Individual Regret in Cooperative Stochastic Multi-Armed Bandits over Communication Graph

Idan Barnea
Tel Aviv University
idanbarnea1@mail.tau.ac.il

Tal Lancewicki
Tel Aviv University
lancewicki@mail.tau.ac.il

Yishay Mansour
Tel Aviv University and Google Research
mansour.yishay@gmail.com

Abstract

We study the regret in stochastic Multi-Armed Bandits (MAB) with multiple agents that communicate over an arbitrary connected communication graph. We show a near-optimal individual regret bound of $\tilde{O}(\sqrt{AT/m} + A)$, where A is the number of actions, T the time horizon, and m the number of agents. In particular, assuming a sufficient number of agents, we achieve a regret bound of $\tilde{O}(A)$, which is independent of the sub-optimality gaps and depends only logarithmically on the time horizon. To the best of our knowledge, our study is the first to show an individual regret bound in cooperative stochastic MAB that is independent of the graph's diameter and applicable to non-fully-connected communication graphs.

1 Introduction

Multi-Armed Bandit (MAB) is a fundamental framework for studying sequential decision making, with an expanding scope of practical applications (see, [Lattimore and Szepesvári, 2020]). Recent research expanded the classic MAB problem into a cooperative setting, sometimes referred to as multiplayer or multi-agent MAB, where multiple agents share the same goal and can communicate with each other.

A significant focus of this research has centered on cooperating agents within a communication graph, often referred to as a communication network. This framework, in which all agents address the same problem, dates back to Landgren et al. [2016a] for stochastic rewards and Cesa-Bianchi et al. [2016] for the nonstochastic case. In this setting, agents transmit their information to adjacent neighbors, from where it continues to propagate throughout the entire network, while encountering delay at each step. Communication graphs arise naturally in many problems. For example, a communication network try to find a suitable configuration. The network elements (e.g., routers) experience the same rewards for the same actions, but they cannot always send information directly to all the other network elements. In this example, the agents are the network elements, the configurations are the actions, and the communication graph is determined by the environment in which the system resides. Other cooperating entities in physical environment can also be modeled with this setting, for example, drones, cloud servers and more. The communication graph setting can also be applied to model problems involving social networks. One such example is when individuals are working towards the same objective but choose to communicate directly only with their friends within the social network.

The literature of cooperative MAB, both in the stochastic and non-stochastic case, distinguishes between group (a.k.a. average) regret and individual regret, where the latter is much stronger and more challenging to achieve. In stochastic setting, group regret was studied by Landgren et al.

17th European Workshop on Reinforcement Learning (EWRL 2024).

[2016a,b, 2018, 2021], Chakraborty et al. [2017], Martínez-Rubio et al. [2019], Wang et al. [2020], Yang et al. [2023], Chen et al. [2024], and individual regret was studied by Dubey and Pentland [2020], Wang et al. [2022]. Both in the stochastic and non-stochastic cases, approaches that explicitly synchronize between agents have often been employed to achieve individual regret bounds. While such synchronization seems to nearly optimize regret in the non-stochastic setting [Bar-On and Mansour, 2019], it appears to introduce an unnecessary artifact in the stochastic case. This artifact causes the regret bound to depends inversely on the degree of each agent in the graph [Dubey and Pentland, 2020], rather than on total number of agents. In a cycle graph, for example, their bound is similar to a single-agent bound. [Wang et al., 2022] also showed an individual regret bound, but their bound has an additive term which is linear in the diameter of the graph. In contrast, our individual regret vanishes even for a cycle, where all the degrees are two, and the diameter is of order of m.

The focus of this work is to bridge the gap between group and individual regret in the stochastic setting. To the best of our knowledge, this is the first paper to show a graph-independent individual regret bound.

1.1 Key contributions

Our key contributions are as follows:

- We present `Coop-SE`, a natural extension of the known Successive Elimination (SE) algorithm [Even-Dar et al., 2006] to the cooperative setting. `Coop-SE` is completely decentralized and each agent plays it independently. We prove that `Coop-SE` achieves a near-optimal individual regret bound of $\tilde{O}(\sqrt{(AT)/m} + A)$, which is independent of the graph and the sub-optimality gaps.
- We show a lower bound for the individual regret of $\Omega(\sqrt{AT/m} + \sqrt{A})$.
- We extend our algorithm for a bounded communication framework. We show that with $O(A \log(ATm))$ bits per message we achieve the same individual regret bound as with unbounded message size. For message of size $O(\log(ATm))$ bits we reach an individual regret bound of $\tilde{O}(A\sqrt{T/m} + A^2)$.

[Kolla et al., 2018] raised the question of whether it is feasible, in a general network, to surpass the performance of well-established single-agent policies, such as UCB [Auer et al., 2002] and SE, when executed independently across the network. Our work addresses this open question by demonstrating that, apart from an additive term of A and logarithmic factors, it is not possible to obtain an individual regret bound that improves upon the bound achieved by the simple `Coop-SE` algorithm.

Main techniques We conducted an analysis similar to SE analysis, where the number of observations for an action is limited to prevent excessive exploration. However, in a single-agent setting, the number of observations directly corresponds to the number of plays, which is not the case in our multi-agent scenario. The challenge is to ensure that a sufficient number of agents provide relevant information. To address this problem, we employ a technique we call *Implicit Synchronization*. This technique is based on the idea that there exist long enough time intervals in which every agent in large neighborhoods plays the same policy. By leveraging this property, we can establish an upper bound on the number of times an agent performs a specific action. Our algorithm is designed in such a way that every agent follows it independently, and the synchronization arises implicitly.

1.2 Additional related work

Communication-aware MABs appear in [Sankararaman et al., 2019, Chawla et al., 2020, Madhushani et al., 2021, Madhushani and Leonard, 2020, 2021a, Agarwal et al., 2022, Pavlovic et al., 2024]. The works address group regret which vanishes, or becomes independent of the number of actions, as the number of agents increase. This this line of work focus on minimizing the number of messages sent. The main difference is that we do not restrict the number of messages being sent.

Directed communication graphs were addressed in Zhu et al. [2021], Zhu and Liu [2021, 2023]. Their instance-dependent regret has an additive term which is linear in the number of agents.

Best action identification using cooperation was studied in Hillel et al. [2013], Tao et al. [2019] where the network is fully-connected and they also minimize the number of messages.

Algorithm 1 Stochastic MAB on Graph. Protocol for agent v

1: **for** $t \in [T]$ **do**
2: Agent v picks an action $a_t^v \in \mathbb{A}$.
3: Environment samples a reward, $r_t^v(a_t^v) \sim \mathcal{D}_{a_t^v}$.
4: Agent v observes reward $r_t^v(a_t^v)$.
5: Agent v sends messages $m_t^{v,u}$ to each neighbor u.
6: Agent v receives messages $m_t^{u,v}$ from each neighbor u.
7: **end for**

Heterogeneous agents which observe their neighbors with different probabilities and minimize the group regret were also addressed. Madhushani and Leonard [2019] derives a group regret based on various properties of the graph and Madhushani and Leonard [2021b] studies group regret in multi-star networks. The case of each agent having a subset of actions that are relevant to them, was studied in Yang et al. [2022], and group regret bound was derived.

Linear contextual MAB with a network of users of similar linear utility was analyzed in [Cesa-Bianchi et al., 2013].

Cooperation in Markov-Decision-Processes (MDPs) has been studied by Lidard et al. [2022] who have shown group regret guarantees in cooperative stochastic MDPs over a general network. Lancewicki et al. [2022] considered both the stochastic and non-stochastic case in cooperative MDPs but only over a fully-connected graph.

2 Model and problem formulation

Stochastic MAB (SMAB): A stochastic Multi-armed bandit problem has A actions, denoted by $\mathbb{A} = \{1, \ldots, A\}$. Each action $a \in \mathbb{A}$ has a reward distribution $\mathcal{D}_a$, whose support is $[0, 1]$, and its expectation is $\mu_a = \mathbb{E}_{r \sim \mathcal{D}_a}[r]$.

An optimal action is denoted with $a^\star \in \arg\max_{a \in \mathbb{A}} \mu_a$, and $\mu^\star = \mu_{a^\star}$. The gap of a sub-optimal action a is $\Delta_a = \mu^\star - \mu_a$.

Multi-player MAB: We have an undirected graph $\mathcal{G}(V, E)$, where V is the set of vertices and E the set of edges. Every vertex represents an agent. An agent u is a neighbor of agent v iff $(v, u) \in E$. Let $N_{\leq d}^v$ be the set of agents at distance at most d from agent v, i.e., $N_{\leq d}^v := \{u \in V | d_{\mathcal{G}}(v, u) \leq d\}$, where $d_{\mathcal{G}}(v, u)$ is the minimal path length (number of edges) from v to u in $\mathcal{G}$.

There are T rounds of play. Each agent $v \in V$, in each round of play $t \in [T]$ does the following: (1) selects an action $a_t^v \in \mathbb{A}$ and observes a reward $r_t^v(a_t^v)$. (2) sends messages to neighboring agents $u \in N_{\leq 1}^v$. (3) receives messages from neighboring agents $u \in N_{\leq 1}^v$. See Algorithm 1.

Regret definition: The individual (pseudo) regret of an agent v is defined by $\mathcal{R}_T^v = \mathbb{E}[\sum_{t=1}^T (\mu^* - r_t^v(a_t^v))]$.[1] In this paper we focus on minimizing the individual (pseudo) regret of every agent.

Events and Messages: An event is a tuple describing reward, or a tuple describing an elimination of an action. A reward event is $(\texttt{rwd}, t, id, a, r)$, where t is the timestep, id is the agent's ID, $a = a_t^{id}$ is the action, and $r = r_t^{id}(a_t^{id})$ is the reward. An elimination event is $(\texttt{elim}, t, id, a)$, where t is the timestep, id is the agent, a is the eliminated action. To denote individual elements within an event tuple, we use subscript notation. For example, if we have an event $event = (\texttt{rwd}, t, id, a, r)$, we denote the action a using $event_a$. A message is a set of events.

3 The `Coop-SE` algorithm and individual regret guarantees

We present `Coop-SE`, our main algorithm, which is a natural extension of the well-known Successive Elimination (SE) algorithm to the cooperative setting. `Coop-SE` is a decentralized algorithm, and it is

[1] The expectation of the pseudo regret is also over the randomness of the algorithm. We will refer to the pseudo regret as the regret for the rest of the article.

Algorithm 2 Cooperative Successive Elimination (`Coop-SE`)

Input: number of rounds T, neighbor agents N, number of actions A, ID of current agent v.
Initialization: $t \leftarrow 1$; Set of *active* actions $\mathcal{A} = \mathbb{A}$; $R_t(a) = 0, n_t(a) = 0$ for every action a; $M_{\texttt{in}} = \emptyset$; $M_{\texttt{updates}} = \emptyset$; $M_{\texttt{sent}} = \emptyset$; $M_{\texttt{seen}} = \emptyset$;
for $t = 1, ..., T$ **do**
  for $event \in M_{\texttt{updates}}$ **do**
    if $event \notin M_{\texttt{seen}}$ **then**
      $M_{\texttt{seen}} = M_{\texttt{seen}} \cup event$
      if $event$ is `elim`-event **then** $\mathcal{A} = \mathcal{A} \setminus event_a$
      else if $event_a \in \mathcal{A}$ **then** $n_t(a) = n_t(a) + 1$, $R_t(a) = R_t(a) + event_r$
      end if
    end if
  end for
  $E = \texttt{ElimStep}(\mathcal{A}, n_t, \hat{\mu}_t, L)$, $\mathcal{A} = \mathcal{A} \setminus E$
  Choose action a_t uniformly from $\mathcal{A}$, and get reward $r_t(a_t)$
  // Send and receive messages
  $M_{\texttt{me}} = \{(\texttt{rwd}, t, v, a_t, r_{a_t})\} \cup \{(\texttt{elim}, t, v, a) | \exists a \in E\}$, $M_t^v = (M_{\texttt{me}} \cup M_{\texttt{in}}) \setminus M_{\texttt{sent}}$
  Send message M_t^v to all neighbors, receive messages $M_t^{v'}$ from each neighbor $v' \in N$
  $M_{\texttt{sent}} = M_{\texttt{sent}} \cup M_t^v$, $M_{\texttt{updates}} = M_{\texttt{me}} \cup_{v' \in N} M_t^{v'}$, $M_{\texttt{in}} = M_{\texttt{in}} \cup_{v' \in N} M_t^{v'}$
end for

Algorithm 3 Elimination Step (`ElimStep`)

Input: active actions $\mathcal{A}$, number of samples $n(a)$ for each active actions a, empirical mean for every active action $\hat{\mu}(a)$.
$E = \emptyset$
for $a \in \mathcal{A}$ **do**

$$\lambda(a) = \sqrt{\frac{2\iota}{n(a) \vee 1}}, \quad UCB(a) = \hat{\mu}(a) + \lambda(a), \quad LCB(a) = \hat{\mu}(a) - \lambda(a) \tag{1}$$

where $\iota := \log(3mTA)$.
end for
for $a \in \mathcal{A}$ **do**
  if exists a' with $UCB(a) < LCB(a')$ **then** $E = E \cup \{a\}$
  end if
end for
Return E

important to note that each agent plays `Coop-SE` independently. In this algorithm, the agent keeps track of a set of active actions and maintains a confidence interval for the mean reward associated with each action. When the upper confidence bound of an action is strictly lower than the lower confidence bound of another action, the agent can be almost certain that the former action is sub-optimal and remove it from her set of active actions, i.e., to eliminate the action. In each round, the agent selects an action with uniform distribution among the active actions. Furthermore, each agent shares all the information she generates and receives from other agents with her neighbors in the communication graph. Our cooperative adaptation of the SE algorithm utilizes all observed samples received by the agent during communication to calculate these confidence bounds. This increased information sharing significantly reduces the regret compared to the non-cooperative setting. The formal description of the algorithm is provided in Algorithm 2.

Our main result is the following theorem.

Theorem 1. *When all the agents play `Coop-SE`, i.e., Algorithm 2, the individual regret of each agent $v \in V$ is bounded by,*

$$\mathcal{R}_T^v \leq 1089\sqrt{\frac{TA\log(3mTA)}{m}} + 138A\log(3mTA) + 1.$$

A problem-specific flavor of an individual regret bound can also be found:

Theorem 2. *When all the agents play* `Coop-SE`*, i.e., Algorithm 2, the individual regret of each agent $v \in V$ is bounded by,*

$$\mathcal{R}_T^v \leq 1088 \left(\sum_{a \in A: \Delta_a > 0} \frac{\log(3mTA)}{m\Delta_a} \right) + 138A \log(3mTA) + 1.$$

To the best of our knowledge, the current state-of-the-art individual regret is achieved by Wang et al. [2022], who obtain a bound of $\tilde{O}(\sum_{a \in A: \Delta_a > 0} \frac{1}{m\Delta_a} + \sum_{a \in A} \Delta_a D)$, where D is the diameter of the communication graph. For graphs such as line graphs, their bound is linear in m, and thus, does not vanish with increasing m as opposed to our regret bound. Landgren et al. [2016a,b, 2018, 2021], Martínez-Rubio et al. [2019] show only an average regret bound of $\tilde{O}(\sum_{a \in A: \Delta_a > 0} (\frac{1}{m\Delta_a} + \Delta_a))$, however, the O-notion here hides various dependencies on the graph parameters. Our individual regret bound matches their average regret bounds, and in addition does not have any dependency on the graph parameters.

In Section 5, we present a lower bound of $\Omega(\sqrt{TA/m} + \sqrt{A})$, which almost matches the upper bound of the individual regret of `Coop-SE`. However, the exact dependency on A in this additive term still remains open.

An important insight that follows from these results is that for a sufficiently large number of agents, the instance-independent bound depends only logarithmically on T. I.e., when $m = T$, we achieve an individual regret bound of $\tilde{O}(A)$. It is interesting to see that in every large graph, even a line graph, the individual regret of every agent on the line depends on the number of actions, and logarithmically depends on T. This is in contrast to previous individual regret bounds that scale linearly with the number of agents, m, for a line graph.

Notably, our natural extension to the Successive Elimination algorithm demonstrates that effective individual regret bounds can be achieved without resorting to complex algorithms or heavy dependencies on the graph structure. Resolving the question in Kolla et al. [2018].

To summarize, our results present a simple algorithm that achieves a near-optimal individual regret bound. Importantly, this bound is independent of the graph structure and simple to understand.

In the following section, we present the key ideas employed in the analysis of the individual regret.

Remark 1. *In* `Coop-SE` *each agent selects a random action from its set of active actions. A natural alternative is to have the actions selected in a round-robin way, deterministically. For the round-robin selection we have obtained a slightly worse $\tilde{O}(\sqrt{\frac{AT}{m}} + A^2)$ regret bound (proof omitted).*

4 Individual regret analysis

In this section we provide a proof sketch that outlines the key steps in our analysis. We analyse the regret of an arbitrary agent v, and all the definitions are reference to this agent, unless explicitly stated otherwise.

The Stages. Our proof heavily rely on a notion we call *stages*. These are the time intervals between the eliminations of agent v. Formally, a stage $j \in [A]$ is the interval $[t_j, t_{j+1})$ where $t_1 = 1$, and t_{j+1} is the timestep of the j'th elimination. We'll also denote by τ_j the length of the j'th step and the number of active actions in the j'th step by $A_j := A - j + 1$.

Section 4.1 bounds the stage length as a function of the number of samples. This bound relies on *Implicit Synchronization* between agents, which we discuses in Section 4.2. Since the number of observed samples from a sub-optimal action cannot be too large before it is eliminated, this induces a restriction on stage lengths. Finally, Section 4.3 bounds the agent's regret in terms of stage length. By combining these results we obtain our main theorem.

4.1 Number of samples in terms of Stage length

Consider the j'th stage and an action a which is still active in that stage. For the sake of intuition, assume that the agents are completely synchronized, i.e., have the same set of active actions. (We will

remove this assumption later). In the first $\tau_j/2$ steps of the stage, each of the closest $\tau_j/4$ agents to v contribute to v's information approximately $\tau_j/(2A_j)$ samples of action a. Moreover, these samples are observed by v with delay of at most $\tau_j/4$ and thus will reach her before the end of the stage.

For the ease of notation, we omit the v from $N^v_{\leq\tau}$ and denote $N_{\leq\tau} := N^v_{\leq\tau}$.

If a is active in the first i stages, we would expect that the number of sameples that reaches v for that action a from the first i stages is at least of order of $\sum_{j=1}^{i} \frac{\tau_j}{A_j} \cdot |N_{\leq\tau_j/4}|$, where $N_{\leq\tau_j/4}$ is v's neighborhood of radius $\tau_j/4$.

However, in general, the agent's policies are *not* completely synchronized, and thus, we need a stronger argument to rigorously establish the above claim. In the the next subsection, we show that under `Coop-SE` the agents have *Implicit Synchronization*. Specifically, while their policies may not be synchronized for the entire stage, they do synchronize during a specific time interval of length $\Theta(\tau_j)$. In Appendix B.2 we define a *"good event"*, which intuitively captures the fact that the observed means are close to their expectations. This allows us to show the following lemma:

Lemma 1. *Under a "good event" defined in appendix B.2 (which holds w.h.p), for every action a that was not eliminated before the end of stage i, the number of samples that v observes by the end of stage i is bounded as,*

$$n_{t_{i+1}-1}(a) \geq \sum_{j=1,\tau_j>16}^{i} \frac{\tau_j}{16A_j}|N_{\leq\tau_j/4}| - 2\log(3mTA),$$

where $n_t(a)$ is the number of samples that v observed by the beginning of time t.

The above lemma implies that the amount of observed feedback from each stage is boosted by a factor $|N_{\leq\tau_j/4}|$ compared to the number of times that v itself choose the action. In particular, if the number of agents is sufficiently large (say, $m = \Theta(T)$), then $|N_{\leq\tau_j/4}| \geq \tau_j/4$ and the number of observed samples from state j is at least from an order of τ_j^2/A_j.

4.2 Implicit synchronization of neighborhoods over intervals

We now prove the *Implicit Synchronization* of our algorithm, which is one of the key components that allows us to show individual regret.

Lemma 2. *Let j be a stage index such that $\tau_j > 16$. Then every agent $u \in N_{\leq\tau_j/4}$ plays the same policy (i.e., has the same set of active actions) at time interval $[t_j + \lceil\tau_j/4\rceil, t_j + \lfloor\tau_j/2\rfloor]$.*

Proof. Let $t \in [t_j + \lceil\tau_j/4\rceil, t_j + \lfloor\tau_j/2\rfloor]$ and let $u \in N_{\leq\tau_j/4}$. Denote the active actions of v in the j'th stage as $\mathcal{A}_j$. We will show that an action a is active for u at t iff $a \in \mathcal{A}_j$.

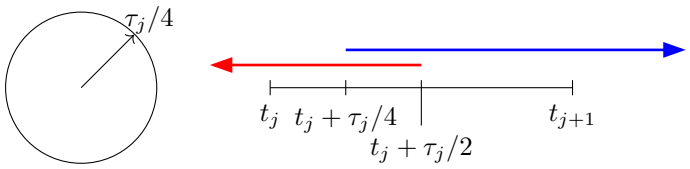
Let a be an active action of u at time t. Since $u \in N_{\leq\tau_j/4}$, we have $d_{\mathcal{G}}(u, v) \leq \tau_j/4$. The distance $d_{\mathcal{G}}(u, v)$ is a natural number, then it is at most $\lfloor\tau_j/4\rfloor$. Therefore u gets all v's eliminations (the first $j-1$ eliminated actions) until the beginning of round $t_j + \lfloor\tau_j/4\rfloor$. By the stage's definition, the agent v does not send any elimination event about one of its active actions until the end of the stage. Therefore, for any $t' \geq t_j + \lceil\tau_j/4\rceil$, u does not have any active action which is not in $\mathcal{A}_j$. Hence, $a \in \mathcal{A}_j$.

Let a be an action in $\mathcal{A}_j$. We will show that a is an active action of u at time t. Assume for contradiction that u, at timestep $t_j + \lfloor\tau_j/2\rfloor$ or before, encounters an elimination of a. The elimination event should arrive to v in no more than $\lfloor\tau_j/4\rfloor$ timesteps, so v should get the elimination event at most at timestep $t_j + \lfloor\tau_j/2\rfloor + \lfloor\tau_j/4\rfloor \leq t_j + \frac{3\tau_j}{4}$. But $\tau_j > 16$, then $\tau_j/4 > 4$, so $t_j + \frac{3\tau_j}{4} < t_{j+1} - 4$. Therefore, the elimination event about an action in $\mathcal{A}_j$ should arrive to v at least 5 timesteps before stage $j+1$ begin. Contradiction. Therefore, a is an active action of u at t.

We get that for every $t \in [t_j + \lceil\tau_j/4\rceil, t_j + \lfloor\tau_j/2\rfloor]$ and for every $u \in N_{\leq\tau_j/4}$, the active actions of u at t are exactly $\mathcal{A}_j$. In other words, we get that in time interval $[t_j + \lceil\tau_j/4\rceil, t_j + \lfloor\tau_j/2\rfloor]$ all agents in $N_{\leq\tau_j/4}$ plays the same policy, i.e., choosing randomly from $\mathcal{A}_j$. □

One can notice that $\tau_j/4$ is both distance in the graph, and a timesteps interval length. The two roles the same $\tau_j/4$ fulfills are depicted in Figure 1.

Figure 1: **Visualization of the proof idea in Lemma 2.** The circle on the left depicts the $\tau_j/4$-neighborhood of v. On the red interval (left-pointing arrow), every relevant agent has at least $\mathcal{A}_j$ as the active actions, and on the blue (right-pointing arrow) interval every relevant agent has at most $\mathcal{A}_j$ as the active actions. The intersection of these intervals is the $\tau_j/4$-length of timesteps interval in which the $\tau_j/4$-neighborhood plays with the same active actions as v.



4.3 Bounding the regret

We start by bounding agent's v regret in terms of stage lengths. Fix a sub-optimal action a and assume that i is the last stage that a was active. Since v samples active actions uniformly, in each stage $j \leq i$, she sampled a approximately τ_j/A_j times. Thus, the total number of times v samples a is approximately $\sum_{j=1}^{i} \frac{\tau_j}{A_j}$ and we can roughly bound the regret with,

$$\mathcal{R}_T \lesssim \sum_{i=1}^{A} \sum_{j=1}^{i} \frac{\tau_j}{A_j} \Delta_i, \tag{2}$$

where we slightly abuse notation and let Δ_i to be the sub-optimality gap of the action that was eliminated at the end of stage i.

Next, we use Lemma 1, to induce constraints on $\{\tau_j\}_{j=1}^{A}$. Using standard concentration bounds, we can show that the number of samples v can see from a sub-optimal action a, without eliminating it, is approximately $1/\Delta_a^2$. Thus, from Lemma 1 we get that for the i'th action v eliminates,

$$\sum_{j=1}^{i} |N_{\leq \tau_j/4}| \frac{\tau_j}{16 A_j} \lesssim \frac{1}{\Delta_i^2}. \tag{3}$$

If $N_{\leq \tau_j/4}$ is not the entire graph (i.e., not of size m) then $|N_{\leq \tau_j/4}| \geq \frac{\tau_j}{4}$. (Note that for a line graph we have approximate equality). Hence, $|N_{\leq \tau_j/4}| \geq \min\{m, \frac{\tau_j}{4}\}$. Denote the indices of the short intervals in which $\tau_j/4 < m$ with S_τ, i.e., $S_\tau = \{j : \tau_j/4 < m\}$. Now, from (3) we get

$$\sum_{j=1, j\in S_\tau}^{i} \frac{\tau_j^2}{64 A_j} + \sum_{j=1, j\notin S_\tau}^{i} m \frac{\tau_j}{16 A_j} \leq \sum_{j=1}^{i} \min\{m, \frac{\tau_j}{4}\} \frac{\tau_j}{16 A_j} \leq \sum_{j=1}^{i} |N_{\leq \tau_j/4}| \frac{\tau_j}{16 A_j} \lesssim \frac{1}{\Delta_i^2}$$

This implies that

$$\sum_{j=1, j\in S_\tau}^{i} \frac{\tau_j^2}{64 A_j} \lesssim \frac{1}{\Delta_i^2}, \qquad \text{and} \qquad \sum_{j=1, j\notin S_\tau}^{i} m \frac{\tau_j}{16 A_j} \lesssim \frac{1}{\Delta_i^2}. \tag{4}$$

With that in hand, we can break the sum in Equation (2) to stages in S_τ and outside S_τ, and use the above two conditions in order to obtain our final regret bound of Theorem 1. The formal details are rather technical and deferred to the appendix.

5 Lower bound

The following section derives the lower bound, which also shows that our algorithm achieves near-optimal individual regret.

Theorem 3. *For every algorithm, and for every T, A, m, there exists a problem instance of the cooperative stochastic MAB over a communication graph such that there exists an agent for which the individual regret is at least,*

$$\Omega\left(\sqrt{\frac{AT}{m}} + \sqrt{A}\right).$$

Proof. We will show two separate lower bounds, $\Omega(\sqrt{AT/m})$ and $\Omega(\sqrt{A})$.

For the $\Omega(\sqrt{AT/m})$, we consider a fully-connected network. This lower bound follows from the lower bound on group regret. The group regret of m agents with T time steps is lower bounded by the optimal regret of a single agent running mT time steps. The standard lower bounds for MAB imply an $\Omega(\sqrt{mTA})$ for the single agent running for mT time steps. Therefore, there exists some agent whose regret is $\Omega(\sqrt{AT/m})$. This is essentially the lower bound in Ito et al. [2020].

For the $\Omega(\sqrt{A})$ bound, we use a line graph. The lower bound appears formally in Theorem 6. The intuition of the proof is the following. We consider a deterministic MAB where one action has reward 1 and all other actions have reward 0. An agent, during the first τ time steps receives $\Theta(\tau^2)$ observations. Therefore, if $\tau \lesssim \sqrt{A}/10$, then an agent receives information about at most $A/100$ of the actions. If the optimal action is selected at random then with probability 0.99 the agent will not observe it, and hence will have an individual regret of at least $\sqrt{A}/10$. For the formal proof see Theorem 6. □

6 Low communication results

In this setting we model a problem in which the communication channels between the agents are bounded. This type of restriction is common when modeling a communication network. Specifically, we limit the size of the messages that the agent can transmit. This model is referred to as the CONGEST model in the distributed literature [Peleg, 2000].

We derived two results. The first is for message size of $O(A\log(ATm))$ bits. The second requires the messages to be logarithmic in all parameters, including the number of actions, and each message is at most $O(\log(ATm))$ bits.

The first step in our solution is to avoid duplicated messages. This can be done by selecting a spanning tree in the graph and limiting all our communication to this spanning tree. Note that limiting the communication to the spanning tree does not affect the individual regret bound of `Coop-SE`. This is unique to our algorithm, since `Coop-SE` promises a regret bound that does not depend on the graph parameters. When we send the messages on the tree, we perform a broadcast, which is done by forwarding the message on all the edges except the edge it was received from. This guarantees that each message is received by each agent only once.

The next observation is that the `Coop-SE` algorithm can aggregate all the events regarding an action a into two events, one for the rewards, `rwd`, and one for elimination, `elim`. The message will contain information about every action a, whether it was eliminated, the number of times it was observed, and the sum of the rewards observed. The size of such message is $O(A\log(ATm))$ bits. This is all the information the agents need to receive from the multiple messages.

Algorithm `Coop-SE-Restricted`, Algorithm 4, uses the above two observations. It creates a spanning tree out of the communication graph, on which it transmits the messages. It also merges multiple message for the same action, as outline above. We derive the following regret bound for Algorithm `Coop-SE-Restricted`.

Theorem 4. *When the message size is bounded by $O(A\log(ATm))$ bits, and when all the agents play `Coop-SE-Restricted` (Algorithm 4), the same bounds for `Coop-SE` hold, i.e., Theorem 1 and Theorem 2 hold. Specifically, the individual regret of each agent $v \in V$ is bounded by,*

$$\mathcal{R}_T^v \leq 1089\sqrt{\frac{TA\log(3mTA)}{m}} + 138A\log(3mTA) + 1.$$

The above model assume that messages are $\tilde{O}(A)$ bits long. In case of a large number of action, it is reasonable to require the messages to be only logarithmic in the problem parameters, including the number of actions A. For this reason we consider the case that the size of the messages is limited to $O(\log(ATm))$ bits. Algorithm `Coop-SE-Low-Comm`, Algorithm 5, works with messages of size at most $O(\log(ATm))$ bits.

Here is the high level idea of the algorithm. Rather than sending messages of size $O(A\log(mTA))$, we break them into A messages of logarithmic size. Specifically, `Coop-SE-Low-Comm` operates over $\lfloor T/A \rfloor$ blocks of A timesteps each. At the beginning of each block, the agent samples an

action and plays it throughout the entire block. To simulate faithfully, the agent fixes the action and ignores the information she produces except for the first round in the block. In each timestep within a block she sends a message that includes only the `rwd` and `elim` events of a single action, aggregating messages from the previous block. In this way, each block effectively simulates a single timestep of `Coop-SE-Restricted` - that is, a play of an action and transmission of messages containing events for all A actions. Since the same action is played throughout the entire block, the regret incurred in each block of `Coop-SE-Low-Comm` corresponds to the regret in each timestep of `Coop-SE-Restricted`, scaled up by a factor of A. Consequently, the total regret experienced over the entire duration of `Coop-SE-Low-Comm` corresponds to the cumulative regret of `Coop-SE-Restricted` over $\lfloor T/A \rfloor$ timesteps, also scaled by a factor of A.

Formally, we obtain the following theorem. A detailed proof is provided in the supplementary material.

Theorem 5. *When the message size is bounded by $O(\log(ATm))$ bits, and when all the agents play `Coop-SE-Low-Comm`, i.e., Algorithm 5, the individual regret of each agent $v \in V$ is bounded by,*

$$\mathcal{R}_T^v \leq 1089A\sqrt{\frac{T\log(3mTA)}{m}} + 140A^2\log(3mTA).$$

7 Summary and future work

In this paper, we introduced `Coop-SE`, a simple extension of the well-known Successive Elimination (SE) algorithm, to address the problem of cooperative stochastic MAB over a communication graph. Our main contribution is the proof that when all agents play `Coop-SE`, the individual regret is bounded by $\tilde{O}(\sqrt{\frac{TA}{m}} + A)$, which is near-optimal and independent of the graph structure. We also provided a lower bound of $\Omega(\sqrt{\frac{TA}{m}} + \sqrt{A})$ for this problem. Although the upper and lower bounds nearly match, obtaining the optimal dependency on A remains an open question.

We also discussed the effect of messages of bounded size. For the case where the message size is restricted to $O(A\log(ATm))$ bits, we presented the `Coop-SE-Restricted` algorithm, which achieves the same individual regret bound as `Coop-SE`. When the message size is restricted to $O(\log(ATm))$ bits, the `Coop-SE-Low-Comm` algorithm achieves an individual regret of $\tilde{O}(A\sqrt{\frac{T}{m}} + A^2)$. An obvious open problem is to reduce the later regret bound.

Future research could investigate how different communication restrictions impact the performance of `Coop-SE` and other algorithms targeting individual regret. For example, bounding not only the message size, but also the number of messages.

It would be very interesting to extend the results to other MAB agorithms, specifically, Upper Confidence Bound (UCB) and Thompson sampling. Unlike the SE algorithm, the UCB algorithm does not possess the property of *Implicit Synchronization* when each agent runs an independent UCB algorithm. We leverage the *Implicit Synchronization* property, which means that agents play the same policy, to ensure the number of times an agent plays an action is related to the number of observations she makes. On the other hand, for agents using UCB to create the same policy, they must have identical empirical bounds. We leave the analysis and adaptation of UCB for the individual regret in the cooperative setting as future work. Similar issues might arise also in the potential adaptation of Thomson sampling to the cooperative setting.

Acknowledgement

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation, the Yandex Initiative for Machine Learning at Tel Aviv University and a grant from the Tel Aviv University Center for AI and Data Science (TAD).

References

M. Agarwal, V. Aggarwal, and K. Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *J. Mach. Learn. Res.*, 23:212:1–212:24, 2022. URL `https://jmlr.org/papers/v23/21-138.html`.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Y. Bar-On and Y. Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

N. Cesa-Bianchi, C. Gentile, and G. Zappella. A gang of bandits. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 737–745, 2013.

N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 605–622. JMLR.org, 2016.

M. Chakraborty, K. Y. P. Chua, S. Das, and B. Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 164–170. ijcai.org, 2017.

R. Chawla, A. Sankararaman, A. Ganesh, and S. Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 3471–3481. PMLR, 2020.

Z. Chen, K. Cai, J. Zhang, and Z. Yu. Fair distributed cooperative bandit learning on networks for intelligent internet of things systems (technical report). *CoRR*, abs/2403.11603, 2024.

A. Cohen, Y. Efroni, Y. Mansour, and A. Rosenberg. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021.

C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Dubey and A. S. Pentland. Cooperative multi-agent bandits with heavy tails. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2730–2739. PMLR, 2020.

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, 2006.

E. Hillel, Z. S. Karnin, T. Koren, R. Lempel, and O. Somekh. Distributed exploration in multi-armed bandits. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 854–862, 2013.

S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4872–4883. Curran Associates, Inc., 2020.

R. K. Kolla, K. P. Jagannathan, and A. Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Trans. Netw.*, 26(4):1782–1795, 2018.

T. Lancewicki, A. Rosenberg, and Y. Mansour. Cooperative online learning in stochastic and adversarial MDPs. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11918–11968. PMLR, 17–23 Jul 2022.

P. Landgren, V. Srivastava, and N. E. Leonard. On distributed cooperative decision-making in multiarmed bandits. In *15th European Control Conference, ECC 2016, Aalborg, Denmark, June 29 - July 1, 2016*, pages 243–248. IEEE, 2016a.

P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *55th IEEE Conference on Decision and Control, CDC 2016, Las Vegas, NV, USA, December 12-14, 2016*, pages 167–172. IEEE, 2016b.

P. Landgren, V. Srivastava, and N. E. Leonard. Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *57th IEEE Conference on Decision and Control, CDC 2018, Miami, FL, USA, December 17-19, 2018*, pages 5239–5244. IEEE, 2018.

P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Autom.*, 125:109445, 2021.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

J. Lidard, U. Madhushani, and N. E. Leonard. Provably efficient multi-agent reinforcement learning with fully decentralized communication. In *2022 American Control Conference (ACC)*, pages 3311–3316. IEEE, 2022.

U. Madhushani and N. E. Leonard. Heterogeneous stochastic interactions for multiple agents in a multi-armed bandit problem. In *17th European Control Conference, ECC 2019, Naples, Italy, June 25-28, 2019*, pages 3502–3507. IEEE, 2019.

U. Madhushani and N. E. Leonard. Distributed learning: Sequential decision making in resource-constrained environments. *CoRR*, abs/2004.06171, 2020.

U. Madhushani and N. E. Leonard. When to call your neighbor? strategic communication in cooperative stochastic bandits. *CoRR*, abs/2110.04396, 2021a.

U. Madhushani and N. E. Leonard. Heterogeneous explore-exploit strategies on multi-star networks. *IEEE Control. Syst. Lett.*, 5(5):1603–1608, 2021b.

U. Madhushani, A. Dubey, N. E. Leonard, and A. Pentland. One more step towards reality: Cooperative bandits with imperfect communication. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 7813–7824, 2021.

D. Martínez-Rubio, V. Kanade, and P. Rebeschini. Decentralized cooperative stochastic bandits. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4531–4542, 2019.

N. Pavlovic, S. Salgia, and Q. Zhao. Order-optimal regret in distributed kernel bandits using uniform sampling with shared randomness. *CoRR*, abs/2402.13182, 2024. doi: 10.48550/ARXIV.2402.13182. URL `https://doi.org/10.48550/arXiv.2402.13182`.

D. Peleg. *Distributed computing: a locality-sensitive approach*. SIAM, 2000.

A. Sankararaman, A. Ganesh, and S. Shakkottai. Social learning in multi agent multi armed bandits. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(3):53:1–53:35, 2019.

C. Tao, Q. Zhang, and Y. Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In D. Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 126–146. IEEE Computer Society, 2019.

P. Wang, A. Proutière, K. Ariu, Y. Jedra, and A. Russo. Optimal algorithms for multiplayer multi-armed bandits. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 4120–4129. PMLR, 2020.

X. Wang, L. Yang, Y.-z. J. Chen, X. Liu, M. Hajiesmaili, D. Towsley, and J. C. Lui. Achieving near-optimal individual regret & low communications in multi-agent bandits. In *The Eleventh International Conference on Learning Representations*, 2022.

L. Yang, Y. J. Chen, M. H. Hajiesmaili, J. C. S. Lui, and D. Towsley. Distributed bandits with heterogeneous agents. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, May 2-5, 2022*, pages 200–209. IEEE, 2022.

L. Yang, X. Wang, M. Hajiesmaili, L. Zhang, J. C. S. Lui, and D. Towsley. Cooperative multi-agent bandits: Distributed algorithms with optimal individual regret and constant communication costs. *CoRR*, abs/2308.04314, 2023.

J. Zhu and J. Liu. A distributed algorithm for multi-armed bandit with homogeneous rewards over directed graphs. In *2021 American Control Conference, ACC 2021, New Orleans, LA, USA, May 25-28, 2021*, pages 3038–3043. IEEE, 2021.

J. Zhu and J. Liu. Distributed multiarmed bandits. *IEEE Trans. Autom. Control.*, 68(5):3025–3040, 2023.

J. Zhu, E. Mulle, C. S. Smith, and J. Liu. Decentralized multi-armed bandit can outperform classic upper confidence bound. *arXiv preprint arXiv:2111.10933*, 2021.

A Summery of notations

For convenience, the table below summarizes most of the notation used throughout the paper.

$\mathcal{D}_a$	The reward distribution of action a
μ_a	The expected reward of action a
$\mu^\star$	The maximal expected reward
$a^\star$	An optimal action
Δ_a	The sub-optimality gap of action a
$N^u_{\leq d}$	The set of agents at distance at most d from agent u
$N_{\leq d}$	For ease of notation $N_{\leq d} := N^v_{\leq d}$; see Remark 2
$d_{\mathcal{G}}(v,u)$	The minimal path length (number of edges) from v to u
t_j	The beginning of stage j of agent v; see Remark 2
τ_j	The length of stage j of agent v; see Remark 2
A_j	The number of active actions in the j'th step of agent v; see Remark 2
ι	$\log(3mTA)$
$n^u_t(a)$	The number of samples that u observed by the beginning of time t
$n_t(a)$	For ease of notation $n_t(a) := n^v_t(a)$; see Remark 2
$b^u_t(a)$	The number of times agent u played action a until the beginning of round t.
$b_t(a)$	For ease of notation $b_t(a) := b^v_t(a)$; see Remark 2
p^u_k	The policy of agent u at time k
A_Δ	The set of elimination indices (with respect to agent v) with gaps larger than $\sqrt{\frac{A\iota}{Tm}}$
a_i	The i'th action being eliminaed by agent v
Δ_i	The sub-optimality gap of a_i
G_τ	The set of "Good Intervals": $\{j \vert \tau_j > 16\}$.
S_τ	The set of "Short Intervals": $\{j \vert j \in G_\tau \ \ \& \ \ \tau_j/4 < m\}$

B Proof of the main theorem

Remark 2. *For the ease of notation, the following proof and definitions focus on a specific agent, named v.*

B.1 Definitions

Definition 1. *A stage is a timestep-interval when its boundaries are the eliminations. The stage's index is usually denoted by j. The time interval is split into A different stages. Assume that the elimination timesteps are $s_1, s_2, \ldots$. The first stage starts at $t = 1$ and ends with the first elimination. I.e., it is the timesteps that are in time interval $[1, s_1)$. The second stage is $[s_1, s_2)$, etc. Denote t_j to be the timestep in which the agent started the j'th stage, where $t_1 = 1$ and $t_{A+1} = T + 1$.*

Definition 2. *Denote τ_j to be the length of the j'th stage (for agent v).*

Definition 3. *Denote $A_j := A - j + 1$ to be the number of remained actions in the j'th stage.*

Definition 4. *Elimination index i of the action a is the stage index in which in its end the action is eliminated. Every action has a unique elimination index.*

If some actions are eliminated in the same timestep, then the stage is of zero length and the elimination index are chosen arbitrary. The elimination index of a is denoted by i_a, and the appropriate action for elimination index i is denoted by a_i.

Definition 5. *Denote with A_Δ the set of elimination indices of large gaps. $A_\Delta = \{i \vert \Delta_{a_i} \geq \sqrt{\frac{A\iota}{Tm}}\}$.*

Definition 6. *For the ease of notation, denote $\Delta_i := \Delta_{a_i}$.*

Definition 7. *Define the set of "Good Intervals" to be the set of long enough intervals: $G_\tau = \{j \vert \tau_j > 16\}$. These are the intervals we will focus in the proofs.*

Definition 8. *Denote the group of indices of short stages with S_τ. Specifically,*

$$S_\tau := \{j \vert j \in G_\tau \ \ \& \ \ \tau_j/4 < m\}$$

Definition 9. *Denote the number of samples an agent u sees for action a until the begging of timestep t with $n_t^u(a)$. For the ease of notation, denote $n_t(a) := n_t^v(a)$.*

Definition 10. *Denote by $b_t^u(a)$ the number of times agent u played action a until the beginning of round t.*

Definition 11. *Denote the upper confidence bound for agent u for action a with $UCB_{n(a)}^u(a) = \hat{\mu}(a) + \sqrt{\frac{2\log(3mTA)}{n(a)\vee 1}}$, where $n(a)$ is the number of times agent u observed action a, and $\hat{\mu}(a)$ is the empirical mean calculated by u for action a. Similarly, let $LCB_{n(a)}^u(a) = \hat{\mu}(a) - \sqrt{\frac{2\log(3mTA)}{n(a)\vee 1}}$ denote the corresponding lower confidence bound. In other words, $UCB_{n(a)}^u(a)$ and $LCB_{n(a)}^u(a)$ are the confidence bounds calculated in Algorithm 3 Equation* (1) *when agent u calls this algorithm with parameters n, a vector containing the number of observations for each action, and $\hat{\mu}$, the vector of empirical means.*

Definition 12. *The policy of agent u at time t is denoted with p_t^u. I.e., $p_t^u(a)$ is the probability that agent u plays action a at time t given her observations up to time t. In all the proposed algorithms it is simply* 1 *divide the number of active actions if the action is active, and zero otherwise.*

Definition 13. *Denote the logarithmic term used in Algorithm 3 with ι, i.e., $\iota = \log(3mTA)$*

B.2 The good event

The first good event G^1 captures the intuition that the true expectation of each action is between the UCB and the LCB.

Definition 14. *Define the good event, G^1, to be the event in which for every agent u, for every action a and for every* `rwd`*-event that was received, the empirical mean is in the confidence interval, i.e.,*

$$\mu_a \in [LCB_{n(a)}^u(a), UCB_{n(a)}^u(a),]$$

where $n(a)$ is the number of `rwd`*-events that were received for this action by the agent u.*

Lemma 3. *Let w be an agent and let $X_t^w(a) := \mathbb{I}(a_t^w = a)$ be the indicator that w plays action a at timestep t. Then for any agent u, timestep t, and action a,*

$$n_t^u(a) = \sum_{k=1}^{t-1} \sum_{w \in N_{\leq t-k}^u} X_k^w(a)$$

Proof. Let w be an agent such that $w \neq u$ and $d_{\mathcal{G}}(w,u) = d$. Every $X_k^w(a)$ reaches u at the end of round $k+d-1$. Therefore, it contributes to $n_{t'}^u(a)$ at timestep $t' = k+d$. We get that for $w \neq u$, $w \in N_{\leq t-k}^u$, $X_k^w(a)$ reaches u until the beginning of timestep t.

Now, let $w = u$ and $k < t$. An agent u uses the information she creates only at the next timestep. Since we do not sum the information for the current timestep t, i.e., $t-k \geq 1$, the information u creates is summed only for timesteps that passed. In other words, for $w = u$, $X_k^w(a)$ is summed only at timesteps $t' < t$, for them the information reaches u until the beginning of t. Therefore, we get that for all $k < t$, $w \in N_{\leq t-k}^u$, $X_k^w(a)$ reaches u until the beginning of timestep t. Summing over all the timesteps at which information on action a can be produced and we obtain the result. □

The second good event G^2 requires that the number of observations of an action is not much less than its expectation.

Definition 15. *Define the good event G^2 to be the event in which for all $u \in V$, action a and timestep $t \in T$ simultaneously,*

$$n_t^u(a) \geq \frac{1}{2} \sum_{k=1}^{t-1} \sum_{w \in N_{\leq t-k}^u} p_k^w(a) - 2\iota.$$

The third good event G^3 requires that the number of plays of an action is not much more than its expectation.

Definition 16. *Define the good event G^3 to be the event in which for all $u \in V$, action a and timestep $t \in T$ simultaneously,*

$$b_t^u(a) \leq 2\sum_{k=1}^{t-1} p_k^u(a) + 12\iota.$$

The following lemma show that with high probability all the good events hold.

Lemma 4. *When all agents play Algorithm 2 the good event, $G := G^1 \cup G^2 \cup G^3$, happens with probability of at least $1 - \frac{1}{T^2}$.*

Proof. We will show that each of the events $\neg G^1, \neg G^2$ and $\neg G^3$ happens with probability of at most $\frac{1}{3T^2}$. Thus, by the union bound, G occur with probability of at least $1 - \frac{1}{T^2}$.

Event $\neg G^1$: Denote $M_a^u(k)$ to be the k'th `rwd` event agent u received for action a. Define $X_n^u(a) := \sum_{k=1}^{n}(M_a^u(k) - \mu_a)$ and $\lambda_n := \sqrt{\frac{2\iota}{n}}$. Note that $X_n^u(a)$ is a martingale. From Azuma's inequality we get

$$Pr\left(\left|\frac{X_n^u(a)}{n} - \mu_a\right| \geq \lambda_n\right) \leq \frac{1}{3m^3T^3A^3}$$

There are at most $m \cdot T$ `rwd` events the agent can get. The same holds for every action and for every message. The upper confidence bound ($UCB_n^u(a)$) is defined as $X_n^u(a) + \lambda_n$ and the lower confidence bound ($LCB_n^u(a)$) is defined as $X_n^u(a) - \lambda_n$. From the union bound we get that with high probability for every agent, for every timestep, for every action and for every `rwd` event message the agent get, the actual mean of the action would be inside the confidence bound. Specifically

$$G^1 := \forall u \in V, \forall a \in A, \forall n \in [m \cdot T](\mu_a \in [LCB_n^u(a), UCB_n^u(a)])$$

$$\mathbb{P}(\neg G^1) \leq \frac{1}{3mT^2A^2} \leq \frac{1}{3T^2}$$

Event $\neg G^2$: Fix an action a and agent u. Let $X_{k,w} = \mathbb{I}\{a_k^w = a\}$ and $\mathcal{F}_{t,w}$ be the sigma algebra induced by the first $t-1$ rounds; and the actions chosen by the first $w-1$ agents in round t (where we assume a linear order on the agents - for example the alphabetic order induced by their IDs). Notice that $\mathcal{F}_{t,1}$ is induced simply by the first $t-1$ rounds. Note that $p_k^w(a)$ is $\mathcal{F}_{k,w}$-measurable, $\mathbb{E}[X_{k,w} \mid \mathcal{F}_{k,w}] = p_k^w(a)$ and that $X_{k,w}$ is $\mathcal{F}_{k,w+1}$-measurable (or if w is the last agent, $X_{k,w}$ is $\mathcal{F}_{k+1,1}$-measurable). By applying Lemma 15, with probability $1 - \frac{1}{9T^2A^2m^2}$ for all $t \in [T]$ simultaneously we have,

$$n_t^u(a) = \sum_{k=1}^{t-1}\sum_{w \in N_{\leq t-k}^u} X_k^w(a) \geq \frac{1}{2}\sum_{k=1}^{t-1}\sum_{w \in N_{\leq t-k}^u} p_k^w(a) - 2\iota,$$

where the equality is from Lemma 3. By taking the union bound over all actions, a, and agents u we get that $\mathbb{P}(\neg G^2) \leq 1/(9mAT^2) \leq 1/(3T^2)$.

Event $\neg G^3$: Fix an action a, agent u and timestep t. Let $X_k = \mathbb{I}\{a_k^u = a\}$ and $\mathcal{F}_t$ be the sigma algebra induced by the first $t-1$ rounds. Note that $p_k^w(a)$ is $\mathcal{F}_k$-measurable, $\mathbb{E}[X_k \mid \mathcal{F}_k] = p_k^u(a)$ and that X_k is $\mathcal{F}_{k+1}$-measurable. By applying Lemma 16, with probability $1 - \frac{1}{27T^3A^3m^3}$,

$$b_t^u(a) = \sum_{k=1}^{t-1} X_k(a) \leq 2\sum_{k=1}^{t-1} p_k^u(a) + 12\iota.$$

By taking the union bound over all time steps t, actions a, and agents u we have $\mathbb{P}(\neg G^3) \leq \frac{1}{27T^2A^2m^2} \leq \frac{1}{3T^2}$.

Taking the union bound over $\neg G^1 \cup \neg G^2 \cup \neg G^3$, we complete the proof. □

B.3 Proof of Theorem 1

Lemma 5. *The complementary event of the good event adds no more than* 1 *to the regret of each agent.*

Proof. From Lemma 4, the complementary event of the good event happens in probability lower than $\frac{1}{T^2}$. Every agent plays T timesteps, and the gaps are bounded by 1, i.e., for every action a we have $\Delta_a \leq 1$. Hence, in expectation, this adds at most $\frac{1}{T} \leq 1$ to the regret. □

In the proof from now on, we assume the good event $G := G^1 \cup G^2 \cup G^3$ holds.

Lemma 6 (restatement of Lemma 1)**.** *For every action a that was not eliminated before the end of stage i, we have*

$$n_{t_{i+1}-1}(a) \geq \sum_{j=1, j\in G_\tau}^{i} \frac{\tau_j}{16A_j} |N_{\leq \tau_j/4}| - 2\iota.$$

Proof. Under the good event G^2,

$$\begin{aligned}
n_{t_{i+1}-1}(a) &\geq \frac{1}{2} \sum_{t=1}^{t_{i+1}-2} \sum_{u\in N_{\leq t_{i+1}-t-2}} p_t^u(a) - 2\iota \\
&\geq \frac{1}{2} \sum_{j=1}^{i} \sum_{t=t_j}^{t_j+\tau_j-2} \sum_{u\in N_{\leq t_{i+1}-t-1}} p_t^u(a) - 2\iota \\
&\geq \frac{1}{2} \sum_{j=1, j\in G_\tau}^{i} \sum_{t=t_j+\lceil \tau_j/4 \rceil}^{t_j+\lfloor \tau_j/2 \rfloor} \sum_{u\in N_{\leq t_{i+1}-t-2}} p_t^u(a) - 2\iota \\
&\geq \frac{1}{2} \sum_{j=1, j\in G_\tau}^{i} \sum_{t=t_j+\lceil \tau_j/4 \rceil}^{t_j+\lfloor \tau_j/2 \rfloor} \sum_{u\in N_{\leq \tau_j/4}} p_t^u(a) - 2\iota.
\end{aligned}$$

The second inequality is by splitting the rounds to stages and summing partially. The third inequality is by summing partially over $j \in G_\tau$ ($\lfloor \tau_j/2 \rfloor \leq \tau_j/2 - 1 \leq \tau_j - 2$). The last inequality is since $N_{\leq \tau_j/4} \subseteq N_{\leq t_{i+1}-t-2}$ as for all $j \in [i] \cap G_\tau$ and $t \leq t_j + \lfloor \tau_j/2 \rfloor$,

$$t_{i+1} - t - 2 \geq t_{j+1} - t_j - \lfloor \tau_j/2 \rfloor - 2 \geq \tau_j - \tau_j/2 - 3 = \tau_j/2 - 3 \geq \tau_j/4.$$

Finally, by Lemma 2, all agents $u \in N_{\leq \tau_j/4}$ play the same policy at time steps $t \in [t_j + \lceil \tau_j/4 \rceil, t_j + \lfloor \tau_j/2 \rfloor]$ which is uniform over the active actions. I.e., $p_t^u(a) = \frac{1}{A_j}$ for active actions in $[t_j + \lceil \tau_j/4 \rceil, t_j + \lfloor \tau_j/2 \rfloor]$. The interval $[t_j + \lceil \tau_j/4 \rceil, t_j + \lfloor \tau_j/2 \rfloor]$ is of size at least $\tau_j/8$, since $t_j + \lceil \tau/4 \rceil - t_j + \lfloor \tau/2 \rfloor \geq \frac{\tau_j}{2} - \frac{\tau_j}{4} - 2 = \frac{\tau_j}{4} - 2 \geq \frac{\tau_j}{8}$, when the last inequality follows from the that for every $j \in G_\tau, \tau_j > 16$. Thus,

$$\frac{1}{2} \sum_{j=1, j\in G_\tau}^{i} \sum_{t=t_j+\lceil \tau_j/4 \rceil}^{t_j+\lfloor \tau_j/2 \rfloor} \sum_{u\in N_{\leq \tau_j/4}} p_t^u(a) \geq \sum_{j=1, j\in G_\tau}^{i} \frac{\tau_j}{16A_j} |N_{\leq \tau_j/4}|,$$

as desired. □

Lemma 7. *For every action a that was not eliminated before the end of stage i,*

$$\sum_{j=1, j\in G_\tau}^{i} \frac{\tau_j}{A_j} |N_{\leq \tau_j/4}| \leq \frac{544\iota}{\Delta_a^2}.$$

Proof. Fix an action a that was not eliminated before the end of stage i. Denote $t' = t_{i+1} - 1$. The action a is still active by agent v at time t', and thus, $UCB_{t'}^v(a) \geq LCB_{t'}^v(a^\star)$. Note the slightly

abuse of notation, when $UCB^v_{t'}(a)$ is actually $UCB^v_{n_{t'}(a)}(a)$, and the same for LCB. Under the good event G^1,

$$\mu_a + 2\lambda^v_{t'}(a) \geq UCB^v_{t'}(a) \geq LCB^v_{t'}(a^\star) \geq \mu_{a^\star} - 2\lambda^v_{t'}(a^\star).$$

Rearranging it we get,

$$\Delta_a \leq 2\sqrt{\frac{2\iota}{n_{t'}(a)}} + 2\sqrt{\frac{2\iota}{n_{t'}(a^\star)}}.$$

Recall that under the good event, $a^\star$ is never eliminated. Thus, we can apply Lemma 6 on both a and $a^\star$ and further bound Δ_a by,

$$\Delta_a \leq 4\sqrt{\frac{2\iota}{\sum_{j=1,j\in G_\tau}^{i} \frac{\tau_j}{16A_j}|N_{\leq\tau_j/4}| - 2\iota}},$$

then

$$\Delta_a^2 \leq 16\frac{2\iota}{\sum_{j=1,j\in G_\tau}^{i} \frac{\tau_j}{16A_j}|N_{\leq\tau_j/4}| - 2\iota},$$

we get

$$\sum_{j=1,j\in G_\tau}^{i} \frac{\tau_j}{16A_j}|N_{\leq\tau_j/4}| - 2\iota \leq \frac{32\iota}{\Delta_a^2},$$

and,

$$\sum_{j=1,j\in G_\tau}^{i} \frac{\tau_j}{16A_j}|N_{\leq\tau_j/4}| \leq \frac{32\iota}{\Delta_a^2} + 2\iota \leq \frac{34\iota}{\Delta_a^2}.$$

By rearranging terms we get the Lemma's statement. □

Lemma 8. *For any* $\tau \geq 0$

$$\min\{\tau, m\} \leq |N_{\leq\tau}|$$

Proof. The graph is connected, so either there exists an agent u at distance $\lfloor\tau\rfloor$ from v, in which case $N_{\leq\tau} \geq \lceil\tau\rceil \geq \tau$, or all the agents are at distance at most τ from v, in which case $N_{\leq\tau} = m$. □

Lemma 9. $\sum_{j=1}^{A} \frac{1}{A_j} \leq \log A + 1$

Proof.

$$\begin{aligned}
\sum_{j=1}^{A} \frac{1}{A_j} &= \sum_{j=1}^{A} \frac{1}{A-j+1} \\
&= \sum_{i=1}^{A} \frac{1}{i} \\
&= 1 + \sum_{i=2}^{A} \frac{1}{i} \\
&\leq 1 + \int_1^A \frac{1}{x}dx \\
&= 1 + \log A
\end{aligned}$$

□

Lemma 10. *Let $\{\tau_j | j \in [A]\}$ be the stage lengths. The regret of agent v (under the good event) is bounded by*

$$\mathcal{R}_T \leq 2 \sum_{i \in A_\Delta} \sum_{j=1, j \in G_\tau}^{i} \frac{\tau_j}{A_j} \Delta_i + \sqrt{\frac{TA\iota}{m}} + 44A\iota \tag{5}$$

Proof. Under the good event,

$$\begin{aligned} b_{t_{i+1}}(a_i) &\leq 2 \sum_{t=1}^{t_{i+1}-1} p_k^v(a_i) + 12\iota \\ &= 2 \sum_{j=1}^{i} \sum_{t=t_j}^{t_j+\tau_j-1} p_k^v(a_i) + 12\iota \\ &= 2 \sum_{j=1}^{i} \frac{\tau_j}{A_j} + 12\iota \end{aligned}$$

Now the regret can be bounded by,

$$\begin{aligned} \mathcal{R}_T &= \sum_{i \in [A]} b_{t_{i+1}}(a_i) \Delta_i \\ &\leq 2 \sum_{i \in [A]} \sum_{j=1}^{i} \frac{\tau_j}{A_j} \Delta_i + 12A\iota \\ &\leq 2 \sum_{i \in A_\Delta} \sum_{j=1}^{i} \frac{\tau_j}{A_j} \Delta_i + \sum_{i \notin A_\Delta} b_{t_{i+1}}(a_i) \sqrt{\frac{A\iota}{Tm}} + 12A\iota \\ &\leq 2 \sum_{i \in A_\Delta} \sum_{j=1}^{i} \frac{\tau_j}{A_j} \Delta_i + T\sqrt{\frac{A\iota}{Tm}} + 12A\iota \\ &\leq 2 \sum_{i \in A_\Delta} \sum_{j=1, j \in G_\tau}^{i} \frac{\tau_j}{A_j} \Delta_i + \sum_{i \in A_\Delta} \sum_{j=1, j \notin G_\tau}^{i} \frac{\tau_j}{A_j} \Delta_i + \sqrt{\frac{TA\iota}{m}} + 12A\iota \\ &\leq 2 \sum_{i \in A_\Delta} \sum_{j=1, j \in G_\tau}^{i} \frac{\tau_j}{A_j} \Delta_i + \sqrt{\frac{TA\iota}{m}} + 44A\iota, \end{aligned} \tag{6}$$

where the last is since,

$$\sum_{i \in A_\Delta} \sum_{j=1, j \notin G_\tau}^{i} \frac{\tau_j}{A_j} \Delta_i \leq \sum_{i \in A_\Delta} \sum_{j=1}^{i} \frac{16}{A_j} \leq A \sum_{j=1}^{A} \frac{16}{A_j} \leq 32A \log A.$$

as $\sum_{j=1}^{A} \frac{1}{A_j} \leq \log A + 1$ by Lemma 9. □

Lemma 11. *For every action elimination index $i \in A_\Delta$, it holds that*

$$\sum_{j=1, j \in S_\tau}^{i} \frac{\tau_j^2}{4A_j} + \sum_{j=1, j \in G_\tau \setminus S_\tau}^{i} \frac{\tau_j}{A_j} m \leq \frac{544\iota}{\Delta_i^2}$$

where $S_\tau := \{j | j \in G_\tau \ \ \& \ \ \tau_j/4 < m\}$, and $\{\tau_j | j \in [A]\}$ are the stage lengths.

Proof. From Lemma 7,

$$\sum_{j=1, j \in G_\tau}^{i} \frac{\tau_j}{A_j} |N_{\leq \tau_j/4}| \leq \frac{544\iota}{\Delta_{a_i}^2}.$$

On the other hand, using Lemma 8

$$\begin{aligned}\sum_{j=1,j\in G_\tau}^{i}\frac{\tau_j}{A_j}|N_{\le\tau_j/4}| &\ge \sum_{j=1,j\in G_\tau}^{i}\frac{\tau_j}{A_j}\min\{m,\tau_j/4\}\\ &= \sum_{j=1,j\in S_\tau}^{i}\frac{\tau_j^2}{4A_j}+\sum_{j=1,j\in G_\tau\setminus S_\tau}^{i}\frac{\tau_j}{A_j}m\end{aligned}$$

□

Lemma 12. $\sum_{i=1}^{A}\sqrt{\sum_{j=1}^{i}\frac{1}{A_j}}\le A$

Proof. Using Cauchy–Schwarz inequality

$$\begin{aligned}\sum_{i=1}^{A}\sqrt{\sum_{j=1}^{i}\frac{1}{A_j}} &\le \sqrt{A}\sqrt{\sum_{i=1}^{A}\sum_{j=1}^{i}\frac{1}{A_j}}\\ &= \sqrt{A}\sqrt{\sum_{j=1}^{A}\sum_{i=j}^{A}\frac{1}{A_j}}\\ &= \sqrt{A}\sqrt{\sum_{j=1}^{A}\frac{A-j+1}{A_j}}\\ &= \sqrt{A}\sqrt{\sum_{j=1}^{A}1}\\ &= A.\end{aligned}$$

□

Proof of Theorem 1. Let us write again the Right-Hand-Side of Equation (5)

$$2\sum_{i\in A_\Delta}\sum_{j=1,j\in G_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i+\sqrt{\frac{TA\iota}{m}}+44A\iota.$$

Note that the bound on the regret that is depicted in Equation (5) assumes that the good event holds, and we will remove this assumption later. Let's assume that the good event hold. We'll break the first sum in the Right-Hand-Side of Equation (5) as

$$\sum_{i\in A_\Delta}\sum_{j=1,j\in S_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i+\sum_{i\in A_\Delta}\sum_{j=1,j\in G_\tau\setminus S_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i. \tag{7}$$

For the first term above, using Lemma 11, for every $i\in A_\Delta$,

$$\begin{aligned}\sum_{j=1,j\in G_\tau\setminus S_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i &= \frac{\Delta_i}{m}\sum_{j=1,\tau_j\in G_\tau\setminus S_\tau}^{i}\frac{\tau_j}{A_j}m\\ &\le \frac{544\iota}{m\Delta_i}\\ &\le 544\sqrt{\frac{T\iota}{mA}}.\end{aligned} \tag{8}$$

where the second inequality is since $i\in A_\Delta$. Summing over all elimination indices in A_Δ we get the the first term in Equation (7) is bounded by $544\sqrt{\frac{TA\iota}{m}}$.

For the second term, for every i, using Cauchy–Schwarz inequality

$$\begin{aligned}\sum_{j=1,j\in S_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i &\le \Delta_i\sqrt{\sum_{j=1,j\in S_\tau}^{i}\frac{\tau_j^2}{A_j}}\sqrt{\sum_{j=1,j\in S_\tau}^{i}\frac{1}{A_j}}\\ &\le \Delta_i\sqrt{\frac{2176\iota}{\Delta_i^2}}\sqrt{\sum_{j=1}^{i}\frac{1}{A_j}}\\ &\le \sqrt{2176\iota}\sqrt{\sum_{j=1}^{i}\frac{1}{A_j}},\end{aligned}$$

where the second inequality is from Lemma 11. Using Lemma 12, summing over all actions we get the second term in Equation (7) is bounded by $47A\sqrt{\iota}$. Combining this with the other terms in Equation (5) yields the part of the bound corresponding to the good event. From Lemma 5, the complementary event of the good events adds no more than 1 to the regret. We get,

$$\begin{aligned}\mathcal{R}_T &\le 2\cdot 544\sqrt{\frac{TA\iota}{m}} + 2\cdot 47A\sqrt{\iota} + \sqrt{\frac{TA\iota}{m}} + 44A\iota + 1\\ &\le 1088\sqrt{\frac{TA\iota}{m}} + 94A\iota + \sqrt{\frac{TA\iota}{m}} + 44A\iota + 1\\ &= 1089\sqrt{\frac{TA\iota}{m}} + 138A\iota + 1\\ &= 1089\sqrt{\frac{TA\log(3mTA)}{m}} + 138A\log(3mTA) + 1.\end{aligned}$$

□

B.4 Instance dependant bound

It is important to note that when the analysis is not split into large and small gaps, a bound specific to the problem instance can also be derived. We can conclude that the individual regret is bounded by,

$$\tilde{O}(\sum_{a:\Delta_a>0}\frac{1}{m\Delta_a})$$

as depicted in Theorem 2.

Despite being a suitable bound for various scenarios, there are cases where it fails to provide a good approximation. For example, two action and the gap is $\Delta_a = 1/T\cdot m$. We will get regret which is linear in T. We have made this distinction between large and short gaps to be problem independent.

Although the changes that yield the instance dependent bound are simple, we provide for clarity the relevant parts where the proof changes.

Lemma 13. *Under the good event, the regret of agent v is bounded by*

$$\mathcal{R}_T \le 2\sum_{i\in[A]}\sum_{j=1,j\in G_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i + 44A\iota. \tag{9}$$

Proof. The proof follows the same steps as Lemma 10, but without splitting the gaps as in Equation (6). □

Lemma 14. *Under the good event, the following holds,*

$$\sum_{i\in[A],\Delta_i>0}\sum_{j=1,j\in G_\tau}^{i}\frac{\tau_j}{A_j}\Delta_i \le \sum_{i\in[A],\Delta_i>0}\frac{544\iota}{m\Delta_i} + 47A\sqrt{\iota}.$$

Proof. The proof follows the same steps as the proof of Theorem 1, but treating all non optimal actions the same, and stopping the analysis in Equation (8), i.e., without bounding the expression with $\sqrt{T\iota/mA}$. □

Proof of Theorem 2. The proof follows by combining the results of Lemma 13 and Lemma 14, and with the fact from Lemma 5 that the complementary event of the good events adds no more than 1 to the regret. We get,

$$\begin{aligned}
\mathcal{R}_T &\leq 2 \sum_{i\in[A]} \sum_{j=1, j\in G_\tau}^{i} \frac{\tau_j}{A_j}\Delta_i + 44A\iota + 1 \\
&\leq 2 \cdot \sum_{i\in[A],\Delta_i>0} \frac{544\iota}{m\Delta_i} + 2 \cdot 47A\sqrt{\iota} + 44A\iota + 1 \\
&\leq \sum_{i\in[A],\Delta_i>0} \frac{1088\iota}{m\Delta_i} + 94A\iota + 44A\iota + 1 \\
&= \sum_{i\in[A],\Delta_i>0} \frac{1088\iota}{m\Delta_i} + 138A\iota + 1 \\
&= \left(1088 \sum_{i\in[A],\Delta_i>0} \frac{\log(3mTA)}{m\Delta_i}\right) + 138A\log(3mTA) + 1.
\end{aligned}$$

□

C Lower bound

Theorem 6. *For any algorithm, there exists an instance of the cooperative MAB over a communication graph problem, for which the individual regret of any agent is bounded from below by*

$$\Omega(\sqrt{A}) \leq \mathcal{R}_T.$$

Proof. Let the graph be a line of length T. Let A be the number of actions such that $\sqrt{A} > 20$. Let $a^\star$ be the only best action. Let $\Delta_a = 1$ for every $a \neq a^\star$. Namely the reward of a^* is 1 and the rewards of the other actions $a \neq a^*$ is 0.

Let v be an agent in the graph. After t timesteps, the maximum number of samples v sees, for all actions together, is no more than $2 \cdot (2 + t + 1)t/2 = (t + 3)t$ (twice the sum of arithmetic series). At timestep $\lfloor\sqrt{A}/20\rfloor$ the agent sees at most $\frac{A+30\sqrt{A}}{400}$ samples for all the actions together.

From the assumption on A, $\frac{3\sqrt{A}}{40} \leq \frac{A}{200}$. It implies that

$$\frac{A + 30\sqrt{A}}{400} \leq \frac{A}{200} + \frac{3\sqrt{A}}{20} \leq \frac{A}{100}.$$

It means that until this timestep, the agent didn't see at least $0.99A$ of the actions.

Let us randomly choose an instantiation of the best action a^*. Define the random variable X that chooses the best action uniformly. I.e., $\mathbb{P}(X = a) = \frac{1}{A}$. Denote the event in which the agent doesn't see the best action until timestep $\lfloor\frac{\sqrt{A}}{20}\rfloor$ with $\mathcal{E}$. From the above, event $\mathcal{E}$ happens with probability at least $\frac{99}{100}$. I.e., $\mathbb{P}(\mathcal{E}) \geq \frac{99}{100}$. Under event $\mathcal{E}$, from the assumption that $\Delta_a = 1$, the regret until this timestep is $\lfloor\frac{\sqrt{A}}{20}\rfloor$, and we get $\frac{\sqrt{A}}{20} - 1 \leq \mathcal{R}_T$.

For any algorithm the agents play,

$$\mathbb{E}_X(\mathcal{R}_T) \geq \frac{99}{100} \cdot (\frac{\sqrt{A}}{20} - 1).$$

Therefore, for any algorithm, there exists an instance such that $\mathcal{R}_T \geq \frac{99}{100} \cdot (\frac{\sqrt{A}}{20} - 1)$. □

D Bounded communication

This section relies on the definitions and theorem that are depicted in Appendix B.

We introduce a new event type, the aggregated event for many rewards.

Definition 17. *A reward-many event is a tuple* $(\texttt{rwdMany}, t, v, a, r, n)$ *that represent an aggregation of many rewards, where* v *is the agent's ID,* t *is the time,* a *is the action,* r *is the reward, and* n *is the number of samples of this event.*

Remark 3. *The good event occurs with probability higher than or equals to* $1 - 1/T^2$*, when all agent play Algorithm 4 or when all agent play Algorithm 5. Although these algorithms uses the* `rwdMany` *events, the same proof of Lemma 4 applies also to them.*

Proof of Theorem 4. In Algorithm 4, we do not have duplicated messages. We achieve this by the tree structure, and by not sending to a neighbor u information that u already sent to v. The tree structure promises that there is only one path from an agent to another. This property ensures that a message originating from one agent will reach all other agents exactly once, as it traverses the tree along the single possible route. Consequently, the combination of the spanning tree structure and the selective forwarding of messages allows for efficient and duplicate-free communication among all agents.

The `Coop-SE-Restricted` algorithm aggregate all events regarding an action a into two events: `rwdMany` for rewards and `elim` for elimination. The message contains information about action a, its elimination status, observation count, and sum of observed rewards, requiring $O(A \log(ATm))$ bits. This is all the information agents need from multiple messages.

Therefore, the agent has exactly the same information if all agents had played Algorithm 2 on that spanning tree. The individual regret bound that is induced from `Coop-SE` does not depend on the structure of the graph, therefore the same regret bound applied for `Coop-SE-Restricted` as well.

The agent sends to each neighbor $2A$ events. Each event has $O(\log(TAm))$ bits. Therefore each message is bounded by $O(A \log(TAm))$ bits. This completes the proof. □

Proof of Theorem 5. In Algorithm 5, on every timestep the agent v sends and receives only 2 events per action, for every neighbor. Every event is of size $O(\log(mTA))$. This means that the communication in Algorithm 5 is bounded by $O(\log(TAm))$.

`Coop-SE-Low-Comm` sends in each block (i.e., A timesteps) the exact same information that `Coop-SE-Restricted` sends after 1 timestep. So `Coop-SE-Low-Comm` simulates `Coop-SE-Restricted`, but it takes it A timesteps to simulate 1 timestep of `Coop-SE-Restricted`. For simulating faithfully `Coop-SE-Restricted`, `Coop-SE-Low-Comm` ignores the information of the timesteps when $t \mod A \neq 1$. Note that `Coop-SE-Low-Comm` is designed in a way that when all agents play `Coop-SE-Low-Comm`, every A timesteps the state of the algorithm is exactly the same as if all agents had played `Coop-SE-Restricted`, and only one timestep had passed. As a result, the regret incurred in each of the $\lfloor T/A \rfloor$ blocks is A times the regret incurred in the corresponding timestep for `Coop-SE-Restricted`. If T is not divisible by A we might have a remainder of at most $A-1$ timesteps, for each the regret is at most 1. That is, if $\tilde{\mathcal{R}}_T$ is the regret of `Coop-SE-Restricted` over T rounds, then the regret of `Coop-SE-Low-Comm` is at most $A\tilde{\mathcal{R}}_{\lfloor T/A \rfloor} + A$. We get

$$\begin{aligned}
\mathcal{R}_T &\leq A\left(1089\sqrt{\frac{AT\log(3mTA)}{Am}} + 138A\log(3mTA) + 1\right) + A \\
&= 1089A\sqrt{\frac{T\log(3mTA)}{m}} + 138A^2\log(3mTA) + 2A \\
&\leq 1089A\sqrt{\frac{T\log(3mTA)}{m}} + 140A^2\log(3mTA).
\end{aligned}$$

□

Algorithm 4 Cooperative Successive Elimination with Restricted Communication (`Coop-SE-Restricted`)

Input: number of rounds T, neighbor agents N, number of actions A, id of current agent v, confidence bound parameter L.
Initialization: $t \leftarrow 1$; Set of *active* actions $\mathcal{A} = \mathbb{A}$; $R_t(a) = 0, n_t(a) = 0$ for every action a; $M_{\texttt{in}} = \emptyset$; $M_{\texttt{updates}} = \emptyset$; $M_{\texttt{sent}} = \emptyset$;
Coordinate a spanning tree, $\mathcal{T}$, out of the communication tree $\mathcal{G}$.
Set $N' \subseteq N$ to be the agent's neighbors in $\mathcal{T}$.
for $t = 1, ..., T$ **do**
    for $u \in N'$ **do**
        for $a \in \mathcal{A}$ **do**
            $n_a^u = 0$, $r_a^u = 0$
        end for
    end for
    $E_{\texttt{received}} = \emptyset$
    for $event \in M_{\texttt{updates}}$ **do**
        if $event$ is `elim`-event **then**
            $\mathcal{A} = \mathcal{A} \setminus event_a$
            $E_{\texttt{received}} = E_{\texttt{received}} \cup event$
        else if $event_a \in \mathcal{A}$ **then** // $event = (\texttt{rwdMany}, t, id, a, r, n)$.
            $n_t(a) = n_t(a) + event_n$, $R_t(a) = R_t(a) + event_r$
            for $u \in N'$ **do**
                if $event_{id} \neq u$ **then**
                    $n_a^u = n_a^u + event_n$; $r_a^u = r_a^u + event_r$
                end if
            end for
        end if
    end for
    $M_{\texttt{updates}} = \emptyset$

    $E = \texttt{ElimStep}(\mathcal{A}, n_t, \hat{\mu}_t, L)$
    $\mathcal{A} = \mathcal{A} \setminus E$
    Choose action a_t uniformly from $\mathcal{A}$, and get reward $r_t(a_t)$
    $n_t(a_t) = n_t(a_t) + 1$, $R_t(a_t) = R_t(a_t) + r_t(a_t)$
    for $u \in N'$ **do**
        $n_{a_t}^u = n_{a_t}^u + 1$, $r_{a_t}^u = r_{a_t}^u + r_t(a_t)$
    end for

    for $u \in N'$ **do**
        $M_{\texttt{elim}}(u) = \{(\texttt{elim}, t, v, a) | \exists a \in E\} \cup \{(\texttt{elim}, t, v, event_a) | \exists event \in E_{\texttt{received}}, event_{id} \neq u\}$
        $M_{\texttt{rwd}}(u) = \{(\texttt{rwdMany}, t, v, a, r_a^u, n_a^u) | a \in \mathcal{A}\}$
        $M_t^v(u) = M_{\texttt{elim}}(u) \cup M_{\texttt{rwd}}(u)$
        Send $M_t^v(u)$ and receive $M_t^u(v)$
        $M_{\texttt{updates}} = M_{\texttt{updates}} \cup M_t^u(v)$
    end for
end for

Algorithm 5 Cooperative Successive Elimination with Low Communication (`Coop-SE-Low-Comm`)

Input: number of rounds T, neighbor agents N, number of actions A, id of current agent v, confidence bound parameter L.
Initialization: $t \leftarrow 1$; Set of *active* actions $\mathcal{A} = \mathbb{A}$; $R_t(a) = 0, n_t(a) = 0$ for every action a; $M_{\texttt{in}} = \emptyset$; $M_{\texttt{updates}} = \emptyset$; $M_{\texttt{sent}} = \emptyset$;
Coordinate a spanning tree, $\mathcal{T}$, out of the communication tree $\mathcal{G}$.
Set $N' \subseteq N$ to be the agent's neighbors in $\mathcal{T}$.
for $t = 1, ..., T$ **do**
  if $t \mod A = 1$ **then**
    for $u \in N'$ **do**
      for $a \in \mathcal{A}$ **do**
        $n_a^u = 0, r_a^u = 0$
      end for
    end for
    $E_{\texttt{received}} = \emptyset$
    for $event \in M_{\texttt{updates}}$ **do**
      if $event$ is `elim`-event **then**
        $\mathcal{A} = \mathcal{A} \setminus event_e$
        $E_{\texttt{received}} = E_{\texttt{received}} \cup event$
      else if $event_a \in \mathcal{A}$ **then**
        $n_t(a) = n_t(a) + event_n$, $R_t(a) = R_t(a) + event_r$
        for $u \in N'$ **do**
          if $event_{id} \neq u$ **then**
            $n_a^u = n_a^u + event_n$, $r_a^u = r_a^u + event_r$
          end if
        end for
      end if
    end for
    $M_{\texttt{updates}} = \emptyset$

    $E = \texttt{ElimStep}(\mathcal{A}, n_t, \hat{\mu}_t, L)$
    $\mathcal{A} = \mathcal{A} \setminus E$
    Choose action a_t uniformly from $\mathcal{A}$, and get reward $r_t(a_t)$
    $n_t(a_t) = n_t(a_t) + 1$, $R_t(a_t) = R_t(a_t) + r_t(a_t)$
    Set $a_{fix} = a_t$
    for $u \in N'$ **do**
      $n_{a_t}^u = n_{a_t}^u + 1$, $r_{a_t}^u = r_{a_t}^u + r_t(a_t)$
    end for
  else
    Choose action a_{fix} , and get reward $r_t(a_{fix})$
  end if
  $a' = (t \mod A) + 1$ // The actions are $[A]$. The agent sends each round all the information regarding one action.
  for $u \in N'$ **do**
    $M_{\texttt{elim}}(u)[a'] = \{(\texttt{elim}, t, v, a') | \exists event \in E_{\texttt{received}}, event_{id} \neq u, event_a = a'\}$
    if $a' \in E$ **then**
      $M_{\texttt{elim}}(u)[a'] = \{(\texttt{elim}, t, v, a')\}$
    end if
    $M_{\texttt{rwd}}(u)[a'] = \{(\texttt{rwdMany}, t, v, a', r_{a'}^u, n_{a'}^u)\}$
    $M_t^v(u)[a'] = M_{\texttt{elim}}(u)[a'] \cup M_{\texttt{rwd}}(u)[a']$
    Send $M_t^v(u)[a']$ and receive $M_t^u(v)[a']$
    $M_{\texttt{updates}} = M_{\texttt{updates}} \cup M_t^u(v)[a']$
  end for
end for

E Auxiliary lemmas

Lemma 15 (Lemma F.4 in Dann et al. [2017])**.** *Let $\{X_t\}_{t=1}^T$ be a sequence of Bernoulli random and a filtration $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq ...\mathcal{F}_T$ with $\mathbb{P}(X_t = 1 \mid \mathcal{F}_t) = P_t$, P_t is $\mathcal{F}_t$-measurable and X_t is $\mathcal{F}_{t+1}$-measurable. Then, for all $t \in [T]$ simultaneously, with probability $1 - \delta$,*

$$\sum_{k=1}^{t} X_k \geq \frac{1}{2} \sum_{k=1}^{t} P_k - \log \frac{1}{\delta}.$$

Lemma 16 (Consequence of Freedman's Inequality, e.g., Lemma E.2 in Cohen et al. [2021])**.** *Let $\{X_t\}_{t \geq 1}$ be a sequence of random variables, supported in $[0, R]$, and adapted to a filtration $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq ...\mathcal{F}_T$. For any T, with probability $1 - \delta$,*

$$\sum_{t=1}^{T} X_t \leq 2\mathbb{E}[X_t \mid \mathcal{F}_t] + 4R \log \frac{1}{\delta}.$$

EWRL Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction refers to Theorems 1 to 5 which we rigorously prove either in the main text or the appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We stated the setting formally in Section 2, and the proofs assume this setting exactly. When we extend the setting for the low-communication setting, we stated the restrictions in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are provided in Section 2 and most of the proofs reside in the appendix, while one lemma's proof appears in the main text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While EWRL does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the EWRL code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the EWRL code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the EWRL Code of Ethics `https://ewrl.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the EWRL Code of Ethics, and the research conducted conforms in every aspect with this code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the EWRL Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper presents a theoretical work, and we do not see any direct societal impact of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not include data or models that pose such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use any existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the EWRL Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the EWRL Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.