step t, the reward $r_k^j(t)$ is generated as for any $j \notin C_G$, $r_k^j(t) = \begin{cases} X & \text{arm 1} \\ \frac{1}{2} & \text{arm 2} \end{cases}$ and for any $j \in C_G$, we have $r_k^j(t) = \begin{cases} \frac{1}{2} & \text{arm 1} \\ \frac{1}{2} & \text{arm 2} \end{cases}$ where the random variable X is independent of everything at client $j \in C_G$ as client $j \in C_G$ only has the information of their own arms. We have $\Delta_2 = \frac{1}{2(M-Q)}$, no matter what value X takes since it only changes the choice of optimal arms. Specifically, when $X = 1$, the global optimal arm is arm 1 and the suboptimality gap is $\Delta_2 = \mu_1 - \mu_2 = (1 - \frac{1}{2})/(M - Q)$. When $X = 0$, the global optimal arm is arm 2 and the suboptimality gap is $\Delta_2 = \mu_2 - \mu_0 = (\frac{1}{2} - 0)/(M - Q)$, the other way around.

Subsequently, we consider the regret at client $j \in C_G$ to obtain

$$
\begin{aligned}
E[R_T^\pi] &= \frac{1}{M}\sum_m \sum_{t=1}^{T}(E[\mu_* - \mu_{a_t^m}]) \\
&\geq \frac{1}{M}\sum_{t=1}^{T}(E[\mu_* - \mu_{a_t^M}]) \\
&= \frac{1}{M}(\frac{1}{2}E[\Delta n_{j,1}(T)|X=0] + \frac{1}{2}E[\Delta(T - n_{j,1}(T))|X=1]) \\
&= \frac{1}{M}(\frac{1}{2}E[\Delta n_{j,1}(T)] + \frac{1}{2}E[\Delta(T - n_{j,1}(T))]) \\
&= \frac{\Delta}{4M(M-Q)}T = \Omega(T)
\end{aligned}
$$

where the first inequality uses the non-negativity of value $\mu_* - \mu_{a_t^m}$ and the third equality leverages the independence between X and client j. □

B.5 Proof of Theorem 6

Proof. We show the mean-gap free regret lower bound starting with complete graphs. Note that a complete graph is equivalent to a centralized problem with M agents. This implies that each client can observe the reward of multiple arms by communicating with $M-1$ neighbors, where the number of observations is thereby upper bounded by M. Henceforth, we consider Theorem 4 in [Shamir, 2014] and obtain

$$R_T^B \geq \sqrt{\frac{KT}{1+M}} = \Omega(\sqrt{T}).$$

This completes the proof of the complete graphs.

Regarding the monotonicity of the regret in the graph complexity, the proof follows the proof of Theorem 3.

□

B.6 Proof of Theorem 8

Proof. Note that the graph structure determines the communication efficiency of the clients. To consider the lower bound, we leverage sparse graphs in the connected graph family to perform the worst-case scenario analysis.

Specifically, we consider the designed graph consisting of clients $1, \ldots, M$ in this order. It takes exactly $O(M)$ time steps for client 1 to obtain the information of client M, which results in a deterministic delay.

If $I_0 = \{1, \ldots, \frac{M}{4}\}$ and $I_1 = \{\frac{3M}{4}, \ldots, M\}$, then the shortest path d_p from I_0 to I_1 meets the condition

$$d_p \geq \Omega(\frac{M+1}{3}).$$

By the choice of M such that $M > \Omega(T^{\frac{1}{3}})$, we obtain

$$d_p \geq \Omega(T^{\frac{1}{3}}). \tag{3}$$

We star with a full-information setting. Following a similar argument and constructing the same instance as in Lemma A.4 in [Yi and Vojnović, 2023], we arrive that in the full-information setting

$$R_T \geq \Omega(\sqrt{d_p \cdot T}).$$

Subsequently, we obtain that

$$
\begin{aligned}
R_T &\geq \Omega(\sqrt{d_p \cdot T}) \\
&= \Omega(\sqrt{T} \cdot \sqrt{d_p}) \\
&\geq \Omega(\sqrt{T} \cdot T^{\frac{1}{6}}) = \Omega(T^{\frac{2}{3}})
\end{aligned}
$$

where the last inequality is by (3). Equivalently, we write it as

$$R_T^F \geq \Omega(T^{\frac{2}{3}}). \tag{4}$$

Meanwhile, by Theorem 1, we have that the regret lower bound in the bandit setting is larger than the regret in the full information setting and thus by (4) we obtain

$$R_T^B \geq \Omega(T^{\frac{2}{3}}).$$

This completes the proof of Theorem 8.

□

B.7 Proof of Theorem 9

Proof. Let $M \, mod \, 4 = 0$ and $T > 8$. Denote expanders of size $\frac{M}{4}$ as two disjoint subsets of nodes $I_0 = \{1, 2, \ldots, \frac{M}{4}\}$ and $I_1 = \{\frac{3}{4}M, \frac{3}{4}M+1, \ldots, M\}$. Note that $|I_0| = |I_1| = \frac{M}{4}$. By the definition of G_t, the shortest path distance between I_0 and I_1 is $d \geq \frac{\eta M}{8}$. We set $\epsilon = \sqrt{\frac{4}{\eta}}\frac{M^2}{2}T^{-\frac{1}{3}}$. It follows $8\epsilon^2 d \leq 1$.

Let B_1 be Bernoulli with probability $\frac{1}{2} + \epsilon$ and B_2 Bernoulli with probability $\frac{1}{2}$. Consider the bandit problem as follows. Let X be a random variable following a uniform distribution on $\{0, 1, \ldots, \frac{M}{4}\}$. For client $X \geq 1$, arm 1 follows B_1 and arm 2 follows B_2. For $i \in I_0 \backslash \{X\}$, let the arms follow B_2. All clients not in I_0 have all rewards 0.

Additionally, we re-sample random variable X every d steps, i.e. we re-specify the client X if $X \geq 1$. If $X = 0$, all clients have reward based on B_2. We denote the number of such re-sampling steps as D, $D = \lfloor \frac{T}{d} \rfloor$, which leads to a sequence $\{X_1, X_2, \ldots, X_D\}$. The following holds for $i \in I_0$. Subsequently, let us define distribution $Q_j^i(arm) = P(arm|X_j = i)$ and $Q_j^{-1}(arm) = P(arm|X_j = 0)$. Note that Q_j^{-1} represents that all clients in I_0 share the same reward distribution. Let $Q_{j,t}^i(arm) = P(arm|\sigma_t, X_j = i)$ and $Q_{j,t}^{-1}(arm) = P(arm|\sigma_t, X_j = 0)$. It is easy to verify that $D_{KL}(Q_{j,t}^{-1}, Q_{j,t}^i) = \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}-\epsilon} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}+\epsilon} = \frac{1}{2}\log(1 + \frac{4\epsilon^2}{1-4\epsilon^2}) \leq \frac{1}{2} \cdot \frac{4\epsilon^2}{1-4\epsilon^2} \leq 4\epsilon^2$, where the first inequality uses the fact that $\log(1+x) \leq x$ and the second inequality holds by the choice of $\epsilon = \frac{M^2}{2}T^{-\frac{1}{3}} \leq \frac{1}{4}$ since $T > 8$.

Therefore, by the chain rule for relative entropy, we obtain $D_{KL}(Q_j^{-1}, Q_j^i) = \sum_{t=jd}^{(j+1)d} D_{KL}(Q_{j,t}^{-1}, Q_{j,t}^i) \leq$

$\sum_{t=jd}^{(j+1)d} 4\epsilon^2 \leq 4\epsilon^2 d$. By the Pinsker's inequality we have that $D_{TV}(Q_j^{-1}, Q_j^i) \leq \sqrt{\frac{D_{KL}(Q_j^{-1}, Q_j^i)}{2}} \leq \epsilon\sqrt{2d}$. (5)

The expected reward of arm 1 is $\frac{1}{8} + \frac{1}{M}\frac{|I_0|}{|I_0|+1}\epsilon$ from

$$\begin{aligned}
\mu_1 &= \frac{1}{M}\sum_{m=1}^{M}\mu_1^m = \frac{1}{M}\sum_{m\in I_0}\mu_1^m + \frac{1}{M}\sum_{m\notin I_0}\mu_1^m \\
&= \frac{1}{M}\sum_{m\in I_0}\Big[E[\mu_1^m|X_1\in I_0]P(X_1\in I_0)+ \\
&\qquad \sum_{m\in I_0} E[\mu_1^m|X_1\notin I_0]P(X_1\notin I_0)\Big] + \frac{1}{M}\sum_{m\notin I_0} 0 \\
&= \frac{1}{M}(\frac{|I_0|}{|I_0|+1}(\frac{1}{2}+\epsilon+\frac{1}{2}(|I_0|-1))+ \\
&\qquad \frac{1}{|I_0|+1}(\frac{1}{2}+\frac{1}{2}(|I_0|-1)) = \frac{1}{8}+\frac{1}{M}\frac{|I_0|}{|I_0|+1}\epsilon
\end{aligned}$$

and of arm 2 is $\frac{1}{8}$ from $\mu_2 = \frac{1}{M}\sum_{m=1}^{M}\mu_2^m = \frac{1}{M}\sum_{m\in I_0}\mu_2^m + \frac{1}{M}\sum_{m\notin I_0}\mu_2^m = \frac{1}{M}\sum_{m\in I_0}\frac{1}{2} + \frac{1}{M}\sum_{m\notin I_0} 0 = \frac{1}{8}$. As a result $\Delta_1 = \frac{\epsilon}{M}\frac{|I_0|}{|I_0|+1} \geq \frac{\epsilon}{2M}$ since $|I_0| \geq 1$. Let us denote by $n_{m,1}(T,j)$ the number of pulls of arm 1 by client m during the j^{th} epoch which is the optimal arm. Therefore, we obtain

$$\begin{aligned}
&E[R_T^B] = E[E[R_T^B|X_1,\ldots,X_D]] \qquad (6)\\
&= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(T-n_{m,1}(T)))|X_1,\ldots,X_D]] \\
&= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(\sum_{j=1}^{D}d-\sum_{j=1}^{D}n_{m,1}(T,j)))|X_1,\ldots,X_D]] \\
&= E[\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}E[(\frac{\epsilon}{2M}(d-n_{m,1}(T,j)))|X_1,\ldots,X_D]] \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}E[E[(\frac{\epsilon}{2M}(d-n_{m,1}(T,j)))|X_j]] \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}\frac{E[(\frac{\epsilon}{2M}(d-n_{m,1}(T,j)))|X_j=i]]}{|I_0|+1} \\
&\geq \frac{1}{2M^2}(\frac{1}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}E[\epsilon\cdot(d-n_{1,1}(T,j))|X_j=i]) \\
&= \frac{1}{2M^2}(\epsilon\cdot T - \frac{\epsilon}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}E_{Q_j^i}[(n_{1,1}(T,j))])
\end{aligned}$$

where the first and fifth equality use the law of total expectation, the third equality is by the fact that $T = \sum_{j=1}^{D} d$ and $\sum_{j=1}^{D} n_{m,1}(T,j) = n_{m,1}(T)$, and the sixth equality uses the distribution of X_j defined by $P(X_j = i) = \frac{1}{|I_0|+1}$ for $i \in I_0 \cup \{0\}$.

Note that $E_{Q_j^i}[(n_{1,1}(T,j))] - E_{Q_j^{-1}}[(n_{1,1}(T,j))] = \sum_{t=jd}^{(j+1)d}(Q_j^i(a_t^1=1) - Q_j^{-1}(a_t^1=1)) \leq d\cdot D_{TV}(Q_j^{-1}, Q_j^i)$ where the last inequality is by the definition of the total variation D_{TV}.

This immediately gives us that

$$\begin{aligned}
&\sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}E_{Q_j^i}[(n_{1,1}(T,j))] \\
&\leq \sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}\sum_{t=jd}^{(j+1)d}(Q_j^{-1}(a_t^1=1) + d\cdot D_{TV}(Q_j^i, Q_j^{-1})) \\
&\leq T + d\sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}D_{TV}(Q_j^i, Q_j^{-1})) \\
&\leq T + d\sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}(\epsilon\sqrt{2d}) \\
&= T + dD\epsilon\sqrt{2d}(|I_0|+1) = T + T\cdot\frac{|I_0|+1}{4}
\end{aligned}$$

where the second inequality uses $\sum_i Q_j^{-1}(a_t^1=1) = 1$ and $dD = T$, and the third inequality uses (5), and the last equality holds by the choices of d and ϵ that satisfy $\epsilon\sqrt{2d}(|I_0|+1) \leq \frac{|I_0|+1}{4}$. Here we also use the lower bound on η.

Consequently, we arrive at $E[R_T^B] \geq \frac{1}{2M^2}(\epsilon T - \frac{\epsilon(T+T\cdot\frac{|I_0|+1}{4})}{|I_0|+1}) \geq \frac{1}{2M^2}\frac{1}{4}\epsilon T = \Omega(T^{\frac{2}{3}})$ where the last inequality uses $|I_0| = \frac{M}{4} \geq 2$ and the equality holds by the choice of ϵ and M. □

References

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances in Neural Information Processing Systems*, 2023.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, pages 38499–38517. PMLR, 2023.

Jialin Yi and Milan Vojnović. Doubly adversarial federated bandits. *arXiv preprint arXiv:2301.09223*, 2023.

Jingxuan Zhu and Ji Liu. Distributed multi-armed bandits. *IEEE Transactions on Automatic Control*, 2023.