Achieving Near-Optimal Individual Regret & Low Communications in Multi-Agent Bandits

Xuchuang Wang[1], Lin Yang[2], Yu-Zhen Janice Chen[3], Xutong Liu[1], Mohammad Hajiesmaili[3], Don Towsley[3], John C.S. Lui[1]

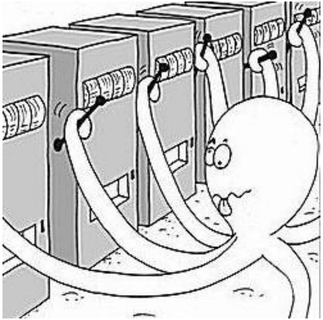
To Appear in ICLR 2023

The Chinese University of Hong Kong[1], Nanjing University[2], University of Massachusetts Amherst[3]

March 5, 2023

Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.



Multi-Armed Bandits

- *K arms*: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1								
Arm 2								
Arm 3								
Arm 4								
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- *K arms*: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$							
Arm 2								
Arm 3								
Arm 4								
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$							
Arm 2		$\surd$						
Arm 3								
Arm 4								
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- *K* arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	…	T
Arm 1	$\surd$							
Arm 2		$\surd$						
Arm 3								
Arm 4			$\surd$					
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- *K* arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$			$\surd$				
Arm 2		$\surd$						
Arm 3								
Arm 4			$\surd$					
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
    - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$			$\surd$				
Arm 2		$\surd$						
Arm 3					$\surd$			
Arm 4			$\surd$					
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$			$\surd$				
Arm 2		$\surd$				$\surd$		
Arm 3					$\surd$			
Arm 4			$\surd$					
Arm 5								
$\vdots$								
Arm K								

Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$			$\surd$				
Arm 2		$\surd$				$\surd$		
Arm 3					$\surd$			
Arm 4			$\surd$					
Arm 5							$\surd$	
$\vdots$								
Arm K								

Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- Given T decision rounds:

Time	1	2	3	4	5	6	...	T
Arm 1	$\surd$			$\surd$				
Arm 2		$\surd$				$\surd$		
Arm 3					$\surd$			
Arm 4			$\surd$					
Arm 5							$\surd$	
$\vdots$								
Arm K								$\surd$

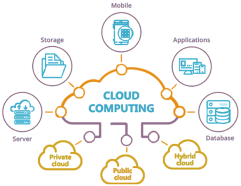
Cooperative Multiple-Agent Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.

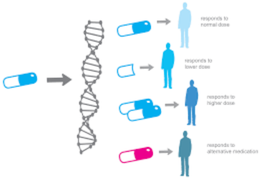
Cooperative Multiple-Agent Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- M Agents in $t = 1, \ldots, T$:
  - Each agent i pulls an arm and collects reward $X_k^{(i)}$ from pulled arms.

(a) Online advertising with multiple servers



(b) Cloud computing with multiple clients



(c) Clincal treatment with multiple hospitals

Cooperative Multiple-Agent Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- M Agents in $t = 1, \ldots, T$:
  - Each agent i pulls an arm and collects reward $X_k^{(i)}$ from pulled arms.
- Group regret:

$$\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{gro}}(\mathcal{A})] := MT\mu(1) - \mathbb{E}\left[\sum_{i\in[M]}\sum_{t\in[T]} X_t^{(i)}(A_t^{(i)})\right]$$

Cooperative Multiple-Agent Multi-Armed Bandits

- *K arms*: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- *M Agents* in $t = 1, \ldots, T$:
  - Each agent i pulls an arm and collects reward $X_k^{(i)}$ from pulled arms.
- Group regret:

$$\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{gro}}(\mathcal{A})] := MT\mu(1) - \mathbb{E}\left[\sum_{i\in[M]}\sum_{t\in[T]} X_t^{(i)}(A_t^{(i)})\right]$$

  - Full communication: $\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{gro}}(\mathcal{A})] = \Theta(K \log T)$
  - No communication: $\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{gro}}(\mathcal{A})] = O(MK \log T)$

Cooperative Multiple-Agent Multi-Armed Bandits

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$.
  - Assume $\mu(1) > \cdots > \mu(K)$.
- M Agents in $t = 1, \ldots, T$:
  - Each agent i pulls an arm and collects reward $X_k^{(i)}$ from pulled arms.
- Group regret:

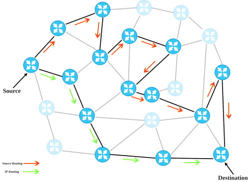
$$\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{gro}}(\mathcal{A})] := MT\mu(1) - \mathbb{E}\left[\sum_{i\in[M]}\sum_{t\in[T]} X_t^{(i)}(A_t^{(i)})\right]$$

  - Full communication: $\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{gro}}(\mathcal{A})] = \Theta(K\log T)$
  - No communication: $\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{gro}}(\mathcal{A})] = O(MK\log T)$
- Communication costs:

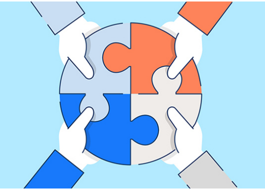
$$\mathbb{E}\left[\mathrm{C}_\mathrm{T}(\mathcal{A})\right] := \mathbb{E}\left[\sum_{i\in[M]}\sum_{t\in[T]} \mathbb{1}\{\text{agent } i \text{ communicates in time } t\}\right].$$

New Objective: Maximum Individual Regret

(a) Drone swarm



(b) Path routing

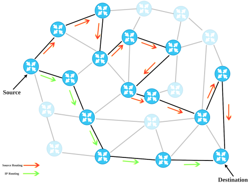


(c) Max-min fairness

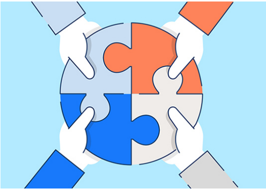
- Overall performance is sensitive to the “bad” agent.
- Max-min fairness is equivalent to minimizing the “bottleneck” agent’s regret.

New Objective: Maximum Individual Regret

(a) Drone swarm



(b) Path routing



(c) Max-min fairness

- Overall performance is sensitive to the “bad” agent.
- Max-min fairness is equivalent to minimizing the “bottleneck” agent’s regret.

$$\mathbb{E}[\mathrm{R}^{\mathrm{ind}}_{\mathrm{T}}(\mathcal{A})] := T\mu(1) - \mathbb{E}\left[\min_{i\in[M]} \sum_{t\in[T]} X_t^{(i)}(A_t^{(i)})\right].$$

Related Works and Contributions

Table 1: A comparison summary of prior literature and this work

	Individual regret	Group regret	Communication cost
`DPE2` [Wang et al., 2020]	$O(K \log T)$	$O(K \log T)$	$O(K^2 M^2)$
`ComEx` [Madhushani and Leonard, 2021]	$O(K \log T)$	$O(K \log T)$	$O(KM \log T)$
`GosInE` [Chawla et al., 2020]	$O((K/M + 2) \log T)$	$O((K + 2M) \log T)$	$\Omega(\log T)$
`Dec_UCB` [Zhu et al., 2021]	$O((K/M) \log T)$	$O(K \log T)$	$O(MT)$
`UCB-TCOM` (our algorithm)	$O((K/M) \log T)$	$O(K \log T)$	$O(KM \log(\log T))$

Related Works and Contributions

Table 1: A comparison summary of prior literature and this work

	Individual regret	Group regret	Communication cost
DPE2 [Wang et al., 2020]	$O(K \log T)$	$O(K \log T)$	$O(K^2 M^2)$
ComEx [Madhushani and Leonard, 2021]	$O(K \log T)$	$O(K \log T)$	$O(KM \log T)$
GosInE [Chawla et al., 2020]	$O((K/M + 2) \log T)$	$O((K + 2M) \log T)$	$\Omega(\log T)$
Dec_UCB [Zhu et al., 2021]	$O((K/M) \log T)$	$O(K \log T)$	$O(MT)$
UCB-TCOM (our algorithm)	$O((K/M) \log T)$	$O(K \log T)$	$O(KM \log(\log T))$

1. The first near-optimal algorithm UCB-TCOM on individual regret with efficient communications.

Related Works and Contributions

Table 1: A comparison summary of prior literature and this work

	Individual regret	Group regret	Communication cost
DPE2 [Wang et al., 2020]	$O(K \log T)$	$O(K \log T)$	$O(K^2 M^2)$
ComEx [Madhushani and Leonard, 2021]	$O(K \log T)$	$O(K \log T)$	$O(KM \log T)$
GosInE [Chawla et al., 2020]	$O((K/M + 2) \log T)$	$O((K + 2M) \log T)$	$\Omega(\log T)$
Dec_UCB [Zhu et al., 2021]	$O((K/M) \log T)$	$O(K \log T)$	$O(MT)$
UCB-TCOM (our algorithm)	$O((K/M) \log T)$	$O(K \log T)$	$O(KM \log(\log T))$

1. The first near-optimal algorithm UCB-TCOM on individual regret with efficient communications.
2. A communication policy TCOM that

Related Works and Contributions

Table 1: A comparison summary of prior literature and this work

	Individual regret	Group regret	Communication cost
`DPE2` [Wang et al., 2020]	$O(K \log T)$	$O(K \log T)$	$O(K^2 M^2)$
`ComEx` [Madhushani and Leonard, 2021]	$O(K \log T)$	$O(K \log T)$	$O(KM \log T)$
`GosInE` [Chawla et al., 2020]	$O((K/M + 2) \log T)$	$O((K + 2M) \log T)$	$\Omega(\log T)$
`Dec_UCB` [Zhu et al., 2021]	$O((K/M) \log T)$	$O(K \log T)$	$O(MT)$
`UCB-TCOM` (our algorithm)	$O((K/M) \log T)$	$O(K \log T)$	$O(KM \log(\log T))$

1. The first near-optimal algorithm `UCB-TCOM` on individual regret with efficient communications.
2. A communication policy `TCOM` that
   (a) **Meta policy:** can be executed on top of any bandit algorithm;

Related Works and Contributions

Table 1: A comparison summary of prior literature and this work

	Individual regret	Group regret	Communication cost
DPE2 [Wang et al., 2020]	$O(K \log T)$	$O(K \log T)$	$O(K^2 M^2)$
ComEx [Madhushani and Leonard, 2021]	$O(K \log T)$	$O(K \log T)$	$O(KM \log T)$
GosInE [Chawla et al., 2020]	$O((K/M + 2) \log T)$	$O((K + 2M) \log T)$	$\Omega(\log T)$
Dec_UCB [Zhu et al., 2021]	$O((K/M) \log T)$	$O(K \log T)$	$O(MT)$
UCB-TCOM (our algorithm)	$O((K/M) \log T)$	$O(K \log T)$	$O(KM \log(\log T))$

1. The first near-optimal algorithm UCB-TCOM on individual regret with efficient communications.
2. A communication policy TCOM that
   (a) **Meta policy:** can be executed on top of any bandit algorithm;
   (b) **Tunable:** can be tuned to trade off communications (0 to $O(T)$) with regrets.

Tunable COMmunication `TCOM` (1/3): $O(\log T)$ Focus on Suboptimal Arms' Observation Sharing

IDEA: Share suboptimal arms' obs., Yes! ~~Share optimal arm, No~~.

- share suboptimal arms' observation $\Rightarrow$ reduce this arm's #pulls $\Rightarrow$ save cost
- share optimal arms' observation $\Rightarrow$ reduce this arm's #pulls $\Rightarrow$ increase cost

Tunable COMmunication TCOM (1/3): $O(\log T)$
Focus on Suboptimal Arms' Observation Sharing

IDEA: Share suboptimal arms' obs., Yes! ~~Share optimal arm, No~~.

- share suboptimal arms' observation $\Rightarrow$ reduce this arm's $\#$pulls $\Rightarrow$ save cost
- share optimal arms' observation $\Rightarrow$ reduce this arm's $\#$pulls $\Rightarrow$ increase cost

DESIGN: Construct a communication arm set $\mathcal{C}_t(\alpha)$

- include the arms that are likely to be suboptimal.
- only share new observations for arms in the set $\mathcal{C}_t(\alpha)$.

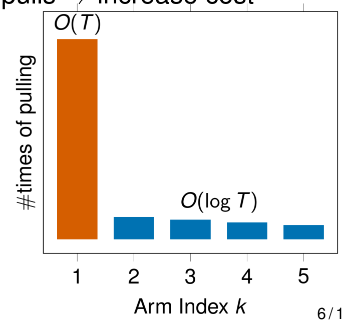
Tunable COMmunication TCOM (1/3): $O(\log T)$
Focus on Suboptimal Arms' Observation Sharing

IDEA: Share suboptimal arms' obs., Yes! ~~Share optimal arm, No~~.

- share suboptimal arms' observation $\Rightarrow$ reduce this arm's #pulls $\Rightarrow$ save cost
- share optimal arms' observation $\Rightarrow$ reduce this arm's #pulls $\Rightarrow$ increase cost

DESIGN: Construct a communication arm set $\mathcal{C}_t(\alpha)$

- include the arms that are likely to be suboptimal.
- only share new observations for arms in the set $\mathcal{C}_t(\alpha)$.



Tunable COMunication `TCOM` (2/3): $O(\log_\beta \log T)$ Dynamically Buffer Observations for Communication

IDEA: Regret only deteriorates up to a constant multiplier when the observation delays increase geometrically [Gao et al., 2019].

Tunable COMunication TCOM (2/3): $O(\log_\beta \log T)$ Dynamically Buffer Observations for Communication

IDEA: Regret only deteriorates up to a constant multiplier when the observation delays increase geometrically [Gao et al., 2019].

DESIGN: Buffer observations and communicate **whenever the buffered $\#$obs increases by a ratio $\beta\,(> 1)$.**

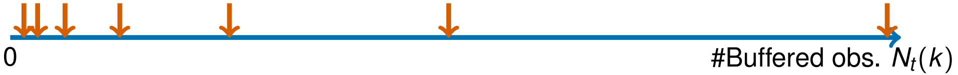
- e.g., if the ratio β is 2, broadcast when $N_t(k) = 2, 4, 8, 16, \ldots$

Tunable COMunication TCOM (2/3): $O(\log_\beta \log T)$ Dynamically Buffer Observations for Communication

IDEA: Regret only deteriorates up to a constant multiplier when the observation delays increase geometrically [Gao et al., 2019].

DESIGN: Buffer observations and communicate **whenever the buffered $\#$obs increases by a ratio $\beta\,(> 1)$.**

- e.g., if the ratio β is 2, broadcast when $N_t(k) = 2, 4, 8, 16, \ldots$



Tunable COMunication TCOM (3/3): $\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{ind}}(\mathcal{A})] = \frac{\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{gro}}(\mathcal{A})]}{M}$

Symmetric Actions for All Agents

IDEA: Minimize maximum individual regret
$\iff$ Evenly divide group regret
$\Longleftarrow$ In each time slot, all agents pull the same arm

Tunable COMunication TCOM (3/3): $\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{ind}}(\mathcal{A})] = \frac{\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{gro}}(\mathcal{A})]}{M}$ Symmetric Actions for All Agents

IDEA: Minimize maximum individual regret
$\Longleftrightarrow$ Evenly divide group regret
$\Longleftarrow$ In each time slot, all agents pull the same arm

DESIGN: Agents run the same arm-pulling policy and use the same set of global observations (communicated to all agents).

Algorithm 1 The `UCB-TCOM` Algorithm (for each agent)

1: **Input:** communication arm set parameter α and buffering ratio β

2: **Initialization:** $\hat{n}_t(k) = 0, N_t(k) = 0, \hat{\mu}_t(k) = 0, \tau_t(k) = 0$

3: **for** each decision round t **do** ▷ *Parallelly run for-loops in Lines 3 and 12.*

12: **for** each newly received message $(\tilde{\mu}_t(k), N_t(k), k)$ from the past round **do**

13: Update the empirical mean $\hat{\mu}_t(k)$, $\hat{n}_t(k)$, and the communication set $\mathcal{C}_t(\alpha)$

Algorithm 1 The `UCB-TCOM` Algorithm (for each agent)

1: **Input:** communication arm set parameter α and buffering ratio β
2: **Initialization:** $\hat{n}_t(k) = 0, N_t(k) = 0, \hat{\mu}_t(k) = 0, \tau_t(k) = 0$
3: **for** each decision round t **do** ▷ *Parallelly run for-loops in Lines 3 and 12.*
4: Pull arm A_t with the highest global UCB
5: Observe arm A_t's reward $X_t(A_t)$

12: **for** each newly received message $(\tilde{\mu}_t(k), N_t(k), k)$ from the past round **do**
13: Update the empirical mean $\hat{\mu}_t(k)$, $\hat{n}_t(k)$, and the communication set $\mathcal{C}_t(\alpha)$

Algorithm 1 The `UCB-TCOM` Algorithm (for each agent)

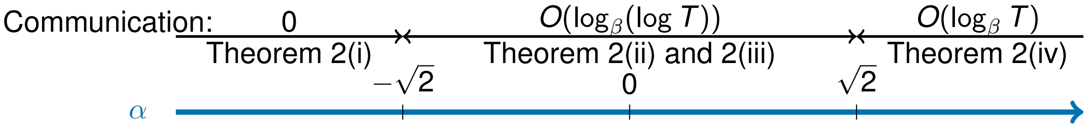
1: **Input:** communication arm set parameter α and buffering ratio β
2: **Initialization:** $\hat{n}_t(k) = 0, N_t(k) = 0, \hat{\mu}_t(k) = 0, \tau_t(k) = 0$
3: **for** each decision round t **do** ▷ *Parallelly run for-loops in Lines 3 and 12.*
4: Pull arm A_t with the highest global UCB
5: Observe arm A_t's reward $X_t(A_t)$
6: **if** $A_t \in \mathcal{C}_t(\alpha)$ **then** ▷ *Pick suboptim arms for observation sharing*
7: Increase $N_t(A_t)$ by 1
8: Update this phase's empirical mean $\tilde{\mu}_t(A_t)$

12: **for** each newly received message $(\tilde{\mu}_t(k), N_t(k), k)$ from the past round **do**
13: Update the empirical mean $\hat{\mu}_t(k)$, $\hat{n}_t(k)$, and the communication set $\mathcal{C}_t(\alpha)$

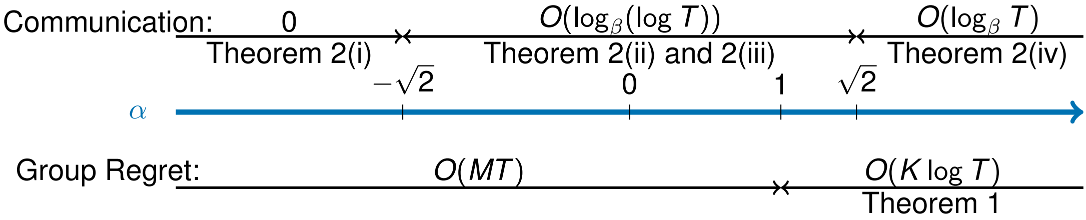
Algorithm 1 The `UCB-TCOM` Algorithm (for each agent)

1: **Input:** communication arm set parameter α and buffering ratio β
2: **Initialization:** $\hat{n}_t(k) = 0, N_t(k) = 0, \hat{\mu}_t(k) = 0, \tau_t(k) = 0$
3: **for** each decision round t **do** ▷ *Parallelly run for-loops in Lines 3 and 12.*
4:     Pull arm A_t with the highest global UCB
5:     Observe arm A_t's reward $X_t(A_t)$
6:     **if** $A_t \in \mathcal{C}_t(\alpha)$ **then** ▷ *Pick suboptim arms for observation sharing*
7:         Increase $N_t(A_t)$ by 1
8:         Update this phase's empirical mean $\tilde{\mu}_t(A_t)$
9:     **if** $N_t(A_t) \geqslant \lceil \beta N_{\tau_t(A_t)}(A_t) \rceil$ **then** ▷ *Buffer size increases geometrically.*
10:         Broadcast the message $(\tilde{\mu}_t(A_t), N_t(A_t), A_t)$
11:         $\tau_t(A_t) \leftarrow t$
12: **for** each newly received message $(\tilde{\mu}_t(k), N_t(k), k)$ from the past round **do**
13:     Update the empirical mean $\hat{\mu}_t(k)$, $\hat{n}_t(k)$, and the communication set $\mathcal{C}_t(\alpha)$

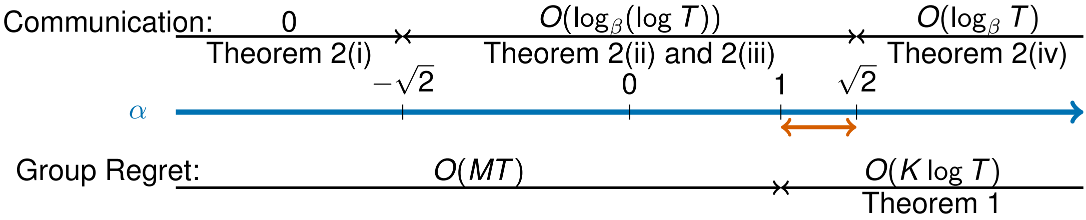
Theoretical Results of `UCB-TCOM`



Theoretical Results of `UCB-TCOM`

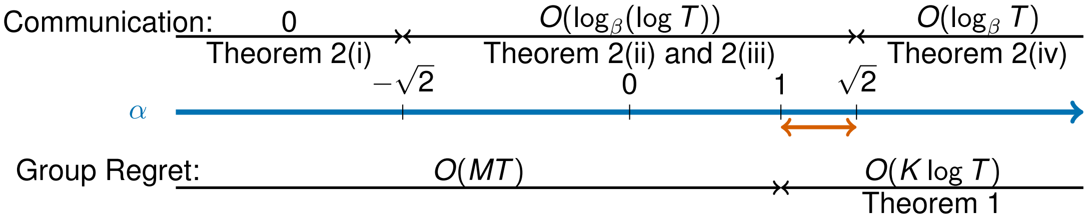


Theoretical Results of UCB−TCOM



- When $\alpha \in \left(1, \sqrt{2}\right)$, `UCB-TCOM` achieves the near-optimal group regret upper bounds with $O(\log(\log T))$ communications.

Theoretical Results of UCB-TCOM



- When $\alpha \in \left(1, \sqrt{2}\right)$, UCB-TCOM achieves the near-optimal group regret upper bounds with $O(\log(\log T))$ communications.
- Symmetric: $\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{ind}}(\mathcal{A})] = \dfrac{\mathbb{E}[\mathrm{R}_\mathrm{T}^{\mathrm{gro}}(\mathcal{A})]}{M}$—near-optimal individual regret.

Simulations (1/3): UCB–TCOM vs. Baselines

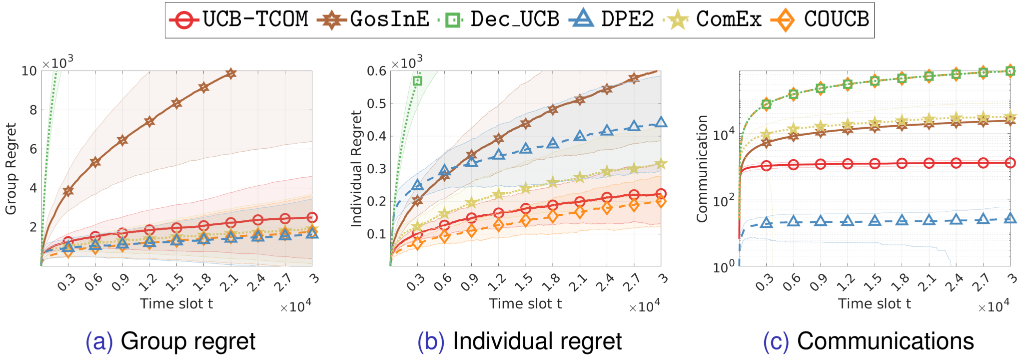
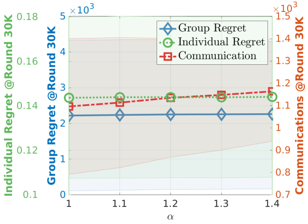
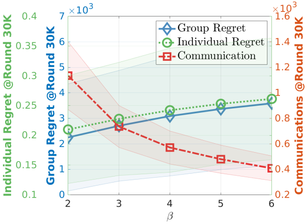


Figure 5: UCB-TCOM vs. Dec_UCB, GosInE, DPE2, ComEx and COUCB

Simulations (2/3): Tunable Parameters α and β



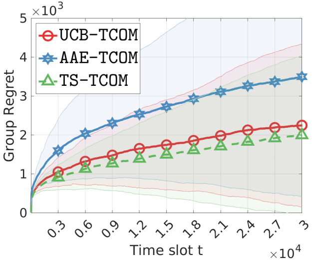
(a) Varying α (given $\beta = 2$)



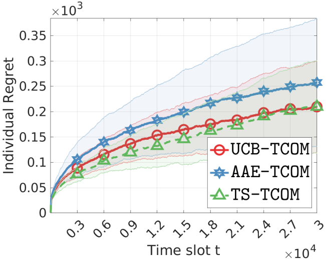
(b) Varying β (given $\alpha = 1.2$)

Figure 6: Impact of communication set parameter α with fixed $\beta = 2$ in Figures 6a; and buffering ratio β with fixed $\alpha = 1.2$ in Figures 6b

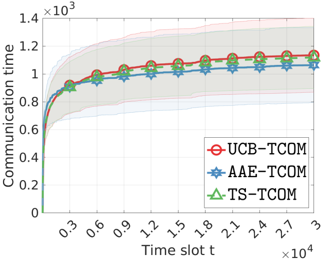
Simulations (3/3): Meta-Policy TCOM to AAE and TS



(a) Group regret



(b) Individual regret



(c) Communications

Figure 7: UCB-TCOM vs. AAE−TCOM, TS−TCOM

Conclusion

1. An algorithm achieves the **near-optimal individual and group regrets with $O(\log\log T)$ communications.**
2. A **meta** and **tunable** communication policy `TCOM`.
   1. share suboptimal action's observations;
   2. geometrical growth buffer;
   3. symmetric design.

Conclusion

1. An algorithm achieves the **near-optimal individual and group regrets with $O(\log \log T)$ communications.**
2. A **meta** and **tunable** communication policy `TCOM`.
   1. share suboptimal action's observations;
   2. geometrical growth buffer;
   3. symmetric design.

Future works:

- Pareto frontier of group/individual regrets vs. communication costs trade-off.
- Remove the time-dependence of the communication costs.

Thank you!

Full paper at openreview.net/forum?id=QTXKTXJKIh

Detail of Communication Arm Set Construction

Given tuning parameter α, communication arm set $\mathcal{C}_t(\alpha)$ of agent i at time t contains all arms identified as suboptimal, i.e.,

$$\mathcal{C}_t(\alpha) := \{k \in [K] : \exists k' \in [K] \setminus \{k\} \text{ such that } \texttt{tUCB}_t(k', \alpha) > \texttt{tLCB}_t(k, \alpha)\}, \quad (1)$$

where $\texttt{tUCB}_t(k, \alpha) := \hat{\mu}_t(k) + \alpha\sqrt{\frac{\log t}{\hat{n}_t(k)}}$, and $\texttt{tLCB}_t(k, \alpha) := \hat{\mu}_t(k) - \alpha\sqrt{\frac{\log t}{\hat{n}_t(k)}}$,

and $\hat{n}_t(k)$ denotes the number of times of the global reward observations of arm k up to time slot t.

Theoretial Results Detail (1/2)

Theorem (Regret upper bounds of UCB-TCOM for $\alpha > 1$)

When the communication arm set parameter $\alpha > 1$[1] and buffering-ratio $\beta > 1$, UCB-TCOM attains a near-optimal group regret upper bound in terms of number of decision rounds T, arms K, and agents M, or formally,

$$\mathbb{E}[\mathrm{R}_{\mathrm{T}}(\mathcal{A})] \leqslant \sum_{k>1} \frac{8\beta \log T}{\Delta(k)} + MK\frac{2\alpha^2 - 1}{\alpha^2 - 1}, \tag{2}$$

and UCB-TCOM also attains a near-optimal individual regret upper bound, or formally,

$$\mathbb{E}[\mathrm{R}_{\mathrm{T}}^{\mathrm{ind}}(\mathcal{A})] \leqslant \sum_{k>1} \frac{8\beta \log T}{M\Delta(k)} + K\frac{2\alpha^2 - 1}{\alpha^2 - 1}. \tag{3}$$

[1] The condition $\alpha > 1$ can be relaxed to $\alpha > 1/\sqrt{2}$ via the peeling technique.

Theoretical Results Detail (2/2)

Theorem (communication costs of `UCB-TCOM` for all α)

The communication costs of `UCB-TCOM` *has the following properties:*

(i) *When $\alpha \leqslant -\sqrt{2}$, no communication occurs among agents.*

(ii) *When $-\sqrt{2} < \alpha < \sqrt{2}$ and $\beta > 1$, the number of broadcasts of observations of the optimal arm by one agent is $O(\log(\log T))$. More rigorously, it is less than*

$$\log_\beta\left(\left(\frac{\sqrt{2}+\alpha}{\sqrt{2}-\alpha}\right)^2\left(\frac{8\log T}{\Delta_2^2} + MK\frac{2\alpha^2-1}{\alpha^2-1}\right)\right). \tag{4}$$

(iii) *When $\alpha > 1$, almost all observations of suboptimal arms—except for a finite number independent of T—are broadcast.*

(iv) *When $\alpha \geqslant \frac{2\sqrt{2}\mu(1)}{\Delta_2}$, almost all observations of the optimal arm—except for a finite number that is independent of T—are broadcast.*

References I

Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3471–3481. PMLR, 2020.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Udari Madhushani and Naomi Leonard. When to call your neighbor? strategic communication in cooperative stochastic bandits. *arXiv preprint arXiv:2110.04396*, 2021.

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

References II

Jingxuan Zhu, Ethan Mulle, Christopher Salomon Smith, and Ji Liu. Decentralized multi-armed bandit can outperform classic upper confidence bound. *arXiv preprint arXiv:2111.10933*, 2021.