Supplementary Material: Decentralized Multi-player Multi-armed Bandits with No Collision Information

Chengshuai Shi
University of Virginia

Wei Xiong
University of Virginia

Cong Shen
University of Virginia

Jing Yang
Pennsylvania State University

A Error Correction Codes for Communication over the Z-channel

More details about the representative coding techniques for the Z-channel are provided in this section.

A.1 Repetition code

Repetition code seems simple but is surprisingly powerful in the Z-channel. Chen et al. (2013) has proved that it is the optimal code for $Q = 1$. The encoding and decoding processes are described as follows.

- Encoding. Repeat bit 0 or bit 1 in message $\boldsymbol{m}$ for A times to generate codeword $\boldsymbol{X}$.
- Decoding. For channel output $\boldsymbol{Y}$, if there exists i such that $\boldsymbol{Y}[i] \neq 0$, then the decoder outputs 0. Otherwise, we have $\boldsymbol{Y}[i] = 0$ for all i, and the decoder outputs 1.

With a crossover probability no larger than $1 - \mu_{\min}$, the bit error probability is:

$$P(Y_i \neq X_i) < (1 - \mu_{\min})^A.$$

For a message length of Q bits, the error probability is:

$$\begin{aligned}
P_e &= P(\exists i, Y_i \neq X_i) \\
&= 1 - P(Y_i = X_i)^Q \\
&\leq 1 - (1 - (1 - \mu_{\min})^A)^Q \\
&\leq Q e^{-\mu_{\min} A}.
\end{aligned}$$

With the choice of $A = \lceil \frac{\log(QT)}{\mu_{\min}} \rceil$, we have $P_e < \frac{1}{T}$. Thus, the total code length for a Q-bit message is:

$$N_{rep} = Q \left\lceil \frac{\log(QT)}{\mu_{\min}} \right\rceil.$$

With $N_{rep} = \Theta(\log(T))$, the regret remains order-optimal.

A.2 Flip code

The flip code is designed by Chen et al. (2013) to better utilize the Z-channel property. The encoding and decoding processes are illustrated with the case of 4 codewords as follows.

- Encoding. Assuming we encode every two bits into a $2A$-bit codeword, the encoding function is:

$$(0,0) \rightarrow (\underbrace{1,...,1}_{A}, \underbrace{1,...,1}_{A});\ (0,1) \rightarrow (\underbrace{1,...,1}_{A}, \underbrace{0,...,0}_{A});$$

$$(1,0) \rightarrow (\underbrace{0,...,0}_{A}, \underbrace{1,...,1}_{A});\ (1,1) \rightarrow (\underbrace{0,...,0}_{A}, \underbrace{0,...,0}_{A}).$$

- Decoding. It is similar to the repetition code. A codeword $\boldsymbol{m}$ of length $2A$ will be divided into $\boldsymbol{m_1}$ of length A and $\boldsymbol{m_2}$ of length A

- if all bits in $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are 1s, decoder outputs $(0,0)$;
- if all bits in $\boldsymbol{m_1}$ are 1s and $\boldsymbol{m_2}$ contains 0, decoder outputs $(0,1)$;
- if $\boldsymbol{m_1}$ contains 0 and all bits in $\boldsymbol{m_2}$ are 1s, decoder outputs $(1,0)$;
- for all other cases, decoder outputs $(1,1)$.

With a crossover probability no larger than $1-\mu_{\min}$, the bit error probability is (Chen et al., 2013):

$$P(Y_i \neq X_i) \leq (1-\mu_{\min})^A - \frac{1}{4}(1-\mu_{\min})^{2A}$$

The inequality holds because the function $q^A - \frac{1}{4}q^{2A}$ monotonically increases for $q \in [0,1]$. For a message length of Q bits (we assume Q is even here, otherwise an additional bit 0 can always be padded to make it even), the error probability is:

$$\begin{aligned}
P_e &= P(\exists i, Y_i \neq X_i) \\
&= 1 - P(Y_i = X_i)^{\frac{Q}{2}} \\
&\leq 1 - (1 - (1-\mu_{\min})^A + \frac{1}{4}(1-\mu_{\min})^{2A})^{\frac{Q}{2}} \\
&= 1 - (1 - \frac{1}{2}(1-\mu_{\min})^A)^Q \\
&\leq \frac{Q}{2}(1-\mu_{\min})^A \\
&\leq \frac{Q}{2}e^{-\mu_{\min}A}.
\end{aligned}$$

With the choice of $A = \lceil \frac{\log(QT/2)}{\mu_{\min}} \rceil$, we have $P_e < \frac{1}{T}$. Thus, the total codeword length for a message of length Q is:

$$N_{flip} = Q\lceil \frac{\log(QT/2)}{\mu_{\min}} \rceil.$$

With $N_{flip} = \Theta(\log(T))$, the regret remains order-optimal.

A.3 Modified Hamming code

As the number of codewords increases to 16 (4 bits), a modified (7,4) Hamming Code can be designed. It is a concatenated code, with the standard (7,4) Hamming code as the inner code and a repetition code as the outer code.

- Encoding. The standard (7,4) Hamming encoding matrix $\boldsymbol{G}$ is first used to encode a 4-bit message to a 7-bit codeword. Then we repeat each bit of the 7-bit codeword A times, leading to a $7A$-bit codeword;

- Decoding. First by using the repetition code's decoding rule, $7A$-bit coded message is decoded into 7 bits. This 7 bits is then decoded with the standard (7,4) Hamming decoding matrix $\boldsymbol{H}$. The final output is a decoded 4-bit message.

$$\boldsymbol{G} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{H} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

The repetition code reduces the crossover probability from q to q^A. With this relatively small crossover probability and the error correction capability of the Hamming Code, a reliable performance can be achieved. As stated by Barbero et al. (2006), with q^A as the crossover probability, we have the following error rate for the Hamming Code over a Z-channel.

$$P(Y_i \neq X_i) = \frac{7}{2}(q^A)^2 + o((q^A)^3).$$

We neglect $o((q^A)^3)$ in the following analysis. The error probability of transmitting Q-bit messages (assuming Q can be divided by 4) using the (7,4) modified Hamming code is:

$$\begin{aligned} P_e &= P(\exists i, Y_i \neq X_i) \\ &= 1 - P(Y_i = X_i)^{\frac{Q}{4}} \\ &= 1 - (1 - \frac{7}{2}q^{2A})^{\frac{Q}{4}} \\ &\leq 1 - (1 - \frac{7}{2}(1-\mu_{\min})^{2A})^{\frac{Q}{4}} \\ &\leq \frac{7Q}{8}(1-\mu_{\min})^{2A} \\ &\leq \frac{7Q}{8}e^{-2\mu_{\min}A}. \end{aligned} \tag{6}$$

By choosing $A = \frac{1}{2}\lceil \frac{\log(7QT/8)}{\mu_{\min}} \rceil$, we have $P_e < \frac{1}{T}$. Thus, the total codeword length for a message of length Q is:

$$N_{ham} = \frac{7Q}{8}\left\lceil \frac{\log(7QT/8)}{\mu_{\min}} \right\rceil,$$

which is still $\Theta(\log(T))$, but the bound in (6) indicates an improvement over the repetition code and flip code.

B Proofs for the Regret Analysis

B.1 Initialization phase

The initialization phase starts with a "Muscial Chair" phase, which assigns a unique external rank in $1, ..., K$ for each of the player. Then the following sequential hopping protocol converts the external rank into a unique internal rank in $1, ..., M$ for each player and estimates the number of players M. The proof for Lemma 1 is the same as Lemma 11 in Boursier and Perchet (2019). We re-state the algorithm and the proof for the sake of completeness.

Algorithm 4 Musical_Chair

```
Input: $[K]$, $T_0$
Output: Fixed (external rank)
Initialize Fixed $\leftarrow -1$
for $T_0$ time steps do
    if Fixed $= -1$ then
        Sample $k$ uniformly at random in $[K]$, play it in round $t$ and receive reward $r(t)$
        if $r(t) > 0$ then Fixed $\leftarrow k$
        end if
    end if
end for
return Fixed
```

Proof. As there is at least one arm that is not played by all the other players, the probability to encounter a positive reward for player j during the Musical Chair process at time t is lower bounded by $\frac{\mu_{\min}}{K}$. Thus with the choice of $T_0 = K\lceil \frac{\log(T)}{\mu_{\min}} \rceil$, the probability for a single player to encounter only zero rewards until time T_0 is:

$$P(\forall t \leq T_0, r^j(t) = 0) \leq (1 - \frac{\mu_{\min}}{K})^{T_0} \leq \exp(-\frac{T_0 \mu_{\min}}{K}) \leq \frac{1}{T}.$$

Applying a union bound over all players, the Musical Chair process is successful with a probability at least $1 - O(\frac{M}{T})$.

Now we analyze the *Estimate_M_NoSensing* protocol. Similar to using repetition code for communication, the probability that a player detects a "collision" while there is none is:

$$P_e \leq (1-\mu_{\min})^{T_c} \leq e^{-\mu_{\min}T_c} \leq \frac{1}{T}.$$

The union bound over the M players and the $2K$ blocks yields that it will be successful with a probability at least $1 - O(\frac{MK}{T})$. Furthermore, the initialization phase lasts $3KT_c$ time steps. Hence the regret satisfies:

$$R^{init} \leq 3MKT_c = 3MK\lceil \frac{\log(T)}{\mu_{\min}} \rceil.$$

$\square$

Algorithm 5 Estimate_M_NoSensing

```
Input: $k$ (external rank), $T_c$
Output: $M$, $j$ (internal rank)
Initialize $M \leftarrow 1$ and $\pi \leftarrow k$
for $n = 1, 2, ..., 2K$ do
    $r \leftarrow 0$
    if $n \leq 2k$ then
        for $T_c$ time steps do
            Pull $\pi$ and get reward $r_\pi(t)$
            Update $r \leftarrow r + r_\pi(t)$
        end for
        if r=0 then $M \leftarrow M + 1$, $j \leftarrow j + 1$
        end if
    else $\pi \leftarrow \pi + 1 (\text{mod } K)$
        for $T_c$ time steps do
            Pull $\pi$ and get reward $r_\pi(t)$
            Update $r \leftarrow r + r_\pi(t)$
        end for
        if r=0 then $M \leftarrow M + 1$
        end if
    end if
end for
return $M$, $j$
```

B.2 Exploration phase

This section aims at proving Lemma 2, which bounds the exploration regret. We start with the required lemmas and then go back to proving Lemma 2.

B.2.1 Proof for Lemma 3

This lemma ensures that event A_2 happens with a high probability. As mentioned before, there are at most $\log_2(T)$ communication phases, which leads to at most $(MK + 2M)\log_2(T)$ instances of transmissions to send arm statistics to the leader and send the acc/rej arm sets to the followers. Since there are at most K arms to be accepted or rejected, no more than MK instances of transmissions are required for sending the acc/rej arm sets.

Proof. Denote $P(\xi_p)$ as the probability that the decoding of a Q-bit message produces a wrong result at round p. With the choice of $N' = \max\{\frac{Q}{C_Z(1-\mu_{\min})}, \frac{\log(T)}{E(R)}\}$, and X_p, Y_p denoting the message before encoding and after decoding at round p, we have

$$P(\xi_p) = P(X_p \neq Y_p) \leq \frac{1}{T}.$$

A simple union bound leads to

$$P_r = 1 - P(\cup p, \xi_p) \geq 1 - \sum_p P(\xi_p) \geq 1 - \frac{(MK + 2M)\log(T) + MK}{T} = 1 - O\left(\frac{MK\log(T)}{T}\right).$$

□

B.2.2 Proof for Lemma 4

Lemma 4 ensures the acceptance and rejection of arms are successful with a high probability, which requires a good estimation of the statistics of arms. The estimation error consists of two parts: the quantization error and the sampling error. We analyze them separately.

Proof. With the choice of $Q \geq \log_2(\frac{1}{\frac{\Delta}{4} - \epsilon})$, the quantization error in phase p can be bounded as:

$$\forall i \in [M], |\bar{\mu}_i^p[k] - \hat{\mu}_i^p[k]| \leq \frac{\Delta}{4} - \epsilon,$$

$$|\bar{\mu}^p[k] - \hat{\mu}^p[k]| = \frac{\left|\sum_{i=1}^{M}(\bar{\mu}_i^p[k] - \hat{\mu}_i^p[k])T_p^i\right|}{T_p} \leq \frac{\Delta}{4} - \epsilon.$$

For any active arm $k \in [K_p]$, the gap between the sample mean $\hat{\mu}^p[k]$ (using all players' samples) and the true mean can be bounded with Hoeffding's inequality:

$$P\left(|\hat{\mu}^p[k] - \mu[k]| \geq \sqrt{\frac{2\log(T)}{T_p}}\right) \leq \frac{2}{T}.$$

Then, the overall gap between the quantized mean and the true mean for any active arm $k \in [K_p]$ can be bounded as:

$$\begin{aligned}
&P\left(|\bar{\mu}^p[k] - \mu[k]| > B_{T_p}\right)\\
=&P\left(|\bar{\mu}^p[k] - \hat{\mu}^p[k] + \hat{\mu}^p[k] - \mu[k]| \geq \sqrt{\frac{2\log(T)}{T_p}} + \frac{\Delta}{4} - \epsilon\right)\\
\leq&P\left(|\bar{\mu}^p[k] - \hat{\mu}^p[k]| + |\hat{\mu}^p[k] - \mu[k]| \geq \sqrt{\frac{2\log(T)}{T_p}} + \frac{\Delta}{4} - \epsilon\right)\\
\leq&P\left(|\hat{\mu}^p[k] - \mu[k]| \geq \sqrt{\frac{2\log(T)}{T_p}}) \cup P(|\bar{\mu}^p[k] - \hat{\mu}^p[k]| \geq \frac{\Delta}{4} - \epsilon\right)\\
=&P\left(|\hat{\mu}^p[k] - \mu[k]| \geq \sqrt{\frac{2\log(T)}{T_p}}\right)\\
\leq&\frac{2}{T}.
\end{aligned}$$

There are at most $\log_2(T)$ iterations of exploration and communication. By using a union bound of all these iterations and K arms, Eqn. (4) is obtained. □

B.2.3 Proof for Lemma 5

Lemma 5 bounds the number of time steps an arm is pulled before being accepted or rejected, and is essential to control the rounds of exploration and communication. The proof is similar to the proof to Proposition 1 in Boursier and Perchet (2019).

Proof. The proof is conditioned on the typical event. We first consider an optimal arm k. Let $\Delta_k = \mu[k] - \mu_{(M+1)}$ be the gap between the arm k and the first sub-optimal arm. Let s_k be the first integer such that $4B_{s_k} \leq \Delta_k$. It satisfies:

$$s_k \geq \frac{32\log(T)}{(\Delta_k - \Delta + 4\epsilon)^2} = \frac{32\log(T)}{\left(\mu[k] - \mu_{(M)} + 4\epsilon\right)^2}.$$

Recall that the number of time steps an active arm is pulled before the p-th exploration is $T_p = \sum_{l=1}^{p} M_l 2^l \lceil \log(T) \rceil$. With a non-increasing M_p, it holds that

$$T_{p+1} \leq 3T_p. \tag{7}$$

For some p such that $T_{p-1} \leq s_k < T_p$, the following inequalities are in order: $\Delta_k \geq 4B_{T_p}$; $|\bar{\mu}^p[k] - \mu[k]| \leq B_{T_p}$; and $|\bar{\mu}^p[i] - \mu[i]| \leq B_{T_p}$ for all sub-optimal arm i. We then have

$$\bar{\mu}^p[k] - B_{T_p} \geq \bar{\mu}^p[i] + B_{T_p} + \mu[k] - \mu[i] - 4B_{T_p} \geq \bar{\mu}^p[i] + B_{T_p}.$$

This suggests arm k will be accepted at time T_p. Eqn. (7) also leads to $T_p = O(s_k) = O\left(\frac{\log(T)}{(\mu[k]-\mu_{(M)}+4\epsilon)^2}\right)$. Thus, arm k will be accepted after at most $O\left(\frac{\log(T)}{(\mu[k]-\mu_{(M)}+4\epsilon)^2}\right)$ pulls. The part of rejecting sub-optimal arms can be similarly proved with $\Delta_k = \mu_{(M)} - \mu[k]$. □

B.2.4 Lemma 7 and its proof

Lemma 7. *In the typical event, the following results hold.*

1) for any sub-optimal arm k, $(\mu_{(M)} - \mu[k])T_k^{expl}(T) = O\left(\frac{\Delta}{4\epsilon}\min\left\{\frac{\log(T)}{\mu_{(M+1)} - \mu[k] + 4\epsilon}, \sqrt{T\log(T)}\right\}\right)$;

2) $\sum_{k \leq M}(\mu_{(k)} - \mu_{(M)})(T^{expl} - T_{(k)}^{expl}) = O\left(\sum_{k>M}\min\left\{\frac{\log(T)}{\mu_{(M+1)} - \mu_{(k)} + 4\epsilon}, \sqrt{T\log(T)}\right\}\right)$.

The proof of the first part in Lemma 7 is as follows.

Proof. For a sub-optimal arm k, Lemma 5 leads to $T_k^{expl}(T) \leq O\left(\min\left\{\frac{\log(T)}{(\mu_{(M+1)}-\mu[k]+4\epsilon)^2}, T\right\}\right)$, and thus

$$
\begin{aligned}
(\mu_{(M)} - \mu[k])T_k^{expl}(T) &= \frac{\mu_{(M)} - \mu[k]}{\mu_{(M+1)} - \mu[k] + 4\epsilon} O\left(\min\left\{\frac{\log(T)}{(\mu_{(M+1)} - \mu[k] + 4\epsilon)}, (\mu_{(M+1)} - \mu[k] + 4\epsilon)T\right\}\right) \\
&\overset{(i)}{\leq} O\left(\frac{\Delta}{4\epsilon} \min\left\{\frac{\log(T)}{\delta}, \delta T\right\}\right) \\
&\overset{(ii)}{\leq} O\left(\frac{\Delta}{4\epsilon} \min\left\{\frac{\log(T)}{(\mu_{(M+1)} - \mu[k] + 4\epsilon)}, \sqrt{T\log(T)}\right\}\right),
\end{aligned}
$$

in which inequality (i) comes from

$$
\frac{\mu_{(M)} - \mu[k]}{\mu_{(M+1)} - \mu[k] + 4\epsilon} = \frac{\mu_{(M)} - \mu[k]}{\mu_{(M)} - \mu[k] + 4\epsilon - \Delta} \leq \frac{\Delta}{4\epsilon},
$$

and $\delta = \mu_{(M+1)} - \mu[k] + 4\epsilon$. Inequality (ii) can be obtained with the observation that the term $\frac{\Delta}{4\epsilon}O(\min\{\frac{\log(T)}{\delta}, \delta T\})$ is maximized by $\delta = \sqrt{\frac{\log(T)}{T}}$. □

The second part of Lemma 7 is based on Lemmas 8 and 9.

Lemma 8. *Define $\hat{t}_k$ as the number of exploratory pulls before accepting/rejecting the arm k and H is the total number of exploration phases. Conditioned on the typical event, we have:*

$$
\sum_{k \leq M} \left(\mu_{(k)} - \mu_{(M)}\right)\left(T^{expl} - T_{(k)}^{expl}\right) \leq \sum_{j>M} \sum_{k \leq M} \sum_{p=1}^{H} 2^p \lceil \log(T) \rceil \left(\mu_{(k)} - \mu_{(M)}\right) \mathbb{1}_{\min\left\{\hat{t}_{(j)}, \hat{t}_{(k)}\right\} \geq T_{p-1}}.
$$

Proof. For an optimal arm k, during phase p, if k has already been accepted, it will be pulled $K_p 2^p\lceil\log(T)\rceil$ times. If it is still active (i.e., $\hat{t}_k > T_{p-1}$), it will be pulled $M_p 2^p\lceil\log(T)\rceil$ times, meaning that this arm is not pulled for $(K_p - M_p)2^p\lceil\log(T)\rceil$ times. Thus, it holds that $T_k^{expl} \geq T^{expl} - \sum_{p=1}^{H} 2^p (K_p - M_p)\lceil\log(T)\rceil \mathbb{1}_{\hat{t}_k > T_{p-1}}$. Notice that $K_p - M_p = \sum_{j>M} \mathbb{1}_{\hat{t}_{(j)} > T_{p-1}}$. We have $T_k^{expl} \geq T^{expl} - \sum_{p=1}^{H} \sum_{j>M} 2^p\lceil\log(T)\rceil \mathbb{1}_{\min\{\hat{t}_{(j)}, \hat{t}_{(k)}\} > T_{p-1}}$, which proves the lemma. □

Lemma 9. *Conditioned on the typical event, we have:*

$$
\sum_{k \leq M} \sum_{p=1}^{H} 2^p \lceil \log(T) \rceil \left(\mu_{(k)} - \mu_{(M)}\right) \mathbb{1}_{\min\left\{\hat{t}_{(j)}, \hat{t}_{(k)}\right\} \geq T_{p-1}} \leq O\left(\min\left\{\frac{\log(T)}{\mu_{(M)} - \mu_{(j)} + 4\epsilon}, \sqrt{T\log(T)}\right\}\right).
$$

Proof. Define $A_j = \sum_{k \leq M} \sum_{p=1}^{H} 2^p\lceil\log(T)\rceil(\mu_{(k)} - \mu_{(M)})\mathbb{1}_{\min\{\hat{t}_{(j)}, \hat{t}_{(k)}\} \geq T_{p-1}}$. Notice that

$$
\hat{t}_{(k)} \leq \min\left\{\frac{c\log(T)}{(\mu_{(k)} - \mu_{(M)} + 4\epsilon)^2}, T\right\},
$$

and denote $\Delta(p) = \sqrt{\frac{c\log(T)}{T_{p-1}}}$. The inequity $\hat{t}_{(k)} > T_{p-1}$ implies $\mu_{(k)} - \mu_{(M)} < \Delta(p) - 4\epsilon$. We also denote N^j as the smallest integer satisfying $\hat{t}_{(j)} \leq T_{N^j}$. Then we have the following:

$$
\begin{aligned}
A_j &\leq \sum_{k \leq M} \sum_{p=1}^{N^j} 2^p\lceil\log(T)\rceil(\Delta(p) - 4\epsilon)\mathbb{1}_{\hat{t}_{(k)} \geq T_{p-1}} \\
&\leq \sum_{p=1}^{N^j} \Delta(p) 2^p \lceil\log(T)\rceil \sum_{k \leq M} \mathbb{1}_{\hat{t}_{(k)} \geq T_{p-1}} \\
&= \sum_{p=1}^{N^j} \Delta(p) 2^p \lceil\log(T)\rceil M_p \\
&\leq \sum_{p=1}^{N^j} \Delta(p)(T_p - T_{p-1}) \\
&= c\log(T) \sum_{p=1}^{N^j} \Delta(p) \left(\frac{1}{\Delta(p+1)} + \frac{1}{\Delta(p)}\right)\left(\frac{1}{\Delta(p+1)} - \frac{1}{\Delta(p)}\right).
\end{aligned}
$$

Since $T_{p+1} \leq 3T_p$, $\Delta(p)\left(\frac{1}{\Delta(p+1)} + \frac{1}{\Delta(p)}\right) = 1 + \sqrt{\frac{T_p}{T_{p-1}}} \leq 1 + \sqrt{3}$. Thus,

$$A_j \leq c\log(T)\sum_{p=1}^{N^j}\left(\frac{1}{\Delta(p+1)} - \frac{1}{\Delta(p)}\right) \leq (1+\sqrt{3})c\log(T)\frac{1}{\Delta(N^j+1)}.$$

With the definition of N^j, we have $\hat{t}_{(j)} \geq T_{N^j-1}$. With inequality $T_{N^j+1} \leq 3T_{N^j}$ we have $\Delta(N^j) \geq \sqrt{\frac{c\log(T)}{3\hat{t}_{(j)}}}$. $A_j \leq (3+\sqrt{3})\sqrt{c\hat{t}_{(j)}\log(T)}$ then holds. With $\hat{t}_{(j)} \leq O\left(\min\{\frac{c\log(T)}{(\mu_{(M+1)}-\mu_{(j)}+4\epsilon)^2}, T\}\right)$, we have

$$A_j \leq (3+\sqrt{3})\min\left\{\frac{c\log(T)}{\mu_{(M+1)} - \mu_{(j)} + 4\epsilon}, \sqrt{cT\log(T)}\right\}.$$

□

B.3 Communication phase

This section presents the proof related to the bound of the communication regret.

B.3.1 Proof for Lemma 6

Proof. Conditioned on the typical event, we denote H as the number of exploration phases. The communication length for sending arm statistics and acc/rej arm sets for $p \in [H]$ is at most $N'(KM+2M)$. Lemma 5 states that H satisfies $T_H = \sum_{l=1}^{H} M_l 2^l \lceil \log(T) \rceil = O(\max_{k\in[K]}\{s_k\}) = O\left(\min\{\frac{\log(T)}{4\epsilon}, T\}\right)$. Thus,

$$H = O\left(\log(\min\{\frac{1}{4\epsilon}, T\})\right),$$

which leads to a regret of $O(N'(KM^2+2M^2)\log(\min\{\frac{1}{4\epsilon}, T\})$. Next, transmitting acc/rej arm sets at most incurs a regret of M^2KN'. Putting them together, the total communication regret is:

$$O\left(N'(KM^2+2M^2)\log\left(\min\left\{\frac{1}{4\epsilon}, T\right\}\right) + N'M^2K\right).$$

□

C Supplementary Materials for Experiments

C.1 Minor changes to SIC-MMAB2

Some minor changes can be made to SIC-MMAB2 (Boursier and Perchet, 2019) to improve its convergence and make it more practical. First, the original length of exploration in phase p is $K_p 2^p T_0$, where $T_0 = \lceil \frac{2400\log(T)}{\mu_{\min}} \rceil$. The factor 2400 is too large and makes it almost impossible to converge (or even finish one round of exploration) in the time horizon of our experiments. We thus change it to $T_0 = \lceil \frac{24\log(T)}{\mu_{\min}} \rceil$, and we have verified that it converges successfully for the instances in the experiments. Another minor change is to add a random selection in the declaration phases. In SIC-MMAB2 (Boursier and Perchet, 2019), players sequentially declare the arms in their own rejected sets. However, with similar exploration time steps across players, the declaration sets are almost identical for all players. With a sequential selection (rather than a random selection in our implementation), all players will declare the same arm with high probability, which is in fact noticed to happen very frequently in the experiment.

C.2 Supplementary experiment results

The impact of the leader-follower protocol and the enhancement in Section 5 are evaluated in a larger game ($M = 10$, $K = 29$). With $\Delta = 0.06$, EC-SIC without the leader-follower protocol (labeled as "EC-SIC-Mesh"), EC-SIC using none of the enhancement in Section 5 (labeled as "EC-SIC-NE") are compared with EC-SIC and SIC-MMAB2. Compared to the stable performance of EC-SIC and SIC-MMAB2, Fig. 8 shows that in practice, due to the unnecessary communication between every pair of players, EC-SIC-Mesh cannot even finish the first communication phase. Although it converges eventually, EC-SIC-NE has a larger regret. These results highlight the benefit of the tree-structured communication and the selection of better communication arms.

In the case of utilizing repetition code in EC-SIC, we carry on experiments to evaluate the dependency on the knowledge of $\mu_{\min}$. For $\mu_{(k)} = 0.3$, we evaluate EC-SIC and SIC-MMAB2 with $\mu_{\min}$ from 0.05 to 0.3, hence creating a mismatched

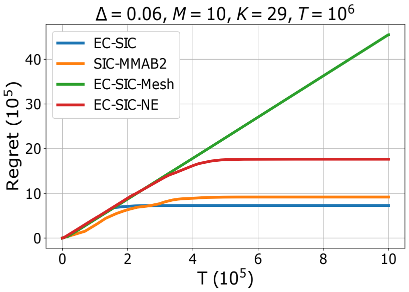

Figure 8: Large game

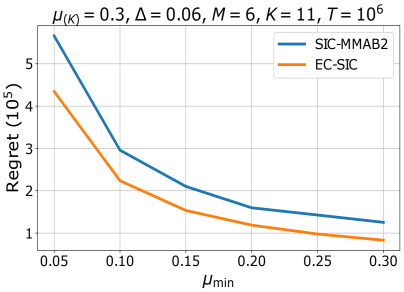

Figure 9: Performance with different $\mu_{\min}$

"estimation" of $\mu_{\min}$. The results shown in Figure 9 state that decreasing $\mu_{\min}$ leads to an increasing regret of both SIC-MMAB2 and EC-SIC, which corroborates the theoretical analysis. Furthermore, EC-SIC has better performance than SIC-MMAB2 across different "estimates" of $\mu_{\min}$.

The knowledge of Δ is assumed in the algorithms and their theoretical analysis. In practice, a precise value of Δ may not always be available. In the last experiment, we demonstrate the robustness of the algorithms by feeding it with inaccurate information Δ_e instead of the true Δ. As shown in Figure 10, with a pessimistic estimation of Δ, the inaccurate information only leads to some additional but acceptable communication loss. The overall regret is still better than SIC-MMAB2. For the optimistic estimations, Figure 11 shows that the algorithm is effective even with $\Delta_e = 2\Delta$. When the estimation error further grows ($\Delta_e = 3\Delta$ or 6Δ), communication errors start to occur, which lead to a few arms that are identified incorrectly. It nevertheless still outperforms SIC-MMAB2 in terms of regret. Thus, practically speaking, EC-SIC has good robustness, and we further comment that it is preferable to have a pessimistic estimation.

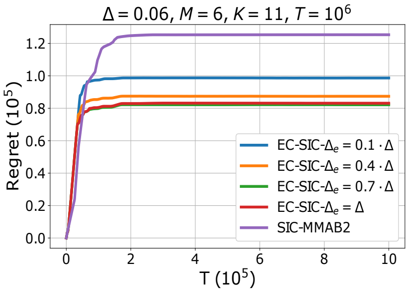

Figure 10: Pessimistic estimation

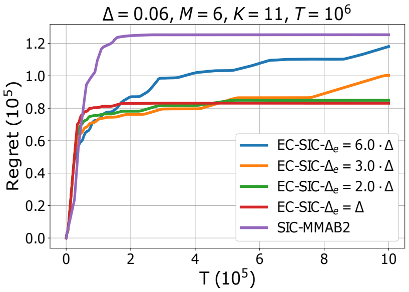

Figure 11: Optimistic estimation