Constant or Logarithmic Regret in Asynchronous Multiplayer Bandits with Limited Communication

Hugo Richard
Criteo AI Lab
FAIRPLAY joint team

Etienne Boursier
INRIA, Université Paris Saclay LMO, Orsay

Vianney Perchet
ENSAE, Crest
Criteo AI Lab
FAIRPLAY joint team

Abstract

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

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 and Szepesvári, 2020, for a recent survey). A fundamental component of MAB is the exploration and exploitation trade-off. Exploration involves trying out different arms to gather information, while exploitation uses the acquired knowledge to favor arms more likely to be the best. It is well known that optimal policies incur a regret scaling as $\mathcal{O}(\sum_{k\neq k^*} \frac{\log(T)}{\mu_{k^*}-\mu_k})$ (Auer et al., 2002).

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

Motivated by Internet of Things networks, we focus

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

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

Notations. Vectors are denoted in bold. If $\mathbf{u} \in \mathbb{R}^n$, 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, the player choices are summarized by the assignment vector $\mathbf{M}(t) = (M_1(t), \dots, M_K(t))$ where $M_k(t)$ is the number of players choosing the arm k at round t; the environment then activates each player with probability[1] p and active players pull the arm they chose in the first phase, each receiving reward $\eta_k(M_k(t))X_k$ with k the pulled arm, $\eta_k = \mathbb{1}\{\text{exactly one player is active on arm } k\}$ and X_k is sampled from an unknown sub-Gaussian distribution with mean μ_k. The arm pulled by player m at time t is denoted $k_m(t) \in [K]$. A player playing arm k observes $X_k\eta_k$ and the collision event η_k. Additionally, 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 also assume:

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

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

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

The goal is to minimize the expected regret defined by:

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

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

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

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 non-adaptive explore-then-commit policy, it yields a regret scaling in $\mathcal{O}(T^{\frac{2}{3}})$. Additionally, Dakdouk (2022) shows there is no random assignment yielding a strictly larger expected reward than the deterministic optimal assignment $\mathbf{M}^*$.

Limited communication setting. We consider the same communication protocol as in Dakdouk (2022). At each round, a player a can decide to send a vector in $\mathbb{R}^{2K}$ to the (central) gateway. It is then successfully received by the gateway with probability p_g, *if only player a sent a message to the gateway* (collision otherwise). Reciprocally at each time step (if it is not receiving any message), the central gateway can send a vector to a *single player*, who successfully receives it with probability p_g. More precisely, agents communicate via a function SEND. $\text{SEND}_{a\to b}(\mathbf{x})$ attempts to send vector $\mathbf{x}$ from agent a (for instance a player) to b (for instance the gateway)[2]. The attempt is said to be successful when b has received the message from a and a is aware of it (*i.e.*, a has received the acknowledgment sent by b). An attempt is successful with a known probability p_g, and after a successful attempt, b can access $\mathbf{x}$ (and calling again $\text{SEND}_{a\to b}(\mathbf{x})$ afterwards does nothing). It follows that if $\text{SEND}_{a\to b}(\mathbf{x})$ is executed $\lceil \frac{\log(\delta)}{\log(1-p_g)} \rceil$ time, then b can access $\mathbf{x}$ with probability δ. If b tries to use $\mathbf{x}$ without a successful transmission, we assume it uses a value chosen uniformly at random in the range of $\mathbf{x}$ instead. We assume agents can call SEND even when they are inactive and at no cost. However, if two agents call SEND simultaneously, nothing gets transmitted.

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

[1] Players cannot know beforehand who will be active, making collisions unavoidable.

[2] The situation where a is the gateway and b a player will also be considered.

Contributions. In the centralized setting, we prove that an adapted version of UCB exhibits a regret in $\mathcal{O}(\sqrt{TK\log(T)\min(Mp,K)})$ where $\mathcal{O}(\cdot)$ hides universal constant factors. More surprisingly, our main contribution shows that, even in the limited communication setting, achieving a constant regret (in T) is sometimes possible with an algorithm called Cautious Greedy using only $\mathcal{O}(\log(T))$ communications. The analysis of UCB is thus postponed to Appendix C and the main text solely focuses on Cautious Greedy that somehow achieves the best of both worlds (very small regret with low communication). In essence, it is a standard greedy algorithm that estimates 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;
- r the infinity norm of the minimal perturbation of the 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, M and p_g.

In particular, Cautious Greedy achieves constant regret if $\nu^* = 0$, i.e., when each arm is assigned at least one player by the optimal policy. As shown by the lower bound in Lemma 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 $(\Delta_{(j)})_{j\leq\nu^*}$.

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 may not reach constant regret when $\nu^* = 0$. On the other hand, Cautious Greedy performs optimally in both cases.

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

2 RELATED WORK

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

Decentralized multiplayer bandits. In decentralized multiplayer bandits, players aim to speed up the collective learning of the arm rewards, while avoiding collisions. Motivated by cognitive radio networks, the decentralized problem of multiplayer bandits recently received a lot of attention (we refer to Boursier and Perchet, 2022, for a review), sometimes assuming a pre-agreement on the ranks of the players (Anandkumar et al., 2010; Liu and Zhao, 2010) or using few collisions to communicate information between players (Avner and Mannor, 2014; Rosenski et al., 2016; Besson and Kaufmann, 2018a). However, Bistritz and Leshem (2018); Boursier and Perchet (2019); Wang et al. (2020) enforce collisions to send a significant number of bits between the players, allowing to reach optimal centralized performance. This idea is also used in many extensions of MMAB (Mehrabian et al., 2020; Shi et al., 2020; Huang et al., 2021; Boursier and Perchet, 2020; Shi et al., 2021). This communication through collision tricks yet highly depends on the synchronicity of the players and becomes costly with a lot of players. In AMMAB, the players are asynchronous ($p < 1$) and nu-

merous ($M \geq K$), making both drawbacks significant. This work thus proposes an asynchronous algorithm with $\mathcal{O}(\log(T))$ communication, leaving open for future work a possible fully decentralized adaptation (see Section 6 for a discussion).

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

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

Resource allocation. Our problem can also be recast as a particular instance of sequential resource allocation with concave utilities (Lattimore et al., 2015; Fontaine et al., 2020; Zuo and Joe-Wong, 2021). Although general resource allocation algorithms could be used in our setting, much better solutions can be obtained by leveraging the very specific structure of the utilities. The utility functions are indeed exactly known here, up to the multiplicative factor 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. They propose an explore and commit algorithm that reaches $\mathcal{O}(T^{\frac{2}{3}})$ regret with constant communications. In our work, we show that under favorable conditions, a constant regret can be reached with $\mathcal{O}(\log(T))$ communications. Quite interestingly in AMMAB, the expected individual reward decreases as more players are assigned to the same arm. This relates the AMMAB model to more advanced collision models for MMAB, where a collision only decreases the reward instead of yielding a 0 reward (Tekin and Liu, 2012; Bande and Veeravalli, 2019; Magesh and Veeravalli, 2019; Boyarski et al., 2021). AMMAB is also related to the problem of online queuing systems (Gaitonde and Tardos, 2020; Sentenac et al., 2021), where packets arrive in a queue (player) with random rates. This setting yet differs from AMMAB, as players are active as long as they hold packets.

3 CAUTIOUS GREEDY, AN EFFICIENT ALGORITHM FOR AMMAB

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

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

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. A player m can compute its own estimate of the mean reward $\hat{\boldsymbol{\mu}}^{(m)}(t)$ with

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

$$\text{where } T_k^{(m)}(t) = \sum_{\rho=1}^{t} \eta_k^\rho(\mathbf{M}(\rho)) \tag{5}$$

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

$$\hat{\mu}_k(t) = \frac{\sum_{m=1}^{M} T_k^{(m)}(t)\hat{\mu}_k^{(m)}(t)}{T_k(t)}. \tag{6}$$

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

For communication purposes, Cautious Greedy is divided into epochs of doubling size and statistics such as $\hat{\boldsymbol{\mu}}$ are only updated at the end of each epoch. Communication happens at each round during the second half of an epoch, increasing the probability of having a successful transmission with the epoch size. This second half is again split in two parts: during the first one, players send their statistics to the gateway; during the second part, the gateway communicate to each individual player the averaged players' statistics.

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

$$\mathbf{M}_{\mathcal{M}}^{\hat{\mu}} = \underset{\mathbf{M}\in\mathcal{M}}{\operatorname{argmax}} \langle \hat{\boldsymbol{\mu}}, g(\mathbf{M})\rangle \tag{7}$$

and $\mathbf{M}_{\mathcal{M}}^{\hat{\mu}}(m)$ is the arm chosen by player m.

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

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

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

$$\begin{aligned}\hat{\boldsymbol{\mu}}^H(t) &= \min(\hat{\boldsymbol{\mu}}(t) + \boldsymbol{\zeta}(t), 1)\\ \hat{\boldsymbol{\mu}}^L(t) &= \max(\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\zeta}(t), 0)\end{aligned} \tag{8}$$

where for all $k\in[K]$, $\zeta_k(t) = \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}$. These bounds are used to eliminate sub-optimal arms. This could suggest a strategy that plays all active arms at each round until enough information is gathered to remove an arm. However, such a strategy yields high regret in the case where two arms that should be eliminated are very close to each other. Therefore, the elimination of several arms at once is allowed. This is done in Cautious Greedy by computing an estimate 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:

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

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 $\mathcal{U}$ among the arms that are not accepted yet. The set of arms put under pressure rotates in a *round-robin* fashion. This mechanism ensures that all active arms are regularly played. After the round-robin rotation is completed, Cautious Greedy reevaluates 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.

Algorithm 1 Cautious Greedy with $\mathcal{O}(\log(T))$ communications

Input : M (number of players), p (activation probability), T (horizon), m (player id), p_g (successful communication probability)
$\nu = 0$, $\mathcal{K} = [K]$, $\mathcal{A} = \varnothing$
$\mathcal{U} = \varnothing$, $n = 0$, $\hat{\boldsymbol{\mu}} = 1$, $\hat{\boldsymbol{\mu}}^L = 0$, $\hat{\boldsymbol{\mu}}^H = 1$
for $s = 0, \dots, \lfloor\log_2(T)\rfloor$ **do**
  for $t = 2^s \dots \min(2^{s+1}-1, T)$ **do**
    if $2^s < \lceil 16M\frac{\log(2MT^2)}{-\log(1-p_g)}\rceil$ **then**
      Play arm $\mathbf{M}_{\mathcal{M}}^{\hat{\boldsymbol{\mu}}}(m)$
    else
      Play arm $\mathbf{M}_{\mathcal{M}_\mathcal{E}}^{\hat{\boldsymbol{\mu}}}(m)$ (7) 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)
      $n = n+1$
      if $n = |\mathcal{K}\setminus\mathcal{A}|$ **then** // end of round robin
        $n = 0$ and compute $\mathbf{M}_{\mathcal{M}}^{\hat{\boldsymbol{\mu}}^L}$ and $\mathbf{M}_{\mathcal{M}_\nu}^{\hat{\boldsymbol{\mu}}^L}$ (7)
        while $\langle\hat{\boldsymbol{\mu}}^L, g(\mathbf{M}_{\mathcal{M}}^{\hat{\boldsymbol{\mu}}^L})\rangle > \langle\hat{\boldsymbol{\mu}}^H, g(\mathbf{M}_{\mathcal{M}_\nu}^{\hat{\boldsymbol{\mu}}^H})\rangle$ **do**
          $\nu = \nu+1$
        end while
        Update $\mathcal{A} = \{k\in[K], \hat{\mu}_{(\nu)}^H < \hat{\mu}_k^L\}$ and $\mathcal{K} = [K]\setminus\{k\in[K], \hat{\mu}_k^H < \hat{\mu}_{(\nu+1)}^L\}$
        Let $\mathcal{U}$ be $\nu - |[K]\setminus\mathcal{K}|$ elements from $\mathcal{K}\setminus\mathcal{A}$
      end if
    end if
    if $2^s \geq 8M$ **then**
      If $t = 2^s + 2^{s-1}$, compute $\boldsymbol{\mu}^{(m)}, \mathbf{T}^{(m)}$ (4) (5)
      If $t \in [2^s + 2^{s-1}, 2^s + 2^{s-1} + 2^{s-2}]$ and $t \bmod M = m$, $\text{SEND}_{m\to\text{gateway}}(\boldsymbol{\mu}^{(m)}, \mathbf{T}^{(m)})$
      If $t \in [2^s + 2^{s-1} + 2^{s-2}, 2^{s+1}]$ and $t \bmod M = m$, $\text{SEND}_{\text{gateway}\to m}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^L, \tilde{\boldsymbol{\mu}}^H)$ with $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^L, \tilde{\boldsymbol{\mu}}^H$ with (6), (8)
    end if
  end for
  $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\mu}}^L = \tilde{\boldsymbol{\mu}}^L$, $\hat{\boldsymbol{\mu}}^H = \tilde{\boldsymbol{\mu}}^H$
end for

3.2 Regret bound

The main result of this section is an upper bound on the expected regret of Cautious Greedy. This bound depends on several data-dependent quantities that we introduce now: $\Delta^{(\nu^*)}$ is the minimum simple regret achieved by an allocation removing exactly ν^*-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 $\mathcal{O}(\frac{1}{r} + \frac{\log(T)}{\Delta^{(\nu^*)}} + \sum_{j \leq \nu^*} \frac{\log(T)}{\Delta_{(j)}})$ where $\mathcal{O}$ hides quantities independent of the data and T.

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

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

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

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 due to the sample complexity of detecting that the optimal policy eliminates ν^* arms. The fourth one is finally due to communication. Interestingly, every term depending on T are null when $\nu^* = 0$, which corresponds to situations where the optimal policy assigns at least one player on every arm. This makes the regret of Cautious Greedy constant in such situations, which happens as soon as arm rewards have a similar order of magnitude (see Lemma B.1).

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

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

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, p_g. The remainder of this section sketches the proof of Proposition 3.1. The precise statement of lemmas and their proofs are deferred to Appendix A.

Proof sketch of Proposition 3.1. We start by an upper bounds R^{UB}_{cCG} on the regret of a quasi-centralized version of Cautious Greedy (see Algorithm 3 in Appendix A) where communication aspects are removed ($p_g = 1$, Lines 20 to 24 are removed, updates Line 26 uses all seen samples). Using classical concentration bounds (Lemma A.1), we can assume that $\boldsymbol{\mu}^H$ and $\boldsymbol{\mu}^L$ (defined in Equation (8)) verify $\boldsymbol{\mu}^H \geq \boldsymbol{\mu} \geq \boldsymbol{\mu}^L$ without affecting the regret bound. Consequently, Algorithm 3 yields 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 all active arms are assigned at least one player regularly, as proven by Lemma A.5. However, because of collisions, assigning at least one player to an arm does not guarantee an observation. Lemma A.6 makes this relation explicit.

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

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

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

Focusing on (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 pro-

vides 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).

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

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

4 LOWER BOUND

The upper bound of Cautious Greedy when $\nu^* = 0$ scales in $\frac{1}{r}$. Lemma 4.1 shows that under mild conditions, this dependency in r cannot be improved:

Lemma 4.1 (Lower bound for $\nu^* = 0$)**.** *Consider $K = 2$ arms and $M = 2N+1$ players for some $N \in \mathbb{N}^*$ and assume $p \leq \frac{1}{M+1}$, $r_0 < \frac{p}{12}$, $T \geq \frac{1}{16g(M)r_0^2}$. For any algorithm A, there exists a choice of rewards $\boldsymbol{\mu}$ such that $r(\boldsymbol{\mu}) = r_0$ and*

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

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 and Chen (2017), we recast this setting as a 2-armed bandit problem where arm k has reward 1 with probability $g(M)\mu_k$, 0 with probability $g(M)(1-\mu_k)$ and $X_k = \perp$ with probability $1 - g(M)$.

The rest of the proof follows closely the proof of Proposition 4 in Mourtada and 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. □

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, for some universal constant $c > 0$,*

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

Proof sketch (see proof in Appendix A.5). Assume for the sketch of proof that $\nu^* = 1$ and that there are 3 arms. We are considering two alternative mean 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. Together, Lemma 4.1 and Lemma 4.2 show that Cautious Greedy is optimal with respect to the data-dependent quantities r and $(\Delta_{(j)})_{j\leq\nu^*}$.

5 EXPERIMENTS

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

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

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

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

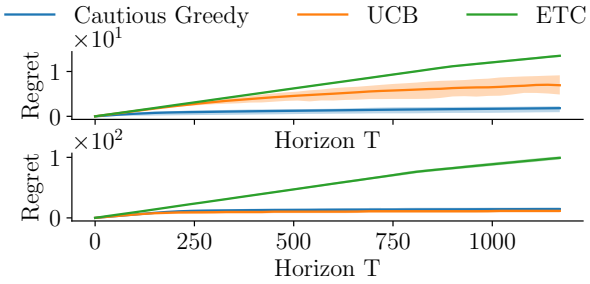

Figure 1: **Benchmark of ETC, UCB and Cautious Cautious Greedy** *(top)* $\nu^*=0$ *(bottom)* $\nu^*=1$.

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

In both experiments, ETC incurs a much larger regret, which is consistent with its $\mathcal{O}(T^{\frac{2}{3}})$ regret. Note however that ETC uses only constant communication costs.

6 CONCLUSION, OPEN PROBLEMS AND FUTURE WORK

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

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

Our algorithm requires several assumptions to perform properly. Most of them are actually very mild, while others would require an involved analysis to get discarded. Without prior knowledge of T, a doubling trick (Besson and Kaufmann, 2018b, Theorem 7) can be used when the horizon T is unknown but would introduce a logarithmic dependency in T. Whether Cautious Greedy can be made anytime is therefore an interesting extension. The activation probabilities of players might be heterogeneous in practice. However, the optimization algorithm of Dakdouk (2022) is only optimal in the homogeneous case. No efficient maximization scheme of the problem in Equation (2) in the heterogeneous case is currently known. If however we were given access to an oracle maximizing this problem, we believe that our algorithms and their bounds

can be adapted although this requires meticulous work. Also, if p is unknown, it can be estimated on the fly. Indeed at every step, players observe a realisation of a Bernoulli(p). Assuming successful communications, agents use the empirical estimate $\hat{p}$ provided by the gateway based on agents' observations. Then, replacing p by $\hat{p}$ just adds a constant term to the regret upper bound. The case of heterogeneous $(p_i)_{i\in[M]}$ is similar. Likewise, if p_g is unknown, notice that agents observe a realization of Bernouilli(p_g) whenever they attempt to communicate. Then, the gateway shares an upper bound of p_g which replaces p_g in the condition Line 5. The time it takes to satisfy the modified condition increases (but remains of the same order) and (despite communication failures) is identical across agents with high probability. The rest of the analysis works without change. The careful analysis of these extensions is left to future work.

Another significant direction is to go beyond the limited communication setting. Being able to handle the decentralized setting where agents are no longer allowed to communicate without cost remains a great challenge and the original motivation of asynchronous multiplayer bandits. A solution to handle the decentralized setting is to use collisions to communicate as done for example in (Bistritz and Leshem, 2018; Boursier and Perchet, 2019; WANG et al., 2020). These previously cited works however only tackle the synchronized case. In the case we study, communicating through collisions remains possible but the length of communication phases would be significantly increased. In the collision sensing setting, if players i and j need to propagate a bit through collision, they roughly need $\frac{\log(T)}{p^2}$ 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. Concerning communication without collisions, in the synchronous case, Dyn-MMAB Boursier and Perchet (2019) and Lugosi and Mehrabian (2022) achieve $\mathcal{O}(T^{2/3})$ regret. The work of Bubeck et al. (2021) even achieves $\mathcal{O}(\sqrt{T})$. Whether these approaches can be extended to the asynchronous setting we consider is an interesting open question.

Acknowledgements

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

References

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

Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. (2011). Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745.

Anantharam, V., Varaiya, P., and Walrand, J. (1987). 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.

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

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

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

Besson, L. and Kaufmann, E. (2018a). Multi-player bandits revisited. In *Algorithmic Learning Theory*, pages 56–92. PMLR.

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

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

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

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

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

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

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

Bubeck, S., Budzinski, T., and Sellke, M. (2021). Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. In *Conference on Learning Theory*, pages 821–822. PMLR.

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

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

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

Chen, Y.-Z. J., Yang, L., Wang, X., Liu, X., Hajiesmaili, M., Lui, J. C., and Towsley, D. (2023). On-demand communication for asynchronous multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3903–3930. PMLR.

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

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

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

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

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

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

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

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

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. (2017). 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.

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

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95.

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

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

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

Lai, L., Jiang, H., and Poor, H. V. (2008). Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 98–102.

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

Landgren, P., Srivastava, V., and Leonard, N. E. (2016). On distributed cooperative decision-making in multi-

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

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

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

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

Lugosi, G. and Mehrabian, A. (2022). Multiplayer bandits without observing collision information. *Mathematics of Operations Research*, 47(2):1247–1265.

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

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

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

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

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

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

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

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

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

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

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

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

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

WANG, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. (2020). Optimal algorithms for multiplayer multi-armed bandits. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4120–4129. PMLR.

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

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

Yang, L., Chen, Y.-Z. J., Pasteris, S., Hajiesmaili, M., Lui, J., and Towsley, D. (2021). Cooperative stochastic bandits with asynchronous agents and constrained feedback. *Advances in Neural Information Processing Systems*, 34:8885–8897.

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

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not

modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

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

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No, will be released over publication]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No, will be released over publication]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [No, will be released over publication]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes (last section in appendix)]
   (b) The license information of the assets, if applicable. [Yes (last section in appendix)]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Analysis of Cautious Greedy

A.1 A useful upper bound

At many places we will have to bound quantity of the form $\langle \boldsymbol{\mu} - \boldsymbol{\mu}', g(\mathbf{M}) - g(\mathbf{M}') \rangle$ 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

$$\langle \boldsymbol{\mu} - \boldsymbol{\mu}', g(\mathbf{M}) - g(\mathbf{M}') \rangle \leq 2Mp \tag{9}$$

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

A.2 A precise description of the Round Robin procedure

Rotating $\mathcal{U}$ 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 $(y_1, \dots, y_{|\mathcal{Y}|})$, $\mathcal{U}$ as $(u_1, \dots, u_s)$. 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 $\mathcal{U} = (y_1, \dots, y_s)$ 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)

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

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

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

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

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

Input : M (number of players), p (probability that a player is active), T (horizon)
$\nu = 0$, $\mathcal{K} = [K]$, $\mathcal{A} = \varnothing$
$\mathcal{U} = \varnothing$, $n = 0$, $\hat{\boldsymbol{\mu}} = 0$, $\hat{\boldsymbol{\mu}}^L = 0$, $\hat{\boldsymbol{\mu}}^H = 1$
for $s = 0, \dots, \lfloor \log_2(T) \rfloor$ **do**
  for $t = 2^s \dots \min(2^{s+1}-1, T)$ **do**
    Play $\mathbf{M}^{\hat{\boldsymbol{\mu}}}_{\mathcal{M}_\mathcal{E}}$ as defined in (7) 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)
    $n = n + 1$
    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 $\mathbf{M}^{\hat{\boldsymbol{\mu}}^L}_{\mathcal{M}_\nu}$ (7)
      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
  Update $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\mu}}^L$ and $\hat{\boldsymbol{\mu}}^H$ using samples from all players
end for

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

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

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

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

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

$$\begin{aligned}
P(|\hat{\mu}_k(t) - \mu_k| \geq \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}) &= \sum_{\tau=1}^{T} P(|\hat{\mu}_k(t) - \mu_k| \geq \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}, T_k(t) = \tau) \\
&\leq \sum_{\tau=1}^{T} P(|\hat{\mu}_{k,\tau} - \mu_k| \geq \sqrt{\frac{\log(2T^3K^2)}{2\tau}}) \\
&\leq \sum_{\tau=1}^{T} 2\exp(-2\tau\sqrt{\frac{\log(2T^3K^2)}{2\tau}}^2) \\
&= \frac{1}{K^2T^2}
\end{aligned}$$

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

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

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

$$\forall t \in [T], \forall k \in [K], |\hat{\mu}_k(t) - \mu_k| \leq \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}} \tag{10}$$

which implies 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,*

$$\mathbb{E}[R] \leq \mathbb{E}[R_G] + 2 \tag{11}$$

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

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 $\hat{\mu}_k^H < \mu_{(\nu+1)}^L$. But since $\nu \leq \nu^*$, this implies $\hat{\mu}_k^H < \mu_{(\nu^*+1)}^L$. Under the good event $\hat{\mu}_k^H > \mu_k$ and $\mu_{(\nu^*+1)}^L < \mu_{(\nu^*+1)}$ so that $\hat{\mu}_k^H < \mu_{(\nu^*+1)}^L$ 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* $\forall t, \tau(t) \geq \tau_{lb}(t) = \lceil \max(\frac{t}{\nu^*+1} - \nu^*, 0)\rceil$. *Furthermore,* $\forall t \geq (\nu^*+1)^2, \tau_{lb}(t) \geq \tau_{lin}(t) = \frac{t}{2(\nu^*+1)}$.

Proof of Lemma A.5. Call t_n the value of t the n-th time where $t = 0 \mod |\mathcal{U}|$. Between t_n and $t_{n+1} - 1$ (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

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

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 $\tau(t) \geq \max(\frac{t}{\nu^*+1} - \nu^*, 0)$.

Furthermore, for any $t \geq 2(\nu^* + 1)^2$, we have that

$$\frac{t}{\nu^* + 1} - \nu^* \geq \frac{t}{\nu^* + 1} - \frac{1}{2}\frac{t}{\nu^* + 1} \geq \frac{1}{2}\frac{t}{\nu^* + 1}$$

□

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

First, we show that $T_k(t)$ stochastically dominates a sum of $\tau_{lb}(t)$ independent Bernouilli random variable with parameter p.

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

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

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

In particular, we have from a multiplicative Chernoff bound and union bound over K that

$$\mathbb{P}(q(t) \leq \frac{1}{3}p\lceil \tau_{lb}(t) \rceil) = \mathbb{P}(\exists k \in [K], T_k(t) \leq \frac{1}{3}p\lceil \tau_{lb}(t) \rceil) \leq K \exp(-\frac{2}{9}p\lceil \tau_{lb}(t) \rceil)$$

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

$$\mathbb{E}[R_G] = \mathbb{E}[R_\nu \mathbb{1}\{GOOD\}] + \mathbb{E}[R_\mathcal{E} \mathbb{1}\{GOOD\}] + \mathbb{E}[R_\mathbf{M} \mathbb{1}\{GOOD\}]$$

where

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

$\mathbf{M}^*_\nu = \mathbf{M}^{\boldsymbol{\mu}}_{\mathcal{M}_\nu}$ 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.

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 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)**.** *Consider assignment* $\mathbf{M}^*_{(k)} = \text{argmax}_{\mathbf{M},\text{support}(\mathbf{M})=K-k}\langle\boldsymbol{\mu}, g(\mathbf{M})\rangle$ *which is the best assignment where k arms assigned zero players.*

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

Proof of Lemma A.7. Call $\mathbf{M}^{*,H}_{(k)}(t) = \text{argmax}_{\mathbf{M}\in\mathcal{M}_k}\langle\boldsymbol{\mu}^H(t), g(\mathbf{M})\rangle$ and for simplicity, call $t' = 2^{\lfloor\log_2(t)\rfloor}$.

$$
\begin{aligned}
& q(t') > \frac{8M^2p^2\log(2K^2T^3)}{(\langle\boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{(k)})\rangle)^2} \\
\implies & q(t') > \frac{8M^2p^2\log(2K^2T^3)}{(\langle\boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{(k)})\rangle)^2} \\
\implies & \min_{a\in\mathcal{K}} T_a(t') > \frac{18M^2p^2\log(2K^2T^3)}{(\langle\boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{(k)})\rangle)^2} \\
\implies & \langle\boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{(k)})\rangle > 8Mp\max_{a\in\mathcal{K}}\sqrt{\frac{\log(2T^3K^2)}{2T_a(t')}} \\
\implies & \langle\boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{(k)})\rangle > 4Mp\max_{a\in\mathcal{K}}\zeta_a(t') \\
\implies & \langle\boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{(k)})\rangle > 2\langle\boldsymbol{\zeta}(t'), g(\mathbf{M}^*) + g(\mathbf{M}^{*,H}_{(k)}(t'))\rangle && \text{(By Equation (9))} \\
\iff & \langle\boldsymbol{\mu} - 2\boldsymbol{\zeta}(t'), g(\mathbf{M}^*)\rangle > \langle\boldsymbol{\mu}, g(\mathbf{M}^*_{(k)})\rangle + 2\langle\boldsymbol{\zeta}(t'), g(\mathbf{M}^{*,H}_{(k)}(t'))\rangle \\
\implies & \langle\boldsymbol{\mu}^L(t'), g(\mathbf{M}^*)\rangle > \langle\boldsymbol{\mu} + 2\boldsymbol{\zeta}(t'), g(\mathbf{M}^{*,H}_{(k)}(t'))\rangle\rangle && \text{(By the GOOD event and optimality of } \mathbf{M}^*_{(k)}) \\
\implies & \max_{\mathbf{M}\in\mathcal{M}}\langle\boldsymbol{\mu}^L(t'), g(\mathbf{M})\rangle > \langle\boldsymbol{\mu}^H(t'), g(\mathbf{M}^{*,H}_{(k)}(t'))\rangle && \text{(By the GOOD event)}
\end{aligned}
$$

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

Note that as long as $\nu \leq \nu^*$, R_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)**.**

$$\mathbb{E}[R_\nu \mathbb{1}\{GOOD\}] \leq \frac{384M^2p\log(2K^2T^3)(\nu^*+1)}{\Delta^{(\nu^*)}} + 42M(\nu^*+1)K$$

Proof of Lemma A.8. Call t_v the last time that $\nu(t) = \nu$ and set $t_{\nu^*} = T + 1$ and $t_{-1} = 0$.

$$
\begin{aligned}
R_\nu &= \sum_{t=1}^{T} \langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu(t)}) \rangle \\
&= \sum_{\nu=0}^{\nu^*} \sum_{t=t_{\nu-1}+1}^{t_\nu} \underbrace{\langle \boldsymbol{\mu}, g(\mathbf{M}^*) - g(\mathbf{M}^*_{\nu}) \rangle}_{A_\nu} \\
&= \sum_{\nu=0}^{\nu^*} (t_\nu - t_{\nu-1}) A_\nu \\
&= \sum_{\nu=0}^{\nu^*} t_\nu A_\nu - \sum_{\nu=0}^{\nu^*} t_{\nu-1} A_\nu \\
&= \sum_{\nu=1}^{\nu^*} t_{\nu-1}(A_{\nu-1} - A_\nu) - t_{-1}A_0 + t_{\nu^*}A_{\nu^*} \\
&= \sum_{\nu=1}^{\nu^*} t_{\nu-1}(A_{\nu-1} - A_\nu) \\
&= \sum_{\nu=1}^{\nu^*} 2 \underbrace{\frac{(t_{\nu-1} - (\nu^*+1))}{2}}_{t'_{\nu-1}} (A_{\nu-1} - A_\nu) + \sum_{\nu=1}^{\nu^*} (\nu^*+1)(A_{\nu-1} - A_\nu) \\
&\leq 2 \sum_{\nu=1}^{\nu^*} t'_{\nu-1}(A_{\nu-1} - A_\nu) + 2Mp(\nu^*+1) \\
&= 2 \sum_{\nu=\{1,\dots,\nu^*\}, t'_{\nu-1} < 2(\nu^*+1)^2} t'_{\nu-1}(A_{\nu-1} - A_\nu) + 2 \sum_{\nu=\{1,\dots,\nu^*\}, t'_{\nu-1} \geq 2(\nu^*+1)^2} t'_{\nu-1}(A_{\nu-1} - A_\nu) + 2Mp(\nu^*+1) \\
&\leq 4 \sum_{\nu=\{1,\dots,\nu^*\}, t'_{\nu-1} < 2(\nu^*+1)^2} (\nu^*+1)^2 (A_{\nu-1} - A_\nu) \\
&\quad + 2 \sum_{\nu=\{1,\dots,\nu^*\}, t'_{\nu-1} \geq 2(\nu^*+1)^2} (\nu^*+1) 2 \lceil \tau_{lin}(t'_{\nu-1}) \rceil (A_{\nu-1} - A_\nu) + 2Mp(\nu^*+1) && \text{(By Lemma A.5)} \\
&\leq 6Mp(\nu^*+1)^2 + 4 \underbrace{\sum_{\nu=\{1,\dots,\nu^*\}, t'_{\nu-1} \geq 2(\nu^*+1)^2} (\nu^*+1) \lceil \tau_{lin}(t'_{\nu-1}) \rceil (A_{\nu-1} - A_\nu)}_{(i)} && \text{(By Equation (9))}
\end{aligned}
$$

Then we have

$$
\begin{aligned}
(i) &= \sum_{\nu=\{1,\dots,\nu^*\},t'_{\nu-1}\geq 2(\nu^*+1)^2} (\nu^*+1)\lceil\tau_{lin}(t'_{\nu-1})\rceil(A_{\nu-1}-A_\nu) \\
&= \sum_{\nu=\{1,\dots,\nu^*\},t'_{\nu-1}\geq 2(\nu^*+1)^2} (\nu^*+1)\lceil\tau_{lin}(t'_{\nu-1})\rceil(A_{\nu-1}-A_\nu)(\mathbb{1}\left\{\lceil\tau_{lin}(t'_{\nu-1})\rceil \leq \frac{6q(t'_{\nu-1})}{p}\right\} \\
&\quad + \mathbb{1}\left\{\lceil\tau_{lin}(t'_{\nu-1})\rceil > \frac{6q(t'_{\nu-1})}{p}\right\}) \\
&\leq \underbrace{\sum_{\nu=\{1,\dots,\nu^*\},t'_{\nu-1}\geq 2(\nu^*+1)^2} (\nu^*+1)\frac{6q(t'_{\nu-1})}{p}(A_{\nu-1}-A_\nu)}_{(ii)} \\
&+ \underbrace{\sum_{\nu=\{1,\dots,\nu^*\},t'_{\nu-1}\geq 2(\nu^*+1)^2} (\nu^*+1)\lceil\tau_{lin}(t'_{\nu-1})\rceil(A_{\nu-1}-A_\nu)\mathbb{1}\left\{\lceil\tau_{lin}(t'_{\nu-1})\rceil > \frac{6q(t'_{\nu-1})}{p}\right\}}_{(iii)}
\end{aligned}
$$

We have that $\nu(2t'_{\nu-1}+\nu^*)=\nu-1$ and therefore by Lemma A.7 we get

$$q(t'_{(\nu-1)}) \leq q(2^{\lfloor \log_2(2t'_{\nu-1})\rfloor}) \leq q_\nu$$

This gives

$$
\begin{aligned}
(ii) &\leq \frac{6(\nu^*+1)}{p}\sum_{\nu=1}^{\nu^*}\frac{4(Mp)^2\log(2K^2T^3)\langle\boldsymbol{\mu}, g(\mathbf{M}^*_\nu)-g(\mathbf{M}^*_{\nu-1})\rangle}{\left(\langle\boldsymbol{\mu}, g(\mathbf{M}^*)-g(\mathbf{M}^*_{\nu-1})\rangle\right)^2} \qquad \text{(By Lemma A.7)} \\
&= 48M^2p\log(2K^2T^3)(\nu^*+1)\sum_{\nu=1}^{\nu^*}\underbrace{\left(\frac{1}{\langle\boldsymbol{\mu}, g(\mathbf{M}^*)-g(\mathbf{M}^*_{\nu-1})\rangle}-\frac{\langle\boldsymbol{\mu}, g(\mathbf{M}^*)-g(\mathbf{M}^*_{\nu})\rangle}{\left(\langle\boldsymbol{\mu}, g(\mathbf{M}^*)-g(\mathbf{M}^*_{\nu-1})\rangle\right)^2}\right)}_{\triangleq l_\nu}
\end{aligned}
$$

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

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

Taking expectations we get

$$\mathbb{E}[R_\nu\mathbb{1}\{GOOD\}] \leq 4\left(\frac{96M^2p\log(2K^2T^3)(\nu^*+1)}{\langle\boldsymbol{\mu}, g(\mathbf{M}^*)-g(\mathbf{M}^*_{\nu^*-1})\rangle}+\mathbb{E}[(iii)]\right)+6Mp(\nu^*+1)^2$$

Focusing on $\mathbb{E}[(iii)]$, we get:

$$
\begin{aligned}
&\mathbb{E}[(iii)] \\
&= \mathbb{E}[\sum_{\nu=\{1,\dots,\nu^*\}, t'_{\nu-1}\geq 2(\nu^*+1)^2} (\nu^*+1)\lceil\tau_{lin}(t'_{\nu-1})\rceil(A_{\nu-1}-A_\nu)\mathbb{1}\left\{\lceil\tau_{lin}(t'_{\nu-1})\rceil > \frac{6q(t'_{\nu-1})}{p}\right\}] \\
&= \mathbb{E}[\sum_{l=1}^{\lceil\log_2(T)\rceil} \mathbb{1}\left\{t'_{\nu-1}\in[2^l,2^{l+1}]\right\}(\nu^*+1)\lceil\tau_{lin}(t'_{\nu-1})\rceil(A_{\nu-1}-A_\nu)\mathbb{1}\left\{\lceil\tau_{lin}(t'_{\nu-1})\rceil > \frac{6q(t'_{\nu-1})}{p}\right\}] \qquad \text{(Peeling)} \\
&\leq 2Mp\sum_{l=1}^{\lceil\log_2(T)\rceil}(\nu^*+1)\lceil\tau_{lin}(2^{l+1})\rceil P(\lceil\tau_{lin}(2^{l+1})\rceil > \frac{6q(2^l)}{p}) \\
&\leq 4Mp\sum_{l=1}^{\lceil\log_2(T)\rceil}(\nu^*+1)\lceil\tau_{lin}(2^{l})\rceil P(\lceil\tau_{lin}(2^{l})\rceil > \frac{3q(2^l)}{p}) \qquad \text{(We use } 2\lceil a\rceil\geq\lceil 2a\rceil\text{)} \\
&\leq 2MpK\int_{t=0}^{\infty} 2^t\exp(-\frac{1}{9(\nu^*+1)}p2^t)dt \\
&\leq \frac{36MpK(\nu^*+1)}{p}\underbrace{\int_{u=1}^{\infty}\frac{\exp(-u)}{u}du}_{\leq 1} \qquad (u=\tfrac{1}{9(\nu^*+1)}p2^t) \\
&\leq 36MK(\nu^*+1)
\end{aligned}
$$

which gives

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

□

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 $\mathcal{E}^*_\nu = \text{support}(\mathbf{M}^*_\nu)$ and let $i\notin\mathcal{E}^*_\nu$ be a sub-optimal arm. For any $t\geq 0$, if $q(2^{\lfloor\log_2(t)\rfloor})\geq q_{E,i}$ with*

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

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

Proof of Lemma A.9. At time $t'=2^{\lfloor\log_2(t)\rfloor}$ we have that

$$
\begin{aligned}
&q(t') > \frac{8\log(2T^3K^2)}{(\mu_{(\nu+1)}-\mu_i)^2} \\
&\implies 4\sqrt{\frac{\log(2T^3K^2)}{2q}} < \frac{\mu_{(\nu+1)}-\mu_i}{2} \\
&\implies \zeta_i+\zeta_{(\nu+1)} < \frac{\mu_{(\nu+1)}-\mu_i}{2} \\
&\implies \mu_i+2\zeta_i < \mu_{(\nu+1)}-2\zeta_{(\nu+1)} \\
&\implies \mu_i^H < \mu^L_{(\nu+1)}
\end{aligned}
$$

Since $i\notin\mathcal{E}^*_\nu$, $\mu_i<\mu_{(\nu+1)}$, the last line means that i will be eliminated at the next update which will happen at the end of the Round Robin phase which can last up to ν^* rounds. □

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 $\mathcal{E}_\nu^* = \text{support}(\mathbf{M}_E^*)$ and let $j \in \mathcal{E}_\nu^*$ an optimal arm. If at time t, arm j and (ν) both been played without collision at least $q_{A,i}$ times with*

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

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

Proof of Lemma A.10. At time $t' = 2^{\lfloor \log_2(t) \rfloor}$ we have that

$$\begin{aligned}
\min(T_j(t'), T_{(\nu)}(t')) > \frac{8\log(2T^3K^2)}{(\mu_j - \mu_{(\nu)})^2} &\implies 2\sqrt{\frac{\log(2T^3K^2)}{2\min(T_j(t'), T_{(\nu)}(t'))}} < \frac{\mu_j - \mu_{(\nu)}}{2} \\
&\implies \zeta_{(\nu)} + \zeta_j < \frac{\mu_j - \mu_{(\nu)}}{2} \\
&\implies \mu_{(\nu)} + 2\zeta_{(\nu)} < \mu_j - 2\zeta_j \\
&\implies \mu_{(\nu)}^H < \mu_j^L
\end{aligned}$$

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

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

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

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

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:

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

□

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

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

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

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

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

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

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 3 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 eliminated. 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

$$
\begin{aligned}
s_i &= (\mu_{(\nu_i+1)} - \mu_{(i)})(T_{\nu_i,i} - T_{\nu_i-1,i}) + \sum_{\nu=i}^{\nu_i-1} (\mu_{(\nu+1)} - \mu_{(i)})(T_{\nu,i} - T_{\nu-1,i}) \\
&= 2 \underbrace{(\mu_{(\nu_i+1)} - \mu_{(i)})(T'_{\nu_i,i} - T'_{\nu_i-1,i})}_{a_i} + 2 \underbrace{\sum_{\nu=i}^{\nu_i-1} (\mu_{(\nu+1)} - \mu_{(i)})(T'_{\nu,i} - T'_{\nu-1,i})}_{b_i}
\end{aligned}
$$

where

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

Furthermore,

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

and we can write

$$
\begin{aligned}
b_i &\leq (\mu_{(\nu_i)} - \mu_{(i)}) \sum_{\nu=i}^{\nu_i-1} (T'_{\nu,i} - T'_{\nu-1,i}) \\
&\leq (\mu_{(\nu_i)} - \mu_{(i)}) T'_{\nu_i-1,i} + \nu^* (\mu_{(\nu_i)} - \mu_{(i)})
\end{aligned}
$$

Let us then notice that

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

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

We therefore get

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

Applying the same steps for a_i yields

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

By Lemma A.9, we have $q(T'_{\nu_i,i})\leq\frac{8\log(2T^2K^2)}{(\mu_{(\nu_i+1)}-\mu_{(i)})^2}$ so that

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

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

$$\begin{aligned}
q(T'_{\nu_i-1,i})&\leq\min(\frac{8M^2p^2\log(2K^2T^2)}{\Delta^2_{\nu_i-1}},\frac{8\log(2T^2K^2)}{(\mu_{(\nu_i)}-\mu_{(i)})^2})\\
&\leq\sqrt{\frac{8M^2p^2\log(2K^2T^2)}{\Delta^2_{\nu_i-1}}\frac{8\log(2T^2K^2)}{(\mu_{(\nu_i)}-\mu_{(i)})^2}}\\
&=\frac{8Mp\log(2K^2T^2)}{\Delta_{\nu_i-1}(\mu_{(\nu_i)}-\mu_{(i)})}
\end{aligned}$$

and therefore

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

so that

$$(i)\leq 12(\nu^*+1)^3p+\mathbb{E}[\sum_{i=1}^{\nu^*}(\nu^*+1)192\underbrace{[\frac{\log(2T^2K^2)}{\mu_{(\nu_i+1)}-\mu_{(i)}}+\frac{Mp\log(2K^2T^2)}{\Delta_{\nu_i-1}}]}_{(a)}+144K(\nu^*+1)^2$$

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

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

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

$$\begin{aligned}
s_i &= \sum_{\nu=i}^{\nu_i} (\mu_{\nu+1} - \mu_{(i)})(T_{\nu,i} - T_{\nu-1,i}) \\
&\leq (\mu_{\nu_i+1} - \mu_{(i)}) \sum_{\nu=i}^{\nu_i} (T_{\nu,i} - T_{\nu-1,i}) \\
&\leq (\mu_{\nu_i+1} - \mu_{(i)}) T_{\nu_i,i} \\
&\leq 2(\nu^*+1)^2 + 2(\nu^*+1)\frac{6q(T'_{\nu_i,i})}{p}(\mu_{\nu_i+1} - \mu_{(i)}) + \frac{36K(\nu^*+1)}{p} \qquad \text{(Similar as the bound of } b_i\text{)} \\
&\leq 2(\nu^*+1)^2 + 2(\nu^*+1)\frac{48M\log(2K^2T^2)}{\Delta_{\nu_i}} + \frac{36K(\nu^*+1)}{p} \\
&\leq 2(\nu^*+1)^2 + 2(\nu^*+1)\frac{48M\log(2K^2T^2)}{\Delta_{\nu^*-1}} + \frac{36K(\nu^*+1)}{p}
\end{aligned}$$

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

So in any case

$$(i) \leq 12(\nu^*+1)^3 p + 180K(\nu^*+1)^2 + \sum_{i=1}^{\nu^*} \frac{288(\nu^*+1)\log(2T^2K^2)}{\mu_{(\nu^*+1)} - \mu_{(i)}} + p\nu^* \frac{192M(\nu^*+1)\log(2K^2T^2)}{\Delta_{\nu^*-1}}$$

Bounding (ii) we have

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

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

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

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

$$(A) = \sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} \underbrace{\sum_{j=\nu+1}^{K} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \le n_{E,\nu'}\} \mathbb{1}\{n \le n_{A,j}\}}_{A_n}$$

Call τ_n the value of τ at the end of phase n, t_n the value of t at the end of phase n and set

$$t'_n = (t_n - (\nu^*+1))/2$$

Notice that if $\tau_n \ge 8(\nu^*+1)^2$, then

$$\begin{aligned}
\tau(t'_n) &\ge \frac{t'_n}{\nu^*+1} - \nu^* \\
&= \frac{t_n - (\nu^*+1)}{2(\nu^*+1)} - \nu^* \\
&= \frac{t_n}{2(\nu^*+1)} - \frac{1}{2} - \nu^* \\
&\ge \frac{\tau_n}{2(\nu^*+1)} - \frac{1}{2} - \nu^* \\
&\ge \frac{\tau_n}{4(\nu^*+1)}
\end{aligned}$$

$$\begin{aligned}
(A) = &\underbrace{\sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} A_n \mathbb{1}\{\tau_{n-1} \ge 8(\nu^*+1)^2\} \mathbb{1}\left\{q(t'_{n-1}) \ge \frac{1}{3} p\tau(t'_{n-1})\right\}}_{A_1} \\
&+ \underbrace{\sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} A_n \mathbb{1}\{\tau_{n-1} < 8(\nu^*+1)^2\}}_{A_2} \\
&+ \underbrace{\sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} A_n \mathbb{1}\{\tau_{n-1} \ge 8(\nu^*+1)^2\} \mathbb{1}\left\{q(t'_{n-1}) < \frac{1}{3} p\tau(t'_{n-1})\right\}}_{A_3}
\end{aligned}$$

Let us start with bounding A_2:

$$\begin{aligned}
A_2 &= \sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} \sum_{j=\nu+1}^{K} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \le n_{E,\nu'}\} \mathbb{1}\{n \le n_{A,j}\} \mathbb{1}\{\tau_{n-1} \le 8(\nu^*+1)^2\} \\
&\le \sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} \sum_{j=\nu+1}^{K} \mathbb{1}\{n \le n_{A,j}\} \mathbb{1}\{\tau_{n-1} \le 8(\nu^*+1)^2\}
\end{aligned}$$

We then use that

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

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.

and the following bound follows:

$$\begin{aligned} A_2 &\leq \sum_{\nu=\nu'}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} (\tau_n - \tau_{n-1}) \mathbb{1}\{\tau_{n-1} \leq 8(\nu^*+1)^2\} \\ &\leq 8(\nu^*+1)^2 \end{aligned}$$

Let us then focus on A_3:

$$\begin{aligned} A_3 &= \sum_{\nu=\nu'}^{\nu^*} \sum_{n \in \mathcal{N}_\nu} \sum_{j=\nu+1}^{K} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \leq n_{E,\nu'}\} \mathbb{1}\{n \leq n_{A,j}\} \mathbb{1}\{\tau_{n-1} \geq 8(\nu^*+1)^2\} \mathbb{1}\left\{q(t'_{n-1}) < \frac{1}{3} p\tau(t'_{n-1})\right\} \\ &\leq \sum_{n \in \mathbb{N}} (\tau_{n-1} - \tau_{n-1}) \mathbb{1}\left\{q(t'_{n-1}) < \frac{1}{3} p\tau(t'_{n-1})\right\} \\ &\leq \sum_{n \in \mathbb{N}} K \mathbb{1}\left\{q(t'_{n-1}) < \frac{1}{3} p\tau(t'_{n-1})\right\} \end{aligned}$$

where we use

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

We have:

$$\begin{aligned} \mathbb{E}[A_3] &\leq \sum_{n \in \mathbb{N}} P(q(t'_{n-1}) < \frac{1}{3} p\tau(t'_{n-1})) \\ &\leq K^2 \sum_{n \in \mathbb{N}, \tau(t'_{n-1}) > 0} \exp(-\frac{2}{9} p\tau(t'_{n-1})) \\ &\leq K^2 \int_0^\infty \exp(-\frac{2}{9} pt) dt \qquad \text{(At each phase, all active arms are played at least one time)} \\ &\leq \frac{9K^2}{2p} \end{aligned}$$

Then we turn to A_1. Call E_n the event

$$E_n = \{q(t'_{n-1}) \geq \frac{1}{3} p\tau(t'_{n-1}), \tau_{n-1} \geq 8(\nu^*+1)^2\}$$

Under E_n, we can write

$$\begin{aligned} n \leq n_{A,j} &\implies q(t'_{n-1}) \leq q_{A,j} \\ &\implies \frac{1}{3} p\tau(t'_{n-1}) \leq q_{A,j} \\ &\implies \frac{1}{3} p \frac{\tau_{n-1}}{4(\nu^*+1)} \leq q_{A,j} \\ &\implies \tau_{n-1} \leq \frac{96(\nu^*+1)\log(2T^3K^2)}{(\mu_{(j)} - \mu_{(\nu)})^2 p} \qquad \text{(By Lemma A.10)} \\ &\implies \mu_{(j)} - \mu_\nu \leq \sqrt{\frac{96(\nu^*+1)\log(2T^3K^2)}{\tau_{n-1}}} \triangleq \delta_n \end{aligned}$$

Calling n_ν the last phase before ν is increased, we have

$$\begin{aligned}
A_1 &= \sum_{\nu=\nu'}^{\nu^*} \sum_{n\in\mathcal{N}_\nu} \sum_{j=\nu+1}^{K} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \le n_{E,\nu'}\} \mathbb{1}\{n \le n_{A,j}\} \mathbb{1}\{E_n\} \\
&= \sum_{\nu=\nu'}^{\nu^*-1} \sum_{j=\nu+1}^{K} \sum_{n\in\mathcal{N}_\nu} (\mu_{(j)} - \mu_{(\nu)}) \mathbb{1}\{n \le n_{E,\nu'}\} \mathbb{1}\{n \le n_{A,j}\} \mathbb{1}\{E_n\} \\
&\quad + \sum_{j=\nu^*+1}^{K} \sum_{n\in\mathcal{N}_{\nu^*}} (\mu_{(j)} - \mu_{(\nu^*)}) \mathbb{1}\{n \le n_{E,\nu'}\} \mathbb{1}\{n \le n_{A,j}\} \mathbb{1}\{E_n\} \\
&\le \sum_{n\in\mathbb{N}} (\tau_n - \tau_{n-1}) \delta_n \Big(\mathbb{1}\{n \le n_{\nu^*-1}\} + \mathbb{1}\{n \le n_{E,\nu^*}\} \Big) \mathbb{1}\{E_n\}
\end{aligned}$$

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

$$\begin{aligned}
\tau_n - \tau_{n-1} &= \frac{96(\nu^*+1)\log(2T^3K^2)}{p} \left(\frac{1}{\delta_{n+1}^2 - \delta_n^2} \right) \\
&= \frac{96(\nu^*+1)\log(2T^3K^2)}{p} \left(\frac{1}{\delta_n} + \frac{1}{\delta_{n+1}} \right) \left(\frac{1}{\delta_{n+1}} - \frac{1}{\delta_n} \right)
\end{aligned}$$

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

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

We can then write:

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

Under E_n, we have

$$\begin{aligned}
n \le n_{E,\nu^*} &\implies q(t'_{n-1}) \le q_{E,\nu^*} \\
&\implies \frac{1}{3} p\tau(t'_{n-1}) \le q_{E,\nu^*} \\
&\implies \tau(t'_{n-1}) \le \frac{24\log(2T^3K^2)}{p(\mu_{(\nu^*+1)} - \mu_{(\nu^*)})^2} && \text{(By Lemma A.9)} \\
&\implies \mu_{(\nu^*+1)} - \mu_{(\nu^*)} \le \sqrt{\frac{96(\nu^*+1)\log(2T^3K^2)}{\tau_{n-1}p}} = \delta_n
\end{aligned}$$

so that

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

Similarly, under E_n, we have:

$$\begin{aligned}
n \le n_\nu &\implies q(t'_{n-1}) \le q_\nu \\
&\implies \frac{1}{3} p\tau(t'_{n-1}) \le q_\nu \\
&\implies \tau(t'_{n-1}) \le \frac{24M^2p^2\log(2T^3K^2)}{(\Delta^{(\nu)})^2 p} && \text{(By Lemma A.7)} \\
&\implies \frac{\Delta^{(\nu)}}{Mp} \le \sqrt{\frac{96(\nu^*+1)\log(2T^3K^2)}{\tau_{n-1}p}} = \delta_n
\end{aligned}$$

so that

$$\frac{\Delta^{(\nu^*)}}{Mp} \leq \frac{\Delta^{(\nu)}}{Mp} \leq \delta_{n_\nu}$$

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

$$A_1 \leq \frac{384(\nu^*+1)\log(2T^3K^2)}{p}\frac{1}{\mu_{(\nu^*+1)}-\mu_{\nu'}} + Mp\frac{384(\nu^*+1)\log(2T^3K^2)}{p}\frac{1}{\Delta_{\nu^*-1}}$$

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

so that

$$(ii) \leq \sum_{\nu'=1}^{\nu^*} \frac{384(\nu^*+1)\log(2T^3K^2)}{\mu_{(\nu^*+1)}-\mu_{\nu'}} + \nu^* Mp\frac{384(\nu^*+1)\log(2T^3K^2)}{\Delta^{(\nu^*)}} + \frac{9K^2}{2} + 8p(\nu^*+1)^2$$

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

$$R_{\mathcal{E}} \leq \sum_{\nu'=1}^{\nu} \frac{672(\nu^*+1)\log(2T^3K^2)}{\mu_{(\nu^*+1)}-\mu_{\nu'}} + \nu^* Mp\frac{576(\nu^*+1)\log(2K^2T^3)}{\Delta^{(\nu^*)}} + 200K(\nu^*+1)^2 + 5K^2$$

□

It remains to bound R_M. Recall that R_M measures the mismatch between the chosen assignment $\mathbf{M}(t)$ and the best possible assignment with the same support. Crucially there is no support mismatch and therefore we are in a setting close to the full information setting which allows us to bound R_M by a quantity independent of the horizon T.

Lemma A.13 (Bound on R_M)**.**

$$\mathbb{E}[R_M] \leq 4Mp(\nu^*+1)^2 + 2MK\frac{6(\nu^*+1)}{r}$$

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

$$\begin{aligned}
\mathbb{E}[R_M] &= \mathbb{E}[\sum_{t=1}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] \\
&= \mathbb{E}[\sum_{t=1}^{2(\nu^*+1)^2}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] + \mathbb{E}[\sum_{t=2(\nu^*+1)^2}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle] \\
&\leq 4Mp(\nu^*+1)^2 + \underbrace{\mathbb{E}[\sum_{t=2(\nu^*+1)^2}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle]}_{(i)} \qquad \text{(By Equation (9))}
\end{aligned}$$

Then we write

$$
\begin{aligned}
(i) &\leq \sum_{t=2(\nu^*+1)^2}^{T} \mathbb{E}[\underbrace{\langle \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t'), g(\mathbf{M}^*_{\mathcal{E}(t)}) - g(\mathbf{M}(t))\rangle}_{Y_t}] && \text{(With } t' = 2^{\lfloor \log_2(t)\rfloor}\text{)} \\
&= \sum_{t=2(\nu^*+1)^2}^{T} \mathbb{E}[Y_t \mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\} + Y_t \mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\}] \\
&\leq \sum_{t=2(\nu^*+1)^2}^{T} (\mathbb{E}[Y_t \mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\}] + 2MpP(q(t') \leq \frac{p}{3}\tau(t'))) \\
&\leq \underbrace{\sum_{t=2(\nu^*+1)^2}^{T} \mathbb{E}[Y_t \mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\}]}_{(a)} + \underbrace{\sum_{t=2(\nu^*+1)^2}^{T} 2KMp\exp(-\frac{2}{9}p\tau(t'))}_{(b)}
\end{aligned}
$$

Bounding (b):

$$
\begin{aligned}
(b) &\leq \sum_{t=2(\nu^*+1)^2}^{T} 2KMp\exp(-\frac{2}{9}p\tau(t/2)) && \text{(Since } t/2 \leq t' \text{ and } \tau \text{ increasing)} \\
&\leq \sum_{t=2(\nu^*+1)^2}^{T} 2KMp\exp(-\frac{2}{9}p(\frac{t}{2(\nu^*+1)} - \nu^*)) && \text{(By Lemma A.5)} \\
&\leq \sum_{t=1}^{T} 2MKp\exp(-\frac{2}{9}p(\frac{t}{2(\nu^*+1)})) \\
&\leq \frac{18MKp(\nu^*+1)}{p}
\end{aligned}
$$

Bounding (a):

$$
\begin{aligned}
(a) &\leq \sum_{t=2(\nu^*+1)^2}^{T} Mp\mathbb{E}\Big[\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t')\|_\infty \mathbb{1}\{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t')\|_\infty \geq r\}\mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\}\Big] \\
&\qquad\qquad \text{(By Equation (9) and definition of } r\text{)} \\
&\leq \sum_{t=2(\nu^*+1)^2}^{T} 2Mp\mathbb{E}\Big[(r\mathbb{P}\{|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t')|_\infty \geq r\} + \int_r^\infty \mathbb{P}\{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t')\|_\infty \geq \varepsilon\}d\varepsilon)\mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\}\Big] \\
&\leq \sum_{t=2(\nu^*+1)^2}^{T} \mathbb{E}[2MpK\Big(r\exp(-2q(t')r^2) + \int_r^\infty \exp(-2q(t')\varepsilon^2)d\varepsilon\Big)\mathbb{1}\Big\{q(t') \leq \frac{p}{3}\tau(t')\Big\}]
\end{aligned}
$$

so we have

$$\begin{aligned}
(i) &\leq \mathbb{E}[\sum_{t=2(\nu^*+1)^2}^{\infty} 2MpK\Big(r\exp(-2q(t')r^2)+\int_r^{\infty}\exp(-2q(t')\varepsilon^2)d\varepsilon\Big)\mathbb{1}\Big\{q(t')\leq\frac{p}{3}\tau(t')\Big\}]\\
&\leq \mathbb{E}[\sum_{t=2(\nu^*+1)^2}^{\infty} 2MpK\Big(r\exp(-2\frac{1}{3}p\tau(t')r^2)+\int_r^{\infty}\exp(-2\frac{1}{3}p\tau(t')\varepsilon^2)d\varepsilon\Big)]\\
&\leq \sum_{t=1}^{\infty} 2MpK\Big(r\exp(-2\frac{1}{3}p\frac{t}{2(\nu^*+1)}r^2)+\int_r^{\infty}\exp(-2\frac{1}{3}p\frac{t}{2(\nu^*+1)}\varepsilon^2)d\varepsilon\Big) \qquad \text{(By Lemma A.5)}\\
&\leq \sum_{t=1}^{\infty} 2MpK\Big(r\exp(-2\frac{1}{3}p\frac{t}{2(\nu^*+1)}r^2)+\int_r^{\infty}\exp(-2\frac{1}{3}p\frac{t}{2(\nu^*+1)}\varepsilon^2)d\varepsilon\Big)\\
&\leq 2MpK\Big(r\frac{6(\nu^*+1)}{2pr^2}+\int_r^{\infty}\frac{6(\nu^*+1)}{2p\varepsilon^2}d\varepsilon\Big)\\
&\leq 2MpK\frac{6(\nu^*+1)}{2p}\Big(\frac{1}{r}+\frac{1}{r}\Big)\\
&= 2MK\frac{6(\nu^*+1)}{r}
\end{aligned}$$

so that

$$\mathbb{E}[R_M]\leq 4Mp(\nu^*+1)^2+2MK\frac{6(\nu^*+1)}{r}$$

□

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

$$\mathbb{E}[R]\leq R_{cCG}^{UB}\triangleq\frac{960M^2p\log(2K^2T^3)(\nu^*+1)}{\Delta^{(\nu^*)}}+\sum_{\nu'=1}^{\nu}\frac{672(\nu^*+1)\log(2T^3K^2)}{\mu_{(\nu^*+1)}-\mu_{\nu'}}+\frac{265MK(\nu^*+1)}{r}$$

where all constants have been put in the term in $\frac{1}{r}$.

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

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

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

Proof. Call E_s the event "$\forall m\in[M]$, at least one call of $SEND_{m\to\text{gateway}}$ and at least one call of $SEND_{\text{gateway}\to m}$ succeeds".

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

$$\forall s\geq s^*\mathbb{P}(\bar{E}_s)\leq 2M(1-p_g)^{\frac{2^{s-2}}{M}}\leq\frac{1}{T^2}$$

and in particular,

$$\mathbb{P}(\cup_{s=s^*}^{\log(T)}\bar{E}_s)\leq\frac{1}{T}$$

The regret R_{CG} can therefore be decomposed as

$$\mathbb{E}[R_{CG}]\leq\underbrace{\sum_{s=1}^{s^*}\sum_{t=2^s}^{2^{s+1}}\mathbb{E}[R_{CG}(t)]}_{(i)}+\underbrace{\sum_{s=s^*+1}^{\lfloor\log_2(T)\rfloor}\sum_{t=2^s}^{2^{s+1}}\mathbb{E}[R_{CG}(t)\mathbb{1}\Big\{\cap_{s=s^*}^{\log(T)}E_s\Big\}]}_{(ii)}+2Mp$$

Under $\cap_{s=s^*}^{\log(T)} E_s$, stage $l > s^*$ uses 2^s samples for each phase $s \leq l-2$ but only 2^{s-1} samples from phase $l-1$. In particular it uses 2^{s-1} samples from each phase up to phase $l-1$. Therefore:

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

For (i), we can write

$$\begin{aligned}(i) &= (i)(\mathbb{1}\{\nu^* = 0\} + \mathbb{1}\{\nu^* \neq 0\}) \\ &\leq (i)\mathbb{1}\{\nu^* = 0\} + 16M^2 p \frac{\log(2MT^2)}{\log(1/(1-p_g))}\mathbb{1}\{\nu^* \neq 0\}\end{aligned}$$

Then we have

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

The condition $s > \log_2(8M)$ is to ensure that at least $2M$ call to SEND can be performed.

Let us then write:

$$\begin{aligned}&\sum_{s=\log_2(8M)}^{s^*} \sum_{t=2^s}^{2^{s+1}} \mathbb{E}[R_{CG}(t)\mathbb{1}\{\nu^* = 0\}\left(\mathbb{1}\{E_s\} + \mathbb{1}\{\bar{E}_s\}\right]) \\ &\leq 2R_{CG}^{UB}\mathbb{1}\{\nu^* = 0\} + \sum_{s=\log_2(8M)}^{s^*} 2Mp2^s\mathbb{E}[\mathbb{1}\{\bar{E}_s\}]\end{aligned}$$

and we have:

$$\begin{aligned}\sum_{s=\log_2(8M)}^{s^*} 2^s\mathbb{E}[\mathbb{1}\{\bar{E}_s\}] &\leq \sum_{s=\log_2(8M)}^{s^*} 2^s \exp(\frac{\log(1-p_g)}{8M}2^s) \\ &\leq \frac{8M}{-\log(1-p_g)}\end{aligned}$$

which concludes the proof □

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

$$\begin{aligned}\mathbb{E}[R_{CG}] &\leq 4R_{cCG}^{UB} + \frac{16M^2p}{-\log(1-p_g)} + 16M^2p\frac{\log(2MT^2)}{-\log(1-p_g))}\mathbb{1}\{\nu^* \neq 0\} + 2Mp + \log_2(8M)2Mp \\ &\leq \frac{3840M^2p\log(2K^2T^3)(\nu^*+1)}{\Delta^{(\nu^*)}} + \sum_{\nu=1}^{\nu^*}\frac{2688(\nu^*+1)\log(2T^3K^2)}{\mu_{(\nu^*+1)} - \mu_\nu} + \frac{1078MK(\nu^*+1)}{r} \\ &+ \frac{16M^2p}{-\log(1-p_g)}(1 + \log(2MT^2)\mathbb{1}\{\nu^* \neq 0\})\end{aligned}$$

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 $x \neq 0$:

$$g(N-x)\frac{1}{2} + g(N+1+x)(\frac{1}{2} + \Delta) \leq g(N)\frac{1}{2} + g(N+1)(\frac{1}{2} + \Delta) \tag{12}$$

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

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

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

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

For $0 < x < N$, we have

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

Therefore if $\Delta \leq p$, Equation (12) is satisfied for $0 < x < N$.

For $-(N+1) < x < 0$, set $y = -x - 1$ so that $x = -y - 1$ and $0 \leq y \leq N$. We can write $g(N-x)\frac{1}{2} + g(N+1+x)(\frac{1}{2}+\Delta) = g(N+y+1)\frac{1}{2} + g(N-y)(\frac{1}{2}+\Delta) < g(N+y+1)(\frac{1}{2}+\Delta) + g(N-y)\frac{1}{2}$ which gives the desired inequality for $y = 0$. For $y > 0$, Equation (12) is satisfied if $\Delta \leq \frac{p}{1-p}$. Therefore if $\Delta \leq \frac{p}{1-p}$, Equation (12) is satisfied and therefore, the optimal assignment if $\boldsymbol{\mu} = \boldsymbol{\mu}_1$ 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 $r = \min_{\boldsymbol{\mu}', \operatorname{argmax}_{\mathbf{M}\in\mathcal{M}}\langle\boldsymbol{\mu}', g(\mathbf{M})\rangle \neq \mathbf{M}^*} \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty$. Call $\boldsymbol{\mu}_r = \operatorname{argmin}_{\boldsymbol{\mu}', \operatorname{argmax}_{\mathbf{M}\in\mathcal{M}}\langle\boldsymbol{\mu}', g(\mathbf{M})\rangle \neq \mathbf{M}^*} \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty$ and $\mathbf{M}_r = \operatorname{argmax}_{\mathbf{M}\in\mathcal{M}}\langle\boldsymbol{\mu}_r, g(\mathbf{M})\rangle$.

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

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

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

r_2 is the minimum value such that

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

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 and 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 and Gaïffas (2019). Call for $i = 1, 2$, let $\mathbb{P}_i$ be the joint probability on $(\mathbf{X}(1), \dots, \mathbf{X}(T))$ 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 and Szepesvári (2020)), we have

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

where KL is the KL-divergence.

More precisely, we have

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

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:

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

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

Equation (13) 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 (14) 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 (15) 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) [3] 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 $\Delta_k(0) = 0$.

[3] 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

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.

$$\mathfrak{C}_k = \min_{m_1,\dots,m_T:\sum_t g(m_t)\geq \frac{\log(T)}{2\Delta_{(k)}^2}} \sum_{t=1}^{T} \Delta_k(m_t) \tag{16}$$

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 (16) 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* $k^* = \text{argmax}_{k\in[K]}\, \mu_k$ *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: $G = \mu_i p - \mu_k(g(M^*_{k^*}) - g(M^*_{k^*} - 1))$.

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 (6) and an upper bound using $\hat{\mu}_k^H(t) = \min(\hat{\mu}_k(t) + \zeta_k(t), 1)$ where $\boldsymbol{\zeta}$ is given by $k \in [K]$, $\zeta_k(t) = \sqrt{\frac{\log(2T^3K^2)}{2T_k(t)}}$ and take

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

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

The code is given in Algorithm 4.

Algorithm 4 UCB

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

C.2 Analysis

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

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

$$\mathbb{E}[R_{UCB}] \leq 2\sqrt{2K\log(2T^3K^2)T\min(K, Mp + \frac{K}{T})} + 2 \tag{17}$$

Proof. Define the GOOD event as in Lemma A.1.

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

Then, under the GOOD event, we have:

$$\begin{aligned}
R_{CUCB} &= \sum_{t=1}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}^*)\rangle - \sum_{t=1}^{T}\langle \boldsymbol{\mu}, g(\mathbf{M}(t))\rangle \\
&= \sum_{t=1}^{T}\langle \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M}^*)\rangle + \langle \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M}^*) - g(\mathbf{M}(t))\rangle + \langle \hat{\boldsymbol{\mu}}^H(t) - \boldsymbol{\mu}, g(\mathbf{M}(t))\rangle \\
&\leq \langle \hat{\boldsymbol{\mu}}^H(t), g(\mathbf{M}^*) - g(\mathbf{M}(t))\rangle + \langle \hat{\boldsymbol{\mu}}^H(t) - \boldsymbol{\mu}, g(\mathbf{M}(t))\rangle && \text{(Since } \hat{\boldsymbol{\mu}}^H \geq \boldsymbol{\mu} \text{ by the GOOD event)} \\
&\leq \sum_{t=1}^{T}\langle \hat{\boldsymbol{\mu}}^H(t) - \boldsymbol{\mu}, g(\mathbf{M}(t))\rangle && \text{(Since } \mathbf{M}(t) = \operatorname{argmax}_{\mathbf{M}\in\mathcal{M}}\langle \hat{\boldsymbol{\mu}}^H, g(\mathbf{M})\rangle\text{)} \\
&= \sum_{k=1}^{K}\sum_{t=1}^{T}\min(1, 2\zeta_k(t))g(M_k(t)) && \text{(Since } \boldsymbol{\mu} \geq \max(\hat{\boldsymbol{\mu}} - \boldsymbol{\zeta}, \mathbf{0}) \text{ by the GOOD event)} \\
&= \sum_{k=1}^{K}\sum_{t=1}^{T}\min(1, \sqrt{2\frac{\log(2T^3K^2)}{T_k(t)}})(g(M_k(t)) - \eta_k(t) + \eta_k(t)) && (*) \\
&\leq \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{T}(g(M_k(t)) - \eta_k(t))}_{(i)} + \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{T}\sqrt{2\frac{\log(2T^3K^2)}{T_k(t)}}\eta_k(t)}_{(ii)}
\end{aligned}$$

(*) 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

$$\begin{aligned}
(ii) &= \sum_{k=1}^{K}\sum_{t=1}^{T}\sqrt{2\frac{\log(2T^3K^2)\eta_k(t)}{T_k(t)}} && \text{(Since } \eta_k(t) = \sqrt{\eta_k(t)} \text{ as } \eta_k(t) \in \{0,1\}\text{)} \\
&= \sum_{k=1}^{K}\sqrt{2\log(2T^3K^2)}\sum_{t=1}^{T}\sqrt{\frac{\eta_k(t)}{\sum_{\rho=1}^{t}\eta_k(\rho)}} \\
&= \sum_{k=1}^{K}\sqrt{2\log(2T^3K^2)}\sum_{i=1}^{\max(T_k(T),1)}\frac{1}{\sqrt{i}} && \text{(Since } \forall\rho\in[t], \eta_k(\rho)\in\{0,1\}\text{)} \\
&\leq \sum_{k=1}^{K} 2\sqrt{2\log(2T^3K^2)\max(T_k(T),1)}
\end{aligned}$$

Then we have trivially:

$$\mathbb{E}[(ii)] \leq 2K\sqrt{2\log(2T^3K^2)T} \tag{18}$$

Otherwise, we write:

$$\begin{aligned}
\mathbb{E}[(ii)] &\leq \mathbb{E}[2\sqrt{2K\log(2T^3K^2)\sum_{k=1}^{K}(T_k(T)+\mathbb{1}\{T_k(T)=0\})}] && \text{(Using } \textstyle\sum_{i=1}^{K}\sqrt{a_i}\leq\sqrt{K\sum_{i=1}^{K}a_i}\text{)}\\
&\leq 2\sqrt{2K\log(2T^3K^2)\sum_{k=1}^{K}(\mathbb{E}[T_k(T)]+\mathbb{P}(T_k(T)=0))} && \text{(By Jensen inequality)}\\
&= 2\sqrt{2K\log(2T^3K^2)\sum_{k=1}^{K}(\sum_{\rho=1}^{T}g(M_k(\rho))+\prod_{\rho=1}^{T}(1-g(M_k(\rho))))}\\
&\leq 2\sqrt{2K\log(2T^3K^2)\sum_{k=1}^{K}(\sum_{\rho=1}^{T}M_kp+1)} && \text{(Since } 0\leq g(M_k)\leq 1 \text{ and } g(M_k)\leq M_kp\text{)}\\
&\leq 2\sqrt{2K\log(2T^3K^2)(TMp+K))}
\end{aligned}$$

and therefore

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

so that

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

□

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

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

We want to solve

$$\operatorname*{argmax}_{\mathcal{M}_\mathcal{E}}\langle g(\mathbf{M}), \mathbf{v}\rangle \tag{19}$$

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 $a_1,\dots,a_M\in[K]$ the arms chosen by the sequential algorithm for players $1,\dots,M$. 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 $a_i, a_1,\dots,a_k$ until $a_{k+1}=a_i$. Then everything happens as if the assignment chosen by A was $a_1,\dots,a_{k+1}$ and therefore the rest of the run is the same as the sequential algorithm.

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

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

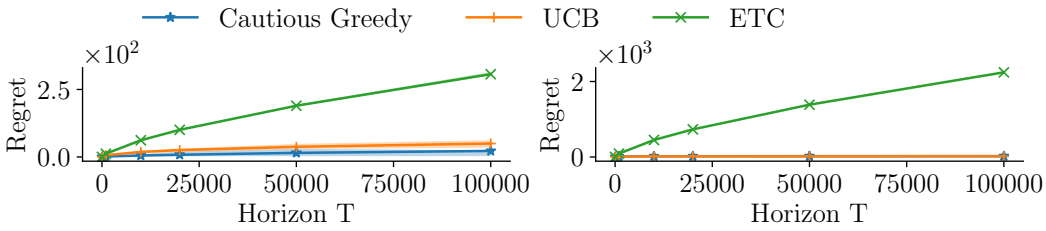

Figure 2: (left) $\nu^* = 0$, (right) $\nu^* = 1$, $T \in \{10, 1e3, 1e4, 2e4, 5e4, 1e5\}$

E Additional experiment

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