Abstract

Multiplayer bandits have recently garnered significant attention due to their relevance in cognitive radio networks. While the existing body of literature predominantly focuses on synchronous players, real-world radio networks, such as those in IoT applications, often feature asynchronous (i.e., randomly activated) devices. This highlights the need for addressing the more challenging asynchronous multiplayer bandits problem. Our first result shows that a natural extension of UCB achieves a minimax regret of $O(\sqrt{T \log(T)})$ in the centralized setting. More significantly, we introduce Cautious Greedy, which uses $O(\log(T))$ communications and whose instance-dependent regret is constant if the optimal policy assigns at least one player to each arm (a situation proven to occur when arm means are sufficiently close). Otherwise, the regret is, as usual, $\log(T)$ times the sum of some inverse sub-optimality gaps. We substantiate the optimality of Cautious Greedy through lower-bound analysis based on data-dependent terms. Therefore, we establish a strong baseline for asynchronous multiplayer bandits, at least with $O(\log(T))$ communications.

1 INTRODUCTION

In the classical multi-armed bandits (MAB) problem, a single player sequentially pulls arms $k_t \in \{1, \ldots, K\} \subseteq [K]$, and receives a reward $X_{k_t}$ sampled from some unknown sub-Gaussian distribution of mean $\mu_k$. This process undergoes repetition for a total of $T$ rounds and the performance of the sampling policy is measured by its regret, the difference between the total expected reward obtained by choosing the best arm $k^*$ at each round and the total expected reward of the player’s actual choices. This setting has been extensively studied (see Lattimore and Szepesvári 2020 for a recent survey). A fundamental component of MAB is the exploration and exploitation trade-off. Exploration involves trying out different arms to gather information, while exploitation uses the acquired knowledge to favor arms more likely to be the best. It is well known that optimal policies incur a regret scaling as $O(\sum_{k \neq k^*} \frac{\log(T)}{\mu_{k^*} - \mu_k})$ (Auer et al., 2002).

Classical applications of MAB include clinical trials, recommendation systems, or ad placements. For many other types of applications, the MAB framework however does not fit the problem at hand. Consider for instance cognitive radios [Lai et al., 2008; Anandkumar et al., 2011; Mitola and Maguire, 1999; Jouini et al., 2010] where arms correspond to communication channels available to radio devices. What differs from standard MAB is that if two radios choose the same communication channel, they interfere. This example motivates the multiplayer multi-armed bandits (MMAB) setting introduced by Liu and Zhao (2010). In MMAB, $M$ players simultaneously pull arms. When a player pulls arm $k$, it receives the reward $\eta_k X_k$ where $\eta_k = 0$ if two or more players collide, meaning they pull the same arm $k$, and $\eta_k = 1$ if a single player pulls $k$. In the centralized setting and $M < K$, MMAB is equivalent to bandits with multiple plays (Komiyama et al., 2015; Anantharam et al., 1987; Chen et al., 2013a; Gopalan et al., 2013), as a central entity decides on the behalf of agents and trivially avoids collisions. Optimal algorithms are then known to yield an asymptotic regret $\sum_{k=1}^{K-M} \frac{\log(T)}{\mu(k-M+1) - \mu(k)}$ (Komiyama et al., 2015), where $\mu(k)$ is the $k$-th smallest mean reward.

Motivated by Internet of Things networks, we focus...
on the asynchronous multiplayer multi-armed setting (AMMAB) where each round is decomposed into three successive steps (see Dakdouk [2022] Bonnefoi et al. [2017]). First, all players decide which arm they would like to play. Second, the environment activates independently each player \(i\) with probability \(p_i\) (and these activations cannot be foreseen before the first step). In the third and last step, activated players pull the arm they chose in the first step. In this model, players correspond to communicating devices, arms to available channels and \(p_i\) is the activation probability of the communicating device \(i\).

**Notations.** Vectors are denoted in bold. If \(u \in \mathbb{R}^n\), \(u_i\) is the \(i\)-th coordinate of \(u\) while \(u_{(i)}\) the \(i\)-th smallest coordinate of \(u\) and \(\text{support}(u) = \{i \in [n], u_i \neq 0\}\). We denote for \(u, v \in \mathbb{R}^n\), \(\langle u, v \rangle = \sum_{i=1}^n u_i v_i\), \(\|u\|_\infty = \max_{i \in [n]} |u_i|\), \(\|u\|_1 = \sum_{i \in [n]} |u_i|\). For a function \(f : \mathbb{R} \rightarrow \mathbb{R}\) and \(u \in \mathbb{R}^n\), \(f(u) \in \mathbb{R}^n\) is defined by \(f(u_i) = f(u_i)\). Lastly, \(E\) denotes the complementary event of \(E\).

**Setting and assumptions.** For simplicity, we follow Bonnefoi et al. (2017) and assume that the probability of being active is the same for all players: \(p_i = p\), for all \(i \in [M]\). This makes players exchangeable and allows for a simplified description of the AMMAB setting. At each round \(t\), the player choices are summarized by the assignment vector \(M(t) = (M_1(t), \ldots, M_K(t))\) where \(M_k(t)\) is the number of players choosing the arm \(k\) at round \(t\); the environment then activates each player with probability\(^1\) \(p\) and active players pull the arm they chose in the first phase, each receiving reward \(\eta_k(M_k(t))X_k\) with \(k\) the pulled arm, \(\eta_k = 1\) (exactly one player is active on arm \(k\)) and \(X_k\) is sampled from an unknown sub-Gaussian distribution with mean \(\mu_k\). The arm pulled by player \(m\) at time \(t\) is denoted \(X_m(t) \in [K]\). A player playing arm \(k\) observes \(X_k|\eta_k\) and the collision event \(\eta_k\). Additionally, the parameters \(M\), \(K\) and \(p\) are assumed to be known beforehand.

At any time \(t\), the assignment \(M(t)\) satisfies the budget constraint \(\sum_{k=1}^K M_k(t) = M\) and we also assume:

**Assumption 1.1.** \(M \geq K\) and for all \(k \in [K]\) and at all stages \(M_k(t) \leq \frac{-1}{\log(1-p)} \approx \frac{1}{p}\).

The second condition is not restrictive (and made for the sake of notations and clarity), as assigning more than \(\frac{-1}{\log(1-p)}\) players on the same arm only decreases the obtained reward and amount of information on that arm. A better policy would then have some players not play at all instead, or equivalently assign players to a dummy arm whose reward is known to be 0. The set of valid assignments is thus denoted by

\[
\mathcal{M} = \left\{ M \in [M]^K \mid \sum_{k=1}^K M_k = M, M_k \leq \frac{-1}{\log(1-p)} \right\}.
\]

The goal is to minimize the expected regret defined by:

\[
\mathbb{E}[R] = \sum_{t=1}^T \sum_{k=1}^K \mathbb{E}[\eta_k(M_k(t))X_k] - \mathbb{E}[\eta_k(M_*(t))X_k] \quad (1)
\]

where \(M^* = (M^*_1, \ldots, M^*_K)\) is an optimal assignment:

\[
M^* \in \arg\max_{M \in \mathcal{M}} \sum_{k=1}^K \eta_k(M_k)X_k. \quad (2)
\]

Bonnefoi et al. (2017) designed an algorithm solving Equation (2) with known \(\mu_k\), and Dakdouk (2022) later proposed a simpler sequential algorithm. In combination with some non-adaptive explore-then-commit policy, it yields a regret scaling in \(O(T^{3/4})\). Additionally, Dakdouk (2022) shows there is no random assignment yielding a strictly larger expected reward than the deterministic optimal assignment \(M^*\).

**Limited communication setting.** We consider the same communication protocol as in Dakdouk (2022). At each round, a player \(a\) can decide to send a vector in \(\mathbb{R}^K\) to the (central) gateway. It is then successfully received by the gateway with probability \(p_g\), if only player \(a\) sent a message to the gateway (collision otherwise). Reciprocally at each time step (if it is not receiving any message), the central gateway can send a vector to a single player, who successfully receives it with probability \(p_g\). More precisely, agents communicate via a function \(\text{SEND}\). \(\text{SEND}_{a \rightarrow b}(x)\) attempts to send vector \(x\) from agent \(a\) (for instance a player) to \(b\) (for instance the gateway). The attempt is said to be successful when \(b\) has received the message from \(a\) and \(a\) is aware of it (i.e., \(a\) has received the acknowledgment sent by \(b\)). An attempt is successful with a known probability \(p_g\), and after a successful attempt, \(b\) can access \(x\) (and calling again \(\text{SEND}_{a \rightarrow b}(x)\) afterwards does nothing). It follows that if \(\text{SEND}_{a \rightarrow b}(x)\) is executed \(t \log(\delta)/\log(1-p_g)\) time, then \(b\) can access \(x\) with probability \(\delta\). If \(b\) tries to use \(x\) without a successful transmission, we assume it uses a value chosen uniformly at random in the range of \(x\) instead. We assume agents can call \(\text{SEND}\) even when they are inactive and at no cost. However, if two agents call \(\text{SEND}\) simultaneously, nothing gets transmitted.

We keep track of the number \(C_A\) of calls of \(\text{SEND}\) when using algorithm \(A\) (ignoring the calls doing nothing).

\(^1\)Players cannot know beforehand who will be active, making collisions unavoidable.

\(^2\)The situation where \(a\) is the gateway and \(b\) a player will also be considered.
Contributions. In the centralized setting, we prove that an adapted version of UCB exhibits a regret in $O(\sqrt{T KM \log(T) \min(M, p, K)})$ where $O(\cdot)$ hides universal constant factors. More surprisingly, our main contribution shows that, even in the limited communication setting, achieving a constant regret (in $T$) is sometimes possible with an algorithm called Cautious Greedy using only $O(\log(T))$ communications. The analysis of UCB is thus postponed to Appendix C and the main text solely focuses on Cautious Greedy that somehow achieves the best of both worlds (very small regret with low communication). In essence, it is a standard greedy algorithm that estimates $\mu_k$ via empirical means, but it is cautious as it avoids assigning zero players to an arm unless, with high confidence, assigning no players to it is optimal. More precisely, Cautious Greedy maintains a lower bound $\nu$ of the number of arms that should be assigned zero players and stops assigning players to the $\nu$ worst arms when confident enough.

The regret of Cautious Greedy depends on several data-dependent quantities defined in Section 3.2:
- $\nu^*$ the number of arms that are assigned zero players in the optimal assignment;
- $\Delta_{(j)} = \mu_{j^*} - \mu_{(j)}$;
- $M_\nu^*$ the optimal assignment when $\nu$ arms are assigned zero players;
- $r$ the infinity norm of the minimal perturbation of the arm means $\mu$ that would modify the sequence $(M_\nu^*)_{\nu=1}^{\nu^*}$.

Proposition 3.1 together with Lemma 3.2 show that the regret of Cautious Greedy is upper bounded by $O\left(\frac{1}{\tau} + \sum_{j \leq \nu^*} \frac{\log(T)}{T}\right)$, where $O$ hides terms depending on $K, p, M$ and $p_g$.

In particular, Cautious Greedy achieves constant regret if $\nu^* = 0$, i.e., when each arm is assigned at least one player by the optimal policy. As shown by the lower bound in Lemma 3.1 under mild conditions, the dependency in $\frac{1}{\tau}$ cannot be improved. In Lemma 3.1 we give a sufficient condition on the dispersion of arm means to get $\nu^* = 0$. In general, Cautious Greedy suffers an additional dependency in $\sum_{j \leq \nu^*} \frac{\log(T)}{T}$. This dependency also appears in bandits with multiple plays (Komiyama et al., 2015) and as shown by Lemma 2.2 cannot be removed. This makes Cautious Greedy optimal with respect to $T$ and with respect to the data-dependent quantities $r$ and $(\Delta_{(j)})_{j \leq \nu^*}$.

The main difficulty of the problem comes from the fact that $\nu^*$ is unknown. A classical Greedy algorithm yields a linear regret when $\nu^* > 0$, while a traditional bandits algorithm may not reach constant regret when $\nu^* = 0$. On the other hand, Cautious Greedy performs optimally in both cases.

Section 5 benchmarks Cautious Greedy against our UCB algorithm and the ETC algorithm of Dakdouk (2022) on synthetic data and shows that Cautious Greedy and UCB perform both significantly better than ETC. Cautious Greedy outperforms UCB when no arms should be assigned zero players while UCB tends to be better when at least one arm should be assigned zero players.

2 RELATED WORK

Centralized setting: multiplay, combinatorial and structured bandits. When $M \leq K$, $p = 1$ and unlimited communication is allowed, AMMAB is equivalent to bandits with multiple plays. A lower bound in $\sum_{j=1}^{\nu^*} \frac{\log(T)}{T}$, where $\nu^* = K - M$ is shown in Anantharam et al. (1987), who also provide an optimal algorithm reaching this bound. Bandits with multiple plays are an instance of combinatorial bandits (Gal et al., 2012; Chen et al., 2013b; Kveton et al., 2015; Combes et al., 2015; Wang and Chen, 2018; Perrault et al., 2020) where an agent chooses an action $a \in S$ and receives reward $r(\mu, a)$. With unlimited communication, when $M \leq K$, AMMAB is an instance of combinatorial bandits with semi-bandit feedback and probabilistically triggered arms (chosen arms are triggered with some probability) (Wang and Chen, 2017; Chen et al., 2016). More generally, it can be viewed as combinatorial bandits or structured bandits (Combes et al., 2017) with semi-bandit feedback and $KM$ possible actions. None of these works yet allow to reach constant regret when $\nu^* = 0$ in the centralized setting, let alone the limited communication setting.

Decentralized multiplayer bandits. In decentralized multiplayer bandits, players aim to speed up the collective learning of the arm rewards, while avoiding collisions. Motivated by cognitive radio networks, the decentralized problem of multiplayer bandits recently received a lot of attention (we refer to Boursier and Perchet, 2022 for a review), sometimes assuming a pre-agreement on the ranks of the players (Anandkumar et al., 2010; Liu and Zhao, 2010) or using few collisions to communicate information between players (Avner and Mannor, 2014; Rosenski et al., 2016; Besson and Kaufmann, 2018a). However, Bistritz and Leskem (2018); Boursier and Perchet (2019); Wang et al. (2020) enforce collisions to send a significant number of bits between the players, allowing to reach optimal centralized performance. This idea is also used in many extensions of MMAB (Mehraban et al., 2020; Shi et al., 2020; Huang et al., 2021; Boursier and Perchet, 2020; Shi et al., 2021). This communication through collision tricks yet highly depends on the synchronicity of the players and becomes costly with a lot of players. In AMMAB, the players are asynchronous ($p < 1$) and nu-
merous \((M \geq K)\), making both drawbacks significant. This work thus proposes an asynchronous algorithm with \(O(\log(T))\) communication, leaving open for future work a possible fully decentralized adaptation (see Section 6 for a discussion).

**Multi-agent multi-armed bandits** In the multi-agent bandit problem considered by Szorenyi et al. (2013); Landgren et al. (2016); Martinez-Rubio et al. (2019); Yang et al. (2021); Chen et al. (2023), no collision happens when several players pull the same arm. The problem is thus different in nature: the main objective of multi-agent bandits is to speed up learning using decentralized communication protocols (e.g. gossip), without consideration of collision.

**Full information.** When each arm is assigned at least one player, it provides information with a strictly positive probability at each time step. Therefore in this regime, the central entity is almost in full information feedback, where information about all arms is received at every round. Bandits with expert advice are examples of problems with full information feedback. The go-to algorithm in the adversarial setting is (variants of) exponential weights or Hedge (Mourtada and Gaïffas 2019). However, in the stochastic setting, a constant regret is achieved by Greedy (aka Follow The Leader) which plays according to the empirical mean estimate of the rewards (Deghene and Perchet 2016). Huang et al. (2017) shows Greedy achieves constant regret in a more structured setting.

**Resource allocation.** Our problem can also be recast as a particular instance of sequential resource allocation with concave utilities (Lattimore et al. 2015; Fontaine et al. 2020; Zuo and Joe-Wong 2021). Although general resource allocation algorithms could be used in our setting, much better solutions can be obtained by leveraging the very specific structure of the utilities. The utility functions are indeed exactly known here, up to the multiplicative factor \(\mu_k\).

**Asynchronous multiplayer bandits.** AMMAB was introduced by Bonnefoi et al. (2017) in the context of cognitive radios. In Dakdouk (2022), players have heterogeneous activation probabilities. They propose an explore and commit algorithm that reaches \(O(T^{2/3})\) regret with constant communications. In our work, we show that under favorable conditions, a constant regret can be reached with \(O(\log(T))\) communications. Quite interestingly in AMMAB, the expected individual reward decreases as more players are assigned to the same arm. This relates the AMMAB model to more advanced collision models for MMAB, where a collision only decreases the reward instead of yielding a 0 reward (Tekin and Liu 2012; Bande and Veeravalli 2019; Magesh and Veeravalli 2019; Boyarski et al. 2021). AMMAB is also related to the problem of online queuing systems (Gaitonde and Tardos 2020; Sentenac et al. 2021), where packets arrive in a queue (player) with random rates. This setting yet differs from AMMAB, as players are active as long as they hold packets.

### 3 CAUTIOUS GREEDY, AN EFFICIENT ALGORITHM FOR AMMAB

Let us first introduce the function \(g(x) = xp(1-p)^{x-1}\), so that the regret in Equation (1) rewrites as

\[
E[R] = \sum_{t=1}^T E[(\mu, g(M^*) - g(M(t)))]
\]

where \(M^* = \text{argmax}_{M \in M} (\mu, g(M))\) is a rewriting of Equation (2).

#### 3.1 Description

Cautious Greedy is based on a standard greedy strategy that plays the best policy according to the estimated mean rewards. A player \(m\) can compute its own estimate of the mean reward \(\hat{\mu}^{(m)}(t)\) with

\[
\hat{\mu}^{(m)}(t) = \frac{\sum_{\rho=1}^t \eta_k(M(\rho))X_k^t \mathbb{I}\{k_m = k\}}{T_k^{(m)}(t)}
\]

and where by convention, we set \(\hat{\mu}^{(m)}(t) = 1\) if \(T_k^{(m)}(t) = 0\). Assuming all players can share the empirical mean reward estimates, \(\hat{\mu}(t)\) is given by

\[
\hat{\mu}(t) = \frac{\sum_{m=1}^M T_k^{(m)}(t)\hat{\mu}^{(m)}(t)}{T_k(t)}
\]

where \(T_k(t) = \sum_{m=1}^M T_k^{(m)}(t)\) and using the convention \(\hat{\mu}_k(t) = 1\) if \(T_k(t) = 0\).

For communication purposes, Cautious Greedy is divided into epochs of doubling size and statistics such as \(\hat{\mu}\) are only updated at the end of each epoch. Communication happens at each round during the second half of an epoch, increasing the probability of having a successful transmission with the epoch size. This second half is again split in two parts: during the first one, players send their statistics to the gateway; during the second part, the gateway communicate to each individual player the averaged players’ statistics.
Given $\tilde{\mu}$, a Greedy algorithm would then choose the assignment $M(t) = M^t_M$ where

$$M^t_M = \arg\max_{M \in M} \langle \tilde{\mu}, g(M) \rangle$$

(7)

and $M^t_M(m)$ is the arm chosen by player $m$.

Such a simple strategy would quickly stop exploring, at the risk of committing to a suboptimal policy. In order to maintain some level of exploration, a natural idea is to impose at least one player per arm. However, in some settings, the optimal solution might assign no players to some arms. The challenging task of Cautious Greedy is then to identify which arms should be assigned zero players. We call such identified arms removed while active arms $K$ are those not removed yet. Cautious Greedy can put a set of arms $\mathcal{U}$ under pressure, meaning that these arms are temporally allowed to be assigned to no player. Arms that are assigned to at least one player are said to be played and note that it is possible that an arm under pressure is played. Formally, the constraints that apply to $M$ in the assignment problem will be described by sets of the form:

$$\mathcal{M}_S = \{ M \in M, \forall k \in S, M_k \geq 1 \}$$

where $S \subset [K]$. In order to identify the arms to remove, Cautious Greedy maintains confidence bounds on the mean of each arm. The upper and lower bounds are given respectively by

$$\tilde{\mu}^H(t) = \min(\tilde{\mu}(t) + \zeta(t), 1) \quad \text{and} \quad \tilde{\mu}^L(t) = \max(\tilde{\mu}(t) - \zeta(t), 0)$$

(8)

where for all $k \in [K]$, $\zeta_k(t) = \sqrt{\log(T/k^2) / 2t_k(t)}$. These bounds are used to eliminate sub-optimal arms. This could suggest a strategy that plays all active arms at each round until enough information is gathered to remove an arm. However, such a strategy yields high regret in the case where two arms that should be eliminated are very close to each other. Therefore, the elimination of several arms at once is allowed. This is done in Cautious Greedy by computing an estimate $\nu$ of the number of arms to remove, which is a lower bound of $\nu^* = |\{k, M_k = 0\}|$ and can be used to eliminate several arms at once without ordering them first. We therefore introduce $\mathcal{M}_\nu$, the set of assignments where $\nu$ arms are under pressure:

$$\mathcal{M}_\nu = \{ M \in M, |\text{support}(M)| \geq K - \nu \}.$$  

The number of arms to remove $\nu$ is then increased when $\langle \tilde{\mu}^L, g(M^t_M) \rangle > \langle \tilde{\mu}^H, g(M^t_M) \rangle$, i.e., when a larger reward is guaranteed by removing more than $\nu$ arms. Cautious Greedy then uses $\nu$ to build a set $A$ of accepted arms which are arms that are not likely to be among the $\nu$ worst arms. Cautious Greedy then puts under pressure a subset of arms $\mathcal{U}$ among the arms that are not accepted yet. The set of arms put under pressure rotates in a round-robin fashion. This mechanism ensures that all active arms are regularly played. After the round-robin rotation is completed, Cautious Greedy reevaluates $\nu$ and updates the sets of accepted arms and active arms. As $\nu$ increases, an arm can be removed from the set of accepted arms. However as $\nu$ never decreases, a removed arm is removed forever. The exact procedure is described in Algorithm 1 below.

**Algorithm 1 Cautious Greedy with $O(\log(T))$ communications**

1: **Input**: $M$ (number of players), $p$ (activation probability), $T$ (horizon), $m$ (player id), $p_g$ (successful communication probability)
2: $\nu = 0$, $K = [K]$, $A = \emptyset$, $\mathcal{U} = \emptyset$, $n = 0$, $\tilde{\mu} = 1$, $\tilde{\mu}^L = 0$, $\tilde{\mu}^H = 1$
3: for $s = 0, \ldots, \lfloor \log_2(T) \rfloor$ do
4:   for $t = 2^s \ldots \min(2^{s+1} - 1, T)$ do
5:     if $2^s < 16M \log(2(2MT^2))$ then
6:       Play arm $M^t_M(m)$
7:     else
8:       Play arm $M^t_M(m)$ \[7\] where $\mathcal{E} = K \setminus \mathcal{U}$
9:       Rotate $\mathcal{U}$ in a round robin fashion over $K \setminus A$ (See Appendix A.2 for details)
10:     end if
11:     if $n = |K \setminus A|$ then // end of round robin
12:       $n = n + 1$
13:       end if
14:     end for
15:   end for
16:   if $2^s \geq 8M$ then
17:     if $t = 2^s + 2^{s-1}$, compute $\mu(t), T(t)$ \[4\] \[5\]
18:     if $t \in [2^s + 2^{k-1}, 2^s + 2^{k-1} + 2^{s-2}]$ and $\mod M = m$, $\text{SEND}_{m \rightarrow \text{gateway}}(\mu(t), T(t))$
19:     if $t \in [2^s + 2^{k-1} + 2^{s-2} + 2^{s-1}]$ and $\mod M = m$, $\text{SEND}_{\text{gateway} \rightarrow m}(\tilde{\mu}, \tilde{\mu}^L, \tilde{\mu}^H)$ with $\mu, \tilde{\mu}^L, \tilde{\mu}^H$ with \[6\], \[8\]
20:     end if
21:   end if
22: end for
23: end if
24: end for
25: $\tilde{\mu} = \mu$, $\tilde{\mu}^L = \mu^L$, $\tilde{\mu}^H = \mu^H$
26: end for
27: end for


3.2 Regret bound

The main result of this section is an upper bound on the expected regret of Cautious Greedy. This bound depends on several data-dependent quantities that we introduce now: $\Delta(\nu^*)$ is the minimum simple regret achieved by an allocation removing exactly $\nu^*-1$ arms, while the number of arms removed by the optimal assignment is equal to $\nu^*$. Denoting $M^*_\nu = M^*_{M^*}$, $\Delta(\nu^*)$ is defined as $\Delta(\nu^*) = \langle \mu, g(M^*_\nu) - g(M^*_{\nu-1}) \rangle$. By convention, we set $\Delta(\nu^*) = \infty$ if $\nu^* = 0$. $\Delta(j) = \mu(j+1) - \mu(j)$ is the difference between the reward of the worst arm not eliminated in the optimal assignment and the reward of the $j$-th worst arm. Lastly, $r$ is the norm of the minimum perturbation of $\mu$ causing $M^*_\nu$ to change for some value of $\nu$. More precisely, define $r(\nu) = \min_{\|M - M^*_\nu\|_\infty} \|\mu - \mu\|_\infty$ then $r = \min_{\|\nu\|_\infty} r(\nu)$. Proposition 3.1 shows that the expected regret of Cautious Greedy is upper bounded by $O\left(\frac{\nu^*}{\Delta(\nu^*)} + \sum_{j \leq \nu^*} \frac{\log(T)}{\Delta(j)}\right)$ where $O$ hides quantities independent of the data and $T$.

**Proposition 3.1** (Upper bound on the regret Cautious Greedy). The regret $R_{CB}$ of Cautious Greedy satisfies

$$
\mathbb{E}[R_{CB}] \leq \frac{3840 \nu^*}{\Delta(\nu^*)} \left(\frac{2K^2T^3}{\nu^*} + 1\right) + \sum_{\nu=1}^{\nu^*} \frac{2688(\nu^* + 1) \log(2T^3K^2)}{\Delta(\nu)} + \frac{1078MK(\nu^* + 1)}{r} + \frac{16M^2p}{\log(1 - p)}(1 + \log(2MT^2))^2 \mathbb{I}\{\nu^* \neq 0\}
$$

where $\mathbb{E}[C_{CB}] \leq \frac{2M^2 \log(T)}{p}$.

The first term is reminiscent of the regret induced by Greedy with full information. The second one comes from the sample complexity of finding the $\nu^*$ worst arms. The third one is due to the sample complexity of detecting that the optimal policy eliminates $\nu^*$ arms. The fourth one is finally due to communication. Interestingly, every term depending on $T$ are null when $\nu^* = 0$, which corresponds to situations where the optimal policy assigns at least one player on every arm. This makes the regret of Cautious Greedy constant in such situations, which happens as soon as arm rewards have a similar order of magnitude (see Lemma 3.1).

At first sight, it seems like the first term in Proposition 3.1 could be arbitrarily larger than the second term. Fortunately, this is untrue as shown below:

**Lemma 3.2.** $\Delta(\nu^*) \geq \langle g(M^*_{\nu^* + 1}) + 1 - g(M^*_{\nu^* + 1}) \rangle \Delta(\nu^*)$

Together with Proposition 3.1, Lemma 3.2 shows that the regret of cautious Greedy is upper bounded by $O\left(\frac{1}{\nu^*} + \sum_{\nu=1}^{\nu^*} \frac{\log(T)}{\Delta(\nu)}\right)$ where $O$ hides terms in $M, p, K, p_T$. The remainder of this section sketches the proof of Proposition 3.1. The precise statement of lemmas and their proofs are deferred to Appendix A.

**Proof sketch of Proposition 3.1**. We start by an upper bounds $R_{UB}$ on the regret of a quasi-centralized version of Cautious Greedy (see Algorithm 3 in Appendix A where communication aspects are removed ($p_T = 1$, Lines 20 to 24 are removed, updates Line 26 uses all seen samples). Using classical concentration bounds (Lemma A.1), we can assume that $\mu^H$ and $\mu^L$ (defined in Equation (8)) verify $\mu^H \geq \mu \geq \mu^L$ without affecting the regret bound. Consequently, Algorithm 3 yields that $\nu$ is only increased if $\nu < \nu^*$ (Lemma A.3) and the update of the set of active arms ensures that optimal arms are never eliminated (Lemma A.4).

We then focus on bounding the number of times each arm is played. The round-robin procedure ensures all active arms are assigned at least one player regularly, as proven by Lemma A.5. However, because of collisions, assigning at least one player to an arm does not guarantee an observation. Lemma A.6 makes this relation explicit.

Denote by $M^*_\nu = M^*_{M^*}$ the optimal assignment of players when at most $\nu$ arms can be assigned zero players and $M^*_\nu(t) = M^*_{M^*_\nu(t)}$ the optimal assignment of players when only arms not in $E(t)$ can be assigned zero players. The regret is the sum of three terms:

$$
\sum_{t=1}^{T} \langle \mu, g(M^*_\nu) - g(M^*_\nu(t)) \rangle + \sum_{t=1}^{T} \langle \mu, g(M^*_\nu(t)) - g(M^*_\nu) \rangle + \sum_{t=1}^{T} \langle \mu, g(M^*_\nu(t)) - g(M(t)) \rangle
$$

These three terms measure different aspects of the regret: (i) the error due to $\nu$ the number of arms under pressure being different from $\nu^*$ the optimal number of players to eliminate; (ii) the error due to $E(t)$ being different from support($M^*_\nu$), the optimal set of arms that must be assigned at least one player by $M^*_\nu$; (iii) the error due to $M(t)$ being different from $M^*_\nu(t)$, the optimal assignment of players among possible assignments in $M^*_\nu(t)$.

Focusing on (i): as the number of samples seen increases, $\nu$ increases to get closer to $\nu^*$. Lemma A.7 bounds the number of samples seen before the algorithm increases $\nu$, which leads to an upper bound on the total regret due to this term shown in Lemma A.8.

Regarding (ii): for a given $\nu$, two things may prevent a sub-optimal choice of arms $E$ on which at least one player must be assigned. Either an arm in $E$ is eliminated or an arm in $[K] \setminus E$ is accepted. Lemma A.9 pro-
vides a lower bound on the number of samples seen be-
fore a sub-optimal arm is eliminated while Lemma [A.10]
provides a lower bound on the number of samples seen
before an optimal arm is accepted. The two previous
lemmas allow to quantify when arms are accepted or
rejected. We then compute the cost of a sub-optimal
choice of arms E in Lemma [A.11] and combine these
three lemmas to bound the total regret due to this
term in Lemma [A.12].

Lastly, the third term (iii) measures the mismatch
between the chosen assignment M(t) and the best
possible assignment with the same support. Crucially,
there is no support mismatch and therefore we are in
a setting close to the full information setting which allows
us to bound the regret due to these terms by a quantity
independent of the horizon T (see Lemma [A.13]).

Moving to Cautious Greedy with limited communi-
cation (Algorithm 1), when the condition in Line 5 does
not hold, all communications succeed with high prob-
ability and the regret is less than $2R_{UB}^B$. Otherwise,
when $\nu^* \neq 0$, a union bound gives the additional term in
$\log(T)$. When $\nu^* = 0$, the regret due to phases with suc-
cessful communication is less than $2R_{UB}^B$ and the other
terms yield the additional constant (see Lemma [A.14]).

Lastly, since it takes $\frac{1}{p_\theta}$ calls of SEND to send a message
successfully in expectation and there is at most $2M$
successful calls by phases, the number of phases being
bounded by $\log(T)$ the proposition follows. □

4 LOWER BOUND

The upper bound of Cautious Greedy when $\nu^* = 0$
scales in $\frac{1}{T}$. Lemma [A.1] shows that under mild condi-
tions, this dependency in $r$ can be improved:

Lemma 4.1 (Lower bound for $\nu^* = 0$). Consider
$K = 2$ arms and $M = 2N + 1$ players for some $N \in \mathbb{N}$
and assume $p \leq \frac{1}{M+1}$, $r_0 < \frac{p}{2}$, $T \geq \frac{1}{\log(M)r_0}$. For any
algorithm A, there exists a choice of rewards $\mu$ such
that $r(\mu) = r_0$ and

$$E[R_A] \geq \frac{1}{256Mr_0}.$$

Proof sketch (see proof in Appendix A.4). We take
parameters $\mu_1$ and $\mu_2$ such that $r = \frac{3}{2}$ and the
optimal solution is $M^* = (N, N + 1)$ if $\mu = \mu_1$ and
$M^* = (N + 1, N)$ if $\mu = \mu_2$. Moreover, we choose them
so that the top two solutions are always $(N, N + 1)$
and $(N + 1, N)$.

First, we augment A so that each arm yields a sample
$X_k$ with probability $g(M)$ instead of $g(M_k(t))$;
moreover A is forced to choose at each step between
$M(t) = (N, N + 1)$ or $M(t) = (N + 1, N)$ (these two
modifications only improve A). Following the proof of
Theorem 3 in [Wang and Chen (2017)], we recast this
setting as a 2-armed bandit problem where arm $k$ has
reward 1 with probability $g(M)\mu_k$, 0 with probability
$g(M)(1 - \mu_k)$ and $X_k = \perp$ with probability $1 - g(M)$.

The rest of the proof follows closely the proof of
Proposition 4 in [Mourtada and Gaillard (2019)] and yields the lower bound
$E[R_A] \geq \frac{\Delta_1(g(M+1) - g(M))}{2} \exp(-4Tg(M)\Delta^2).$ As
the regret increases with $T$, taking $T = \lfloor \frac{1}{g(M)\Delta^2} \rfloor$
concludes.

Next, we investigate the case $\nu^* > 0$ and show a lower
bound inspired by the classical results of [Lai et al.
(1985)]

Let us first introduce the notion of a consistent algo-
rithm. Let $T_i$ be the number of times with at least
one player on the $i$-th worst arm. An algorithm is
consistent if $\forall \alpha > 0$, $\forall j > \nu^* \in \mathbb{E}[T - T_j] = \mathcal{O}(T^\alpha)$
and $\forall j \leq \nu^*$, $\mathbb{E}[T_j] = \mathcal{O}(T^\alpha)$.

Lemma 4.2 (Lower bound for $\nu^* > 0$). For any
integers $M \geq 5$, $\nu^* > 0$, $p \leq \frac{1}{M+1}$, any gaps
$\Delta_1, \ldots, \Delta_{\nu^*} \leq \frac{p^2}{16(M-4)}$, and for any consistent algo-
rithm A, there exists a set of parameters $(\mu_1, \ldots, \mu_{\nu^* + 2})$
such that $\mu_{\nu^* + 1} - \mu_{\nu^*} = \Delta_{\nu^*}$ for all $\nu \in [\nu^*]$ and the
regret of A satisfies, for some universal constant $c > 0$,

$$\lim_{T \to \infty} \frac{E[R_A]}{\log(T)} \geq \sum_{\nu=1}^{\nu^*} c \Delta_{\nu}.$$  

Proof sketch (see proof in Appendix A.5). Assume for
the sketch of proof that $\nu^* = 1$ and that there are 3
arms. We are considering two alternative mean param-
eters $(\mu_0, \mu_1, \mu_1 + \Delta)$ and $(\mu_0, \mu_1, \mu_1 - \Delta)$ chosen so
that the optimal allocation is either $(M - 1, 1, 0)$ or
$(M - 1, 0, 1)$. Moreover, we choose $\mu_0$ and $\mu_1$ such
that in both worlds, the top two allocations are always
the aforementioned ones. This might give the impression
that there exists a trivial reduction to some standard
2-arm bandits (where those arms are the tentative two
optimal allocations). A consistent algorithm would
indeed need $N^* := \Omega(\frac{\log(T)}{\Delta^2})$ samples of sub-optimal
arms to distinguish between the two worlds. In partic-
ular, with the second set of parameters, this requires
putting one player on the third arm $N^*/p$ times (in
expectation), each one incurring a cost of $p\Delta$. This
would give the result for $\nu^* = 1$ and this technique
can be immediately generalized to $\nu^* > 1$. It is how-
ever not that simple, as putting more players on some
(suboptimal) arm gives faster feedback, yet at a higher
cost. We yet show that the best trade-off (in feedback
received vs. suboptimality cost) for an algorithm to
distinguish between the two worlds is indeed to allocate
a single player on arm 2 or 3. The aforementioned intuition is thus actually correct but requires a cautious argument.

Lemma 4.2 shows that the dependency in \( \sum_{j \leq \nu^*} \log(T) \) in the upper bound of Proposition 3.1 cannot be improved. Together, Lemma 4.1 and Lemma 4.2 show that Cautious Greedy is optimal with respect to the data-dependent quantities \( r \) and \( (\Delta(j))_{j \leq \nu^*} \).

5 EXPERIMENTS

The code is in python and available on https://github.com/hugorichard/mmab We use matplotlib (Hunter 2007) for plotting, and numpy (Harris et al., 2020) for array manipulations. The above libraries use open-source licenses. Computations were run on a cluster with 10 cpus and 100 GB of RAM. Our experiments compare the expected regret of Cautious Greedy (Algorithm 1), UCB (Algorithm 4), and ETC (Dakdouk 2022 Algorithm 8). In all these algorithms, maximization problems of the form \( \max_{\Sigma_{M \in \mathcal{M}}} g(M, v) \) are solved using the sequential algorithm of Dakdouk (2022 Algorithm 5). In Cautious Greedy, the sequential algorithm is also adapted to solve \( \max_{M \in \mathcal{M}_E} g(M, v) \) for some set \( \mathcal{E} \subset [K] \) by assigning the first \( |\mathcal{E}| \) players to a different arm in \( \mathcal{E} \) and then running the sequential algorithm for the rest of the players. The optimality of this approach is detailed in Appendix E.

For a given horizon \( T \), assignments \( (M(t))_{t=1}^{T} \) are played based on the rewards seen during the execution of algorithms. We vary \( T \) uniformly between 30 and 1200 using steps of 30 and record \( \sum_{t=1}^{T} (\mu_t g(M^*_t) - g(M(t))) \). Each experiment is run 50 times, we plot the mean value of the regret as a function of \( T \). Error bars represent the first and last decile. We also performed experiments with larger values of \( T \) available in Appendix F.

The first experiment in Figure 1 (top), there are \( M = 30 \) players, \( K = 2 \) arms, \( \mu = (0.8, 0.5) \), \( p = 0.01 \) and \( p_g = 1 \). The optimal assignment is \( \mathbf{M}^* = (26, 4) \) so that \( \nu^* = 0 \). In this example, Cautious Greedy clearly outperforms the other methods as expected when \( \nu^* = 0 \).

In the second experiment in Figure 1 (bottom), there are \( M = 3 \) players, \( K = 2 \) arms, \( \mu = (0.99, 0.01) \), \( p = 0.1 \) and \( p_g = 1 \). The optimal solution is \( \mathbf{M}^* = (3, 0) \) so that \( \nu^* = 1 \). This experiment highlights that Cautious Greedy takes slightly longer time than UCB to assign no player to a suboptimal arm. This is expected for 2 reasons. First UCB is a centralized algorithm so it communicates much more than Cautious Greedy which only communicates \( \log_2(T) \) times in expectation.

Second, the fact that Cautious Greedy is biased towards having good performance when \( \nu^* = 0 \) necessarily means a loss of performance when it is not the case.

In both experiments, ETC incurs a much larger regret, which is consistent with its \( \mathcal{O}(T^{3}) \) regret. Note however that ETC uses only constant communication costs.

6 CONCLUSION, OPEN PROBLEMS AND FUTURE WORK

We proposed an asynchronous multiplayer multi-armed bandits algorithm called Cautious Greedy, achieving a regret of order \( \mathcal{O}(1/r + \sum_{\nu^*} \log(T)/\Delta(\nu)) \) (ignoring data-independent terms) with \( \mathcal{O}(\log(T)) \) communications. In particular, its regret does not scale with \( T \) when \( \nu^* = 0 \). We also prove lower bounds suggesting that the dependency in both \( r \) and \( \sum_{\nu^*} \log(T)/\Delta(\nu) \) is optimal.

A future open question is whether the dependency on parameters \( K, M, p, p_g \) can be enhanced. The analysis in [Degene and Perchet 2016] for the full information setting implies an upper bound on Greedy as \( \mathcal{O}(\log(K)/\Delta) \) when all arms have a sub-optimal gap \( \Delta \), indicating the potential for improvement.

Our algorithm requires several assumptions to perform properly. Most of them are actually very mild, while others would require an involved analysis to get discarded. Without prior knowledge of \( T \), a doubling trick [Besson and Kaufmann 2018] Theorem 7) can be used when the horizon \( T \) is unknown but would introduce a logarithmic dependency in \( T \). Whether Cautious Greedy can be made anytime is therefore an interesting extension. The activation probabilities of players might be heterogeneous in practice. However, the optimization algorithm of Dakdouk (2022) is only optimal in the homogeneous case. No efficient maximization scheme of the problem in Equation 2 in the heterogeneous case is currently known. If however we were given access to an oracle maximizing this problem, we believe that our algorithms and their bounds
can be adapted although this requires meticulous work. Also, if \( p \) is unknown, it can be estimated on the fly. Indeed at every step, players observe a realisation of a Bernoulli(\( p \)). Assuming successful communications, agents use the empirical estimate \( \hat{p} \) provided by the gateway based on agents’ observations. Then, replacing \( p \) by \( \hat{p} \) just adds a constant term to the regret upper bound. The case of heterogeneous \( (p_i)_{i \in [M]} \) is similar. Likewise, if \( p_g \) is unknown, notice that agents observe a realization of Bernoulli(\( p_g \)) whenever they attempt to communicate. Then, the gateway shares an upper bound of \( p_g \) which replaces \( p_g \) in the condition Line 5. The time it takes to satisfy the modified condition increases (but remains of the same order) and (despite communication failures) is identical across agents with high probability. The rest of the analysis works without change. The careful analysis of these extensions is left to future work.

Another significant direction is to go beyond the limited communication setting. Being able to handle the decentralized setting where agents are no longer allowed to communicate without cost remains a great challenge and the original motivation of asynchronous multiplayer bandits. A solution to handle the decentralized setting is to use collisions to communicate as done for example in \cite{BistritzL2018} \cite{BoursierP2019} \cite{WANG2020}. These previously cited works however only tackle the synchronized case. In the case we study, communicating through collisions remains possible but the length of communication phases would be significantly increased. In the collision sensing setting, if players \( i \) and \( j \) need to propagate a bit through collision, they roughly need \( \log(T) / p^2 \) time-steps to send a single bit with high probability. Whether there exist quicker communication schemes (e.g. using random phase length) for the asynchronous case is an open problem. Concerning communication without collisions, in the synchronous case, Dyn-MMAB \cite{BoursierP2019} and Lugosi and Mehrabian \cite{LugosiM2022} achieve \( O(T^{2/3}) \) regret. The work of \cite{BubeckL2021} even achieves \( O(\sqrt{T}) \). Whether these approaches can be extended to the asynchronous setting we consider is an interesting open question.

Acknowledgements

Vianney Perchet acknowledges support from the French National Research Agency (ANR) under grant number ANR-19-CE23-0026 as well as the support grant, as well as from the grant “Investissements d’Avenir” (LabEx Ecodec/ANR-11-LABX-0047). This work was completed while E. Boursier was a member of TML Lab, EPFL, Lausanne, Switzerland.

References


Constant or Logarithmic Regret in Asynchronous Multiplayer Bandits with Limited Communication


armed bandits. In 2016 European Control Conference (ECC), pages 243–248. IEEE.


Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not
modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won’t result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No, will be released over publication]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No, will be released over publication]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [No, will be released over publication]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes (last section in appendix)]

(b) The license information of the assets, if applicable. [Yes (last section in appendix)]
(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
(d) Information about consent from data providers/curators. [Not Applicable]
(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]
A Analysis of Cautious Greedy

A.1 A useful upper bound

At many places we will have to bound quantity of the form $\langle \mu - \mu', g(M) - g(M') \rangle$ where $\mu, \mu' \in [0, 1]^K$ and $M, M' \in M$. We have

$$\langle \mu - \mu', g(M) - g(M') \rangle \leq \langle |\mu - \mu'|, g(M) - g(M') \rangle$$

$$\leq \langle |\mu - \mu'|, g(M) + g(M') \rangle$$

$$\leq \sum_{k=1}^K (g(M_k) + g(M'_k))$$

$$\leq \sum_{k=1}^K (M_k + M'_k) p$$

so that we have

$$\langle \mu - \mu', g(M) - g(M') \rangle \leq 2Mp$$

(9)

Note that since $M_k \leq \frac{1}{\log(1-p)} \leq \frac{1}{p}$, we have $Mp \leq K$.

A.2 A precise description of the Round Robin procedure

Rotating $U$ in a round-robin fashion over $Y \supset U$ means that $U$ undergoes one iteration of the Round Robin (RR) procedure. See $Y$ as $(y_1, \ldots, y_{|Y|})$, $U$ as $(u_1, \ldots, u_s)$. At each iteration, an element from $Y \setminus U$ is added to $U$ and an element of $U$ is dropped in such a way that after $|Y|$ iterations, all elements of $U$ have been added and dropped from $U$ exactly once.

A possible implementation of the RR procedure is the following. Initialize $U = (y_1, \ldots, y_s)$ and $t = s + 1$. Then, performing one iteration of the RR procedure means following Algorithm 2.

**Algorithm 2** Rotate $U$ in a round robin fashion over $Y$ (one iteration)

1: **Input:** $t$ (iteration number), $U = (u_1, \ldots, u_{|U|})$, $Y = (y_1, \ldots, y_{|Y|})$
2: Remove $u_1$ from $U$
3: $\forall i \in [|U| - 1]$, set $u_i \leftarrow u_{i+1}$
4: Set $u_{|U|} = y_t \mod |Y|$.

A.3 Proof of Proposition 3.1 and Lemma 3.2

A.3.1 Proof of Lemma 3.2

**Proof.** Assume $\nu^* \geq 1$. $\Delta^{(\nu^*)}$ is defined as $\Delta^{(\nu^*)} = (\mu, g(M^*) - g(M^*_{\nu^* - 1}))$ and $\Delta^{(\nu^*)} = \mu_{(\nu^* - 1)} - \mu_{(\nu^*)}$.

Call $(i)$ the index of the $i$-th worst arm. $M^*_{\nu^* - 1}$ can be constructed from $M^*_{\nu^*}$. To do so, remove a player from the arm $j$ such that

$$j = \arg\min_{i \in \text{supp}(M^*_{\nu^*}), M^*_i \geq 2} \mu_i (g(M^*_i) - g(M^*_{i - 1}))$$

where $M^*_i$ denotes the $i$-th coordinate of $M^*_{\nu^*}$ and place it on arm $(\nu^*)$.

We then have

$$\Delta^{(\nu^*)} = \mu_j (g(M_j) - g(M_j - 1)) - \mu_{(\nu^*)} p$$

If $j \neq (\nu^* + 1)$, taking a player from arm $j$ in $M^*_{\nu^*}$ to put it on arm $\nu^* + 1$ would yield to a worse assignment, we have that $\mu_j (g(M_j) - g(M_j - 1)) \geq \mu_{\nu^* + 1} (g(M^*_{\nu^* + 1} + 1) - g(M^*_{\nu^* + 1}))$. This inequality is also true if $j = (\nu^* + 1)$.
This implies that

\[
\Delta^{(\nu^*)} \geq \mu_{\nu^*+1}(g(M_{\nu^*+1}^*) + 1) - g(M_{\nu^*+1}^*) - \mu_{(\nu^*)}p \\
\geq \Delta^{(\nu^*)}(g(M_{\nu^*+1}^*) + 1) - g(M_{\nu^*+1}^*) \\
\geq \Delta^{(\nu^*)}(g(M) - g(M - 1)) \\
= \Delta^{(\nu^*)}p(1 - p)^{M - 2}(1 - Mp)
\]

\[\blacksquare\]

### A.3.2 Proof of Proposition 3.1

In this section, we first provide an upper bound \(R_{\text{UB}}^{cCG}\) on the regret \(R\) of a quasi-centralized version of Cautious Greedy described Algorithm 3. It is assumed to have access at the end of a phase to the samples collected by all players during all previous phases. Then, we provide an upper bound on the regret \(R_{\text{CG}}\) of Cautious Greedy in the limited communication setting, that builds upon the upper bound \(R_{\text{UB}}^{cCG}\) on the regret of Algorithm 3.

**Algorithm 3** Cautious Greedy with \(\log(T)\) communications

1: Input : \(M\) (number of players), \(p\) (probability that a player is active), \(T\) (horizon)
2: \(\nu = 0, K = [K], A = \emptyset\)
3: \(\mathcal{U} = \emptyset, n = 0, \hat{\mu} = 0, \hat{\mu}^L = 0, \hat{\mu}^H = 1\)
4: for \(s = 0, \ldots, \lfloor \log_2(T) \rfloor\) do
5: for \(t = 2^s \ldots \min(2^{s+1} - 1, T)\) do
6: Play \(\hat{\mu}_{M}^{\nu}\) as defined in (7) where \(E = K \setminus \mathcal{U}\)
7: Rotate \(\mathcal{U}\) in a round robin fashion over \(K \setminus A\) (See Appendix A.2 for details)
8: \(n = n + 1\)
9: if \(n = |K \setminus A|\) then // end of round robin
10: \(n = 0\) and compute \(\hat{\mu}_{M}^{\nu}\) and \(\hat{\mu}_{M}^{\nu}\) (7)
11: while \(\langle \hat{\mu}^L, g(M_{\nu}^{\mu}) \rangle > \langle \hat{\mu}^H, g(M_{\nu}^{\mu}) \rangle\) do
12: \(\nu = \nu + 1\)
13: end while
14: Update \(A = \{k \in [K], \hat{\mu}_{\nu}^k < \hat{\mu}_{\nu}^L\}\) and \(K = [K] \setminus \{k \in [K], \hat{\mu}_k^H < \hat{\mu}_{(\nu+1)}^L\}\)
15: Let \(\nu\) be \(\nu - |K \setminus A|\) elements from \(K \setminus A\)
16: end if
17: Update \(\hat{\mu}, \hat{\mu}^L\) and \(\hat{\mu}^H\) using samples from all players
18: end for

The analysis heavily builds upon confidence bounds. We first establish a concentration lemma on the mean reward of each arm.

**Lemma A.1** (Concentration of mean rewards). Let GOOD be the event

\[\forall k \in [K], \forall t \in [T], |\hat{\mu}_k(t) - \mu_k| \leq \zeta_{kt}\]

Then, \(P(\text{GOOD}) \leq \frac{1}{TK}\)
Proof of Lemma A.1. Fix $k \in [K]$, by Hoeffding, we have

$$P(|\hat{\mu}_k(t) - \mu_k| \geq \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}) = \sum_{\tau=1}^{T} P(|\hat{\mu}_k,\tau - \mu_k| \geq \sqrt{\frac{\log(2T^3K^2)}{2\tau}})$$

$$\leq \sum_{\tau=1}^{T} 2 \exp(-2\tau \sqrt{\frac{\log(2T^3K^2)^2}{2\tau}}) = \frac{1}{K^2T^2}$$

and with a union bound on $\tau \in [T]$ and a second on $k \in [K]$, we obtain:

$$P(\exists t \in [T], \exists k \in [K], |\hat{\mu}_k(t) - \mu_k| \geq \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}) \leq 1$$

Rearranging, we get with probability $1 - \frac{1}{KT}$:

$$\forall t \in [T], \forall k \in [K], |\hat{\mu}_k(t) - \mu_k| \leq \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}$$

(10)

which implies the desired result. \hfill \square

A consequence of Lemma A.1 is that up to a small additive constant in the regret, we can assume that the GOOD event holds.

Lemma A.2 (Confidence bounds). Define $R_G = R_1\{\text{GOOD}\}$, then,

$$\mathbb{E}[R] \leq \mathbb{E}[R_G] + 2$$

(11)

Proof of Lemma A.2. $\mathbb{E}[R] = \mathbb{E}[R_1\{\text{GOOD}\}] + \mathbb{E}[R_1\{\text{GOOD}\}]$ and $R_1\{\text{GOOD}\} \leq 2KT_1\{\text{GOOD}\}$, we then conclude from Lemma A.1. \hfill \square

Working under the GOOD event makes the analysis much easier. We begin by showing that $\nu$ is a lower bound on the optimal number of arms to eliminate:

Lemma A.3. Under the GOOD event, $\nu \leq \nu^*$ at any time $t$.

Proof of Lemma A.3. $\nu$ is only increased in the while loop. We want to show that if $\nu = \nu^*$, then the condition in the while loop cannot be met. Assume by contradiction that $\nu = \nu^*$ and $\max_{M \in \mathcal{M}} \langle \hat{\mu}_L^*, g(M) \rangle > \max_{M \in \mathcal{M}_\nu} \langle \hat{\mu}_H^*, g(M) \rangle$. By the good event, we have

$$\max_{M \in \mathcal{M}} \langle \hat{\mu}_L, g(M) \rangle < \max_{M \in \mathcal{M}} \langle \mu, g(M) \rangle$$

and

$$\max_{M \in \mathcal{M}_\nu} \langle \hat{\mu}_H^*, g(M) \rangle > \max_{M \in \mathcal{M}_\nu} \langle \mu, g(M) \rangle$$

Therefore

$$\max_{M \in \mathcal{M}} \langle \hat{\mu}_L, g(M) \rangle > \max_{M \in \mathcal{M}_\nu} \langle \mu, g(M) \rangle$$

$$\Rightarrow \max_{M \in \mathcal{M}} \langle \mu, g(M) \rangle > \max_{M \in \mathcal{M}_\nu} \langle \mu, g(M) \rangle$$
and since \( \nu = \nu^* \),
\[
\max_{M \in M_{\nu}} \langle \mu, g(M) \rangle = \max_{M \in M} \langle \mu, g(M) \rangle.
\]

This yields the following contradiction:
\[
\max_{M \in M} \langle \mu, g(M) \rangle > \max_{M \in M} \langle \mu, g(M) \rangle.
\]

More generally, under the GOOD event, Cautious Greedy never eliminates an optimal arm:

**Lemma A.4** (Optimal arms are never eliminated). Under the GOOD event, the set of optimal arms is always included in the set of active arms: \( \text{support}(M^*) \subseteq K(t) \)

**Proof of Lemma A.4** Elimination may happen when the set of active arms is updated. An arm \( k \) is eliminated at this stage if \( \mu^H_k < \mu^L_{(\nu^*+1)} \). But since \( \nu \leq \nu^* \), this implies \( \mu^H_k < \mu^L_{(\nu^*+1)} \). Under the good event \( \mu^H_k > \mu_k \) and \( \mu^L_{(\nu^*+1)} < \mu(\nu^*+1) \) so that \( \mu^H_k < \mu^L_{(\nu^*+1)} \) implies \( \mu_k < \mu(\nu^*+1) \) and therefore \( k \notin E_{\nu^*} \).

Since Cautious Greedy never eliminates any optimal arm and since \( \nu \) increases, \( \nu \) will eventually reach \( \nu^* \) and bad arms will no longer remain. But as long as \( \nu < \nu^* \), Cautious Greedy will pay a non-zero cost. This source of error as well as others strongly depends on the number of times arms are pulled without collisions. Indeed, as the number of pulls without collision increases, the reward estimates \( \hat{\mu} \) become more accurate, making Cautious Greedy’s decisions better. Therefore, we introduce \( q(t) = \min_{k \in K(t)} T_k(t) \), the number of times each active arm has been played without collision.

To be able to understand how \( q(t) \) scales with \( t \), a pre-requisite is to count the number of times that arms are assigned at least one player. Denote \( \tau_k(t) \) the number of times arm \( k \) has been assigned at least one player at time \( t \) and \( \tau(t) = \min_{k \in K(t)} \tau_k(t) \). The next Lemma exhibits a lower bound on \( \tau(t) \):

**Lemma A.5** (Scaling of \( \tau \) with \( t \)). We have \( \forall t, \tau(t) \geq \tau_0(t) = [\max(\nu^* + \nu^*, 0)] \). Furthermore, \( \forall t \geq (\nu^* + 1)^2, \tau_0(t) \geq \tau_{t_0}(t) = \frac{u_n}{2(\nu^* + 1)} \).

**Proof of Lemma A.5** Call \( t_n \) the value of \( t \) the \( n \)-th time where \( t = 0 \mod |\mathcal{U}| \). Between \( t_n \) and \( t_{n+1} \) (included) all arms have been played \( |\mathcal{U}_n| - u_n \) times where \( \mathcal{U}_n \) and \( u_n \) are the set of active but not yet accepted arms \( \mathcal{U} \) and the number of arms under pressure \( u \) after the updates at time \( t = t_n \). \( \tau \) increases linearly between time \( t_n \) and \( t_{n+1} \) except for \( u_n \) time steps where \( u_n = \nu_n - |K| \setminus K \) is the number of arms that need to be put under pressure during phase \( n \) but that are not yet eliminated and \( \nu_n \) is the value of \( \nu \) during phase \( n \).

We have that for \( t_n \leq t < t_{n+1} : \)
\[
\tau(t) \geq \tau(t_n - 1) + \max(t - (t_n - 1) - u_n, 0)
\]
\[
\tau(t_n - 1) + \max((t - (t_n - 1)) \cdot \frac{t_{n+1} - (t_n - 1) - u_n}{t_{n+1} - (t_n - 1)} - (t_{n+1} - t) \cdot \frac{u_n}{t_{n+1} - (t_n - 1)}, 0)
\]
and
\[
\tau(t_{n+1} - 1) - \tau(t_n - 1) = t_{n+1} - (t_n - 1) - u_n
\]
\[
= (t_{n+1} - (t_n - 1)) \cdot \frac{t_{n+1} - (t_n - 1) - u_n}{t_{n+1} - (t_n - 1)}
\]

Since \( t_{n+1} - (t_n - 1) = |K_n \setminus A_n| \) we have:
\[
\frac{t_{n+1} - (t_n - 1) - u_n}{t_{n+1} - (t_n - 1)} = \frac{|K_n| - |A_n| - u_n}{|K_n| - |A_n|} \geq \frac{1}{u_n + 1} \geq \frac{1}{\nu^* + 1}
\]
It follows that for all \( n \geq 1 \),
\[
\tau(t_n - 1) \geq \frac{t_n - 1}{\nu^* + 1}
\]

Therefore, we obtain \( t_n \leq t \leq t_{n+1} \)
\[
\tau(t) \geq \frac{t_n - 1}{\nu^* + 1} + \max((t - (t_n - 1)), \frac{1}{\nu^* + 1} - (t_{n+1} - t), 0)
\]
\[
\geq \frac{t_n - 1}{\nu^* + 1} + \max((t - (t_n - 1)), \frac{1}{\nu^* + 1} - \nu^*, 0)
\]
\[
\geq \max\left(\frac{t}{\nu^* + 1} - \nu^*, \frac{t_n - 1}{\nu^* + 1}\right)
\]
\[
\geq \max\left(\frac{t}{\nu^* + 1} - \nu^*, 0\right)
\]

Since this last line holds for all \( n \), we have for any \( t \) that \( \tau(t) \geq \max\left(\frac{t}{\nu^* + 1} - \nu^*, 0\right) \).

Furthermore, for any \( t \geq 2(\nu^* + 1)^2 \), we have that
\[
\frac{t}{\nu^* + 1} - \nu^* \geq \frac{t}{2(\nu^* + 1)} \geq \frac{1}{2} \frac{t}{\nu^* + 1}
\]

The next step is to link \( \tau_b(t) \) to \( q(t) \). By noting that \( g(M_k) \geq p \), we expect \( q(t) \) to scale approximately with \( p\tau_b(t) \).

First, we show that \( T_k(t) \) stochastically dominates a sum of \( \tau_b(t) \) independent Bernoulli random variables with parameter \( p \).

**Lemma A.6.** The number of times arm \( k \) has been played without collision \( T_k(t) \) stochastically dominates \( B_{[\tau_b(t)],p} \) where \( B_{[\tau_b(t)],p} \) is a binomial random variable with parameters \( n = [\tau_b(t)] \) and \( p = p \).

**Proof of Lemma A.6.** We have \( T_k(t) = \sum_{s=1}^t \eta_k(M_k(s)) \geq \sum_{s \in [t], M_k(s) \geq 1} \eta_k(M_k(s)) \) where \( \eta_k(M_k(s)) = 1 \{ \text{Exactly 1 player (among } M_k(s) \text{) is active on arm } k \} \).

Then notice that \( \sum_{s \in [t], M_k(s) \geq 1} \eta_k(M_k(s)) \) stochastically dominates \( \sum_{s \in [\tau_b(t)], M_k(s) \geq 1} \eta_k(M_k(s)) \) where \( (\eta_k(M_k(s)))_{s=1}^t \) are independent Bernoulli random variables with mean \( p \). Since the number of terms in the sum is greater than \( [\tau_b(t)] \), the lemma follows.

In particular, we have from a multiplicative Chernoff bound and union bound over \( K \) that
\[
P(q(t) \leq \frac{1}{3} p[\tau_b(t)]) = P(\exists k \in [K], T_k(t) \leq \frac{1}{3} p[\tau_b(t)]) \leq K \exp\left(-\frac{2}{9} p[\tau_b(t)]\right)
\]

We can now focus on upper-bounding the different sources of errors. First, \( E[R_G] \) can trivially be written as:
\[
E[R_G] = E[R_a 1\{GOOD\}] + E[R_e 1\{GOOD\}] + E[R_m 1\{GOOD\}]
\]

where
\[
\begin{align*}
R_a & = \sum_{t=1}^T \langle \mu, g(M^*) - g(M^*_t) \rangle \\
R_e & = \sum_{t=1}^T \langle \mu, g(M^*_t) - g(M^*_{\xi(t)}) \rangle \\
R_m & = \sum_{t=1}^T \langle \mu, g(M^*_{\xi(t)}) - g(M(t)) \rangle
\end{align*}
\]
Under the GOOD event, if

These three terms measure a different aspect of the regret: \( R_\nu \) measures the error due to \( \nu \) the number of arms under pressure being different from \( \nu^* \) the optimal number of players to eliminate, \( R_\mathcal{E} \) measures the error due to \( \mathcal{E}(t) \) being different from support(\( M^*_t \)) the optimal set of arms that must be assigned at least one player when up to \( \nu \) players can be assigned zero players and \( R_M \) measures the error due to \( M(t) \) being different from \( M^*_\mathcal{E}(t) \) the optimal assignment of players among possible assignments in \( M_\mathcal{E}(t) \).

Let us start with the first term \( R_\nu \). As the number of samples seen increases, \( \nu \) increases to get closer to \( \nu^* \). The following Lemma provides a maximum on the number of samples seen before the algorithm detects that \( \nu \) should increase.

**Lemma A.7** (Number of iterations before \( \nu \) increases). Consider assignment \( M^*_t = \arg\max_{M, \text{support}(M) = K - k} \langle \mu, g(M) \rangle \) which is the best assignment where \( k \) arms assigned zero players.

Under the GOOD event, if \( t \) is such that \( q(2^{\log_2(t)}) \geq q_k = \frac{8M^2p^2\log(2K^2T^3)}{((\mu, g(M^*) - g(M^*_k)))^2} \), then \( \nu(t + \nu^*) > k \).

**Proof of Lemma A.7** Call \( M^*_t(t) = \arg\max_{M \in M_k} \langle \mu', g(M) \rangle \) and for simplicity, call \( t' = 2^{\log_2(t)} \).

\[ q(t') > \frac{8M^2p^2\log(2K^2T^3)}{((\mu, g(M^*) - g(M^*_k)))^2} \]

\[ \implies q(t') > \frac{8M^2p^2\log(2K^2T^3)}{((\mu, g(M^*) - g(M^*_k)))^2} \]

\[ \implies \min_{a \in K} T_\alpha(t') > \frac{18M^2p^2\log(2K^2T^3)}{((\mu, g(M^*) - g(M^*_k)))^2} \]

\[ \implies \langle \mu, g(M^*) - g(M^*_k) \rangle > 8Mp \max_{a \in K} \sqrt{\frac{\log(2T^3K^2)}{2T_\alpha(t')}} \]

\[ \implies \langle \mu, g(M^*) - g(M^*_k) \rangle > 4Mp \max_{a \in K} \zeta_\nu(t') \]

\[ \implies \langle \mu, g(M^*) - g(M^*_k) \rangle > 2\zeta(t'), g(M^*) + g(M^*_k(t')) \] (By Equation [9])

\[ \iff \langle \mu - 2\zeta(t'), g(M^*) \rangle > \langle \mu, g(M^*_k(t')) \rangle + 2\zeta(t'), g(M^*_k(t')) \]

\[ \iff \langle \mu^L(t'), g(M^*) \rangle > \langle \mu + 2\zeta(t'), g(M^*_k(t')) \rangle \] (By the GOOD event and optimality of \( M^*_t \))

\[ \implies \max_{M \in M} \langle \mu^L(t'), g(M) \rangle > \langle \mu^H(t'), g(M^*_k(t')) \rangle \] (By the GOOD event)

The last line is the while condition of Cautious Greedy for \( \nu = k \). It is execute after the end of Round Robin which can take up to \( \nu^* \) rounds. Also note that since \( q_k \geq q_{k-1} \geq q_{k-1} \cdots q_1 \), after \( t' + \nu^* \) iterations, the while condition has necessarily been executed at least \( k \) times which means \( \nu(t + \nu^*) \geq k + 1 \).

Note that as long as \( \nu \leq \nu^* \), \( R_\nu \) increases by \( \langle \mu, g(M^*) - g(M^*_k) \rangle \). **Lemma A.7** then allows to bound \( R_\nu \):

**Lemma A.8** (Bound on \( R_\nu \)).

\[ \mathbb{E}[R_\nu 1\{GOOD\}] \leq \frac{384M^2p\log(2K^2T^3)(\nu^* + 1)}{\Delta(\nu^*)} + 42M(\nu^* + 1)K \]
Proof of Lemma A.8. Call $t_{\nu}$ the last time that $\nu(t) = \nu$ and set $t_{\nu^*} = T + 1$ and $t_{-1} = 0$.

$$R_{\nu} = \sum_{t=1}^{T} \langle \mu, g(M^*) - g(M^*_{\nu(t)}) \rangle$$

$$= \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_{\nu}} \langle \mu, g(M^*) - g(M^*_{\nu}) \rangle$$

$$= \sum_{\nu=0}^{\nu^*} (t_{\nu} - t_{\nu-1}) A_{\nu}$$

$$= \sum_{\nu=0}^{\nu^*} t_{\nu} A_{\nu} - \sum_{\nu=0}^{\nu^*} t_{\nu-1} A_{\nu}$$

$$= \sum_{\nu=1}^{\nu^*} t_{\nu-1} (A_{\nu-1} - A_{\nu}) - t_{-1} A_{0} + t_{\nu^*} A_{\nu^*}$$

$$= \sum_{\nu=1}^{\nu^*} t_{\nu-1} (A_{\nu-1} - A_{\nu})$$

$$= \sum_{\nu=1}^{\nu^*} 2 \left( \frac{t_{\nu-1} - (\nu^* + 1)}{t_{\nu-1}} \right) (A_{\nu-1} - A_{\nu}) + \sum_{\nu=1}^{\nu^*} (\nu^* + 1) (A_{\nu-1} - A_{\nu})$$

$$\leq 2 \sum_{\nu=1}^{\nu^*} t_{\nu-1}^\prime (A_{\nu-1} - A_{\nu}) + 2Mp(\nu^* + 1)$$

$$= 2 \sum_{\nu=1}^{\nu^*} t_{\nu-1}^\prime (A_{\nu-1} - A_{\nu}) + 2 \sum_{\nu=1}^{\nu^*} t_{\nu-1}^\prime (A_{\nu-1} - A_{\nu}) + 2Mp(\nu^* + 1)$$

$$\leq 4 \sum_{\nu=1}^{\nu^*} (\nu^* + 1)^2 (A_{\nu-1} - A_{\nu})$$

$$+ 2 \sum_{\nu=1}^{\nu^*} (\nu^* + 1)^2 [\tau_{lin}(t_{\nu-1}^\prime)] (A_{\nu-1} - A_{\nu}) + 2Mp(\nu^* + 1)$$

$$\leq 6Mp(\nu^* + 1)^2 + 4 \sum_{\nu=1}^{\nu^*} (\nu^* + 1)^2 [\tau_{lin}(t_{\nu-1}^\prime)] (A_{\nu-1} - A_{\nu})$$

(By Lemma A.5)

(By Equation 9)
Then we have

\[(i) = \sum_{\nu=\{1,\ldots,\nu\}} (\nu^* + 1) [\tau_{\text{lin}}(t_{\nu-1}')] (A_{\nu-1} - A_{\nu}) \]

\[= \sum_{\nu=\{1,\ldots,\nu\}} (\nu^* + 1) [\tau_{\text{lin}}(t_{\nu-1}')] (A_{\nu-1} - A_{\nu}) \left\{ [\tau_{\text{lin}}(t_{\nu}')] \leq \frac{6q(t_{\nu-1}')}{p} \right\} \]

\[+ \sum_{\nu=\{1,\ldots,\nu\}} (\nu^* + 1) [\tau_{\text{lin}}(t_{\nu-1}')] (A_{\nu-1} - A_{\nu}) \left\{ [\tau_{\text{lin}}(t_{\nu}')] > \frac{6q(t_{\nu-1}')}{p} \right\} \]

\[\leq \sum_{\nu=\{1,\ldots,\nu\}} (\nu^* + 1) [\tau_{\text{lin}}(t_{\nu-1}')] (A_{\nu-1} - A_{\nu}) \left\{ [\tau_{\text{lin}}(t_{\nu}')] > \frac{6q(t_{\nu-1}')}{p} \right\} \]

We have that \(\nu(2t_{\nu-1} + \nu^*) = \nu - 1\) and therefore by Lemma A.7, we get

\[q(t_{\nu-1}') \leq q(2^{\log_2(2t_{\nu-1}')}) \leq q_{\nu} \]

This gives

\[(ii) \leq \frac{6(\nu^* + 1)}{p} \sum_{\nu=1}^{\nu^*} 4(Mp)^2 \log(2K^2T^3) \left( \frac{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle}{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle^2} \right) \]

\[= 48M^2p \log(2K^2T^3) (\nu^* + 1) \sum_{\nu=1}^{\nu^*} \left( \frac{1}{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle} - \frac{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle}{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle^2} \right) \]

\[\leq \frac{6q_{\nu}}{p_0} \]

where \(q_{\nu}\) is given in A.7 and we used \(A_{\nu^*} = 0\). From there, we have the following inequalities

\[\sum_{\nu=1}^{\nu^*} l_{\nu} \leq \sum_{\nu=1}^{\nu^*} \left( \frac{1}{\langle \mu, g(M^*) - g(M_{\nu}^*) \rangle} - \frac{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle}{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle^2} \right) \]

\[= \sum_{\nu=1}^{\nu^*} \left( \frac{1}{\langle \mu, g(M^*) - g(M_{\nu}^*) \rangle} - \frac{1}{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle} \right) \left( 1 + \frac{\langle \mu, g(M^*) - g(M_{\nu}^*) \rangle}{\langle \mu, g(M^*) - g(M_{\nu}^*) \rangle} \right) \]

\[\leq 2 \sum_{\nu=1}^{\nu^*} \left( \frac{1}{\langle \mu, g(M^*) - g(M_{\nu}^*) \rangle} - \frac{1}{\langle \mu, g(M^*) - g(M_{\nu-1}^*) \rangle} \right) \]

\[\leq \frac{1}{\langle \mu, g(M^*) - g(M_{\nu^*}^*) \rangle} \]

Taking expectations we get

\[E[R_{\nu}1\{GOOD\}] \leq 4 \frac{6q_{\nu}}{p} (\nu^* + 1) \frac{1}{\langle \mu, g(M^*) - g(M_{\nu^*}^*) \rangle} + E[(iii)] + 6Mp(\nu^* + 1)^2 \]
Focusing on $\mathbb{E}((iii))$, we get:

\[
\mathbb{E}((iii)) = \mathbb{E} \left[ \sum_{\nu=1}^{\nu^*} (\nu^* + 1)[t_{\nu-1}^*(t_{\nu-1}^* - 1)](A_{\nu-1} - A_{\nu}) \mathbb{I} \left\{ \tau_{\nu-1}^*(t_{\nu-1}^* - 1) > \frac{6q(t_{\nu-1}^*)}{\nu} \right\} \right]
\]

\[
\leq 2Mp \sum_{l=1}^{[\log_2(T)]} (\nu^* + 1)[t_{\nu-1}^*(t_{\nu-1}^* - 1)](A_{\nu-1} - A_{\nu}) \mathbb{I} \left\{ \tau_{\nu-1}^*(t_{\nu-1}^* - 1) > \frac{6q(t_{\nu-1}^*)}{\nu} \right\} \quad \text{(Peeling)}
\]

\[
\leq 4Mp \sum_{l=1}^{[\log_2(T)]} (\nu^* + 1)[t_{\nu-1}^*(t_{\nu-1}^* - 1)]P(\tau_{\nu-1}^*(t_{\nu-1}^* - 1) > \frac{6q(t_{\nu-1}^*)}{\nu})
\]

\[
\leq 2MpK \int_{t=0}^{\infty} 2^t \exp \left( -\frac{1}{q(t^* + 1)^\nu} p^2 t^\nu \right) dt
\]

\[
\leq \frac{36MpK(\nu^* + 1)}{p^2} \int_{u=1}^{\infty} \frac{\exp(-u)}{u} du \quad \text{(We use } 2[a] \geq [2a])
\]

\[
\leq 36MK(\nu^* + 1)
\]

which gives

\[
\mathbb{E}[R_\nu] \leq \frac{384M^2p \log(2K^2T^3)(\nu^* + 1)}{(\mu, g(M^*) - g(M_{\nu-1}^*))} + 42M(\nu^* + 1)K
\]

We now focus on the second term $R_\nu$. For a given $\nu$, two things may prevent a sub-optimal choice of arms $\mathcal{E}$ on which at least one player must be assigned. Either an arm in $\mathcal{E}$ is eliminated or an arm in $[K] \setminus \mathcal{E}$ is accepted. Lemma A.9 shows a condition under which a sub-optimal arm $i$ is eliminated:

**Lemma A.9** (Number of samples seen before a sub-optimal arm is eliminated). Fix $\nu$, let $\mathcal{E}_\nu^* = \text{support}(M_{\nu}^*)$ and let $i \notin \mathcal{E}_\nu^*$ be a sub-optimal arm. For any $t \geq 0$, if $q(2^{[\log_2(t)\nu]}) \geq q_{E,i}$ with

\[
q \geq q_{E,i} = \frac{8 \log(2K^2T^3)}{(\mu(\nu+1) - \mu_i)^2}
\]

then arm $i$ has necessarily been eliminated before time $t + \nu^*$.

**Proof of Lemma A.9** At time $t' = 2^{[\log_2(t)\nu]}$ we have that

\[
q(t') > \frac{8 \log(2K^2T^3)}{(\mu(\nu+1) - \mu_i)^2}
\]

\[
\implies 4\sqrt{\frac{\log(2K^2T^3)}{2q}} < \frac{\mu(\nu+1) - \mu_i}{2}
\]

\[
\implies \zeta_i + \zeta(\nu+1) < \frac{\mu(\nu+1) - \mu_i}{2}
\]

\[
\implies \mu_i + 2\zeta_i < \mu(\nu+1) - 2\zeta(\nu+1)
\]

\[
\implies \mu_i^R < \mu(\nu+1)
\]

Since $i \notin \mathcal{E}_\nu^*$, $\mu_i < \mu(\nu+1)$, the last line means that $i$ will be eliminated at the next update which will happen at the end of the Round Robin phase which can last up to $\nu^*$ rounds. \qed
Lemma A.10 shows a condition under which an optimal arm \( j \) is accepted:

**Lemma A.10** (Number of samples seen before an optimal arm is accepted). Fix \( \nu \), let \( \mathcal{E}_\nu^* = \text{support}(\mathbf{M}_E^*) \) and let \( j \in \mathcal{E}_\nu^* \) an optimal arm. If at time \( t \), arm \( j \) and \( (\nu) \) both been played without collision at least \( q_{A,i} \) times with

\[
q_{A,i} = \frac{8 \log(2T^3K^2)}{(\mu_j - \mu_{(\nu)})^2}
\]

then arm \( j \) has been accepted before time \( t + \nu^* \).

**Proof of Lemma A.10.** At time \( t' = 2^{\lceil \log_2(t) \rceil} \) we have that

\[
\min(T_j(t'), T_{(\nu)}(t')) > \frac{8 \log(2T^3K^2)}{(\mu_j - \mu_{(\nu)})^2} \implies 2 \sqrt{\frac{\log(2T^3K^2)}{2 \min(T_j(t'), T_{(\nu)}(t'))}} < \frac{\mu_j - \mu_{(\nu)}}{2}
\]

\[
\implies \zeta_{(\nu)} + \zeta_j < \frac{\mu_j - \mu_{(\nu)}}{2}
\]

\[
\implies \mu_{(\nu)} + 2\zeta_{(\nu)} < \mu_j - 2\zeta_j
\]

\[
\implies \mu_{(\nu)}^H < \mu_j^L
\]

Since \( j \in \mathcal{E}_\nu^*, \mu_j > \mu_{(\nu)}, \) the last line means that \( j \) will be accepted at the next update of \( \mathcal{A} \) which will happen at the end of the Round Robin phase which can last up to \( \nu^* \) rounds.

The two previous lemmas allow to quantify when arms are accepted or rejected. The next lemma measures the cost of choosing a sub-optimal set of arms on which at least one player must be assigned.

**Lemma A.11** (Cost of choosing a sub-optimal \( E \)). Let \( E \) a set of arms of size \( K - \nu \) such that \( \mathcal{E} \neq \mathcal{E}_\nu^* = \text{support}(\mathbf{M}_E^*) \). Then, we have:

\[
\langle \mu, g(\mathbf{M}_E^*) - g(\mathbf{M}_E^*) \rangle \leq\]

\[
p \left( \sum_{i \in \mathcal{E} \setminus \mathcal{E}_\nu^*} \mu_{(\nu+1)} - \mu_i \right) + \sum_{j \notin \mathcal{E}_\nu^* \setminus \mathcal{E}} (\mu_j - \mu_{(\nu)})
\]

**Proof of Lemma A.11.** Let \( \mathcal{E} \neq \mathcal{E}_\nu^* \) and define indexes \( i_1, \ldots, i_n \) by

\[
\mathcal{E} \setminus \mathcal{E}_\nu^* = \{i_1, \ldots, i_n\}
\]

and indexes \( j_1, \ldots, j_n \) by

\[
\mathcal{E}_\nu^* \setminus \mathcal{E} = \{j_1, \ldots, j_n\}
\]

We now construct \( \mathbf{M}_E \). Arms that are in \( \mathcal{E} \) but not in \( \mathcal{E}_\nu^* \) are assigned 1 player the corresponding players are taken from arms in \( \mathcal{E}_\nu^* \) but not in \( \mathcal{E} \). Formally

\[
\forall k \in [n], \mathbf{M}_E[i_k] = 1
\]

and

\[
\forall k \in [n], \mathbf{M}_E[j_k] = \mathbf{M}^*[j_k] - 1
\]

and other arms are untouched:

\[
\forall k \in \mathcal{E}_\nu^* \cap \mathcal{E}, \mathbf{M}_E[k] = \mathbf{M}^*[k]
\]
The cost is given by:

\[
\langle \mu, g(M^*_\nu) - g(M^*_\xi) \rangle \leq \langle \mu, g(M^*_\nu) - g(M^*E) \rangle \\
= \sum_{k=1}^{n} (\mu_{jk}[g(M^*[j_k]) - g(M^*[j_k] - 1)] - \mu_{ik}p) \\
\leq \sum_{k=1}^{n} (\mu_{jk} - \mu_{ik})p \\
\leq \sum_{k=1}^{n} (\mu_{jk} - \mu_{(\nu)} + \mu_{(\nu+1)} - \mu_{ik})p \\
\leq p \left( \sum_{i \in \mathcal{E}\setminus \mathcal{E}_{\nu}} \mu_{(\nu+1)} - \mu_i + \sum_{j \in \mathcal{E} \setminus \mathcal{E}_{(\nu)}} \mu_j - \mu_{(\nu)} \right)
\]

We can now bound \( R_\mathcal{E} \):

**Lemma A.12** (Bound on \( R_\mathcal{E} \)).

\[
\mathbb{E}[R_\mathcal{E}] \leq \sum_{\nu=1}^{\nu^*} \frac{672(\nu^* + 1) \log(2T^3K^2)}{\mu_{(\nu+1)} - \mu_{(\nu)}} + \nu^* Mp \frac{576(\nu^* + 1) \log(2K^2T^3)}{\Delta^{(\nu^*)}} + 200(\nu^* + 1)^2 + 5K^2
\]

**Proof of Lemma A.12**. Call \( t_\nu \) the last time that \( \nu(t) = \nu \) and set \( t_{\nu^*} = T + 1 \) and \( t_{-1} = 0 \). We can write

\[
R_\mathcal{E} = \sum_{t=1}^{T} \langle \mu, g(M^*_\nu(t)) - g(M^*_\xi(t)) \rangle
\]

\[
= \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} \langle \mu, g(M^*_\nu) - g(M^*_\xi(t)) \rangle \\
\leq \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \left( \sum_{i \in \mathcal{E}(t) \setminus \mathcal{E}_{\nu}} (\mu_{(\nu+1)} - \mu_i) + \sum_{j \in \mathcal{E} \setminus \mathcal{E}(t)} \mu_j - \mu_{(\nu)} \right) \quad \text{(Using Lemma A.11)}
\]

\[
= \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \left( \sum_{i \in \mathcal{E}(t) \setminus \mathcal{E}_{\nu}} (\mu_{(\nu+1)} - \mu_i) \right) + \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \left( \sum_{j \in \mathcal{E} \setminus \mathcal{E}(t)} \mu_j - \mu_{(\nu)} \right)
\]

Let us cut the execution of the algorithms in phases where phase \( n \) starts when it is the \( n \)-th time that the condition Line 3 in Algorithm 1 is satisfied. Note again that updates of \( \mathcal{A} \), \( \mathcal{K} \), and \( \nu \) occur at the beginning of each phase. Denote \( \mathcal{N}_{\nu} \) the phases between \( t_{\nu-1} + 1 \) and \( t_{\nu} \).
Bounding (i)  Denote \( \tau_n \) the number of pulls of active arms at the end of phase \( n \).

\[
(i) \leq \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_{\nu}} p \sum_{i \notin \mathcal{E}'_t} (\mu_{(\nu+1)} - \mu_i) \mathbb{I} \{ i \in \mathcal{E}(t) \}
\]

\[
= \sum_{\nu=1}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_{\nu}} p \sum_{i \notin \mathcal{E}'_t} (\mu_{(\nu+1)} - \mu_i) \mathbb{I} \{ i \in \mathcal{E}(t) \}
\]

\[
= \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} \sum_{t=t_{\nu-1}+1}^{t_{\nu}} p \sum_{i \notin \mathcal{E}'_t} (\mu_{(\nu+1)} - \mu_i) \mathbb{I} \{ i \in \mathcal{E}(t) \} \mathbb{I} \{ t \in \text{belong to phase } n \}
\]

\[
\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{i=1}^{\nu} \sum_{\nu', \nu' \in N_\nu} (\mu_{(\nu+1)} - \mu_i)( \text{Number of times arm } i \text{ is pulled during phase } n )
\]

\[
= \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} \sum_{i=1}^{\nu} \sum_{\nu', \nu' \in N_\nu} (\mu_{(\nu+1)} - \mu_i)( \text{Number of times arm } (i) \text{ is pulled during phase } n )
\]

\[
= \sum_{i=1}^{\nu^*} \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} (\mu_{(\nu+1)} - \mu_i)( \text{Number of times arm } (i) \text{ is pulled during phase } n )
\]

where \( (i) \) the index of the arm with reward \( \mu_{(i)} \).

Let \( T_{\nu, i} \) be the number of times arm \( (i) \) has been pulled in total at the end of the epoch where \( \nu(t) = \nu \). This means

\[
\sum_{\nu \in N_\nu} \text{Number of times arm } (i) \text{ is pulled during phase } n = T_{\nu, i} - T_{\nu-1, i}
\]

Call \( n_{E, (i)} \) the phase at which arm \( (i) \) is eliminated. Call \( \nu_i \) the epoch where arm \( (i) \) is eliminated. This means \( n_{E, (i)} \in N_{\nu_i} \).

Call \( s_i = \sum_{\nu=1}^{\nu^*} (\mu_{(\nu+1)} - \mu_i)(T_{\nu, i} - T_{\nu-1, i}) \).

We have

\[
s_i = (\mu_{(\nu+1)} - \mu_i)(T_{\nu, i} - T_{\nu-1, i}) + \sum_{\nu=1}^{\nu^*} (\mu_{(\nu+1)} - \mu_i)(T_{\nu, i} - T_{\nu-1, i})
\]

\[
= 2 (\mu_{(\nu+1)} - \mu_i)(T'_{\nu, i} - T'_{\nu-1, i}) + 2 \sum_{\nu=1}^{\nu^*} (\mu_{(\nu+1)} - \mu_i)(T'_{\nu, i} - T'_{\nu-1, i})
\]

where

\[
T'_{\nu, i} = \frac{T_{\nu, i} - (\nu^* + 1)}{2}
\]

Furthermore,

\[
a_i = (\mu_{(\nu+1)} - \mu_i)(T'_{\nu, i} - T'_{\nu-1, i}) \leq (\mu_{(\nu+1)} - \mu_i)T'_{\nu, i} + (\mu_{(\nu+1)} - \mu_i)(\nu^* + 1)
\]

and we can write

\[
b_i \leq (\mu_{(\nu)} - \mu_i) \sum_{\nu=1}^{\nu^*} (T'_{\nu, i} - T'_{\nu-1, i})
\]

\[
\leq (\mu_{(\nu)} - \mu_i)T'_{\nu-1, i} + \nu^* (\mu_{(\nu)} - \mu_i)
\]
Let us then notice that

\[
\begin{align*}
\mathbb{E}[(\mu_{(\nu^*)} - \mu_{(i)})T'_{\nu^*,i-1}] & \\ 
\leq 2\mathbb{E}[(\nu^* + 1)^2 + 2(\mu_{(\nu^*)} - \mu_{(i)}) (\nu^* + 1) [\tau_{\nu^*}(T'_{\nu^*,i-1},i)] 1 \{ T'_{\nu^*,i-1} \geq (\nu^* + 1)^2 \} ] \quad \text{(By Lemma A.5)} \\
& \leq 2(\nu^* + 1)^2 + 2\mathbb{E}[(\mu_{(\nu^*)} - \mu_{(i)}) (\nu^* + 1) [\tau_{\nu^*}(T'_{\nu^*,i-1},i)] 1 \{ T'_{\nu^*,i-1} \geq (\nu^* + 1)^2 \} ] \\
& \leq 2(\nu^* + 1)^2 + 2\mathbb{E}[(\mu_{(\nu^*)} - \mu_{(i)}) (\nu^* + 1) \frac{6q(T'_{\nu^*,i-1})}{p} 1 \{ T'_{\nu^*,i-1} \geq (\nu^* + 1)^2 \} ] \\
& + 2\mathbb{E}[(\mu_{(\nu^*)} - \mu_{(i)}) (\nu^* + 1) [\tau_{\nu^*}(T'_{\nu^*,i-1},i)] 1 \{ [\tau_{\nu^*}(T'_{\nu^*,i-1},i)] > \frac{6}{p} q(T'_{\nu^*,i-1}) \} ] \\
& \leq 2(\nu^* + 1)^2 + \mathbb{E}[(\mu_{(\nu^*)} - \mu_{(i)}) (\nu^* + 1) \frac{12q(T'_{\nu^*,i-1})}{p} 1 \{ T'_{\nu^*,i-1} \geq (\nu^* + 1)^2 \} ] + 36K(\nu^* + 1) \\
& + \frac{36K(\nu^* + 1)}{p}
\end{align*}
\]

where the last inequality follows from the same steps used to bound \((iii)\) in the proof of Lemma A.8.

We therefore get

\[
\mathbb{E}[p \sum_{i=1}^{\nu^*} b_i] \leq 3(\nu^* + 1)^3 p + \mathbb{E}[\sum_{i=1}^{\nu^*} (\mu_{(\nu^*)} - \mu_{(i)}) (\nu^* + 1) 12q(T'_{\nu^*,i}) 1 \{ T'_{\nu^*,i} \geq (\nu^* + 1)^2 \} ] + 36K(\nu^* + 1)^2
\]

Applying the same steps for \(a_i\) yields

\[
\mathbb{E}[p \sum_{i=1}^{\nu^*} a_i] \leq 3(\nu^* + 1)^3 p + \mathbb{E}[\sum_{i=1}^{\nu^*} (\mu_{(\nu^*)} + 1) - \mu_{(i)}) (\nu^* + 1) 12q(T'_{\nu^*,i}) 1 \{ T'_{\nu^*,i} \geq (\nu^* + 1)^2 \} ] + 36K(\nu^* + 1)^2
\]

By Lemma A.9 we have \(q(T'_{\nu^*,i}) \leq \frac{8\log(2T^2K^2)}{(\mu_{(\nu^*)} - \mu_{(i)})^2}\) so that

\[
\mathbb{E}[p \sum_{i=1}^{\nu^*} a_i] \leq 3(\nu^* + 1)^3 p + \mathbb{E}[\sum_{i=1}^{\nu^*} (\nu^* + 1) 12 \frac{8\log(2T^2K^2)}{\mu_{(\nu^*)} + 1) - \mu_{(i)}} + 36K(\nu^* + 1)^2
\]

By Lemma A.9 \(q(T'_{\nu^*,i}) \leq \frac{8\log(2T^2K^2)}{(\mu_{(\nu^*)} - \mu_{(i)})^2}\) and by Lemma A.7 \(q(T'_{\nu^*,i}) \leq \frac{8M^2p^2\log(2T^2K^2)}{\Delta_{\nu^*,i-1}^2}\) so that

\[
q(T'_{\nu^*,i}) \leq \min \left( \frac{8M^2p^2\log(2T^2K^2)}{\Delta_{\nu^*,i-1}^2}, \frac{8\log(2T^2K^2)}{(\mu_{(\nu^*)} - \mu_{(i)})^2} \right)
\]

\[
\leq \sqrt{\frac{8M^2p^2\log(2T^2K^2)}{\Delta_{\nu^*,i-1}^2}} \frac{8\log(2T^2K^2)}{(\mu_{(\nu^*)} - \mu_{(i)})^2}
\]

\[
= \frac{8Mp\log(2T^2K^2)}{\Delta_{\nu^*,i-1}(\mu_{(\nu^*)} - \mu_{(i)})}
\]

and therefore

\[
\mathbb{E}[p \sum_{i=1}^{\nu^*} b_i] \leq 3(\nu^* + 1)^3 p + \mathbb{E}[\sum_{i=1}^{\nu^*} (\nu^* + 1) 96 \frac{M\log(2T^2K^2)}{\Delta_{\nu^*,i-1}^2} + 36K(\nu^* + 1)^2
\]

so that

\[
(i) \leq 12(\nu^* + 1)^3 p + \sum_{i=1}^{\nu^*} (\nu^* + 1) 192 \left[ \frac{\log(2T^2K^2)}{(\mu_{(\nu^*)} + 1) - \mu_{(i)}} + \frac{M\log(2T^2K^2)}{\Delta_{\nu^*,i-1}} \right] + 144K(\nu^* + 1)^2
\]
Then either \( \nu_i = \nu^* \) and
\[
(a) \leq \frac{\log(2T^2K^2)}{\mu(\nu^*+1) - \mu(i)} + \frac{M \log(2K^2)T}{\Delta_{\nu^* - 1}}
\]
or \( \nu_i < \nu^* \) and then,
\[
s_i = \sum_{\nu=i}^{\nu_i} (\mu_{\nu+1} - \mu(i)) (T_{\nu,i} - T_{\nu-1,i})
\]
\[
\leq (\mu_{\nu+1} - \mu(i)) \sum_{\nu=i}^{\nu_i} (T_{\nu,i} - T_{\nu-1,i})
\]
\[
\leq (\mu_{\nu+1} - \mu(i)) T_{\nu,i}
\]
\[
\leq 2(\nu^* + 1)^2 + 2(\nu^* + 1) \frac{6q(T'_{\nu,i})}{p} (\mu_{\nu+1} - \mu(i)) + \frac{36K(\nu^* + 1)}{p} \tag{Similar as the bound of \( b_i \)}
\]
\[
\leq 2(\nu^* + 1)^2 + 2(\nu^* + 1) \frac{48M \log(2K^2T^2)}{\Delta_{\nu,i}} + \frac{36K(\nu^* + 1)}{p}
\]
\[
\leq 2(\nu^* + 1)^2 + 2(\nu^* + 1) \frac{48M \log(2K^2T^2)}{\Delta_{\nu^* - 1}} + \frac{36K(\nu^* + 1)}{p}
\]

where at the last line we used again Lemma \[A.9\] and Lemma \[A.7\]

So in any case
\[
(i) \leq 12(\nu^* + 1)^3p + 180K(\nu^* + 1)^2 + \sum_{i=1}^{\nu^*} \frac{288(\nu^* + 1) \log(2T^2K^2)}{\mu(\nu^*+1) - \mu(i)} + p \nu^* \frac{192M(\nu^* + 1) \log(2K^2T^2)}{\Delta_{\nu^* - 1}}
\]

**Bounding (ii)** we have
\[
(ii) \leq \sum_{\nu=0}^{\nu^*} \sum_{n \in N_\nu} p \sum_{j \in E_\nu^*} (\mu_j - \mu(\nu)) \mathbb{I} \{ j \notin E(t) \} \mathbb{I} \{ t \text{ belong to phase } n \}
\]
\[
\leq \sum_{\nu=0}^{\nu^*} \sum_{n \in N_\nu} p \sum_{j \in E_\nu^*} (\mu_j - \mu(\nu))( \text{ Number of times arm } j \text{ is not pulled during phase } n )
\]
\[
= \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{j \in E_\nu^*} (\mu_j - \mu(\nu))( \text{ Number of times arm } j \text{ is not pulled during phase } n )
\]
\[
\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{j \in E_\nu^*} (\mu_j - \mu(\nu)) n_{\nu A,j} \mathbb{I} \{ n \leq n_{A,j} \} \mathbb{I} \{ \text{ Last phase where arm } j \text{ is not accepted } \}
\]
\[
\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{j \in E_\nu^*} (\mu_j - \mu(\nu)) \sum_{\nu' = 1}^{\nu} \mathbb{I} \{ n \leq n_{E,j,\nu'} \} \mathbb{I} \{ \text{ Last phase where arm } \nu' \text{ is not rejected } \}
\]
\[
\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{j \in E_\nu^*} (\mu_j - \mu(\nu)) \sum_{\nu' = 1}^{\nu} \mathbb{I} \{ n \leq n_{E,\nu'} \} \mathbb{I} \{ n \leq n_{A,j} \}
\]
\[
= \sum_{\nu'=1}^{\nu^*} p \sum_{\nu'=1}^{\nu^*} \sum_{n \in N_{\nu'}} (\mu_j - \mu(\nu)) \mathbb{I} \{ n \leq n_{E,\nu} \} \mathbb{I} \{ n \leq n_{A,j} \}
\]
Call $n_\nu$ the last phase before $\nu$ is increased. We have that
\[
(A) = \sum_{\nu=\nu'} \sum_{n \in \mathcal{N}_\nu} \sum_{j=\nu+1}^K (\mu(j) - \mu(\nu)) \mathbb{1}\{n \leq n_{E,\nu'}\} \mathbb{1}\{n \leq n_{A,j}\}
\]

Call $\tau_n$ the value of $\tau$ at the end of phase $n$, $t_n$ the value of $t$ at the end of phase $n$ and set
\[
t'_n = (t_n - (\nu^* + 1))/2
\]

Notice that if $\tau_n \geq 8(\nu^* + 1)^2$, then
\[
\tau(t'_n) \geq \frac{t'_n}{\nu^* + 1} - \nu^* = \frac{t_n - (\nu^* + 1)}{2(\nu^* + 1)} - \nu^* = \frac{t_n}{2(\nu^* + 1)} - \frac{1}{2} - \nu^* \geq \frac{\tau_n}{2(\nu^* + 1)} - \frac{1}{2} - \nu^* \geq \frac{\tau_n}{4(\nu^* + 1)}
\]

We then use that
\[
\sum_{j=\nu+1}^K \mathbb{1}\{n \leq n_{A,j}\} = \text{Number of arms not yet accepted at phase } n
\]

\[
= \tau_n - \tau_{n-1}
\]
We have:

\[ A_2 \leq \sum_{\nu=\nu'}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} (\tau_n - \tau_{n-1}) \mathbb{1}\{\tau_{n-1} \leq 8(\nu^* + 1)^2\} \leq 8(\nu^* + 1)^2 \]

Let us then focus on \( A_3 \):

\[
A_3 = \sum_{\nu=\nu'}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} \sum_{j=\nu+1}^{K} (\mu_{ij}) - \mu_{ij}) \mathbb{1}\{n \leq n_{E,\nu'}\} \mathbb{1}\{n \leq n_{A,j}\} \mathbb{1}\{\tau_{n-1} \geq 8(\nu^* + 1)^2\} \mathbb{1}\{q(t_{n-1}') < \frac{1}{3} \} \mathbb{1}\{p\tau(t_{n-1}')\}
\]

\[
\leq \sum_{n \in \mathbb{N}} (\tau_n - \tau_{n-1}) \mathbb{1}\{q(t_{n-1}') < \frac{1}{3} \} \mathbb{1}\{p\tau(t_{n-1}')\}
\]

\[
\leq \sum_{n \in \mathbb{N}} K \mathbb{1}\{q(t_{n-1}') < \frac{1}{3} \} \mathbb{1}\{p\tau(t_{n-1}')\}
\]

where we use

\[
\tau_n - \tau_{n-1} = |\mathcal{K}_n \setminus A_n| - (\nu_n - 1) = |K| - |\mathcal{K}_n| - |A_n| - \nu_n + K - |\mathcal{K}_n| = K - |A_n| - \nu_n \leq K
\]

We have:

\[
E[A_3] \leq \sum_{n \in \mathbb{N}} P(q(t_{n-1}') < \frac{1}{3} \) \mathbb{1}\{p\tau(t_{n-1}')\} \leq K^2 \sum_{n \in \mathbb{N}, \tau(t_{n-1}') > 0} \exp\left(-\frac{2}{9} (p\tau(t_{n-1}')\right) \leq K^2 \int_{0}^{\infty} \exp(-\frac{2}{9} pt) dt \quad \text{(At each phase, all active arms are played at least one time)} \leq \frac{9K^2}{2p}
\]

Then we turn to \( A_1 \). Call \( E_n \) the event

\[
E_n = \{q(t_{n-1}') \geq \frac{1}{3} \) \mathbb{1}\{p\tau(t_{n-1}')\}, \tau_{n-1} \geq 8(\nu^* + 1)^2\}
\]

Under \( E_n \), we can write

\[
n \leq n_{A,j} \implies q(t_{n-1}') \leq q_{A,j} \implies \frac{1}{3} \) \mathbb{1}\{p\tau(t_{n-1}')\} \leq q_{A,j} \implies \frac{1}{3} \) \mathbb{1}\{p\tau_{n-1} \geq q_{A,j} \implies \frac{1}{3} \) \mathbb{1}\{p(\nu^* + 1) \leq q_{A,j} \implies \tau_{n-1} \leq \frac{96(\nu^* + 1) \log(2T^3K^2)}{(\mu_{ij} - \mu_{ij})^2 p} \leq \frac{96(\nu^* + 1) \log(2T^3K^2)}{\tau_{n-1}} \leq \delta_n \quad \text{(By Lemma A.10)}
\]

\[
\implies \mu_{ij} - \mu_{ij} \leq \sqrt{\frac{96(\nu^* + 1) \log(2T^3K^2)}{\tau_{n-1}}} \leq \delta_n
\]
Calling \( n_\nu \) the last phase before \( \nu \) is increased, we have

\[
A_1 = \sum_{\nu=\nu'}^{\nu^*} \sum_{n \in N_\nu, \nu = j + 1}^{n^* - 1} (\mu(j) - \mu(\nu)) \mathbb{I}\{n \leq n_{E,\nu'}\} \mathbb{I}\{n \leq n_{A,j}\} \mathbb{I}\{E_n\}
\]

\[
= \sum_{\nu=\nu'}^{n^* - 1} \sum_{n \in N_\nu, j = \nu + 1}^{n^*} (\mu(j) - \mu(\nu)) \mathbb{I}\{n \leq n_{E,\nu'}\} \mathbb{I}\{n \leq n_{A,j}\} \mathbb{I}\{E_n\}
\]

\[
+ \sum_{j = \nu + 1}^{n^*} \sum_{n \in N_\nu, \nu = j}^{n^*} (\mu(j) - \mu(\nu)) \mathbb{I}\{n \leq n_{E,\nu'}\} \mathbb{I}\{n \leq n_{A,j}\} \mathbb{I}\{E_n\}
\]

\[
\leq \sum_{n \in \mathbb{N}} (\tau_n - \tau_{n-1}) \delta_n \left( \mathbb{I}\{n \leq n_{\nu^* - 1}\} + \mathbb{I}\{n \leq n_{E,\nu}\} \right) \mathbb{I}\{E_n\}
\]

Using the identity \( \frac{\sqrt{96(\nu^* + 1) \log(2T^3K^2)}}{\tau_{n-1}p} = \delta_n \), we get

\[
\tau_n - \tau_{n-1} = \frac{96(\nu^* + 1) \log(2T^3K^2)}{p} \left( \frac{1}{\delta_{n+1}^2 - \delta_n^2} \right)
\]

\[
= \frac{96(\nu^* + 1) \log(2T^3K^2)}{p} \left( \frac{1}{\delta_n} + \frac{1}{\delta_{n+1}} \right) \left( \frac{1}{\delta_{n+1}} - \frac{1}{\delta_n} \right)
\]

Note that \( \tau_n - \tau_{n-1} \leq K \) and \( \tau_0 = K \) (since all arms are active at the first iteration) so that \( 2\tau_{n-1} \geq \tau_{n-1} + K \geq \tau_n \). This implies

\[
\frac{\delta_{n+1} - \delta_n}{\delta_n} = \sqrt{\frac{\tau_n}{\tau_{n-1}}} \leq \sqrt{2}.
\]

We can then write:

\[
A_1 \leq \frac{96(\nu^* + 1) \log(2T^3K^2)}{p} \left( \sqrt{2} + 1 \right) \sum_{n \in \mathbb{N}} \left( \frac{1}{\delta_{n+1}} - \frac{1}{\delta_n} \right) \left( \mathbb{I}\{n \leq n_{\nu^* - 1}\} + \mathbb{I}\{n \leq n_{E,\nu}\} \right) \mathbb{I}\{E_n\}
\]

Under \( E_n \), we have

\[
n \leq n_{E,\nu} \implies q(t_{n-1}') \leq q_{E,\nu^*}
\]

\[
\implies \frac{1}{3} \mu_{\nu}(t_{n-1}') \leq q_{E,\nu^*}
\]

\[
\implies \mu(t_{n-1}') \leq \frac{24 \log(2T^3K^2)}{p(\mu(\nu^* + 1) - \mu(\nu))^2}
\]

\[
\implies \mu(\nu^* + 1) - \mu(\nu) \leq \sqrt{\frac{96(\nu^* + 1) \log(2T^3K^2)}{\tau_{n-1}p}} = \delta_n
\]

so that

\[
\mu(\nu^* + 1) - \mu(\nu) \leq \delta_{n,\nu,\nu}
\]

Similarly, under \( E_n \), we have:

\[
n \leq n_\nu \implies q(t_{n-1}') \leq q_\nu
\]

\[
\implies \frac{1}{3} \mu_{\nu}(t_{n-1}') \leq q_\nu
\]

\[
\implies \mu(t_{n-1}') \leq \frac{24M^2p^2 \log(2T^3K^2)}{(\Delta(\nu))^2p}
\]

\[
\implies \frac{\Delta(\nu)}{Mp} \leq \sqrt{\frac{96(\nu^* + 1) \log(2T^3K^2)}{\tau_{n-1}p}} = \delta_n
\]
so that
\[
\frac{\Delta(\nu^*)}{M_p} \leq \frac{\Delta(\nu)}{M_p} \leq \delta_n
\]
Using again that \( \frac{1}{\delta_n} \leq \sqrt{2} \frac{1}{\delta_{n-1}} \), we get
\[
A_1 \leq \frac{384(\nu^* + 1) \log(2T^3K^2)}{p} \frac{1}{\mu(\nu^*+1) - \mu_{\nu'}} + M_p \frac{384(\nu^* + 1) \log(2T^3K^2)}{p} \frac{1}{\Delta_{\nu^* - 1}}
\]
where we used \( 2 + \sqrt{2} \leq 4 \)
so that
\[
(ii) \leq \sum_{\nu'=1}^{\nu^*} \frac{384(\nu^* + 1) \log(2T^3K^2)}{\mu(\nu^*+1) - \mu_{\nu'}} + \nu^* M_p \frac{384(\nu^* + 1) \log(2T^3K^2)}{\Delta(\nu^*)} + \frac{9K^2}{2} + 8p(\nu^* + 1)^2
\]
From the bound of (i) and (ii), we get:
\[
R_\mathcal{E} \leq \sum_{\nu'=1}^{\nu^*} \frac{672(\nu^* + 1) \log(2T^3K^2)}{\mu(\nu^*+1) - \mu_{\nu'}} + \nu^* M_p \frac{576(\nu^* + 1) \log(2K^2T^3)}{\Delta(\nu^*)} + 200K(\nu^* + 1)^2 + 5K^2
\]
It remains to bound \( R_M \). Recall that \( R_M \) measures the mismatch between the chosen assignment \( \mathbf{M}(t) \) and the best possible assignment with the same support. Crucially there is no support mismatch and therefore we are in a setting close to the full information setting which allows us to bound \( R_M \) by a quantity independent of the horizon \( T \).

**Lemma A.13** (Bound on \( R_M \)).
\[
\mathbb{E}[R_M] \leq 4M_p(\nu^* + 1)^2 + 2MK \frac{6(\nu^* + 1)}{r}
\]

**Proof of Lemma A.13** The proof of Lemma A.13 follows similar techniques as \cite{Huang:2017}.
Then we write

\[
(i) \leq \sum_{t=2(\nu^*+1)^2}^T \mathbb{E}[\langle \mu - \hat{\mu}(t'), g(M^*_E(t)) - g(M(t)) \rangle_{Y_t} \bigg| Y_t]
\]

(With \( t' = 2^{\lceil \log_2(t) \rceil} \))

\[
= \sum_{t=2(\nu^*+1)^2}^T \mathbb{E}[Y_t \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\} + Y_t \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\}]
\]

\[
\leq \sum_{t=2(\nu^*+1)^2}^T \mathbb{E}[Y_t \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\}] + 2MpP(q(t') \leq \frac{p}{3}\tau(t'))
\]

\[
\leq \sum_{t=2(\nu^*+1)^2}^T \mathbb{E}[Y_t \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\}] + \sum_{t=2(\nu^*+1)^2}^T 2KMp\exp(-\frac{2}{9}p\tau(t'))
\]

Bounding \((b)\):

\[
(b) \leq \sum_{t=2(\nu^*+1)^2}^T 2KMp\exp(-\frac{2}{9}p(t/2))
\]

(Since \( t/2 \leq t' \) and \( \tau \) increasing)

\[
\leq \sum_{t=2(\nu^*+1)^2}^T 2KMp\exp(-\frac{2}{9}p\left(\frac{t}{2(\nu^*+1)} - \nu^*\right))
\]

(By Lemma A.5)

\[
\leq \sum_{t=1}^T 2MKp\exp(-\frac{2}{9}p\left(\frac{t}{2(\nu^*+1)}\right))
\]

\[
\leq \frac{18MKp(\nu^* + 1)}{p}
\]

Bounding \((a)\):

\[
(a) \leq \sum_{t=2(\nu^*+1)^2}^T Mp\mathbb{E}\left[\|\mu - \hat{\mu}(t')\|_{\infty} \mathbf{1}\{\|\mu - \hat{\mu}(t')\|_{\infty} \geq r\} \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\}\right]
\]

(By Equation [9] and definition of \( r \))

\[
\leq \sum_{t=2(\nu^*+1)^2}^T 2Mp\mathbb{E}\left[r\mathbb{P}\{\|\mu - \hat{\mu}(t')\|_{\infty} \geq r\} + \int_r^\infty \mathbb{P}\{\|\mu - \hat{\mu}(t')\|_{\infty} \geq \varepsilon\}d\varepsilon \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\}\right]
\]

\[
\leq \sum_{t=2(\nu^*+1)^2}^T \mathbb{E}[2MpK\left(r\exp(-2q(t')r^2) + \int_r^\infty \exp(-2q(t')\varepsilon^2)d\varepsilon \mathbf{1}\{q(t') \leq \frac{p}{3}\tau(t')\}\right]
\]
The regret

then we move on to Cautious Greedy with limited communication and show

the upper bound of quasi-centralized Cautious Greedy in Proposition 3.1 follows by combining the previous

Lemma A.14. The regret of Cautious Greedy with limited communication is given by:

\[ \mathbb{E}[R_{MG}] \leq 4M^2 p \frac{2^{K^2 T^2}}{D^2(\nu^* \nu^* + 1)} + \sum_{\nu' = 1}^{\nu^*} 672(\nu^* + 1) \log(2T^3 K^2) + 265MK(\nu^* + 1) \]

where all constants have been put in the term in \( \nu \).

Then we move on to Cautious Greedy with limited communication and show

Lemma A.14. The regret of Cautious Greedy with limited communication is given by:

\[ \mathbb{E}[R_{CG}] \leq 4R_{UB}^{CG} + \frac{16M^2 p \log(2K^2 T^2)}{D^2(\nu^* \nu^* + 1)} + \sum_{\nu' = 1}^{\nu^*} 672(\nu^* + 1) \log(2MT^2)K^2 \]

Proof. Call \( E_m \) the event "\( \forall \nu \in [M], \) at least one call of \( SEND_m \) and at least one call of \( SEND_{\text{gateway}} \) succeeds".

Call \( s^* = \min\{s, 2^s \geq \frac{8M^2 p \log(2MT^2)}{D^2(1-p_g)} \} \), we have that

\[ \forall s \geq s^* \mathbb{P}(E_s) \leq 2M(1-p_g)^{2^s-2} \leq \frac{1}{T^2} \]

and in particular,

\[ \mathbb{P}(\cup_{s=s^*} E_s) \leq \frac{1}{T} \]

The regret \( R_{CG} \) can therefore be decomposed as

\[ \mathbb{E}[R_{CG}] \leq \sum_{s=1}^{s^*} \sum_{t=2}^{2^{s+1}} \mathbb{E}[R_{CG}(t)] + \sum_{s=s^* + 1}^{\left\lfloor \frac{\log(T)}{2} \right\rfloor} \sum_{t=2^s}^{2^{s+1}} \mathbb{E}[R_{CG}(t) \mathbb{1}\{s=s^* \} + 2Mp] \]
Under \( \cap_{s=s^*}^{\log(T)} E_s \), stage \( l > s^* \) uses \( 2^s \) samples for each phase \( s \leq l - 2 \) but only \( 2^{s-1} \) samples from phase \( l - 1 \). In particular it uses \( 2^{s-1} \) samples from each phase up to phase \( l - 1 \). Therefore:

\[
(ii) \leq 2R_{CG}^{UB}
\]

For (i), we can write

\[
(i) = (i)(\mathbb{1}\{\nu^* = 0\} + \mathbb{1}\{\nu^* \neq 0\}) \\
\leq (i)\mathbb{1}\{\nu^* = 0\} + 16M^2 p \frac{\log(2M T^2)}{\log(1/(1-p_g))} \mathbb{1}\{\nu^* \neq 0\}
\]

Then we have

\[
(i)\mathbb{1}\{\nu^* = 0\} \leq \sum_{s=\log_2(8M)}^{s^*} \sum_{t=2^s}^{2^{s+1}} \mathbb{E}[R_{CG}(t) \mathbb{1}\{\nu^* = 0\} (\mathbb{1}\{E_s\} + \mathbb{1}\{\bar{E}_s\})]
\]

\[
\leq 2R_{CG}^{UB} \mathbb{1}\{\nu^* = 0\} + \sum_{s=\log_2(8M)}^{s^*} 2M p 2^s \mathbb{E}[\mathbb{1}\{\bar{E}_s\}]
\]

and we have:

\[
\sum_{s=\log_2(8M)}^{s^*} 2^s \mathbb{E}[\mathbb{1}\{\bar{E}_s\}] \leq \sum_{s=\log_2(8M)}^{s^*} 2^s \exp\left(\frac{\log(1-p_g)}{8M} 2^s\right)
\]

\[
\leq \frac{8M}{\log(1-p_g)}
\]

which concludes the proof.

Summing up, the upper bound of cautious greedy with limited communication is given by:

\[
\mathbb{E}[R_{CG}] \leq 4R_{CG}^{UB} + \frac{16M^2 p}{-\log(1-p_g)} + 16M^2 p \frac{\log(2M T^2)}{\log(1/(1-p_g))} \mathbb{1}\{\nu^* \neq 0\} + 2Mp + \log_2(8M)2Mp
\]

\[
\leq \frac{3840M^2 p \log(2K^2 T^3)(\nu^* + 1)}{\Delta(\nu^*)} + \sum_{\nu=1}^{\nu^*} \frac{2688(\nu^* + 1) \log(2T^3 K^2)}{\mu(\nu^* + 1) - \mu_\nu} + \frac{1078MK(\nu^* + 1)}{r}
\]

\[
+ \frac{16M^2 p}{-\log(1-p_g)} (1 + \log(2M T^2) \mathbb{1}\{\nu^* \neq 0\})
\]

A.4 Proof of Lemma 4.1

Proof. We assume \( M = 2N + 1 \). Take \( m_1 = \frac{1}{2}, m_2 = \frac{1}{2} + \Delta, \mu_1 = (m_1, m_2) \) and \( \mu_2 = (m_2, m_1) \).

Condition on \( \Delta \) such that \( \mathbf{M}^* = (N, N + 1) \) if \( \mu = \mu_1 \) and \( \mathbf{M}^* = (N + 1, N) \) if \( \mu = \mu_2 \). Let us first find \( \Delta \) such that the optimal assignment is \( (N, N + 1) \) when \( \mu = \mu_1 \) and \( (N + 1, N) \) when \( \mu = \mu_2 \). Assume \( \mu = \mu_1 \), the reasoning is symmetric for \( \mu = \mu_2 \). We want to find \( \Delta \) such that for any \( -(N+1) \leq x \leq N \) such that \( x \neq 0 \):

\[
g(N-x)\frac{1}{2} + g(N+1+x)(\frac{1}{2} + \Delta) \leq g(N)\frac{1}{2} + g(N+1)(\frac{1}{2} + \Delta)
\]

(12)
First for \( x = N \) we look for \( \Delta \) in the form \( \Delta = O(p) \)
\[
g(2N + 1)\left(\frac{1}{2} + \Delta\right) \leq g(N)\left(\frac{1}{2} + \Delta\right) + g(N + 1)\left(\frac{1}{2} + \Delta\right)
\]
\[
\iff (2N + 1)(1 - p)2N\left(\frac{1}{2} + \Delta\right) \leq N\left(\frac{1}{2} + (N + 1)(1 - p)\left(\frac{1}{2} + \Delta\right)\right)
\]
\[
\iff (2N + 1)(1 - p)\left(\frac{1}{2} + \Delta\right) \leq N\left(\frac{1}{2} + (N + 1)(1 - p)\right)\left(\frac{1}{2} + \Delta\right)
\]
(Using \( (1 - p)2N \leq (1 - p) \))
\[
\iff N(1 - p)\left(\frac{1}{2} + \Delta\right) \leq N\left(\frac{1}{2} + (N + 1)(1 - p)\right)\left(\frac{1}{2} + \Delta\right)
\]
\[
\iff \Delta \leq \left(\frac{1}{2} - \frac{1}{1 - p}\right)
\]
\[
\iff \Delta \leq \frac{p}{2(1 - p)}
\]

Then if \( \Delta \leq \frac{p}{1 - p} \), Equation \((12)\) is satisfied for \( x = N \).

For \( x = -(N + 1) \), the left-hand side of Equation \((12)\) is \( g(2N + 1)\left(\frac{1}{2} + \Delta\right) \) so if \( \Delta \leq \frac{p}{1 - p} \), Equation \((12)\) is satisfied for \( x = -(N + 1) \).

For \( 0 < x < N \), we have
\[
g(N - x)\left(\frac{1}{2} + \Delta\right) \leq g(N)\left(\frac{1}{2} + \Delta\right) + g(N + 1)\left(\frac{1}{2} + \Delta\right)
\]
\[
\iff (N - x)\left(\frac{1}{2} + (N + 1 + x)(1 - p)1 + (1 - p)^{2x}\left(\frac{1}{2} + \Delta\right) \leq N(1 - p)^{x} + (N + 1)(1 - p)^{x + 1}\left(\frac{1}{2} + \Delta\right)
\]
(Using \( (1 - p)^{2x + 1} \leq (1 - p)^{x + 1} \))
\[
\iff (x(1 - p)^{x + 1})\left(\frac{1}{2} + \Delta\right) \leq x\left(\frac{1}{2} - \frac{1}{1 - p}\right)\]
(Using \( (1 - p)^{x} \leq \frac{1}{2(1 - p)} \))
\[
\iff \Delta \leq \frac{1}{2}\left(-\frac{1}{1 - p}\right)
\]
\[
\iff \Delta \leq \frac{p}{2(1 - p)}
\]

Therefore if \( \Delta \leq \frac{p}{1 - p} \), Equation \((12)\) is satisfied for \( 0 < x < N \).

For \( -(N + 1) < x < 0 \), set \( y = -x - 1 \) so that \( x = y - 1 \) and \( 0 \leq y \leq N \). We can write \( g(N - x)\left(\frac{1}{2} + \Delta\right) = g(N + y + 1)\left(\frac{1}{2} + \Delta\right) = g(N + y + 1)\left(\frac{1}{2} + \Delta\right) \) which gives the desired inequality for \( y = 0 \). For \( y > 0 \), Equation \((12)\) is satisfied if \( \Delta \leq \frac{p}{1 - p} \). Therefore if \( \Delta \leq \frac{p}{1 - p} \), Equation \((12)\) is satisfied and therefore the optimal assignment if \( \mu = \mu_1 \) is \( \mathbf{M}^* = (N, N + 1) \).

**Computing \( r \)** Let us now compute \( r \). Assume again \( \mu = \mu_1 \) and the reasoning is symmetric for \( \mu = \mu_2 \). We have \( \mathcal{M}_1 \cup \mathcal{M}_0 = \mathcal{M} \) and we know \( \mathbf{M}^* = (N, N + 1) \) so that \( r = \min_{\mu' \in \mathcal{M}} ||\mu' - \mu||_\infty \). Call \( \mu_r = \arg\min_{\mu' \in \mathcal{M}} ||\mu' - \mu||_\infty \) and \( \mu_r = \arg\max_{\mu \in \mathcal{M}} ||\mu - g(M)|| \).

Since the number of players assigned to an arm increases with the reward of this arm, we have either \( \mu_r = \mu + (r_1, -r_1) \) and then \( \mathbf{M}_r = \mathbf{M}^* + (1, -1) \) or \( \mu_r = (-r_2, r_2) \) and then \( \mathbf{M}_r = \mathbf{M}^* + (-1, 1) \).

\( r_1 \) is the minimum value such that
\[
g(N + 1)\left(\frac{1}{2} + r_1\right) + g(N)\left(\frac{1}{2} + \Delta - r_1\right) \geq g(N)\left(\frac{1}{2} + r_1\right) + g(N + 1)\left(\frac{1}{2} + \Delta - r_1\right)
\]
\[
\iff (g(N + 1) - g(N))\left(\frac{1}{2} + r_1\right) \geq (g(N + 1) - g(N))\left(\frac{1}{2} + \Delta - r_1\right)
\]
and therefore \( r_1 = \frac{3}{2} \).

\( r_2 \) is the minimum value such that

\[
g(N-1)\left(\frac{1}{2} - r_2\right) + g(N+2)\left(\frac{1}{2} + \Delta + r_2\right) \geq g(N)\left(\frac{1}{2} - r_2\right) + g(N+1)\left(\frac{1}{2} + \Delta + r_2\right)\]

\( \iff \)

\[
(g(N+2) - g(N+1))(\frac{1}{2} + \Delta + r_2) \geq (g(N) - g(N-1))(\frac{1}{2} - r_2)
\]

\( \iff \)

\[
r_2(g(N+2) - g(N+1) + g(N) - g(N-1)) \geq (g(N+1) - g(N+2))(\frac{1}{2} + \Delta) + (g(N) - g(N-1))\frac{1}{2}
\]

\( \iff \)

\[
r_2 \geq \frac{(g(N+1) - g(N+2))(\frac{1}{2} + \Delta) + (g(N) - g(N-1))\frac{1}{2}}{2(g(N) - g(N-1))}
\]

\( \iff \)

\[
r_2 \geq \frac{((N+1)(1-p)^2 - (N+2)(1-p)^3)(\frac{1}{2} + \Delta) + (N(1-p) - (N-1))\frac{1}{2}}{2(N(1-p) - N - 1)}
\]

\( \iff \)

\[
r_2 \geq \frac{(1-p)^2((N+2)p - 1)(\frac{1}{2} + \Delta) + (1 - Np)\frac{1}{2}}{2(1 - Np)}
\]

\( \iff \)

\[
r_2 \geq \frac{2p(1-p)^2\frac{1}{2} + (1 - p)^2((N+2)p - 1)\Delta}{2(1 - Np)}\]

\( \text{(Using \((1-p)^2 \leq 1\))} \)

\( \iff \)

\[
r_2 \geq \frac{\frac{1}{2}(p - \Delta)}{2(1 - Np)}\]

\( \text{(Using \((1-p)^2 \geq \frac{1}{4} \text{ since } p \leq \frac{1}{2} \))} \)

\( \iff \)

\[
r_2 \geq \frac{1}{4}(p - \Delta)
\]

\( \text{(Using } p \leq \frac{1}{2N} \text{)} \)

Therefore, we choose \( \Delta \leq \frac{p}{6} \) so that \( \frac{\Delta}{2} < \frac{1}{4}(p - \Delta) \) meaning \( r = r_1 = \frac{3}{2} \).

**Improve the power of the algorithm** Let \( A \) be any algorithm that we run on data \( \mu \) such that either \( \mu = \mu_1 \) or \( \mu = \mu_2 \) (the choice is made by an adversary). Let us increase the amount of information available to \( A \). \( A \) is told that the optimal solution is either \( \mu_1 \) or \( \mu_2 \). Furthermore, at each time step, \( A \) chooses \( M(t) \) and observes a sample from arm 1 with probability \( g(M) \) and similarly for arm 2. However \( A \) does not observe the rewards. Note that this problem is simpler than the original problem since in the original problem \( A \) observes a sample from arm \( k \) with probability \( g(M_k(t)) \leq g(M) \). Therefore, at each time step, \( A \) should play either \((N, N+1)\) or \((N+1, N)\) since any other play would lead to a higher regret.

**Link with classical 2-arms bandit problem** With the additional information \( A \) can be seen as playing a 2 arm bandits with probabilistic triggered arms: playing arm 1 means playing \( M(t) = (N, N+1) \) and playing arm 2 means playing \( M(t) = (N+1, N) \). Call \( i^* \) the optimal arm.

We follow the technique used in [Wang and Chen (2017)](2017) to rewrite a bandit problem with probabilistically triggered arms into a classical bandit problem with well chosen discrete random variables: at each time step \( t \), \( A \) chooses an arm \( i_t \in \{1, 2\} \) and observes \( X(t) = (X_{1t}, X_{2t}) \) where \( X_{1t} \) is 1 with probability \( g(M)\mu_i \), \( X_{1t} = 0 \) with probability \( g(M)(1 - \mu_i) \) and \( X_{2t} = 1 \) with probability \( 1 - g(M) \).

However, the regret of \( A \) is computed as in the original problem (and this information is known to \( A \)):

\[
E[R_A] = E[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\}(\mu, g(M^*) - g(M(t)))]
\]

\[
= E[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\}(\frac{1}{2} + \Delta)(g(N+1) - g(N)) + (\frac{1}{2})(g(N) - g(N+1))]
\]

\[
= E[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\}(\Delta)(g(N+1) - g(N))]
\]

Then, the rest of the proof is then identical to [Mourtada and Gaïffas (2019)](2019). Call for \( i = 1, 2 \), let \( P_i \) be the joint probability on \((X(1), \ldots, X(T))\) when \( \mu = \mu_i \).
The regret incurred by $A$ on the worst choice of $\mu$ is higher than the regret incurred by choosing the worst between $\mu_1$ and $\mu_2$.

$$
\mathbb{E}[R_A] \geq \max_{i^* \in \{1, 2\}} \mathbb{E}_{i^*}[\sum_{t=1}^T \mathbbm{1}_{\{i_t \neq i^*\}} (\Delta)(g(N+1) - g(N))]
$$

$$
\geq \frac{1}{2} \sum_{i^* = 1}^{2} \mathbb{E}_{i^*}[\sum_{t=1}^T \mathbbm{1}_{\{i_t \neq i^*\}} (\Delta)(g(N+1) - g(N))]
$$

$$
= \frac{\Delta(g(N+1) - g(N))}{2} \sum_{i^* = 1}^{2} \mathbb{E}_{i^*}[T - \mathbb{E}_{\sum_{t=1}^T 1_{i_t = i^*}}]
$$

$$
\geq \frac{\Delta(g(N+1) - g(N)) T}{2} \sum_{i^* = 1}^{2} \mathbb{P}_{i^*}[T \geq N_{i^*}]
$$

$$
\geq \frac{\Delta(g(N+1) - g(N)) T}{2} (\mathbb{P}_1(N_1 \geq \frac{T}{2}) + \mathbb{P}_2(N_2 \geq \frac{T}{2}))
$$

Then by Bretagnolle–Huber inequality (Th 14.2 in Lattimore and Szepesvári (2020)), we have

$$
\mathbb{P}_1(N_1 \geq \frac{T}{2}) + \mathbb{P}_2(N_2 \geq \frac{T}{2}) \geq \frac{1}{2} \exp(-KL(\mathbb{P}_1, \mathbb{P}_2))
$$

where $KL$ is the KL-divergence.

More precisely, we have

$$
KL(\mathbb{P}_1, \mathbb{P}_2) \leq T g(M)(KL(B((\frac{1}{2} + \Delta)), B(\frac{1}{2}))) + KL(B(\frac{1}{2}), B(\frac{1}{2} + \Delta))
$$

$$
\leq 4T g(M) \Delta^2
$$

and therefore

$$
\mathbb{E}[R_A] \geq \frac{\Delta(g(N+1) - g(N)) T}{2} \exp(-4T g(M) \Delta^2)
$$

and since the regret increases with $T$ (see (a)), we can assume without loss of generality that $T = \lfloor \frac{1}{4T g(M) \Delta^2} \rfloor \geq \frac{1}{8g(M)\Delta^2}$ and obtain

$$
\mathbb{E}[R_A] \geq \frac{g(N+1) - g(N)}{64g(M)\Delta} \exp(-1)
$$

$$
\geq \frac{g(N+1) - g(N)}{64M\rho\Delta} \exp(-1)
$$

$$
= \frac{(N+1)(1-p) - N}{64M\Delta} \exp(-1)
$$

$$
= \frac{1 - (N+1)p}{64M\Delta} \exp(-1)
$$

$$
\geq \frac{1}{128M\Delta} \exp(-1) \quad \text{(Using $p \leq \frac{1}{2(N+1)}$)}
$$

\[\square\]

**A.5 Proof of Lemma 4.2**

Take $K = \nu^* + 2$ arms, $M$ players and $\mu = (\mu_1, \mu_0, \mu_0 + \Delta_1 - \Delta_2, \ldots, \mu_0 + \Delta_1 - \Delta_3, \mu_0 + \Delta_2)$. For simplicity denote $\Delta = \Delta_1$. 
Let us choose $\mu_1, \mu_0$ and $\Delta$ such that the $\nu^* + 1$-st best assignments are to put $M - 1$ player on the first arm and one player on a different arm.

For this we need to ensure the three conditions:

$$g(M - 1) \mu_1 + g(1)(\mu_0 + \Delta) \geq g(M - 2) \mu_1 + 2g(1)(\mu_0 + \Delta)$$  \hspace{1cm} (13)  
$$g(M - 1) \mu_1 + g(1) \mu_0 \geq g(M) \mu_1$$  \hspace{1cm} (14)  
$$g(M) \mu_1 \geq g(M - 2) \mu_1 + g(2)(\mu_0 + \Delta)$$  \hspace{1cm} (15)  

Equation (13) ensures that putting strictly less than $M - 1$ players on the first arm is sub-optimal. Equation (14) ensures that putting $M$ players on the first arm is worse than any assignment that puts exactly $M - 1$ players on the first arm. Equation (15) ensures that putting strictly less than $M - 1$ players on the first arm is worse than putting all players on the first arm.

Equation (13) yields

$$g(M - 1) \mu_1 + g(1)(\mu_0 + \Delta) \geq g(M - 2) \mu_1 + 2g(1)(\mu_0 + \Delta)$$  \hspace{1cm} \iff \hspace{1cm} $(\mu_0 + \Delta) \leq \frac{g(M - 1) - g(M - 2)}{g(1)} \mu_1$  

Equation (14) yields

$$g(M - 1) \mu_1 + g(1) \mu_0 \geq g(M) \mu_1$$  \hspace{1cm} \iff \hspace{1cm} $\mu_0 \geq \frac{g(M) - g(M - 1)}{g(1)} \mu_1$  

Equation (15) yields

$$g(M) \mu_1 \geq g(M - 2) \mu_1 + g(2)(\mu_0 + \Delta)$$  \hspace{1cm} \iff \hspace{1cm} $\frac{g(M) - g(M - 2)}{g(2)} \mu_1 \geq (\mu_0 + \Delta)$  

We have $h_1 > h_2$ and

$$h_3 = \frac{g(M) - g(M - 1) + g(M - 1) - g(M - 2)}{g(2)} \mu_1$$  
$$> \frac{2(g(M) - g(M - 1))}{2g(1)} \mu_1$$  
$$= h_2.$$  

We therefore choose $\mu_1 = 1$, $\mu_0 = \frac{h_2 + \min(h_1, h_3)}{2}$ and need $\Delta \leq \frac{\min(h_1, h_3) - h_2}{4}$.

Since $g(M) - g(M - 1) = p(1 - p)^{M-2}(1 - Mp)$ and

$$g(M) - g(M - 2) = Mp(1 - p)^{M-1} - (M - 2)p(1 - p)^{M-3}$$  
$$= p(1 - p)^{M-3}(M(1 - p)^2 - (M - 2))$$  
$$= p(1 - p)^{M-3}(M(1 - 2p + p^2) - M + 2)$$  
$$= p(1 - p)^{M-3}(2 - Mp + Mp^2)$$
we get

\[
\begin{align*}
    h_1 - h_2 &= (1 - p)^{M-3} (1 - (M-1)p) - (1 - p)^{M-2} (1 - Mp) \\
    &= (1 - p)^{M-3} (1 - (M-1)p) - (1 - p)(1 - Mp) \\
    &= (1 - p)^{M-3} (1 - (M-1)p) - (1 - p)(1 - Mp) \\
    &= (1 - p)^{M-3} (1 - Mp + p - (1 - Mp + M p^2)) \\
    &= (1 - p)^{M-3} (2p - Mp^2) \\
    &\geq (1 - p)^{M-3} p \\
    &\geq (1 - p)^{M-3} p \\
    &\geq \frac{p}{M - 3} 
\end{align*}
\]

(Using \( p \leq \frac{1}{M} \))

(Using \( p \leq \frac{1}{M} \))

(Using \( \min_{x \in [M]} g(x) = p \))

and

\[
\begin{align*}
    h_3 - h_2 &= \frac{1}{2} (1 - p)^{M-4} (2 - 2Mp + Mp^2) - (1 - p)^{M-2} (1 - Mp) \\
    &= \frac{1}{2} (1 - p)^{M-4} (2 - 2Mp + Mp^2 - 2(1 - Mp)(1 - p)^2) \\
    &= \frac{1}{2} (1 - p)^{M-4} (2 - 2Mp + Mp^2 - (1 - Mp)(2 - 4p + 2p^2)) \\
    &= \frac{1}{2} (1 - p)^{M-4} (2 - 2Mp + Mp^2 - 2 - 4p + 2p^2 - 2Mp + 4p^2 - 2Mp^3) \\
    &= \frac{1}{2} (1 - p)^{M-4} (4p - 2p^2 + 2Mp^3 - 3M p^2) \\
    &\geq \frac{1}{2} (1 - p)^{M-4} (4p - (3M + 2)p^2) \\
    &\geq \frac{1}{2} (1 - p)^{M-4} (p) \\
    &\geq \frac{p}{2(M - 4)} 
\end{align*}
\]

(Using \( p \leq \frac{1}{M+1} \))

(Using \( \min_{x \in [M]} g(x) = p \))

Noting that \( 2(M - 4) \geq M - 3 \iff M \geq 5 \), we obtain that \( \Delta \leq \frac{\min(h_1, h_3) - h_2}{4} \) is implied by \( \Delta \leq \frac{p}{8(M-4)} \).

Let \( N_k(T) \) be the number of samples of arm \( k + 1 \) observed by the consistent algorithm \( A \). Using arguments similar to Lai & Robbins result \cite{Lai} \cite{Robbins} we can prove that

\[
\liminf_T \frac{\mathbb{E}[N_k(T)]}{\log(T)} \geq \frac{1}{2\Delta(k)}
\]

If \( m_t \) denotes the number of players put on arm \( k + 1 \) at stage \( t \), then \( \mathbb{E}[N_k(T)] = \sum_{t=1}^{T} g(m_t) \). Denote by \( \Delta_k(m) \) the cost of the best assignment with \( m > 0 \) players on arm \( k + 1 \), i.e.,

\[
\Delta_k(m) := (g(M - 1)\mu_1 + g(1)(\mu_0 + \Delta)) - (g(M - m)\mu_1 + g(m)(\mu_0 + \Delta - \Delta(k))) \\
\geq (g(M - m)\mu_1 + g(m)(\mu_0 + \Delta)) - (g(M - m)\mu_1 + g(m)(\mu_0 + \Delta - \Delta(k))) \\
= g(m)\Delta(k)
\]

and \( \Delta_k(0) = 0. \)

\(^3\)Consider for any sub-optimal arm \( k \) the two possibilities \( \mu \) and \( \mu' \) such that \( \mu'_i = \mu_i \) for all \( i \) except for \( i = k \) where \( \mu'_k = \mu_0 + \Delta_1 + \epsilon \) and use the same arguments as in Lai & Robbins
Then consider \( C_k \) the cost of the assignment putting the optimal number of players on arm \( k + 1 \) and the rest on arm 1, under the constraint that arm \( k + 1 \) has been played sufficiently often.

\[
C_k = \min_{m_1, \ldots, m_T, \sum_i g(m_i) \geq \frac{\log(T)}{2\Delta_k}} \sum_{i=1}^{T} \Delta_k(m_i) \tag{16}
\]

It is clear that

\[
\liminf_T \frac{\mathbb{E}[R(T)]}{\log(T)} \geq \liminf_T \sum_{k=1}^{\nu^*} \frac{C_k}{\log(T)}
\]

The solution of Equation (16) has a specific form: for \( t \in [\tau] \), \( m_t \) is constant, equal to \( m_\tau \), and defined by

\[
\tau g(m_\tau) \geq \frac{\log(T)}{2\Delta^2(\tau)}
\]

and \( m_t = 0 \) afterwards (with a cost also equal to zero).

As a consequence, one gets that, for a specific value of \( \tau^* \),

\[
C_k = \tau^* \Delta_k(m_{\tau^*}) \geq \frac{\log(T)}{2\Delta^2(\tau^*)} \Delta_k(m_{\tau^*}) \geq \frac{\log(T)}{2\Delta_k}
\]

as \( \Delta_k(m) \geq g(m)\Delta_k(k) \).

This implies that, for any consistent algorithm, one must have

\[
\liminf_T \frac{\mathbb{E}[R(T)]}{\log(T)} \geq \sum_{\nu=1}^{\nu^*} \frac{1}{2\Delta_{(\nu)}}
\]

### B Arms elimination when rewards are close

**Lemma B.1** (Necessary conditions for arm elimination). Let \( k^* = \arg\max_{k \in [K]} \mu_k \) and \( \alpha = \frac{M_p}{K} \). If \( p \leq 0.1 \), \( \alpha \in (2p, 1) \), and \( \min_{k' \in [K]} \frac{\mu_{k'}}{\mu_{k^*}} \geq 1.3 \exp(-\alpha) (1 - \alpha) \), then \( \nu^* = 0 \).

**Proof of Lemma B.1.** From Bonnefoi et al. (2017), \( g \) is concave if \( x \leq \frac{2}{\log(1-p)} \) and so this is also the case for \( x \leq \frac{1}{\log(1-p)} \). Therefore, we have that for any \( x \leq \frac{1}{\log(1-p)} \), \( g(x) - g(x - 1) \leq g(y) - g(y - 1) \) for any \( y \leq x \).

Assume \( \nu^* > 0 \) and consider the optimal policy \( M^* \). Then take an eliminated arm \( i \) and consider \( M' \) constructed from \( M^* \) by taking one player from \( k^* \) and putting it on the eliminated arm \( i \). Using \( M' \) instead of \( M^* \) increase the utility by: 

\[
G = \mu_i p - \mu_k (g(M^*_k) - g(M^*_k, 1))
\]

Note that \( M^*_k, g \) for any \( k \neq k^* \) since \( k^* \) is the best arm. In particular \( M^*_k, g \geq \frac{M}{K} \) and by the hypothesis on the range of \( \alpha \), we have \( \frac{M}{K} > 2 \). Also note that by definition of \( \Delta_{\max}, \mu_i \geq \rho \mu_{k^*} \).

We can then write:

\[
G = \mu_i p - \mu_k (g(M^*_k) - g(M^*_k, 1))
\]

\[
\geq \mu_k \left[ pp - (g(M^*_k) - g(M^*_k, 1)) \right] \tag{Since \( \mu_i \geq \rho \mu_{k^*} \)}
\]

\[
\geq \mu_k \left[ pp - (g(\alpha) - g(\alpha, 1)) \right] \tag{By concavity of \( g \) and \( M^*_k, g \geq \frac{M}{K} = \frac{\alpha}{p} \)}
\]

\[
= \mu_k \left[ pp - p(1-p)^{\frac{\alpha}{p}} (1-\alpha) \right]
\]

\[
= \mu_k \left[ pp - p(1-p)^{\frac{\alpha}{p}} (1-\alpha) \right]
\]
The gain is positive if $\rho \geq 1.3 \exp(-\alpha)(1 - \alpha)$ since $\exp(-\alpha) \geq \exp(-\frac{\log(1-p)}{p} \alpha) = (1-p)^{\frac{\alpha}{p}}$ and $1.3 \geq \frac{1}{1-p^p} \geq \frac{1}{(1-p)^2}$.

Therefore, $M^*$ cannot be an optimal policy. This shows that $\nu^* = 0$.

C Centralized UCB

C.1 Description

At time $t \in [T]$, for all $k \in [K]$, compute an estimate $\hat{\mu}_k(t)$ of $\mu_k$ using (6) and an upper bound using $\hat{\mu}^H_k(t) = \min(\hat{\mu}_k(t) + \zeta_k(t), 1)$ where $\zeta$ is given by $k \in [K]$, $\zeta_k(t) = \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}$ and take

$$M(t+1) = \arg\max_{M \in M} \langle \hat{\mu}^H(t), g(M) \rangle$$

where $\hat{\mu}^H[k] = \hat{\mu}^H_k$.

The code is given in Algorithm 4.

**Algorithm 4 UCB**

1: **Input** : $M$ (number of players), $K$ (number of arms), $p$ (probability that a player is active), $T$ (horizon)
2: Initialize estimated rewards: $\hat{\mu}^H = 1$
3: for $t$ from 1 to $T$ do
4: Play $\arg\max_{M \in M} \langle \hat{\mu}^H, g(M) \rangle$
5: Compute $\hat{\mu}$ according to [6]
6: Compute $\zeta$ according to $k \in [K]$, $\zeta_k(t) = \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}$
7: Set $\hat{\mu}^H = \min(\hat{\mu} + \zeta, 1)$
8: end for

C.2 Analysis

The next Lemma gives an upper bound on the regret of UCB:

**Lemma C.1 (Regret of UCB).** The regret of UCB satisfies

$$\mathbb{E}[R_{UCB}] \leq 2\sqrt{2K \log(2T^3K^2)T \min(K, Mp + \frac{K}{T})} + 2$$

(17)

**Proof.** Define the GOOD event as in Lemma [A.1]

From Lemma [A.2] we have $\mathbb{E}[R_{CUCB}] = \mathbb{E}[R_{UCB}\mathbb{1}\{GOOD\}] + 2$. 

Then, under the GOOD event, we have:

\[
R_{CUCB} = \sum_{t=1}^{T} \langle \mu, g(M_t^*) \rangle - \sum_{t=1}^{T} \langle \mu, g(M(t)) \rangle
\]

\[
= \sum_{t=1}^{T} (\mu - \hat{\mu}^H(t), g(M^*) + \langle \hat{\mu}^H(t), g(M^*) - g(M(t)) \rangle + \langle \hat{\mu}^H(t) - \mu, g(M(t)) \rangle
\]

\[
\leq \sum_{t=1}^{T} (\hat{\mu}^H(t), g(M^*) - g(M(t))) + \langle \hat{\mu}^H(t) - \mu, g(M(t)) \rangle
\]

\[
\leq \sum_{t=1}^{T} (\hat{\mu}^H(t) - \mu, g(M(t)))
\]

\[
= \sum_{k=1}^{K} \sum_{t=1}^{T} \min(1, 2\zeta_k(t)) g(M_k(t))
\]

\[
= \sum_{k=1}^{K} \sum_{t=1}^{T} \min(1, 2\zeta_k(t)) g(M_k(t))
\]

\[
\leq \sum_{k=1}^{K} \sum_{t=1}^{T} (g(M_k(t)) - \eta_k(t)) + \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{2\log(2T^3K^2)}{T_k(t)} \eta_k(t)
\]

\[
(i) = \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{2\log(2T^3K^2)}{T_k(t)} \eta_k(t)
\]

\[
(ii) = \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{2\log(2T^3K^2)}{T_k(t)} \eta_k(t)
\]

\[
(\ast) \text{Recall the convention that } \hat{\mu}_k = 1 \text{ if } T_k(t) = 0. \text{ In order to ease the notation, we do not make the distinction and write } \frac{1}{T_k(t)} \text{ instead of } \frac{1}{T_k(t) \neq 0} + 1 \{T_k(t) = 0\}.
\]

We have that \(E[(i)] = 0\) since

\[
E[g(M_k(t)) - \eta_k(t)] = E[g(M_k(t))] - E[E[\eta_k(t)|M_k(t)]]
\]

\[
= E[g(M_k(t))] - E[g(M_k(t)]
\]

\[
= 0
\]

and

\[
(ii) = \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{2\log(2T^3K^2)}{T_k(t)} \eta_k(t)
\]

\[
= \sum_{k=1}^{K} \frac{\sqrt{2\log(2T^3K^2)}}{T_k(T)} \sum_{t=1}^{T} \frac{1}{T_k(T)} \eta_k(t)
\]

\[
\leq \sum_{k=1}^{K} 2\sqrt{2\log(2T^3K^2)} \max(T_k(T), 1)
\]

Then we have trivially:

\[
E[(i)] \leq 2K \sqrt{2\log(2T^3K^2)} T
\]
Otherwise, we write:

\[
\mathbb{E}[(ii)] \leq \mathbb{E}[2 \sqrt{2K \log(2T^3K^2) \sum_{k=1}^{K} (T_k(T) + 1 \{T_k(T) = 0\})}]
\]

(Using \(\sum_{i=1}^{K} \sqrt{a_i} \leq \sqrt{K \sum_{i=1}^{K} a_i}\))

\[
\leq 2 \sqrt{2K \log(2T^3K^2) \sum_{k=1}^{K} (\mathbb{E}[T_k(T)] + \mathbb{P}(T_k(T) = 0))}
\]

(By Jensen inequality)

\[
= 2 \sqrt{2K \log(2T^3K^2) \sum_{k=1}^{K} (\sum_{\rho=1}^{T} g(M_k(\rho)) + \prod_{\rho=1}^{T} (1 - g(M_k(\rho))))}
\]

\[
\leq 2 \sqrt{2K \log(2T^3K^2) \sum_{k=1}^{K} (\sum_{\rho=1}^{T} M_k + 1)}
\]

(Since \(0 \leq g(M_k) \leq 1\) and \(g(M_k) \leq M_k\))

\[
\leq 2 \sqrt{2K \log(2T^3K^2)(TMp + K)}
\]

and therefore

\[
\mathbb{E}[(ii)] \leq 2 \sqrt{2K \log(2T^3K^2)T \min(K, Mp + \frac{K}{T})}
\]

so that

\[
\mathbb{E}[R_{UCB}] \leq 2 \sqrt{2K \log(2T^3K^2)T \min(K, Mp + \frac{K}{T})}
\]

\(\mathbb{E}[R_{UCB}] \leq 2K \sqrt{2\log(2T^3K^2)}T\) also holds in the case where players have different probability of activation \((p_i)_{i \in [M]}\). This is shown by following the same proof and stopping at Equation (18).

### D Solving \(\arg\max_{M_\mathcal{E}} \langle g(M), \nu \rangle\) via a sequential algorithm

We want to solve

\[
\arg\max_{M_\mathcal{E}} \langle g(M), \nu \rangle
\]

(19)

where \(\mathcal{E} \subset [K]\).

The sequential algorithm of \cite{Dakdouk2022} (Algorithm 5) is optimal if \(\mathcal{E} = \emptyset\) and \(\frac{Mp}{1 - p} \leq K\) (Th 4.2). At each time step, the sequential algorithm chooses a new player to assign to an arm based on some arm-specific criterion that decreases with the number of players assigned to this arm (Lemma 4.2).

Call \(a_1, \ldots, a_M \in [K]\) the arms chosen by the sequential algorithm for players 1, \ldots, \(M\). The first thing to note is that if the first player is assigned to \(a_i\) and then the sequential algorithm is run. The resulting algorithm that we call \(A\) reaches the same solution as the sequential algorithm (ignoring the order).

Indeed as adding a player to some arm can only decrease its criterion, the assignment chosen by \(A\) is \(a_i, a_1, \ldots, a_k\) until \(a_{k+1} = a_i\). Then everything happens as if the assignment chosen by \(A\) was \(a_1, \ldots, a_k, a_{k+1}\) and therefore the rest of the run is the same as the sequential algorithm.

Consider \(A^*\) is the algorithm that assigns the first \(|\mathcal{E}|\) players to a different arm in \(\mathcal{E}\) and then follow the sequential algorithm. Call \(\mathcal{E}'\) the set of arms in \(\mathcal{E}\) such that for any arm \(k \in \mathcal{E}'\) there exists an index \(i\) such that \(a_i = k\). Then from the previous argument \(A^*\) behaves as if one player was assigned to every arm in \(\mathcal{E}'' = \mathcal{E} \setminus \mathcal{E}'\) and then the sequential algorithm is run. But since none of the arms in \(\mathcal{E}''\) are equal to \(a_1, \ldots, a_M\) and again because the arm specific criterion decreases with the number of players, the run of \(A^*\) after arms in \(\mathcal{E}''\) are assigned one player is \(a_1, \ldots, a_M - |\mathcal{E}''|\) which is the optimal solution with \(M - |\mathcal{E}''|\) players. This implies that \(A^*\) produces the optimal solution.

In addition, we note that this algorithm applied in a fully decentralized setting (typically when a communication phase fails). Still ensures that the final assignment belongs to \(M_\mathcal{E}\) even if we loose the optimality property in this case.
E Additional experiment

We rerun the experiments in Figure 1 with a larger range of values for $T$ (see Figure 2). Cautious Greedy and UCB have a better scaling in $T$ than ETC.