Constant or logarithmic regret in asynchronous multiplayer bandits

Hugo Richard[*]
Criteo, Paris

Etienne Boursier
INRIA, Université Paris Saclay, LMO, Orsay, France

Vianney Perchet
ENSAE, Paris
Criteo, Paris

Abstract

Multiplayer bandits have recently been extensively studied because of their application to cognitive radio networks. While the literature mostly considers synchronous players, radio networks (e.g. for IoT) tend to have asynchronous devices. This motivates the harder, asynchronous multiplayer bandits problem, which was first tackled with an explore-then-commit (ETC) algorithm (see Dakdouk, 2022), with a regret upper-bound in $\mathcal{O}(T^{\frac{2}{3}})$. Before even considering decentralization, understanding the centralized case was still a challenge as it was unknown whether getting a regret smaller than $\Omega(T^{\frac{2}{3}})$ was possible. We answer positively this question, as a natural extension of UCB exhibits a $\mathcal{O}(\sqrt{T\log(T)})$ minimax regret. More importantly, we introduce Cautious Greedy, a centralized algorithm that yields constant instance-dependent regret if the optimal policy assigns at least one player on each arm (a situation that is proved to occur when arm means are close enough). Otherwise, its regret increases as the sum of $\log(T)$ over some sub-optimality gaps. We provide lower bounds showing that Cautious Greedy is optimal in the data-dependent terms. Therefore, we set up a strong baseline for asynchronous multiplayer bandits and suggest that learning the optimal policy in this problem might be easier than thought, at least with centralization.

1 Introduction

In the classical multi-armed bandits (MAB) problem, a single player sequentially pulls arms $k_t \in \{1,\dots,K\} \triangleq [K]$, and receives a reward X_{k_t} sampled from some unknown sub-Gaussian distribution of mean μ_{k_t}. 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 & Szepesvári, 2020, for a survey). A fundamental component of MAB is the exploration and exploitation trade-off, as a good policy should balance between both. Exploration involves trying out different arms to gather information, while exploitation uses the acquired knowledge to favor arms more likely to be the best. Optimal policies are known to have a regret scaling as $\mathcal{O}(\sum_{k\neq k^*}\frac{\log(T)}{\mu_{k^*}-\mu_k})$ (Auer et al., 2002).

Classical applications of MAB include clinical trials, recommendation systems or ad placements. For many 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 & 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 in Liu & 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

16th European Workshop on Reinforcement Learning (EWRL 2023).

same arm k, and $\eta_k = 1$ if a single player pulls k. In the centralized setting with $M \leq K$, this setting is equivalent to bandits with multiple plays (Komiyama et al., 2015; Anantharam et al., 1987; Chen et al., 2013b; Gopalan et al., 2013), where 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 $\mu_{(i)}$ is the i-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 player i with probability p_i. In the third 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.

We focus on the centralized setting (equivalently, agents' communication is free). Although the decentralized setting might be more fitted to communication network applications, the centralized setting is already challenging enough to warrant a study (furthermore, centralized algorithm performances are benchmarks for decentralized ones, hence it is crucial to investigate the former). Possible extensions, including the decentralized case, are discussed in Section 6.

Notations. Vectors are denoted in bold. If $\mathbf{u} \in \mathbb{R}^n$, u_i is the i-th coordinate of $\mathbf{u}$ while $u_{(i)}$ the i-th smallest coordinate of u and $\text{support}(\mathbf{u}) = \{i \in [n], u_i \neq 0\}$. We denote for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^n u_i v_i$, $\|\mathbf{u}\|_\infty = \max_{i\in[n]} |u_i|$, $\|\mathbf{u}\|_1 = \sum_{i\in[n]} |u_i|$. For a function $f : \mathbb{R} \to \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^n$, $f(\mathbf{u}) \in \mathbb{R}^n$ is defined by $f(\mathbf{u})_i = f(u_i)$. Lastly, $\overline{E}$ 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, a central entity chooses a (possibly random) assignment $\mathbf{M}(t) = (M_1(t), \ldots, M_K(t))$ where $M_k(t)$ is the number of players assigned to arm k at round t; the environment then activates each player with probability[1] p and active players pull the arm to which they are assigned, each receiving reward $\eta_k(M_k(t))X_k$ where $\eta_k = \mathbb{1}\{\text{exactly one player is active on arm } k\}$. Active players communicate to the central entity their earned reward $X_k\eta_k$. Additionally, the central entity observes the collision events $(\eta_k)_{k\in[K]}$, and the parameters M, K and p are assumed to be known beforehand.

At any time t, the assignment $\mathbf{M}(t)$ satisfies the budget constraint $\sum_{k=1}^K M_k(t) = M$ and we assume:
Assumption 1.1. $M \geq K$ and for all $k \in [K]$ and at all stages $M_k(t) \leq \frac{-1}{\log(1-p)} \simeq \frac{1}{p}$.

The second condition is not restrictive, 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

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

The goal is to minimize the expected regret defined by:

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

where $\mathbf{M}^* = (M_1^*, \ldots, M_K^*)$ is the optimal assignment:

$$\mathbf{M}^* = \underset{\mathbf{M}\in\mathcal{M}}{\text{argmax}}\, \mathbb{E}\left[\sum_{k=1}^K \eta_k(M_k)X_k\right]. \tag{2}$$

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

[1]The central entity does not know beforehand which players will be active, making collisions unavoidable.

Contributions. We prove that an adapted version of UCB exhibits a regret in $\mathcal{O}(\sqrt{TK\log(T)\min(Mp,K)})$ where $\mathcal{O}(\cdot)$ hides universal constant factors. This result largely outperforms the $\mathcal{O}(T^{\frac{2}{3}})$ regret bound of the ETC algorithm by Dakdouk (2022). Contrary to the lower bound by Wang & Chen (2017, Theorem 2), the $1/p$ term does not appear in this bound, as the two settings are slightly different: although rewards are observed with probability p, rewards are also scaled by p in our case (thus canceling out the terms in p). More surprisingly, our main contribution shows that achieving a constant regret (in T) is sometimes possible with an algorithm called Cautious Greedy. The analysis of centralized UCB is thus postponed to Appendix C and the main text solely focuses on Cautious Greedy. In essence, it is a standard greedy algorithm that estimates μ_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 ν of the number of arms that should be assigned zero players and stops assigning players to the ν worst arms when confident enough.

The regret of Cautious Greedy depends on several data-dependent quantities defined in Section 3.2:

- ν^* the number of arms that are assigned zero players in the optimal assignment,
- $\Delta_{(j)} = \mu_{(\nu^*+1)} - \mu_{(j)}$,
- $\mathbf{M}^*_\nu$ the optimal assignment when ν arms are assigned zero players and
- r the infinity norm of the minimum perturbation of the true arm means $\boldsymbol{\mu}$ that would modify the sequence $(\mathbf{M}^*_\nu)_{\nu=1}^{\nu^*}$.

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

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 4.1, under mild conditions, the dependency in $\frac{1}{r}$ cannot be improved. In Lemma B.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 $\sum_{j\leq\nu^*}\frac{\log(T)}{\Delta_{(j)}}$. This dependency also appears in bandits with multiple plays (Komiyama et al., 2015) and as shown by Lemma 4.2 cannot be removed. This makes Cautious Greedy optimal with respect to T and with respect to the data-dependent quantities r and $\Delta_{(j)}$.

The main difficulty of the problem comes from the fact that ν^* is unknown. A classical Greedy algorithm yields a linear regret when $\nu^* > 0$, while a traditional bandits algorithm never reaches 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 centralized ETC algorithm of Dakdouk (2022) on synthetic data and show 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. As already highlighted, when $M \leq K$ and $p = 1$, AMMAB is equivalent to bandits with multiple plays. A lower bound in $\sum_{j=1}^{\nu^*}\frac{\log(T)}{\Delta_{(j)}}$ where $\nu^* = K - M$ is shown in Anantharam et al. (1987). This lower bound is reached by a Thompson sampling-based algorithm (Komiyama et al., 2015). Bandits with multiple plays are an instance of combinatorial bandits (Gai et al., 2012; Chen et al., 2013a; Kveton et al., 2015; Combes et al., 2015; Wang & Chen, 2018; Perrault et al., 2020) where an agent chooses an action $\mathbf{a} \in \mathcal{S}$ and receives reward $r(\boldsymbol{\mu}, \mathbf{a})$. When $M \leq K$, AMMAB is an instance of combinatorial bandits with semi-bandit feedback and probabilistically triggered arms (meaning that chosen arms are triggered with some probability) (Wang & Chen, 2017; Chen et al., 2016) with the difference that in these works, rewards are scaled by $1/p$. More generally, AMMAB 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 $\nu^* = 0$.

Decentralized multiplayer bandits. In decentralized multiplayer bandits, players aim at speeding 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 & Perchet, 2022, for a review), sometimes assuming a pre-agreement on the ranks of the players (Anandkumar et al., 2010; Liu & Zhao, 2010) or using few collisions to communicate information between players (Avner & Mannor, 2014; Rosenski et al., 2016; Besson & Kaufmann, 2018b). However, Bistritz & Leshem (2018); Boursier & 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 (Mehrabian et al., 2020; Shi et al., 2020; Huang et al., 2021; Boursier & Perchet, 2020; Shi et al., 2021). This communication through collision trick 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 numerous ($M \geq K$), making both drawbacks significant. This work thus proposes a centralized asynchronous algorithm, leaving open for future work a possible decentralized adaptation (see Section 6 for a discussion).

In the multi-agent bandit problem considered by Szorenyi et al. (2013); Landgren et al. (2016); Martínez-Rubio et al. (2019), 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 & 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 (Degenne & Perchet, 2016). Huang et al. (2017) show Greedy achieves constant regret in a more structured setting.

Resource allocation. Our problem is also a particular instance of sequential resource allocation with concave utilities (Lattimore et al., 2015; Fontaine et al., 2020; Zuo & 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 up to the multiplicative factor μ_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. In addition, the authors keep track of the amount of communication between agents. They propose an explore and commit algorithm that reaches a sub-optimal $\mathcal{O}(T^{\frac{2}{3}})$ regret. In contrast, we show that under favorable conditions, a constant regret can be reached. Extending our results to the heterogeneous setting of Dakdouk (2022) remains open for future work. 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 & Liu, 2012; Bande & Veeravalli, 2019; Magesh & Veeravalli, 2019; Boyarski et al., 2021). AMMAB is also related to the problem of online queuing systems (Gaitonde & 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 centralized algorithm for AMMAB

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

$$\mathbb{E}[R] = \sum_{t=1}^{T} \mathbb{E}\big[\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}(t))\rangle\big] \tag{3}$$

where $\mathbf{M}^* = \operatorname{argmax}_{\mathbf{M}\in\mathcal{M}} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle$ is a rewritting 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. Cautious Greedy therefore computes mean rewards estimates $\hat{\boldsymbol{\mu}}(t)$:

$$\hat{\mu}_k(t) = \frac{\sum_{\rho=1}^{t} \eta_k^{\rho}(\mathbf{M}(\rho)) X_k^{\rho}}{T_k(t)} \tag{4}$$

where $T_k(t) = \sum_{\rho=1}^{t} \eta_k^{\rho}(\mathbf{M}(\rho))$ is the number of samples gathered from arm k at time t or equivalently, the number of times that arm k has been played by exactly one player up to time t. By convention, we set $\hat{\mu}_k(t) = 1$ if $T_k(t) = 0$.

A Greedy algorithm would then choose the assignment $\mathbf{M}(t) = \mathbf{M}_{\mathcal{M}}^{\hat{\mu}(t)}$ where

$$\mathbf{M}_{\mathcal{M}}^{\hat{\mu}} = \underset{\mathbf{M} \in \mathcal{M}}{\operatorname{argmax}} \langle \hat{\boldsymbol{\mu}}, g(\mathbf{M}) \rangle. \tag{5}$$

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* are those not removed yet. Cautious Greedy can put a set of arms $\mathcal{S}$ *under pressure*, meaning that these arms are temporally allowed to be assigned no player. Arms that are assigned 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 $\mathbf{M}$ in the assignment problem will be described by sets of the form:

$$\mathcal{M}_{\mathcal{K}} = \{\mathbf{M} \in \mathcal{M}, \forall k \in \mathcal{K}, M_k \geq 1\}$$

where $\mathcal{K} \subset [K]$. $\mathcal{M}_{\mathcal{K}}$ is the set of assignments that put under pressure $[K] \setminus \mathcal{K}$. 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

$$\hat{\boldsymbol{\mu}}^H(t) = \min(\hat{\boldsymbol{\mu}}(t) + \boldsymbol{\zeta}(t), 1) \quad \text{and} \quad \hat{\boldsymbol{\mu}}^L(t) = \max(\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\zeta}(t), 0) \tag{6}$$

$$\text{where for all } k \in [K], \qquad \zeta_k(t) = \sqrt{\frac{\log(2T^2K^2)}{2T_k(t)}}. \tag{7}$$

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 ν 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 ν arms are under pressure:

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

The number of arms to remove ν is then increased when $\langle \hat{\boldsymbol{\mu}}^L, g(\mathbf{M}_{\mathcal{M}}^{\hat{\boldsymbol{\mu}}^L}) \rangle > \langle \hat{\boldsymbol{\mu}}^H, g(\mathbf{M}_{\mathcal{M}_\nu}^{\hat{\boldsymbol{\mu}}^H}) \rangle$, i.e. when a larger reward is guaranteed by removing more than ν arms. Cautious Greedy then uses ν to build a set $\mathcal{A}$ of *accepted arms* which are arms that are not likely to be among the ν worst arms. Cautious Greedy then puts under pressure a subset of arms 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 ν and updates the sets of accepted arms and active arms. As ν increases, an arm can be removed from the set of accepted arms. However as ν never decreases, a removed arm is removed forever. The exact procedure is described in Algorithm 1 below.

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 now define precisely. $\Delta^{(\nu^*)}$ is the

Algorithm 1 Cautious Greedy

Input : M (number of players), p (probability that a player is active), T (horizon)
$\nu = 0$ // Estimate of ν^*
Initialize the set of active arms $\mathcal{K} = [K]$; the set of accepted arms $\mathcal{A} = \varnothing$
Initialize the set of arms under pressure $\mathcal{U} = \varnothing$; the round-robin counter $n = 0$
for $t = 1, \dots, T$ **do**
  Play $\mathbf{M}^{\hat{\boldsymbol{\mu}}}_{\mathcal{M}_{\mathcal{E}}}$ as defined in (5) where $\mathcal{E} = \mathcal{K} \setminus \mathcal{U}$
  Rotate $\mathcal{U}$ in a round robin fashion over $\mathcal{K} \setminus \mathcal{A}$ (See Appendix A.2 for details)
  Update $\hat{\boldsymbol{\mu}}$ according to (4); $n = n + 1$
  if $n = |\mathcal{K} \setminus \mathcal{A}|$ **then** // end of round robin
    $n = 0$ and compute $\mathbf{M}^{\hat{\boldsymbol{\mu}}^L}_{\mathcal{M}}$ and $\mathbf{M}^{\hat{\boldsymbol{\mu}}^L}_{\mathcal{M}_\nu}$ following (5)
    while $\langle \hat{\boldsymbol{\mu}}^L, g(\mathbf{M}^{\hat{\boldsymbol{\mu}}^L}_{\mathcal{M}})\rangle > \langle \hat{\boldsymbol{\mu}}^H, g(\mathbf{M}^{\hat{\boldsymbol{\mu}}^H}_{\mathcal{M}_\nu})\rangle$ **do**
      $\nu = \nu + 1$
    end while
    Update $\mathcal{A} = \{k \in [K], \hat{\mu}^H_{(\nu)} < \hat{\mu}^L_k\}$ and $\mathcal{K} = [K] \setminus \{k \in [K], \hat{\mu}^H_k < \hat{\mu}^L_{(\nu+1)}\}$
    Let $\mathcal{U}$ be $\nu - |[K] \setminus \mathcal{K}|$ elements from $\mathcal{K} \setminus \mathcal{A}$
  end if
end for

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 ν^*. Denoting $\mathbf{M}^*_\nu = \mathbf{M}^{\boldsymbol{\mu}}_{\mathcal{M}_\nu}$, $\Delta^{(\nu^*)}$ is defined as $\Delta^{(\nu^*)} = \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu^*-1})\rangle$. By convention, we set $\Delta^{(\nu^*)} = \infty$ if $\nu^* = 0$. $\Delta_{(j)} = \mu_{(\nu^*+1)} - \mu_{(j)}$ 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 $\boldsymbol{\mu}$ causing $\mathbf{M}^*_\nu$ to change for some value of ν. More precisely, define $r_\nu = \min_{\hat{\mu}, \mathbf{M}^{\hat{\mu}}_{\mathcal{M}_\nu} \neq \mathbf{M}^*_\nu} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty$, then $r = \min_{\nu \in [\nu^*]} r_\nu$. Proposition 3.1 shows that the expected regret of Cautious Greedy is upper bounded by $\mathcal{O}(\frac{\nu^*+1}{r} + \nu^* \frac{\log(T)}{\Delta^{(\nu^*)}} + \sum_{j \leq \nu^*} \frac{\log(T)}{\Delta_{(j)}})$ where $\mathcal{O}$ hides quantities independent of the data and T.

Proposition 3.1 (Upper bound on the regret Cautious Greedy)**.** *The expected regret of Cautious Greedy satisfies*

$$\mathbb{E}[R] \leq 20 \frac{KM(\nu^*+1)}{r} + \sum_{\nu=1}^{\nu^*} \frac{120 \log(2T^2K^2)}{\Delta_{(\nu)}} + \left[72M \min(Mp, K)(\nu^*+1) + 120\right] \frac{\log(2K^2T^2)}{\Delta^{(\nu^*)}}.$$

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 ν^* worst arms. The third one is finally due to the sample complexity of detecting that the optimal policy eliminates ν^* arms. Interestingly, the two last terms 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 when arm rewards have a similar order of magnitude (see Lemma B.1).

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

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

Together with Proposition 3.1, Lemma 3.2 shows that the regret of cautious Greedy is upper bounded by $\mathcal{O}(\frac{1}{r} + \sum_{\nu=1}^{\nu^*} \frac{\log(T)}{\Delta_{(\nu)}})$ where $\mathcal{O}$ hides terms in M, p, K. 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. Using classical concentration bounds (Lemma A.1), we can assume that $\boldsymbol{\mu}^H$ and $\boldsymbol{\mu}^L$ (defined in Equation (6)) verify $\boldsymbol{\mu}^H \geq \boldsymbol{\mu} \geq \boldsymbol{\mu}^L$ without affecting the regret bound.

Consequently, Algorithm 1 ensures that ν 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 that 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. Now knowing how many observations are gathered on each arm, we can focus on upper bounding the regret.

Denote $\mathbf{M}^*_\nu = \mathbf{M}^{\boldsymbol{\mu}}_{\mathcal{M}_\nu}$ the optimal assignment of players when at most ν arms can be assigned zero players. For $\mathcal{E}(t) \subset [K]$, call $\mathbf{M}^*_{\mathcal{E}(t)} = \mathbf{M}^{\boldsymbol{\mu}}_{\mathcal{M}_{\mathcal{E}(t)}}$ the optimal assignment of players when only arms not in $\mathcal{E}(t)$ can be assigned zero players. We can write the cost of the chosen assignment at time t $\mathbf{M}(t)$ as the sum of three terms:

$$\underbrace{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle}_{(i)} + \underbrace{\langle \mu, g(\mathbf{M}^*_\nu) - g(\mathbf{M}^*_{\mathcal{E}(t)})\rangle}_{(ii)} + \underbrace{\langle \mu, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle}_{(iii)}$$

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

Let us start with (i). As the number of samples seen increases, ν increases to get closer to ν^*. Lemma A.7 bounds the number of samples seen before the algorithm increases ν, which leads to an upper bound on the total regret due to this term shown in Lemma A.8.

Regarding (ii), for a given ν, 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 provides a lower bound on the number of samples seen before 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 $\mathcal{E}$ 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 $\mathbf{M}(t)$ 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 the regret due to these terms by a quantity independent of the horizon T (see Lemma A.13).

Adding the upper bounds due to the terms (i), (ii) and (iii) and reorganizing concludes the proof. □

4 Lower bound

The next lemma lower bounds the best possible constant term in the regret. Under mild conditions, it shows there exists a choice of rewards $\boldsymbol{\mu}$ such that any algorithm has a regret scaling in $\mathcal{O}(\frac{1}{r})$.

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 $p \leq \frac{1}{M+1}$, $r_0 < \frac{p}{12}$, $T \geq \frac{1}{16g(M)r_0^2}$. For any algorithm A, there exists a choice of rewards $\boldsymbol{\mu}$ such that $r(\boldsymbol{\mu}) = r_0$ and*

$$\mathbb{E}[R_A] \geq \frac{1}{256 M r_0}.$$

Proof sketch (see proof in Appendix A.4) . We take parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ such that $r = \frac{\Delta}{2}$ and the optimal solution is $\mathbf{M}^* = (N, N+1)$ if $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $\mathbf{M}^* = (N+1, N)$ if $\boldsymbol{\mu} = \boldsymbol{\mu}_2$. 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 chose at each step between $\mathbf{M}(t) = (N, N+1)$ or $\mathbf{M}(t) = (N+1, N)$ (these two modifications only improve A). Following the proof of Theorem 3 in Wang & 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 & Gaïffas (2019) and yields the lower bound $\mathbb{E}[R_A] \geq \frac{\Delta(g(M+1)-g(M))}{2}\frac{T}{4}\exp(-4Tg(M)\Delta^2)$. As the regret increases with T, taking $T = \lfloor \frac{1}{4g(M)\Delta^2} \rfloor$ concludes. □

The upper bound of Cautious Greedy when $\nu^* = 0$ is given by $\mathbb{E}[R] \leq \frac{20KM}{r}$, i.e., the dependency in r cannot be improved. 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 algorithm. 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 $\Delta_{(1)}, \dots, \Delta_{(\nu^*)} \leq \frac{p}{8(M-4)}$, and for any consistent algorithm A, there exists a set of parameters $(\mu_1, \dots, \mu_{\nu^*+2})$ such that $\mu_{(\nu^*+1)} - \mu_{(\nu)} = \Delta_{(\nu)}$ for all $\nu \in [\nu^*]$ and the regret of A satisfies*

$$\liminf_{T\to\infty} \frac{\mathbb{E}R_A}{\log(T)} \geq \sum_{\nu=1}^{\nu^*} \frac{c}{\Delta_{(\nu)}}$$

for some universal constant $c > 0$.

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 parameters $(\mu_0, \mu_1, \mu_1+\Delta)$ 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 μ_0 and μ_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 particular, 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 however 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 $\sum_{j\leq\nu^*} \frac{\log(T)}{\Delta_{(j)}}$ in the upper bound of Proposition 3.1 cannot be improved. The results in Lemma 4.1 and Lemma 4.2 show that Cautious Greedy is optimal with respect to its dependency in r, $\Delta_{(\nu)}$ and T. However, we do not claim that the dependency in M, p, or K is optimal. Improving the dependency with respect to these parameters would be an interesting although challenging direction for future work.

5 Experiments

Our experiments compare the expected regret of Cautious Greedy (Algorithm 1), UCB (Algorithm 3), and ETC (Dakdouk, 2022, Algorithm 8). In all these algorithms, maximization problems of the form $\max_{\mathbf{M}\in\mathcal{M}} \langle g(\mathbf{M}), \mathbf{v}\rangle$ are solved using the sequential algorithm of Dakdouk (2022, Algorithm 5). In Cautious Greedy, the sequential algorithm is also adapted to solve $\max_{\mathbf{M}\in\mathcal{M}_\mathcal{E}} \langle g(\mathbf{M}), \mathbf{v}\rangle$ for some set $\mathcal{E} \subset [K]$. This is done by first assigning one player to each 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 D.

At each step $t \in [T]$, algorithms compute a player assignment $\mathbf{M}(t) \in \mathcal{M}$ based on the rewards they have seen so far. We record $\sum_{\tau=1}^{t} \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}(\tau))\rangle$ which we call cumulative regret (instead of pseudo regret). Each experiment is run 50 times, we plot the mean value of the cumulative regret as a function of $t \in [T]$. Error bars represent the first and last decile.

The first experiment in Figure 1 (left), there are $M = 30$ players, $K = 2$ arms, $\boldsymbol{\mu} = (0.8, 0.5)$, $p = 0.01$ and $T = 10^4$. The optimal assignment is $\mathbf{M}^* = (26, 4)$ and note that $\nu^* = 0$. In this example, Cautious Greedy clearly outperforms the other methods as expected when $\nu^* = 0$.

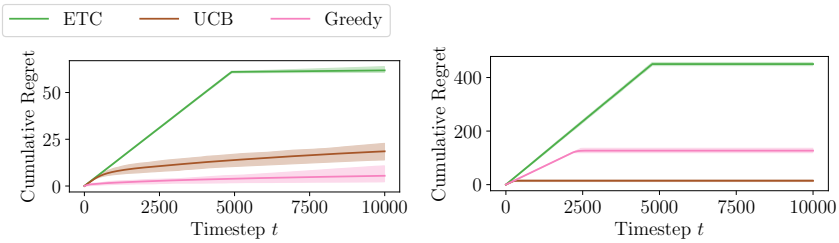

Figure 1: **Benchmark of ETC, UCB and Cautious Cautious Greedy on synthetic data** *(left)* $\nu^* = 0$ *(right)* $\nu^* = 1$.

The second experiment in Figure 1 (right) highlights that Cautious Greedy takes a longer time than UCB to assign no player to a suboptimal arm. We have $M = 3$ players, $K = 2$ arms, $\boldsymbol{\mu} = (0.99, 0.01)$, $p = 0.1$ and $T = 10^4$. The optimal solution is $\mathbf{M}^* = (3, 0)$ so that $\nu^* = 1$. In this example, UCB largely outperforms Cautious Greedy. In both experiments, ETC incurs a much larger regret, which is consistent with its suboptimal $\mathcal{O}(T^{\frac{2}{3}})$ regret.

6 Conclusion, open problems and future work

We proposed an asynchronous multiplayer multi-armed bandits algorithm called Cautious Greedy, achieving a regret of order $1/r + \sum_{\nu=1}^{\nu^*} \log(T)/\Delta_{(\nu)}$. 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 $\sum_{\nu=1}^{\nu^*} \log(T)/\Delta_{(\nu))}$ is optimal. A first remark is that the last term cannot be too large, since Lemma B.1 shows that $\nu^* = 0$ when the gaps $\Delta_{(j)}$ are small.

An open question for future work is whether the dependency with respect to other parameters K, M, p can be improved. In particular, the analysis in Degenne & Perchet (2016) shows that in the stochastic setting with full information, Greedy is upper bounded by $\mathcal{O}(\frac{\log(K)}{\Delta})$ when all arms have sub-optimal gap Δ. This suggests the dependency w.r.t. M and K of the first term in our bound can be improved.

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 & Kaufmann, 2018a, Theorem 7) can be used when the horizon T is unknown. 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. The analysis in the heterogeneous case would not need any change for centralized UCB. Concerning Cautious Greedy, adapting its analysis is more difficult and requires cautious work. Also, if p was unknown beforehand, it can easily be estimated on the fly. However, additional errors would come from this estimation and should be handled carefully.

Another significant direction is to go beyond the centralized 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 first possibility is to track the number of communications, similarly to Dakdouk (2022). Although Cautious Greedy's number of communication steps is linear in T, a simple low-communication extension would proceed in epochs of doubling size, where communication and updates occur at the end of each epoch. This would make the number of communication steps sub-logarithmic in T, with the same regret bound (up to a 2 factor). A second possibility is to directly use collisions to communicate as done for example in (Bistritz & Leshem, 2018; Boursier & Perchet, 2019). These works however only tackle the synchronized case. In our case, 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 $\frac{\log(T)}{p^2}$ 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.

References

Anandkumar, A., Michael, N., and Tang, A. Opportunistic spectrum access with multiple users: Learning under competition. In *2010 Proceedings IEEE INFOCOM*, pp. 1–9. IEEE, 2010.

Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, April 2011. ISSN 1558-0008. doi: 10.1109/JSAC.2011.110406.

Anantharam, V., Varaiya, P., and Walrand, J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, November 1987. ISSN 0018-9286. doi: 10.1109/TAC.1987.1104491.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Avner, O. and Mannor, S. Concurrent bandits and cognitive radio networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 66–81. Springer, 2014.

Bande, M. and Veeravalli, V. V. Multi-user multi-armed bandits for uncoordinated spectrum access. In *2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 653–657. IEEE, 2019.

Besson, L. and Kaufmann, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018a.

Besson, L. and Kaufmann, E. Multi-player bandits revisited. In *Algorithmic Learning Theory*, pp. 56–92. PMLR, 2018b.

Bistritz, I. and Leshem, A. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

Bonnefoi, R., Besson, L., Moy, C., Kaufmann, E., and Palicot, J. Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings. In *International Conference on Cognitive Radio Oriented Wireless Networks*, pp. 173–185. Springer, 2017.

Boursier, E. and Perchet, V. SIC-MMAB: Synchronisation Involves Communication in Multiplayer Multi-Armed Bandits, November 2019.

Boursier, E. and Perchet, V. Selfish robustness and equilibria in multi-player bandits. In *Conference on Learning Theory*, pp. 530–581. PMLR, 2020.

Boursier, E. and Perchet, V. A survey on multi-player bandits. *arXiv preprint arXiv:2211.16275*, 2022.

Boyarski, T., Leshem, A., and Krishnamurthy, V. Distributed learning in congested environments with partial information. *arXiv preprint arXiv:2103.15901*, 2021.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pp. 151–159. PMLR, 2013a.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159. PMLR, 2013b.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

Combes, R., Magureanu, S., and Proutiere, A. Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Dakdouk, H. *Massive Multi-Player Multi-Armed Bandits for Internet of Things Networks*. PhD thesis, Ecole nationale supérieure Mines-Télécom Atlantique, 2022.

Degenne, R. and Perchet, V. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pp. 1587–1595. PMLR, 2016.

Fontaine, X., Mannor, S., and Perchet, V. An adaptive stochastic optimization algorithm for resource allocation. In *Algorithmic Learning Theory*, pp. 319–363. PMLR, 2020.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

Gaitonde, J. and Tardos, É. Stability and learning in strategic queuing systems. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 319–347, 2020.

Gopalan, A., Mannor, S., and Mansour, Y. Thompson sampling for complex bandit problems. *arXiv preprint arXiv:1311.0466*, 2013.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. Following the leader and fast rates in online linear prediction: Curved constraint sets and other regularities. *The Journal of Machine Learning Research*, 18(1):5325–5355, 2017.

Huang, W., Combes, R., and Trinh, C. Towards optimal algorithms for multi-player bandits without collision sensing information. *arXiv preprint arXiv:2103.13059*, 2021.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. Upper confidence bound based decision making strategies and dynamic spectrum access. In *2010 IEEE International Conference on Communications*, pp. 1–5. IEEE, 2010.

Komiyama, J., Honda, J., and Nakagawa, H. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pp. 1152–1161. PMLR, 2015.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543. PMLR, 2015.

Lai, L., Jiang, H., and Poor, H. V. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 98–102, October 2008. doi: 10.1109/ACSSC.2008.5074370.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Landgren, P., Srivastava, V., and Leonard, N. E. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pp. 243–248. IEEE, 2016.

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

Lattimore, T., Crammer, K., and Szepesvári, C. Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems*, 28, 2015.

Liu, K. and Zhao, Q. Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing*, 58(11):5667–5681, 2010.

Magesh, A. and Veeravalli, V. V. Multi-user mabs with user dependent rewards for uncoordinated spectrum access. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 969–972. IEEE, 2019.

Martínez-Rubio, D., Kanade, V., and Rebeschini, P. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Mehrabian, A., Boursier, E., Kaufmann, E., and Perchet, V. A practical algorithm for multiplayer bandits when arm means vary among players. In *International Conference on Artificial Intelligence and Statistics*, pp. 1211–1221. PMLR, 2020.

Mitola, J. and Maguire, G. Q. Cognitive radio: making software radios more personal. *IEEE personal communications*, 6(4):13–18, 1999.

Mourtada, J. and Gaïffas, S. On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20:1–28, 2019.

Perrault, P., Boursier, E., Valko, M., and Perchet, V. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *Advances in Neural Information Processing Systems*, 33:5429–5440, 2020.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pp. 155–163. PMLR, 2016.

Sentenac, F., Boursier, E., and Perchet, V. Decentralized learning in online queuing systems. *Advances in Neural Information Processing Systems*, 34:18501–18512, 2021.

Shi, C., Xiong, W., Shen, C., and Yang, J. Decentralized multi-player multi-armed bandits with no collision information. In *International Conference on Artificial Intelligence and Statistics*, pp. 1519–1528. PMLR, 2020.

Shi, C., Xiong, W., Shen, C., and Yang, J. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in neural information processing systems*, 34:22392–22404, 2021.

Szorenyi, B., Busa-Fekete, R., Hegedus, I., Ormándi, R., Jelasity, M., and Kégl, B. Gossip-based distributed stochastic bandit algorithms. In *International conference on machine learning*, pp. 19–27. PMLR, 2013.

Tekin, C. and Liu, M. Online learning in decentralized multi-user spectrum access with synchronized explorations. In *MILCOM 2012-2012 IEEE Military Communications Conference*, pp. 1–6. IEEE, 2012.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Wang, Q. and Chen, W. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Zuo, J. and Joe-Wong, C. Combinatorial multi-armed bandits for resource allocation. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–4. IEEE, 2021.

A Analysis of Cautious Greedy

A.1 A useful upper bound

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

$$\begin{aligned}
\langle \boldsymbol{\mu} - \boldsymbol{\mu}', g(\mathbf{M}) - g(\mathbf{M}')\rangle &\leq \langle |\boldsymbol{\mu} - \boldsymbol{\mu}'|, |g(\mathbf{M}) - g(\mathbf{M}')|\rangle \\
&\leq \langle |\boldsymbol{\mu} - \boldsymbol{\mu}'|, g(\mathbf{M}) + g(\mathbf{M}')\rangle \\
&\leq \sum_{k=1}^{K} (g(M_k) + g(M'_k)) \\
&\leq \sum_{k=1}^{K} (M_k + M'_k)p \\
&\leq 2Mp
\end{aligned}$$

so that we have

$$\langle \boldsymbol{\mu} - \boldsymbol{\mu}', g(\mathbf{M}) - g(\mathbf{M}')\rangle \leq 2Mp \tag{8}$$

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

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

Algorithm 2 Rotate $\mathcal{U}$ in a round robin fashion over $\mathcal{Y}$ (one iteration)

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

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 $\Delta^{(\nu^*)} = \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu^*-1})\rangle$ and $\Delta_{(\nu^*)} = \mu_{(\nu^*+1)} - \mu_{(\nu^*)}$.

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

$$j = \underset{i \in \mathrm{supp}(\mathbf{M}^*_{\nu^*}), M^*_i \geq 2}{\mathrm{argmin}} \mu_i (g(M^*_i) - g(M^*_{i-1}))$$

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

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 $\mathbf{M}^*_{\nu^*}$ 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

$$\begin{aligned}\Delta^{(\nu^*)} &\geq \mu_{\nu^*+1}(g(M^*_{(\nu^*+1)}+1) - g(M^*_{(\nu^*+1)})) - \mu_{(\nu^*)}p \\ &\geq \Delta_{(\nu^*)}(g(M^*_{(\nu^*+1)}+1) - g(M^*_{(\nu^*+1)})) \\ &\geq \Delta_{(\nu^*)}(g(M) - g(M-1)) \\ &= \Delta_{(\nu^*)}p(1-p)^{M-2}(1-Mp)\end{aligned}$$

□

A.3.2 Proof of Proposition 3.1

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*

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

Then, $P(\overline{\text{GOOD}}) \leq \frac{1}{TK}$

Proof of Lemma A.1. Fix $k \in [K]$, by Hoeffding, we have

$$P(|\hat{\mu}_k(t) - \mu_k| \geq \epsilon | T_k(t) = \tau) \leq 2\exp(-2\tau\epsilon^2)$$

so that we have

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

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

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

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

$$\forall k \in [K], |\hat{\mu}_k(t) - \mu_k| \leq \sqrt{\frac{\log(2T^2K^2)}{2T_k(t)}} \tag{9}$$

which is the desired result. □

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* $R_G = R\mathbb{1}\{\text{GOOD}\}$*, then,*

$$\mathbb{E}[R] \leq \mathbb{E}[R_G] + 2 \tag{10}$$

Proof of Lemma A.2. $\mathbb{E}[R] = \mathbb{E}[R\mathbb{1}_{\text{GOOD}}] + \mathbb{E}[R\mathbb{1}_{\overline{\text{GOOD}}}]$ and $R\mathbb{1}_{\overline{\text{GOOD}}} \leq 2KT\mathbb{1}_{\overline{\text{GOOD}}}$, we then conclude from Lemma A.1. □

Working under the GOOD event makes the analysis much easier. We begin by showing that ν 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. ν 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 $\max_{\mathbf{M}\in\mathcal{M}}\langle\hat{\boldsymbol{\mu}}^L, g(\mathbf{M})\rangle > \max_{\mathbf{M}\in\mathcal{M}_\nu}\langle\hat{\boldsymbol{\mu}}^H, g(\mathbf{M})\rangle$. By the good event, we have

$$\max_{\mathbf{M}\in\mathcal{M}}\langle\hat{\boldsymbol{\mu}}^L, g(\mathbf{M})\rangle < \max_{\mathbf{M}\in\mathcal{M}}\langle\boldsymbol{\mu}, g(\mathbf{M})\rangle$$

and

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

Therefore

$$\begin{aligned}
&\max_{\mathbf{M}\in\mathcal{M}} \langle \hat{\boldsymbol{\mu}}^L, g(\mathbf{M})\rangle > \max_{\mathbf{M}\in\mathcal{M}_\nu} \langle \hat{\boldsymbol{\mu}}^H, g(\mathbf{M})\rangle \\
&\implies \max_{\mathbf{M}\in\mathcal{M}} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle > \max_{\mathbf{M}\in\mathcal{M}_\nu} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle
\end{aligned}$$

and since $\nu = \nu^*$,

$$\max_{\mathbf{M}\in\mathcal{M}_\nu} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle = \max_{\mathbf{M}\in\mathcal{M}} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle.$$

This yields the following contradiction:

$$\max_{\mathbf{M}\in\mathcal{M}} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle > \max_{\mathbf{M}\in\mathcal{M}} \langle \boldsymbol{\mu}, g(\mathbf{M})\rangle.$$

□

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:* $\mathrm{support}(\mathbf{M}^*) \subseteq \mathcal{K}(t)$

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 $\hat{\mu}_k^H < \mu^L_{(\nu+1)}$. But since $\nu \leq \nu^*$, this implies $\hat{\mu}_k^H < \mu^L_{(\nu^*+1)}$. Under the good event $\hat{\mu}_k^H > \mu_k$ and $\mu^L_{(\nu^*+1)} < \mu_{(\nu^*+1)}$ so that $\hat{\mu}_k^H < \mu^L_{(\nu^*+1)}$ implies $\mu_k < \mu_{(\nu^*+1)}$ and therefore $k \notin \mathcal{E}^*_{\nu^*}$. □

Since Cautious Greedy never eliminates any optimal arm and since ν increases, ν will eventually reach ν^* 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 $\hat{\boldsymbol{\mu}}$ become more accurate, making Cautious Greedy's decisions better. Therefore, we introduce $q(t) = \min_{k\in\mathcal{K}(t)} T_k(t)$, 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 $\tau(t) = \min_{k\in\mathcal{K}(t)} \tau_k(t)$. The next Lemma exhibits a lower bound on $\tau(t)$:

Lemma A.5 (Scaling of τ with t)**.** *We have* $\tau(t) \geq \max(\frac{t}{\nu^*+1} - \nu^*, 0)$. *Furthermore, if the condition Line 8 in Algorithm 1 is satisfied, we have* $\tau(t-1) \geq \frac{t-1}{\nu^*+1}$.

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 $t_{n+1} - 1$ (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 U and the number of arms under pressure u after the updates at time $t = t_n$. τ increases linearly between time t_n and t_{n+1} except for u_n time steps where $u_n = \nu_n - |[K] \setminus \mathcal{K}|$ is the number of arms that need to be put under pressure during phase n but that are not yet eliminated and ν_n is the value of ν during phase n.

We have that for $t_n \leq t < t_{n+1}$:

$$\begin{aligned}
\tau(t) &\geq \tau(t_n - 1) + \max(t - (t_n - 1) - u_n, 0) \\
&= \tau(t_n - 1) + \max((t - (t_n - 1))\frac{t_{n+1} - (t_n - 1) - u_n}{t_{n+1} - (t_n - 1)} - (t_{n+1} - t)\frac{u_n}{t_{n+1} - (t_n - 1)}, 0)
\end{aligned}$$

and

$$\tau(t_{n+1}-1)-\tau(t_n-1) = t_{n+1}-(t_n-1)-u_n$$
$$= (t_{n+1}-(t_n-1))\frac{t_{n+1}-(t_n-1)-u_n}{t_{n+1}-(t_n-1)}$$

Since $t_{n+1}-(t_n-1) = |\mathcal{K}_n \setminus \mathcal{A}_n|$ we have:

$$\frac{t_{n+1}-(t_n-1)-u_n}{t_{n+1}-(t_n-1)} = \frac{|\mathcal{K}_n|-|\mathcal{A}_n|-u_n}{|\mathcal{K}_n|-|\mathcal{A}_n|} \geq \frac{1}{u_n+1} \geq \frac{1}{\nu^*+1}$$

It follows that for all $n \geq 1$,

$$\tau(t_n-1) \geq \frac{t_n-1}{\nu^*+1}$$

Therefore, we obtain $t_n \leq t \leq t_{n+1}$

$$\begin{aligned}
\tau(t) &\geq \frac{t_n-1}{\nu^*+1} + \max((t-(t_n-1))\frac{1}{\nu^*+1} - (t_{n+1}-t)\frac{u_n}{t_{n+1}-(t_n-1)}, 0) \\
&\geq \frac{t_n-1}{\nu^*+1} + \max((t-(t_n-1))\frac{1}{\nu^*+1} - \nu^*, 0) \\
&\geq \max(\frac{t}{\nu^*+1} - \nu^*, \frac{t_n-1}{\nu^*+1}) \\
&\geq \max(\frac{t}{\nu^*+1} - \nu^*, 0)
\end{aligned}$$

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

The next step is to link $\tau(t)$ to $q(t)$ by taking collisions into account. By noting that $g(M_k) \geq p$, a first easy observation is that $\mathbb{E}[T_k \mid \tau_k] \geq \tau_k p$. We, therefore, expect q to scale approximately with $p\tau$. The next Lemma shows a more precise statement:

Lemma A.6. *Define* $R_q = \sum R_{q,t}$ *where* $R_{q,t} = R_t \mathbb{1}\{\text{GOOD}\}\, \mathbb{1}\big\{q(t) \geq \frac{1}{3}p\tau(t)\big\}$,

$$\mathbb{E}[R_G] = \mathbb{E}[R_q] + 11(\nu^*+1)KM$$

Proof of Lemma A.6. Rename $T_k(t) = T_k(\tau(t))$, to make the dependence on τ obvious. We have the lower bound $\mathbb{E}[T_k(\tau(t))] \geq p\tau(t)$.

We can write

$$\begin{aligned}
\mathbb{E}[R_G] &= \mathbb{E}[\sum_{t=1}^{(\nu^*+1)^2} R_{G,t} + \sum_{t=(\nu^*+1)^2}^{T} R_{G,t}] \\
&\leq (\nu^*+1)^2 2Mp + \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} R_{G,t}] \qquad \text{(By Equation (8))}
\end{aligned}$$

Then for $R_t \triangleq \sum_{k=1}^{K} \mathbb{E}[\eta_k(\mathbf{M}^*)X_k] - \mathbb{E}[\eta_k^t(\mathbf{M}(t))X_k^t]$,

$$\mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} R_{G,t}]$$

$$= \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} R_t(\mathbb{1}\{\exists k \in \mathcal{K}(t), T_k(\tau(t)) < (1-\rho)\mathbb{E}[T_k(\tau(t))|\mathbf{M}_{[1:t]}]\} + \mathbb{1}\{\forall k \in \mathcal{K}(t), T_k(\tau(t)) \geq (1-\rho)\mathbb{E}[T_k(\tau(t))|\mathbf{M}$$

$$\leq \underbrace{\mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} 2Mp\mathbb{E}[\mathbb{1}\{\exists k \in \mathcal{K}(t), T_k(\tau(t)) < (1-\rho)\mathbb{E}[T_k(\tau(t))|\mathbf{M}_{[1:t]}]\}|\mathbf{M}_{[1:t]}]]}_{(i)} \quad \text{(By Equation (8))}$$

$$+ \underbrace{\sum_{t=(\nu^*+1)^2}^{T} \mathbb{E}[R_t\mathbb{1}\{\forall k \in \mathcal{K}(t), T_k(\tau(t)) \geq (1-\rho)\mathbb{E}[T_k(\tau(t))|\mathbf{M}_{[1:t]}]\}]}_{(ii)}$$

Bounding (i):

$$\begin{aligned}
(i) &\leq \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} 2Mp \sum_{k=1}^{\mathcal{K}(t)} P(T_k(\tau(t)) < (1-\rho)\mathbb{E}[T_k(\tau(t))|\mathbf{M}_{[1:t]}]|\mathbf{M}_{[1:t]})] \\
&\leq \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} 2Mp \sum_{k=1}^{\mathcal{K}(t)} \mathbb{E}[\exp(-\frac{\rho^2}{2}\mathbb{E}[T_k(\tau(t))|\mathbf{M}_{[1:t]}])|\mathbf{M}_{[1:t]}]] \\
&\leq \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T} 2KMp\exp(-\frac{\rho^2}{2}p\tau(t))] \\
&\leq \sum_{t=1}^{+\infty} 2KMp\exp(-\frac{\rho^2}{2}p\frac{t}{\nu^*+1}) && \text{(By Lemma A.5)} \\
&\leq 2KMp\frac{2(\nu^*+1)}{\rho^2 p} \\
&= 2KM\frac{2(\nu^*+1)}{\rho^2}
\end{aligned}$$

Bounding (ii)

$$\begin{aligned}
(ii) &= \sum_{t=1}^{T} \mathbb{E}[R_t\mathbb{1}\{\forall k \in \mathcal{K}(t), T_k(\tau(t)) \geq (1-\rho)\mathbb{E}[T_k(\tau(t))|\mathbf{M}_{[1:t]}]\}] \\
&\leq \sum_{t=1}^{T} \mathbb{E}[R_t\mathbb{1}\{\forall k \in \mathcal{K}(t), T_k(\tau(t)) \geq (1-\rho)p\tau(t)\}]
\end{aligned}$$

Setting $\rho = \frac{2}{3}$ gives

$$\mathbb{E}[R_G] \leq \mathbb{E}[R_q] + 2(\nu^*+1)^2Mp + 9KM(\nu^*+1) \leq \mathbb{E}[R_q] + 11KM(\nu^*+1)$$

$\square$

We can now focus on upper-bounding the different sources of errors. First, $\mathbb{E}[R_q]$ can trivially be written as:

$$\mathbb{E}[R_q] = \mathbb{E}[R_\nu] + \mathbb{E}[R_\mathcal{E}] + \mathbb{E}[R_\mathbf{M}]$$

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

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

$$\text{and } R_\mathbf{M} = \sum_{t=1}^{T} \langle \mu, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t)) \rangle.$$

$\mathbf{M}^*_\nu = \mathbf{M}^{\boldsymbol{\mu}}_{\mathcal{M}_\nu}$ is the optimal assignment of players when at most ν arms can be assigned zero players and $\mathbf{M}^*_{\mathcal{E}(t)} = \mathbf{M}^{\boldsymbol{\mu}}_{\mathcal{M}_{\mathcal{E}(t)}}$ is the optimal assignment of players when only arms not in $\mathcal{E}(t)$ can be assigned zero players. Here and in the rest of the analysis, we have dropped the factors $\mathbb{1}\{\text{GOOD}\}\,\mathbb{1}\big\{q(t) \geq \frac{1}{3}p\tau(t)\big\}$ to simplify the notations.

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

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

Lemma A.7 (Number of iterations before ν increases)**.** *If each active arm has been played without collision at least q times with $q \geq q_k = \frac{8M^2p^2\log(2K^2T^2)}{(\Delta^{(k)})^2}$, then $\nu \geq k$.*

Proof of Lemma A.7. Call $\mathbf{M}^{*,H}_\nu = \text{argmax}_{\mathbf{M}\in\mathcal{M}_\nu} \langle \boldsymbol{\mu}^H, g(\mathbf{M}) \rangle$.

$$\begin{aligned}
&\max_{\mathbf{M}\in\mathcal{M}} \langle \boldsymbol{\mu}^L, g(\mathbf{M}) \rangle > \langle \boldsymbol{\mu}^H, g(\mathbf{M}^{*,H}_\nu) \rangle \\
&\Longleftarrow \langle \boldsymbol{\mu}^L, g(\mathbf{M}^*) \rangle > \langle \boldsymbol{\mu} + 2\boldsymbol{\zeta}, g(\mathbf{M}^{*,H}_\nu) \rangle && \text{(By the GOOD event)} \\
&\Longleftarrow \langle \boldsymbol{\mu} - 2\boldsymbol{\zeta}, g(\mathbf{M}^*) \rangle > \langle \boldsymbol{\mu}, g(\mathbf{M}^*_\nu) \rangle + 2\langle \boldsymbol{\zeta}, g(\mathbf{M}^{*,H}_\nu) \rangle \\
& && \text{(By the GOOD event and optimality of } \mathbf{M}^*_\nu) \\
&\Longleftarrow \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu) \rangle > 2\langle \boldsymbol{\zeta}, g(\mathbf{M}^*) + g(\mathbf{M}^{*,H}_\nu) \rangle \\
&\Longleftarrow \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu) \rangle > 4Mp \max_{k\in\mathcal{K}} \zeta_k && \text{(By Equation (8))} \\
&\Longleftarrow \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu) \rangle > 4Mp \max_{k\in\mathcal{K}} \sqrt{\frac{\log(2T^2K^2)}{2T_k(t)}} \\
&\Longleftarrow \frac{8M^2p^2\log(2K^2T^2)}{(\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu) \rangle)^2} < \min_{k\in\mathcal{K}} T_k \\
&\Longleftarrow \frac{8M^2p^2\log(2K^2T^2)}{(\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu) \rangle)^2} < q
\end{aligned}$$

□

Note that as long as $\nu \leq \nu^*$, R_ν increases by $\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu) \rangle$. Lemma A.7 then allows to bound R_ν:

Lemma A.8 (Bound on R_ν)**.**

$$\mathbb{E}[R_\nu] \leq 72M \min(Mp, K)(\nu^* + 1)\frac{\log(2K^2T^2)}{\Delta^{(\nu^*)}}$$

Proof of Lemma A.8. Call t_v the last time that $\nu(t) = \nu$ and set $t_{\nu^*} = T + 1$ and $t_{-1} = 0$. Note that $t_\nu + 1$ necessarily verifies $t = 0 \mod |\mathcal{U}|$ so that $\tau(t_\nu) \geq \frac{t_\nu}{\nu^*+1}$ according to Lemma A.5.

$$
\begin{aligned}
\mathbb{E}[R_\nu] &= \mathbb{E}[\sum_{t=1}^{T} \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu(t)})\rangle] \\
&= \mathbb{E}[\sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} \underbrace{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle}_{A_\nu}] \\
&= \mathbb{E}[\sum_{\nu=0}^{\nu^*} (t_\nu - t_{\nu-1}) A_\nu] \\
&= \mathbb{E}[\sum_{\nu=0}^{\nu^*} t_\nu A_\nu - \sum_{\nu=0}^{\nu^*} t_{\nu-1} A_\nu] \\
&= \mathbb{E}[\sum_{\nu=1}^{\nu^*} t_{\nu-1}(A_{\nu-1} - A_\nu) - t_{-1}A_0 + t_{\nu^*}A_{\nu^*}] \\
&= \mathbb{E}[\sum_{\nu=1}^{\nu^*} t_{\nu-1}(A_{\nu-1} - A_\nu)] \\
&\leq \mathbb{E}[\sum_{\nu=1}^{\nu^*} (\nu^*+1)\tau(t_{\nu-1})(A_{\nu-1} - A_\nu)] && \text{(By Lemma A.5)} \\
&\leq \mathbb{E}[\sum_{\nu=1}^{\nu^*} \frac{3(\nu^*+1)}{p} q_{\nu-1}(A_{\nu-1} - A_\nu)] && \text{(By Lemma A.6)} \\
&= \frac{3(\nu^*+1)}{p} \sum_{\nu=1}^{\nu^*} \frac{8(Mp)^2 \log(2K^2T^2)\langle \boldsymbol{\mu}, g(\mathbf{M}^*_\nu) - g(\mathbf{M}^*_{\nu-1})\rangle}{\left(\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle\right)^2} && \text{(By Lemma A.7)} \\
&= 24M^2p\log(2K^2T^2)(\nu^*+1) \sum_{\nu=1}^{\nu^*} \underbrace{\left(\frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle} - \frac{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle}{\left(\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle\right)^2} \right)}_{\triangleq l_\nu} \\
&= 24M^2p\log(2K^2T^2)(\nu^*+1) \Bigg[\frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle} \\
&+ \sum_{\nu=1}^{\nu^*-1} \underbrace{\left(\frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle} - \frac{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle}{\left(\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle\right)^2} \right)}_{\triangleq l_\nu} \Bigg]
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{\nu=1}^{\nu^*-1} l_\nu &\leq \sum_{\nu=1}^{\nu^*-1} \left(\frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle} - \frac{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle}{\left(\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle\right)^2} \right) \\
&= \sum_{\nu=1}^{\nu^*-1} \left(\frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle} - \frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle} \right) \left(1 + \frac{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle}\right) \\
&\leq 2 \sum_{\nu=1}^{\nu^*-1} \left(\frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_\nu)\rangle} - \frac{1}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu-1})\rangle} \right) \\
&\leq \frac{2}{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu^*-1})\rangle}
\end{aligned}
$$

We conclude the proof by using $Mp \leq \min(Mp, K)$. □

We now focus on the second term $R_{\mathcal{E}}$. For a given ν, 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 ν, let $\mathcal{E}^*_\nu = \mathrm{support}(\mathbf{M}^*_\nu)$ and let $i \notin \mathcal{E}^*_\nu$ be a sub-optimal arm. When each arm has been played without collision at least q times with*

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

then arm i has necessarily been eliminated.

Proof of Lemma A.9. Since $i \notin \mathcal{E}^*_\nu$, $\mu_i < \mu_{(\nu+1)}$, the algorithm notices that i must be eliminated if

$$\begin{aligned}
&\mu_i^H < \mu_{(\nu+1)}^L \\
&\Longleftarrow \mu_i + 2\zeta_i < \mu_{(\nu+1)} - 2\zeta_{(\nu+1)} \\
&\Longleftarrow \zeta_i + \zeta_{(\nu+1)} < \frac{\mu_{(\nu+1)} - \mu_i}{2} \\
&\Longleftarrow 2\sqrt{\frac{\log(2T^2K^2)}{2q}} < \frac{\mu_{(\nu+1)} - \mu_i}{2} \\
&\Longleftarrow q > \frac{8\log(2T^2K^2)}{(\mu_{(\nu+1)} - \mu_i)^2}
\end{aligned}$$

□

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 ν, let $\mathcal{E}^*_\nu = \mathrm{support}(\mathbf{M}^*_E)$ and let $j \in \mathcal{E}^*_\nu$ an optimal arm. When each arm has been played without collision at least q times with*

$$q \geq q_{A,i} = \frac{8\log(2T^2K^2)}{(\mu_j - \mu_{(\nu)})^2}$$

then arm j has necessarily been accepted.

Proof of Lemma A.10. Since $j \in \mathcal{E}^*_\nu$, $\mu_j > \mu_{(\nu)}$, the algorithm notices that j must be accepted if

$$\mu_{(\nu)}^H < \mu_j^L \tag{11}$$

$$\Longleftarrow \mu_{(\nu)} + 2\zeta_{(\nu)} < \mu_j - 2\zeta_j \tag{12}$$

$$\Longleftarrow \zeta_{(\nu)} + \zeta_j < \frac{\mu_j - \mu_{(\nu)}}{2} \tag{13}$$

$$\Longleftarrow 2\sqrt{\frac{\log(2T^2K^2)}{2q}} < \frac{\mu_j - \mu_{(\nu)}}{2} \tag{14}$$

$$\Longleftarrow q > \frac{8\log(2T^2K^2)}{(\mu_j - \mu_{(\nu)})^2} \tag{15}$$

□

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 $\mathcal{E}$)**.** *Let $\mathcal{E}$ a set of arms of size $K - \nu$ such that $\mathcal{E} \neq \mathcal{E}^*_\nu = \mathrm{support}(\mathbf{M}^*_\nu)$. Then, we have:*

$$\langle \boldsymbol{\mu}, g(\mathbf{M}^*_\nu) - g(\mathbf{M}^*_{\mathcal{E}}) \rangle \leq \\ p\Big(\sum_{i \in \mathcal{E} \setminus \mathcal{E}^*_\nu} \mu_{(\nu+1)} - \mu_i + \sum_{j \in \mathcal{E}^*_\nu \setminus \mathcal{E}} \mu_j - \mu_{(\nu)}\Big)$$

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

$$\mathcal{E} \setminus \mathcal{E}^*_\nu = \{i_1, \dots, i_n\}$$

and indexes $j_1, \dots, j_n$ by

$$\mathcal{E}^*_\nu \setminus \mathcal{E} = \{j_1, \dots, j_n\}$$

We now construct $\mathbf{M}_\mathcal{E}$. 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

$$\forall k \in [n], \mathbf{M}_\mathcal{E}[i_k] = 1$$

and

$$\forall k \in [n], \mathbf{M}_\mathcal{E}[j_k] = \mathbf{M}^*[j_k] - 1$$

and other arms are untouched:

$$\forall k \in \mathcal{E}^*_\nu \cap \mathcal{E}, \mathbf{M}_\mathcal{E}[k] = \mathbf{M}^*[k]$$

The cost is given by:

$$\begin{aligned}
\langle \boldsymbol{\mu}, g(\mathbf{M}^*_\nu) - g(M^*_\mathcal{E})\rangle &\leq \langle \boldsymbol{\mu}, g(\mathbf{M}^*_\nu) - g(\mathbf{M}_\mathcal{E})\rangle \\
&= \sum_{k=1}^{n} (\mu_{j_k} [g(\mathbf{M}^*[j_k]) - g(\mathbf{M}^*[j_k] - 1)] - \mu_{i_k} p) \\
&\leq \sum_{k=1}^{n} (\mu_{j_k} - \mu_{i_k}) p \\
&\leq \sum_{k=1}^{n} (\mu_{j_k} - \mu_{(\nu)} + \mu_{(\nu+1)} - \mu_{i_k}) p \\
&\leq p(\sum_{i \in \mathcal{E} \setminus \mathcal{E}^*_\nu} \mu_{(\nu+1)} - \mu_i + \sum_{j \in \mathcal{E}^*_\nu \setminus \mathcal{E}} \mu_j - \mu_{(\nu)})
\end{aligned}$$

□

We can now bound $R_\mathcal{E}$:

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

$$\mathbb{E}[R_\mathcal{E}] \leq \frac{120 \log(2K^2T^2)}{\Delta^{(\nu^*)}} + \sum_{i=1}^{\nu^*} \frac{120 \log(2T^2K^2)}{\Delta_{(i)}}$$

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

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

$$\begin{aligned}
&= \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} \langle \boldsymbol{\mu}, g(\mathbf{M}^*_\nu) - g(\mathbf{M}^*_{\mathcal{E}(t)})\rangle \\
&\leq \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p(\sum_{i \in \mathcal{E}(t) \setminus \mathcal{E}^*_\nu} (\mu_{(\nu+1)} - \mu_i) + \sum_{j \in \mathcal{E}^*_\nu \setminus \mathcal{E}(t)} (\mu_j - \mu_{(\nu)})) \qquad \text{(Using Lemma A.11)} \\
&= \underbrace{\sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \sum_{i \in \mathcal{E}(t) \setminus \mathcal{E}^*_\nu} (\mu_{(\nu+1)} - \mu_i)}_{(i)} + \underbrace{\sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \sum_{j \in \mathcal{E}^*_\nu \setminus \mathcal{E}(t)} (\mu_j - \mu_{(\nu)})}_{(ii)}
\end{aligned}$$

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 1 in Algorithm 1 is satisfied. Note again that updates of $\mathcal{A}$, $\mathcal{K}$, and ν occur at the beginning of each phase. Denote $\mathcal{N}_\nu$ the phases between $t_{\nu-1} + 1$ and t_ν.

Bounding (i) Denote τ_n the number of pulls of active arms at the end of phase n.

$$
\begin{aligned}
(i) &\leq \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \sum_{i \notin \mathcal{E}_\nu^*} (\mu_{(\nu+1)} - \mu_i) \mathbb{1}\{i \in \mathcal{E}(t)\} \\
&= \sum_{\nu=1}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \sum_{i \notin \mathcal{E}_\nu^*} (\mu_{(\nu+1)} - \mu_i) \mathbb{1}\{i \in \mathcal{E}(t)\} \\
&= \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} \sum_{t=t_{\nu-1}+1}^{t_\nu} p \sum_{i \notin \mathcal{E}_\nu^*} (\mu_{(\nu+1)} - \mu_i) \mathbb{1}\{i \in \mathcal{E}(t)\} \, \mathbb{1}\{ \, t \text{ belong to phase } n \, \} \\
&\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{i \notin \mathcal{E}_\nu^*} (\mu_{(\nu+1)} - \mu_i) (\text{ Number of times arm } i \text{ is pulled during phase } n \,) \\
&= \sum_{\nu=1}^{\nu^*} \sum_{n \in N_\nu} p \sum_{i=1}^{\nu} (\mu_{(\nu+1)} - \mu_{(i)}) (\text{ Number of times arm } (i) \text{ is pulled during phase } n \,) \\
&= p \sum_{i=1}^{\nu^*} \underbrace{\sum_{\nu=i}^{\nu^*} \sum_{n \in N_\nu} (\mu_{(\nu+1)} - \mu_{(i)}) (\text{ Number of times arm } (i) \text{ is pulled during phase } n \,)}_{s_i}
\end{aligned}
$$

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 rejected. Call ν_i the epoch where arm (i) is eliminated. This means $n_{E,(i)} \in N_{\nu_i}$. Call $s_i = \sum_{\nu=i}^{\nu_i} (\mu_{\nu+1} - \mu_{(i)})(T_{\nu,i} - T_{\nu-1,i})$.

We have

$$
s_i = \underbrace{(\mu_{\nu_i+1} - \mu_{(i)})(T_{\nu_i,i} - T_{\nu_i-1,i})}_{(iii)} + \underbrace{\sum_{\nu=i}^{\nu_i-1} (\mu_{(\nu+1)} - \nu_{(i)})(T_{\nu,i} - T_{\nu-1,i})}_{(iv)}
$$

By Lemma A.6 and Lemma A.9, $T_{\nu_i,i} \leq \frac{24 \log(2T^2K^2)}{p(\mu_{(\nu_i+1)} - \mu_{(i)})^2}$ so that

$$
(iii) \leq \frac{24 \log(2T^2K^2)}{p(\mu_{(\nu_i+1)} - \mu_{(i)})}
$$

We also have

$$
\begin{aligned}
(iv) &\leq (\mu_{(\nu_i)} - \nu_{(i)}) \sum_{\nu=i}^{\nu_i-1} (T_{\nu,i} - T_{\nu-1,i}) \\
&\leq (\mu_{(\nu_i)} - \nu_{(i)}) T_{\nu_i-1,i}
\end{aligned}
$$

By Lemma A.6 and Lemma A.9, $T_{\nu_i-1,i} \leq \frac{24\log(2T^2K^2)}{p(\mu_{(\nu_i)}-\mu_{(i)})^2}$ and by Lemma A.7, $T_{\nu_i-1,i} \leq \frac{24M^2p\log(2K^2T^2)}{\Delta^2_{\nu_i-1}}$ so that

$$\begin{aligned}
T_{\nu_i-1,i} &\leq \min\left(\frac{24M^2p\log(2K^2T^2)}{\Delta^2_{\nu_i-1}}, \frac{24\log(2T^2K^2)}{p(\mu_{(\nu_i)}-\mu_{(i)})^2}\right) \\
&\leq \sqrt{\frac{24M^2p\log(2K^2T^2)}{\Delta^2_{\nu_i-1}}\frac{24\log(2T^2K^2)}{p(\mu_{(\nu_i)}-\mu_{(i)})^2}} \\
&= \frac{24M\log(2K^2T^2)}{\Delta_{\nu_i-1}(\mu_{(\nu_i)}-\mu_{(i)})}
\end{aligned}$$

and therefore

$$(iv) \leq \frac{24M\log(2K^2T^2)}{\Delta_{\nu_i-1}}$$

Then either $\nu_i = \nu^*$ and

$$s_i \leq \frac{24\log(2T^2K^2)}{\mu_{(\nu^*+1)}-\mu_{(i)}} + \frac{24M\log(2K^2T^2)}{\Delta_{\nu^*-1}}$$

or $\nu_i < \nu^*$ and then,

$$\begin{aligned}
s_i &= \sum_{\nu=i}^{\nu_i}(\mu_{\nu+1}-\mu_{(i)})(T_{\nu,i}-T_{\nu-1,i}) \\
&\leq (\mu_{\nu_i+1}-\mu_{(i)})\sum_{\nu=i}^{\nu_i}(T_{\nu,i}-T_{\nu-1,i}) \\
&\leq (\mu_{\nu_i+1}-\mu_{(i)})T_{\nu_i,i} \\
&\leq \frac{24M\log(2K^2T^2)}{\Delta_{\nu_i}} \\
&\leq \frac{24M\log(2K^2T^2)}{\Delta_{\nu^*-1}}
\end{aligned}$$

where at the last line we used again Lemma A.9 and Lemma A.7.
So in any case

$$(i) \leq \sum_{i=1}^{\nu^*}\frac{24\log(2T^2K^2)}{\mu_{(\nu^*+1)}-\mu_{(i)}} + p\nu^*\frac{24M\log(2K^2T^2)}{\Delta_{\nu^*-1}}$$

Bounding (ii) we have

$$(ii) \leq \sum_{\nu=0}^{\nu^*}\sum_{t=t_{\nu-1}+1}^{t_\nu} p\sum_{j\in\mathcal{E}^*_\nu}(\mu_j-\mu_{(\nu)})\mathbb{1}\{j\notin\mathcal{E}(t)\}$$

We call $u_n = \nu_n - |[K] \setminus \mathcal{K}_n|$ the number of arms put under pressure during phase n. We have:

$$
\begin{aligned}
(ii) &\leq \sum_{\nu=0}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} p \sum_{j \in \mathcal{E}_\nu^*} (\mu_{(j)} - \mu_{(\nu)}) \sum_{t=t_{\nu-1}+1}^{t_\nu} \mathbb{1}\{j \notin \mathcal{E}(t)\} \, \mathbb{1}\{ \, t \text{ belong to phase } n \, \} \\
&\leq \sum_{\nu=0}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} p \sum_{j \in \mathcal{E}_\nu^*} (\mu_{(j)} - \mu_{(\nu)}) (\text{ Number of times arm } j \text{ is not pulled during phase } n) \\
&= \sum_{\nu=1}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} p \sum_{j \in \mathcal{E}_\nu^*} (\mu_{(j)} - \mu_{(\nu)}) (\text{ Number of times arm } j \text{ is not pulled during phase } n) \\
&\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} p \sum_{j \in \mathcal{E}_\nu^*} (\mu_{(j)} - \mu_{(\nu)}) u_n \mathbb{1} \left\{ n \leq \underbrace{n_{A,j}}_{\text{Last phase where arm } j \text{ is not accepted}} \right\} \\
&\leq \sum_{\nu=1}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} p \sum_{j \in \mathcal{E}_\nu^*} (\mu_{(j)} - \mu_{(\nu)}) \sum_{\nu'=1}^{\nu} \mathbb{1} \left\{ n \leq \underbrace{n_{E,\nu'}}_{\text{Last phase where arm } \nu' \text{ is not rejected}} \right\} \mathbb{1}\{n \leq n_{A,j}\} \\
&= \sum_{\nu'=1}^{\nu^*} p \underbrace{\sum_{\nu=\nu'}^{\nu^*} \sum_{j=\nu+1}^{K} \sum_{n \in \mathcal{N}_\nu} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \leq n_{E,\nu'}\} \, \mathbb{1}\{n \leq n_{A,j}\}}_{(A)}
\end{aligned}
$$

Call n_ν the last phase before ν is increased. We have that

$$
\begin{aligned}
(A) &= \sum_{\nu=\nu'}^{\nu^*-1} \sum_{j=\nu+1}^{K} \sum_{n \in \mathcal{N}_\nu} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \leq n_{E,\nu'}\} \, \mathbb{1}\{n \leq n_{A,j}\} \\
&\quad + \sum_{j=\nu^*+1}^{K} \sum_{n \in \mathcal{N}_{\nu^*}} (\mu_{(j)} - \mu_{(\nu^*)}) \mathbb{1}\{n \leq n_{E,\nu'}\} \, \mathbb{1}\{n \leq n_{A,j}\} \\
&\leq \sum_{n \in \mathbb{N}} \sum_{j=\nu+1}^{K} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \leq n_{A,j}\} \, \mathbb{1}\{n \leq n_{\nu^*-1}\} + \sum_{j=\nu^*+1}^{K} \sum_{n \in \mathbb{N}} (\mu_{(j)} - \mu_{(\nu^*)}) \mathbb{1}\{n \leq n_{E,\nu^*}\} \, \mathbb{1}\{n \leq n_{A,j}\}
\end{aligned}
$$

Call τ_n the value of τ at the end of phase n. First notice that

$$
\begin{aligned}
n \leq n_{A,j} &\implies q_{n-1} \leq q_{A,j} && (q_{n-1}\text{: value of } q \text{ at the end of phase } n-1) \\
&\implies \frac{1}{3} p \tau_{n-1} \leq q_{A,j} && \text{(By Lemma A.6)} \\
&\implies \tau_{n-1} \leq \frac{24 \log(2T^2K^2)}{(\mu_{(j)} - \mu_{(\nu)})^2 p} && \text{(By Lemma A.10)} \\
&\implies \mu_{(j)} - \mu_\nu \leq \sqrt{\frac{24 \log(2T^2K^2)}{\tau_{n-1} p}} \triangleq \delta_n
\end{aligned}
$$

Next, we write:

$$
\begin{aligned}
\sum_{j=\nu+1}^{K} \mathbb{1}\{n \leq n_{A,j}\} &= \text{ Number of arms not yet accepted at phase } n \\
&= \tau_n - \tau_{n-1}
\end{aligned}
$$

so we have:

$$(A) \leq \sum_{n\in\mathbb{N}} (\tau_n - \tau_{n-1})\delta_n \Big(\mathbb{1}\{n \leq n_{\nu^*-1}\} + \mathbb{1}\{n \leq n_{E,\nu^*}\} \Big)$$

Note that $\tau_n - \tau_{n-1}$ is the number of pulls during phase n which is equal to $|\mathcal{K}_n \setminus \mathcal{A}_n| - u_n$ and therefore equal to the number of arms that should be accepted but are not yet accepted.

Using the identity $\sqrt{\frac{24\log(2T^2K^2)}{\tau_{n-1}p}} = \delta_n$, we get

$$\begin{aligned}
\tau_n - \tau_{n-1} &= \frac{24\log(2T^2K^2)}{p}\left(\frac{1}{\delta_{n+1}^2 - \delta_n^2}\right) \\
&= \frac{24\log(2T^2K^2)}{p}\left(\frac{1}{\delta_n} + \frac{1}{\delta_{n+1}}\right)\left(\frac{1}{\delta_{n+1}} - \frac{1}{\delta_n}\right)
\end{aligned}$$

Let us now argue that for any $n >= 1$, $\tau_n \leq 2\tau_{n-1}$. First, let us notice that

$$\begin{aligned}
\tau_n - \tau_{n-1} &= |\mathcal{K}_{n-1} \setminus \mathcal{A}_{n-1}| - (\nu_{n-1} - |[K] \setminus \mathcal{K}_{n-1}| \\
&= |\mathcal{K}_{n-1}| - |\mathcal{A}_{n-1}| - \nu_{n-1} + K - |\mathcal{K}_{n-1}| \\
&= K - |\mathcal{A}_{n-1}| - \nu_{n-1} \\
&\leq K
\end{aligned}$$

Then note that $\tau_0 = K$ (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

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

We can then write:

$$(A) \leq \frac{24\log(2T^2K^2)}{p}(\sqrt{2}+1)\sum_{n\in\mathbb{N}}\left(\frac{1}{\delta_{n+1}} - \frac{1}{\delta_n}\right)\Big(\mathbb{1}\{n \leq n_{\nu^*-1}\} + \sum_{n\in\mathbb{N}} \mathbb{1}\{n \leq n_{E,\nu}\}\Big)$$

We have

$$\begin{aligned}
n \leq n_{E,\nu^*} &\implies q_n \leq q_{E,\nu^*} \\
&\implies \frac{1}{3}p\tau_{n-1} \leq q_{E,\nu^*} && \text{(By Lemma A.6)} \\
&\implies \tau_{n-1} \leq \frac{24\log(2T^2K^2)}{(\mu_{(\nu^*+1)} - \mu_{(\nu^*)})^2 p} && \text{(By Lemma A.9)} \\
&\implies \mu_{(\nu^*+1)} - \mu_{(\nu^*)} \leq \sqrt{\frac{24\log(2T^2K^2)}{\tau_{n-1}p}} = \delta_n
\end{aligned}$$

so that

$$\mu_{(\nu^*+1)} - \mu_{(\nu)} \leq \delta_{n_{E,\nu}}$$

Similarly

$$\begin{aligned}
n \leq n_{\nu} &\implies q_n \leq q_{\nu} \\
&\implies \frac{1}{3}p\tau_{n-1} \leq q_{\nu} && \text{(By Lemma A.6)} \\
&\implies \tau_{n-1} \leq \frac{24M^2p^2\log(2T^2K^2)}{\Delta_\nu^2 p} && \text{(By Lemma A.7)} \\
&\implies \frac{\Delta_\nu}{Mp} \leq \sqrt{\frac{24\log(2T^2K^2)}{\tau_{n-1}p}} = \delta_n
\end{aligned}$$

so that

$$\frac{\Delta_{\nu^*-1}}{Mp} \leq \frac{\Delta_\nu}{Mp} \leq \delta_{n_\nu}$$

Using again that $\frac{1}{\delta_n} \leq \sqrt{2}\frac{1}{\delta_{n-1}}$, we get

$$(A) \leq \frac{24\log(2T^2K^2)}{p}(\sqrt{2}+1)\sqrt{2}\frac{1}{\mu_{(\nu^*+1)}-\mu_{\nu'}} + Mp\frac{24\log(2T^2K^2)}{p}(\sqrt{2}+1)\sqrt{2}\frac{1}{\Delta_{\nu^*-1}}$$

so that

$$(ii) \leq \sum_{\nu'=1}^{\nu^*} \frac{96\log(2T^2K^2)}{\mu_{(\nu^*+1)}-\mu_{\nu'}} + \nu^* Mp\frac{96\log(2T^2K^2)}{\Delta_{\nu^*-1}}$$

where we used $2+\sqrt{2} \leq 4$

From the bound of (i) and (ii), we get:

$$R_{\mathcal{E}} \leq \sum_{\nu'=1}^{\nu} \frac{120\log(2T^2K^2)}{\mu_{(\nu^*+1)}-\mu_{\nu'}} + \nu^* Mp\frac{120\log(2K^2T^2)}{\Delta_{\nu^*-1}}$$

Since $Mp \leq \frac{1}{K}$ we have the result. □

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)**.**

$$\mathbb{E}[R_M] \leq 2Mp(\nu^*+1)^2 + 2MK\frac{3(\nu^*+1)}{r}$$

Proof of Lemma A.13. The proof of Lemma A.13 follows similar techniques as Huang et al. (2017).

$$\begin{aligned}
\mathbb{E}[R_M] &= \mathbb{E}[\sum_{t=1}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] \\
&= \mathbb{E}[\sum_{t=1}^{(\nu^*+1)^2}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] + \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] \\
&\leq 2Mp(\nu^*+1)^2 + \underbrace{\mathbb{E}[\sum_{t=(\nu^*+1)^2}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle]}_{(i)} \qquad \text{(By Equation (8))}
\end{aligned}$$

Then we write

$$\begin{aligned}
&\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle \leq \mathbb{E}[\langle \boldsymbol{\mu}-\hat{\boldsymbol{\mu}}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] \\
&\leq 2Mp\mathbb{E}[|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}|_\infty \mathbb{1}\{|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}|_\infty \geq r\}] \qquad \text{(By Equation (8) and definition of } r\text{)} \\
&\leq 2Mp\mathbb{E}[\Big(r\mathbb{P}\{|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}|_\infty \geq r\} + \int_r^\infty \mathbb{P}\{|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}|_\infty \geq \varepsilon\}d\varepsilon\Big)] \\
&\leq \mathbb{E}[2Mp\sum_{k\in\mathcal{K}(t)}\Big(r\exp(-2q(t)r^2) + \int_r^\infty \exp(-2q(t)\varepsilon^2)d\varepsilon\Big)]
\end{aligned}$$

so we have

$$
\begin{aligned}
(i) &\leq \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{\infty} 2MpK\Big(r\exp(-2q(t)r^2) + \int_r^{\infty}\exp(-2q(t)\varepsilon^2)d\varepsilon\Big)] \\
&\leq \mathbb{E}[\sum_{t=(\nu^*+1)^2}^{\infty} 2MpK\Big(r\exp(-2\frac{1}{3}p\tau(t)r^2) + \int_r^{\infty}\exp(-2\frac{1}{3}p\tau(t)\varepsilon^2)d\varepsilon\Big)] \quad \text{(By Lemma A.6)} \\
&\leq \sum_{t=(\nu^*+1)^2}^{\infty} 2MpK\Big(r\exp(-2\frac{1}{3}p\max(\frac{t}{\nu^*+1}-\nu^*,0)r^2) + \int_r^{\infty}\exp(-2\frac{1}{3}p\max(\frac{t}{\nu^*+1}-\nu^*,0)\varepsilon^2)d\varepsilon\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By Lemma A.5)} \\
&\leq \sum_{t=1}^{\infty} 2MpK\Big(r\exp(-2\frac{1}{3}p\frac{t}{\nu^*+1}r^2) + \int_r^{\infty}\exp(-2\frac{1}{3}p\frac{t}{\nu^*+1}\varepsilon^2)d\varepsilon\Big) \\
&\leq 2MpK\Big(r\frac{3(\nu^*+1)}{2pr^2} + \int_r^{\infty}\frac{3(\nu^*+1)}{2p\varepsilon^2}d\varepsilon\Big) \\
&\leq 2MpK\frac{3(\nu^*+1)}{2p}\Big(\frac{1}{r}+\frac{1}{r}\Big) \\
&= 2MK\frac{3(\nu^*+1)}{r}
\end{aligned}
$$

so that

$$
\mathbb{E}[R_M] \leq 2Mp(\nu^*+1)^2 + 2MK\frac{3(\nu^*+1)}{r}
$$

□

The upper bound of Cautious Greedy in Proposition 3.1 follows by combining the previous lemmas. We use $2Mp(\nu^*+1)^2 \leq \frac{2MK(\nu^*+1)}{r}$, $11(\nu^*+1)KM \leq \frac{11KM(\nu^*+1)}{r}$ and $2 \leq \frac{KM(\nu^*+1)}{r}$ to bound the additive terms.

A.4 Proof of Lemma 4.1

Proof. We assume $M = 2N+1$. Take $m_1 = \frac{1}{2}, m_2 = \frac{1}{2}+\Delta$, $\boldsymbol{\mu}_1 = (m_1, m_2)$ and $\boldsymbol{\mu}_2 = (m_2, m_1)$.

Condition on Δ such that $\mathbf{M}^* = (N, N+1)$ if $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $\mathbf{M}^* = (N+1, N)$ if $\boldsymbol{\mu} = \boldsymbol{\mu}_2$ Let us first find Δ such that the optimal assignment is $(N, N+1)$ when $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $(N+1, N)$ when $\boldsymbol{\mu} = \boldsymbol{\mu}_2$. Assume $\boldsymbol{\mu} = \boldsymbol{\mu}_1$, the reasoning is symmetric for $\boldsymbol{\mu} = \boldsymbol{\mu}_2$. We want to find Δ such that for any $-(N+1) \leq x \leq N$ such that $x \neq 0$:

$$
g(N-x)\frac{1}{2} + g(N+1+x)(\frac{1}{2}+\Delta) \leq g(N)\frac{1}{2} + g(N+1)(\frac{1}{2}+\Delta) \tag{16}
$$

First for $x = N$ we look for Δ in the form $\Delta = \mathcal{O}(p)$

$$
\begin{aligned}
& g(2N+1)(\frac{1}{2}+\Delta) \leq g(N)\frac{1}{2} + g(N+1)(\frac{1}{2}+\Delta) \\
& \iff (2N+1)(1-p)^{2N}(\frac{1}{2}+\Delta) \leq N\frac{1}{2} + (N+1)(1-p)(\frac{1}{2}+\Delta) \\
& \impliedby (2N+1)(1-p)(\frac{1}{2}+\Delta) \leq N\frac{1}{2} + (N+1)(1-p)(\frac{1}{2}+\Delta) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using } (1-p)^{2N} \leq (1-p)) \\
& \iff N(1-p)(\frac{1}{2}+\Delta) \leq N\frac{1}{2} \\
& \iff \Delta \leq \frac{1}{2}(\frac{1}{1-p} - 1) \\
& \iff \Delta \leq \frac{p}{2(1-p)}
\end{aligned}
$$

Then if $\Delta \leq \frac{p}{1-p}$, Equation (16) is satisfied for $x = N$.

For $x = -(N+1)$, the left-hand side of Equation (16) is $g(2N+1)\frac{1}{2} \leq g(2N+1)(\frac{1}{2}+\Delta)$ so if $\Delta \leq \frac{p}{1-p}$ Equation (16) is satisfied for $x = -(N+1)$.

For $0 < x < N$, we have

$$
\begin{aligned}
& g(N-x)\frac{1}{2} + g(N+1+x)(\frac{1}{2}+\Delta) \leq g(N)\frac{1}{2} + g(N+1)(\frac{1}{2}+\Delta) \\
& \iff (N-x)\frac{1}{2} + (N+1+x)(1-p)^{1+2x}(\frac{1}{2}+\Delta) \leq N(1-p)^x + (N+1)(1-p)^{x+1}(\frac{1}{2}+\Delta) \\
& \impliedby \Big(x(1-p)^{x+1}\Big)(\frac{1}{2}+\Delta) \leq N(1-p)^x - (N-x)\frac{1}{2} \quad \text{(Using } (1-p)^{2x+1} \leq (1-p)^{x+1}) \\
& \impliedby \Big(x(1-p)^{x+1}\Big)(\frac{1}{2}+\Delta) \leq \frac{x}{2} \qquad \text{(Using } (1-p)^x \geq \tfrac{1}{\sqrt{e}} \geq \tfrac{1}{2} \text{ since } x \leq \tfrac{1}{-2\log(1-p)}) \\
& \impliedby \Big(x(1-p)\Big)(\frac{1}{2}+\Delta) \leq \frac{x}{2} \qquad \text{(Using } (1-p)^{x+1} \leq (1-p)) \\
& \iff \Delta \leq \frac{1}{2}(\frac{1}{1-p} - 1) \\
& \iff \Delta \leq \frac{p}{2(1-p)}
\end{aligned}
$$

Therefore if $\Delta \leq p$, Equation (16) 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 $g(N-x)\frac{1}{2} + g(N+1+x)(\frac{1}{2}+\Delta) = g(N+y+1)\frac{1}{2} + g(N-y)(\frac{1}{2}+\Delta) < g(N+y+1)(\frac{1}{2}+\Delta) + g(N-y)\frac{1}{2}$ which gives the desired inequality for $y = 0$. For $y > 0$, Equation (16) is satisfied if $\Delta \leq \frac{p}{1-p}$. Therefore if $\Delta \leq \frac{p}{1-p}$, Equation (16) is satisfied and therefore, the optimal assignment if $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ is $\mathbf{M}^* = (N, N+1)$.

Computing r Let us now compute r. Assume again $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and the reasoning is symmetric for $\boldsymbol{\mu} = \boldsymbol{\mu}_2$. We have $\mathcal{M}_1 \cup \mathcal{M}_0 = \mathcal{M}$ and we know $\mathbf{M}^* = (N, N+1)$ so that $r = \min_{\boldsymbol{\mu}', \text{argmax}_{\mathbf{M}\in\mathcal{M}}\langle \boldsymbol{\mu}', g(\mathbf{M})\rangle \neq \mathbf{M}^*} \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty$. Call $\boldsymbol{\mu}_r = \text{argmin}_{\boldsymbol{\mu}', \text{argmax}_{\mathbf{M}\in\mathcal{M}}\langle \boldsymbol{\mu}', g(\mathbf{M})\rangle \neq \mathbf{M}^*} \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty$ and $\mathbf{M}_r = \text{argmax}_{\mathbf{M}\in\mathcal{M}}\langle \boldsymbol{\mu}_r, g(\mathbf{M})\rangle$.

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

r_1 is the minimum value such that

$$g(N+1)(\frac{1}{2}+r_1)+g(N)(\frac{1}{2}+\Delta-r_1) \geq g(N)(\frac{1}{2}+r_1)+g(N+1)(\frac{1}{2}+\Delta-r_1)$$
$$\iff (g(N+1)-g(N))(\frac{1}{2}+r_1) \geq (g(N+1)-g(N))(\frac{1}{2}+\Delta-r_1)$$

and therefore $r_1 = \frac{\Delta}{2}$

r_2 is the minimum value such that

$$g(N-1)(\frac{1}{2}-r_2)+g(N+2)(\frac{1}{2}+\Delta+r_2) \geq g(N)(\frac{1}{2}-r_2)+g(N+1)(\frac{1}{2}+\Delta+r_2)$$
$$\iff (g(N+2)-g(N+1))(\frac{1}{2}+\Delta+r_2) \geq (g(N)-g(N-1))(\frac{1}{2}-r_2)$$
$$\iff r_2(g(N+2)-g(N+1)+g(N)-g(N-1)) \geq (g(N+1)-g(N+2))(\frac{1}{2}+\Delta)+(g(N)-g(N-1))\frac{1}{2}$$
$$\implies r_2 \geq \frac{(g(N+1)-g(N+2))(\frac{1}{2}+\Delta)+(g(N)-g(N-1))\frac{1}{2}}{2(g(N)-g(N-1))}$$
$$\iff r_2 \geq \frac{((N+1)(1-p)^2-(N+2)(1-p)^3)(\frac{1}{2}+\Delta)+(N(1-p)-(N-1))\frac{1}{2}}{2(N(1-p)-N-1)}$$
$$\iff r_2 \geq \frac{(1-p)^2((N+2)p-1)(\frac{1}{2}+\Delta)+(1-Np)\frac{1}{2}}{2(1-Np)}$$
$$\implies r_2 \geq \frac{2p(1-p)^2\frac{1}{2}+(1-p)^2((N+2)p-1)\Delta}{2(1-Np)} \qquad \text{(Using } (1-p)^2 \leq 1\text{)}$$
$$\implies r_2 \geq \frac{\frac{1}{4}(p-\Delta)}{2(1-Np)} \qquad \text{(Using } (1-p)^2 \geq \tfrac{1}{4} \text{ since } p \leq \tfrac{1}{2}\text{)}$$
$$\implies r_2 \geq \frac{1}{4}(p-\Delta) \qquad \text{(Using } p \leq \tfrac{1}{2N}\text{)}$$

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

Improve the power of the algorithm Let A be any algorithm that we run on data $\boldsymbol{\mu}$ such that either $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ or $\boldsymbol{\mu} = \boldsymbol{\mu}_2$ (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 $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$. Furthermore, at each time step, A chooses $\mathbf{M}(t)$ 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 $\mathbf{M}(t) = (N, N+1)$ and playing arm 2 means playing $\mathbf{M}(t) = (N+1, N)$. Call i^* the optimal arm.

We follow the technique used in Wang & Chen (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 $\mathbf{X}(t) = (X_{1t}, X_{2t})$ where $X_{it} = 1$ with probability $g(M)\mu_i$, $X_{it} = 0$ with probability $g(M)(1-\mu_i)$ and $X_{it} = \perp$ with probability $1 - g(M)$.

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

$$\begin{aligned}
\mathbb{E}[R_A] &= \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\} \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}(t))\rangle] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\} (\frac{1}{2} + \Delta)(g(N+1) - g(N)) + (\frac{1}{2})(g(N) - g(N+1))] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\} (\Delta)(g(N+1) - g(N))
\end{aligned}$$

Then, the rest of the proof is then identical to Mourtada & Gaïffas (2019). Call for $i = 1, 2$, let $\mathbb{P}_i$ be the joint probability on $(\mathbf{X}(1), \ldots, \mathbf{X}(T))$ when $\boldsymbol{\mu} = \boldsymbol{\mu}_i$.

The regret incurred by A on the worst choice of $\boldsymbol{\mu}$ is higher than the regret incurred by choosing the worst between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

$$\begin{aligned}
\mathbb{E}[R_A] &\geq \max_{i^* \in \{1,2\}} \mathbb{E}_{i^*}[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\} (\Delta)(g(N+1) - g(N))] \\
&\geq \frac{1}{2} \sum_{i^*=1}^{2} \mathbb{E}_{i^*}[\sum_{t=1}^{T} \mathbb{1}\{i_t \neq i^*\} (\Delta)(g(N+1) - g(N))] \\
&= \frac{\Delta(g(N+1) - g(N))}{2} \sum_{i^*=1}^{2} \mathbb{E}_{i^*}[T - \underbrace{N_{i^*}}_{\triangleq \sum_{t=1}^{T} \mathbb{1}\{i_t = i^*\}}] \\
&\geq \frac{\Delta(g(N+1) - g(N))}{2} \frac{T}{2} \sum_{i^*=1}^{2} \mathbb{P}_{i^*}[\frac{T}{2} \geq N_{i^*}] \\
&\geq \frac{\Delta(g(N+1) - g(N))}{2} \frac{T}{2} (\mathbb{P}_1(N_1 \geq \frac{T}{2}) + \mathbb{P}_2(N_2 \geq \frac{T}{2}))
\end{aligned}$$

Then by Bretagnolle–Huber inequality (Th 14.2 in Lattimore & 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

$$\begin{aligned}
KL(\mathbb{P}_1, \mathbb{P}_2) &\leq Tg(M)(KL(\mathcal{B}((\frac{1}{2} + \Delta)), \mathcal{B}(\frac{1}{2})) + KL(\mathcal{B}((\frac{1}{2})), \mathcal{B}(\frac{1}{2} + \Delta))) \\
&\leq 4Tg(M)\Delta^2
\end{aligned}$$

and therefore

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

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

$$
\begin{aligned}
\mathbb{E}[R_A] &\geq \frac{(g(N+1)-g(N))}{64g(M)\Delta}\exp(-1) \\
&\geq \frac{(g(N+1)-g(N))}{64Mp\Delta}\exp(-1) \\
&= \frac{((N+1)(1-p)-N)}{64M\Delta}\exp(-1) \\
&= \frac{(1-(N+1)p)}{64M\Delta}\exp(-1) \\
&\geq \frac{1}{128M\Delta}\exp(-1) && \text{(Using } p \leq \tfrac{1}{2(N+1)}\text{)}
\end{aligned}
$$

□

A.5 Proof of Lemma 4.2

Take $K = \nu^* + 2$ arms, M players and $\boldsymbol{\mu} = (\mu_1, \mu_0, \mu_0 + \Delta_{(1)} - \Delta_{(2)}, \ldots, \mu_0 + \Delta_{(1)} - \Delta_{(\nu^*)}, \mu_0 + \Delta_{(1)})$. For simplicity denote $\Delta = \Delta_{(1)}$.

Let us choose μ_1, μ_0 and Δ 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:

$$
g(M-1)\mu_1 + g(1)(\mu_0 + \Delta) \geq g(M-2)\mu_1 + 2g(1)(\mu_0 + \Delta) \tag{17}
$$

$$
g(M-1)\mu_1 + g(1)\mu_0 \geq g(M)\mu_1 \tag{18}
$$

$$
g(M)\mu_1 \geq g(M-2)\mu_1 + g(2)(\mu_0 + \Delta) \tag{19}
$$

Equation (17) ensures that putting strictly less than $M - 1$ players on the first arm is sub-optimal. Equation (18) 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 (19) ensures that putting stricly less than $M - 1$ players on the first arm is worse than putting all players on the first arm.

Equation (17) yields

$$
\begin{aligned}
&g(M-1)\mu_1 + g(1)(\mu_0 + \Delta) \geq g(M-2)\mu_1 + 2g(1)(\mu_0 + \Delta) \\
&\iff (\mu_0 + \Delta) \leq \underbrace{\frac{g(M-1) - g(M-2)}{g(1)}}_{h_1}\mu_1
\end{aligned}
$$

Equation (18) yields

$$
\begin{aligned}
&g(M-1)\mu_1 + g(1)\mu_0 \geq g(M)\mu_1 \\
&\iff \mu_0 \geq \underbrace{\frac{g(M) - g(M-1)}{g(1)}}_{h_2}\mu_1
\end{aligned}
$$

Equation (19) yields

$$
\begin{aligned}
&g(M)\mu_1 \geq g(M-2)\mu_1 + g(2)(\mu 0 + \Delta) \\
&\iff \underbrace{\frac{g(M) - g(M-2)}{g(2)}}_{h_3}\mu_1 \geq (\mu_0 + \Delta)
\end{aligned}
$$

We have $h_1 > h_2$ and

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

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

$$\begin{aligned} g(M) - g(M-2) &= Mp(1-p)^{M-1} - (M-2)p(1-p)^{M-3} \\ &= p(1-p)^{M-3}(M(1-p)^2 - (M-2)) \\ &= p(1-p)^{M-3}(M(1-2p+p^2) - M + 2) \\ &= p(1-p)^{M-3}(2 - 2Mp + Mp^2) \end{aligned}$$

we get

$$\begin{aligned} h_1 - h_2 &= (1-p)^{M-3}(1 - (M-1)p) - (1-p)^{M-2}(1 - Mp) \\ &= (1-p)^{M-3}(1 - (M-1)p - (1-p)(1-Mp)) \\ &= (1-p)^{M-3}(1 - (M-1)p - (1-p)(1-Mp)) \\ &= (1-p)^{M-3}(1 - Mp + p - (1 - Mp - p + Mp^2)) \\ &= (1-p)^{M-3}(2p - Mp^2) \\ &\geq (1-p)^{M-3}p && \text{(Using } p \leq \tfrac{1}{M}\text{)} \\ &\geq (1-p)^{M-3}p && \text{(Using } p \leq \tfrac{1}{M}\text{)} \\ &\geq \frac{p}{M-3} && \text{(Using } \min_{x\in[M]} g(x) = p\text{)} \end{aligned}$$

and

$$\begin{aligned} h_3 - h_2 &= \frac{1}{2}(1-p)^{M-4}(2 - 2Mp + Mp^2) - (1-p)^{M-2}(1 - Mp) \\ &= \frac{1}{2}(1-p)^{M-4}(2 - 2Mp + Mp^2 - 2(1-Mp)(1-p)^2) \\ &= \frac{1}{2}(1-p)^{M-4}(2 - 2Mp + Mp^2 - (1-Mp)(2 - 4p + 2p^2)) \\ &= \frac{1}{2}(1-p)^{M-4}(2 - 2Mp + Mp^2 - (2 - 4p + 2p^2 - 2Mp + 4Mp^2 - 2Mp^3)) \\ &= \frac{1}{2}(1-p)^{M-4}(4p - 2p^2 + 2Mp^3 - 3Mp^2) \\ &\geq \frac{1}{2}(1-p)^{M-4}(4p - (3M+2)p^2) \\ &\geq \frac{1}{2}(1-p)^{M-4}(p) && \text{(Using } p \leq \tfrac{1}{M+1}\text{)} \\ &\geq \frac{p}{2(M-4)} && \text{(Using } \min_{x\in[M]} g(x) = p\text{)} \end{aligned}$$

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 Lai et al. (1985) [2] we can prove that

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

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.,

$$\begin{aligned}\Delta_k(m) &:= \Big(g(M-1)\mu_1 + g(1)(\mu_0+\Delta)\Big) - \Big(g(M-m)\mu_1 + g(m)(\mu_0+\Delta-\Delta_{(k)}\Big) \\ &\geq \Big(g(M-m)\mu_1 + g(m)(\mu_0+\Delta)\Big) - \Big(g(M-m)\mu_1 + g(m)(\mu_0+\Delta-\Delta_{(k)}\Big) \\ &= g(m)\Delta_{(k)}\end{aligned}$$

and $\Delta_k(0) = 0$.

Then consider $\mathfrak{C}_k$ 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.

$$\mathfrak{C}_k = \min_{m_1,\dots,m_T : \sum_t g(m_t) \geq \frac{\log(T)}{2\Delta_{(k)}^2}} \sum_{t=1}^T \Delta_k(m_t) \tag{20}$$

It is clear that

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

The solution of Equation (20) has a specific form: for $t \in [\tau]$, m_t is constant, equal to m_τ, and defined by

$$\tau g(m_\tau) \geq \frac{\log(T)}{2\Delta_{(k)}^2}$$

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

As a consequence, one gets that, for a specific value of τ^*,

$$\mathfrak{C}_k = \tau^* \Delta_k(m_{\tau^*}) \geq \frac{\log(T)}{2\Delta_{(k)}^2} \frac{\Delta_k(m_{\tau^*})}{g(m_{\tau^*})} \geq \frac{\log(T)}{2\Delta_{(k)}}$$

as $\Delta_k(m) \geq g(m)\Delta_{(k)}$.

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* $k^* = \operatorname{argmax}_{k\in[K]} \mu_k$ *and* $\alpha = \frac{Mp}{K}$. *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 $\mathbf{M}^*$. Then take an eliminated arm i and consider $\mathbf{M}'$ constructed from $\mathbf{M}^*$ by taking one player from k^* and putting it on the eliminated arm i. Using $\mathbf{M}'$ instead of $\mathbf{M}^*$ increase the utility by: $G = \mu_i p - \mu_k(g(M_{k^*}^*) - g(M_{k^*}^* - 1))$.

[2] Consider for any sub-optimal arm k the two possibilities $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ 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

Note that $M^*_{k^*} \geq M_k$ for any $k \neq k^*$ since k^* is the best arm. In particular $M_{k^*} \geq \frac{M}{K}$ and by the hypothesis on the range of α, we have $\frac{M}{K} > 2$. Also note that by definition of Δ_{max}, $\mu_i \geq \rho\mu_{k^*}$.

We can then write :

$$\begin{aligned}
G &= \mu_i p - \mu_k(g(M^*_{k^*}) - g(M^*_{k^*} - 1)) \\
&\geq \mu_{k^*}\left[\rho p - (g(M^*_{k^*}) - g(M^*_{k^*} - 1))\right] && \text{(Since } \mu_i \geq \rho\mu_{k^*}\text{)} \\
&\geq \mu_{k^*}\left[\rho p - (g(\frac{\alpha}{p}) - g(\frac{\alpha}{p} - 1))\right] && \text{(By concavity of } g \text{ and } M^*_{k^*} \geq \tfrac{M}{K} = \tfrac{\alpha}{p}\text{)} \\
&= \mu_{k^*}\left[\rho p - p(1-p)^{\frac{\alpha}{p}-2}(1-\alpha)\right] \\
&= \mu_{k^*}\left[\rho p - \frac{p(1-p)^{\frac{\alpha}{p}}(1-\alpha)}{(1-p)^2}\right]
\end{aligned}$$

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 $1.3 \geq \frac{1}{0.9^2} \geq \frac{1}{(1-p)^2}$.

Therefore, $\mathbf{M}^*$ 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 μ_k using (4) and an upper bound using $\hat{\mu}^H_k(t) = \min(\hat{\mu}_k(t) + \zeta_k(t), 1)$ where $\boldsymbol{\zeta}$ is given by Equation (7) and take

$$\mathbf{M}(t+1) = \underset{\mathbf{M}\in\mathcal{M}}{\text{argmax}}\langle \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M})\rangle$$

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

The code is given in Algorithm 3.

Algorithm 3 UCB

1: **Input :** M (number of players), K (number of arms), p (probability that a player is active), T (horizon)
2: Initialize estimated rewards: $\hat{\boldsymbol{\mu}}^H = \mathbf{1}$
3: **for** t from 1 to T **do**
4: Play $\text{argmax}_{\mathbf{M}\in\mathcal{M}}\langle \hat{\boldsymbol{\mu}}^H, g(\mathbf{M})\rangle$
5: Compute $\hat{\boldsymbol{\mu}}$ according to (4)
6: Compute $\boldsymbol{\zeta}$ according to (7)
7: Set $\hat{\boldsymbol{\mu}}^H = \min(\hat{\boldsymbol{\mu}} + \boldsymbol{\zeta}, \mathbf{1})$
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*

$$\mathbb{E}[R_{UCB}] \leq 2\sqrt{2K\log(2T^2K^2)T\min(K, Mp + \frac{K}{T})} + 2 \tag{21}$$

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:

$$
\begin{aligned}
R_{CUCB} &= \sum_{t=1}^{T} \langle \boldsymbol{\mu}, g(\mathbf{M}^*) \rangle - \sum_{t=1}^{T} \langle \boldsymbol{\mu}, g(\mathbf{M}(t)) \rangle \\
&= \sum_{t=1}^{T} \langle \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M}^*) \rangle + \langle \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M}^*) - g(\mathbf{M}(t)) \rangle + \langle \hat{\boldsymbol{\mu}}^H(t) - \boldsymbol{\mu}, g(\mathbf{M}(t)) \rangle \\
&\leq \langle \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M}^*) - g(\mathbf{M}(t)) \rangle + \langle \hat{\boldsymbol{\mu}}^H(t) - \boldsymbol{\mu}, g(\mathbf{M}(t)) \rangle \\
&\qquad\qquad \text{(Since } \hat{\boldsymbol{\mu}}^H \geq \boldsymbol{\mu} \text{ by the GOOD event)} \\
&\leq \sum_{t=1}^{T} \langle \hat{\boldsymbol{\mu}}^H(t) - \boldsymbol{\mu}, g(\mathbf{M}(t)) \rangle \qquad \text{(Since } \mathbf{M}(t) = \operatorname{argmax}_{\mathbf{M} \in \mathcal{M}} \langle \hat{\boldsymbol{\mu}}^H, g(\mathbf{M}) \rangle) \\
&= \sum_{k=1}^{K} \sum_{t=1}^{T} \min(1, 2\zeta_k(t)) g(M_k(t)) \qquad \text{(Since } \boldsymbol{\mu} \geq \max(\hat{\boldsymbol{\mu}} - \boldsymbol{\zeta}, \mathbf{0}) \text{ by the GOOD event)} \\
&= \sum_{k=1}^{K} \sum_{t=1}^{T} \min(1, \sqrt{2 \frac{\log(2T^2K^2)}{T_k(t)}})(g(M_k(t)) - \eta_k(t) + \eta_k(t)) \qquad (*) \\
&\leq \underbrace{\sum_{k=1}^{K} \sum_{t=1}^{T} (g(M_k(t)) - \eta_k(t))}_{(i)} + \underbrace{\sum_{k=1}^{K} \sum_{t=1}^{T} \sqrt{2 \frac{\log(2T^2K^2)}{T_k(t)}} \eta_k(t)}_{(ii)}
\end{aligned}
$$

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

We have that $\mathbb{E}[(i)] = 0$ since

$$
\begin{aligned}
\mathbb{E}[g(M_k(t)) - \eta_k(t)] &= \mathbb{E}[g(M_k(t)] - \mathbb{E}[\mathbb{E}[\eta_k(t) | M_k(t)]] \\
&= \mathbb{E}[g(M_k(t)] - \mathbb{E}[g(M_k(t)] \\
&= 0
\end{aligned}
$$

and

$$
\begin{aligned}
(ii) &= \sum_{k=1}^{K} \sum_{t=1}^{T} \sqrt{2 \frac{\log(2T^2K^2)\eta_k(t)}{T_k(t)}} \qquad \text{(Since } \eta_k(t) = \sqrt{\eta_k(t)} \text{ as } \eta_k(t) \in \{0,1\}) \\
&= \sum_{k=1}^{K} \sqrt{2\log(2T^2K^2)} \sum_{t=1}^{T} \sqrt{\frac{\eta_k(t)}{\sum_{\rho=1}^{t} \eta_k(\rho)}} \\
&= \sum_{k=1}^{K} \sqrt{2\log(2T^2K^2)} \sum_{i=1}^{\max(T_k(T),1)} \frac{1}{\sqrt{i}} \qquad \text{(Since } \forall \rho \in [t], \eta_k(\rho) \in \{0,1\}) \\
&\leq \sum_{k=1}^{K} 2\sqrt{2\log(2T^2K^2)\max(T_k(T),1)}
\end{aligned}
$$

Then we have trivially:

$$
\mathbb{E}[(ii)] \leq 2K\sqrt{2\log(2T^2K^2)T} \tag{22}
$$

Otherwise, we write:

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

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

$$\leq 2\sqrt{2K\log(2T^2K^2)\sum_{k=1}^{K}(\mathbb{E}[T_k(T)]+\mathbb{P}(T_k(T)=0))} \quad \text{(By Jensen inequality)}$$

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

$$\leq 2\sqrt{2K\log(2T^2K^2)\sum_{k=1}^{K}(\sum_{\rho=1}^{T}M_kp+1)} \quad \text{(Since } 0\leq g(M_k)\leq 1 \text{ and } g(M_k)\leq M_kp\text{)}$$

$$\leq 2\sqrt{2K\log(2T^2K^2)(TMp+K))}$$

and therefore

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

so that

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

□

$\mathbb{E}[R_{UCB}] \leq 2K\sqrt{2\log(2T^2K^2)T}$ 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 (22).

D Solving $\text{argmax}_{\mathcal{M}_\mathcal{E}}\langle g(\mathbf{M}), \mathbf{v}\rangle$ via a sequential algorithm

We want to solve

$$\underset{\mathcal{M}_\mathcal{E}}{\text{argmax}}\langle g(\mathbf{M}), \mathbf{v}\rangle \tag{23}$$

where $\mathcal{E} \subset [K]$.

The sequential algorithm of (Dakdouk, 2022, Algorithm 5) is optimal if $\mathcal{E} = \varnothing$ 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 $a_1, \ldots, a_{k+1}$ and therefore the rest of the run is the same as the sequential algorithm.

Consider A^* is the algorithm that starts by assigning one player to every arm in $\mathcal{E}$ and then follow the sequential algorithm. Cal $\mathcal{E}'$ 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.