Coordination without communication: optimal regret in two players multi-armed bandits

Sébastien Bubeck SEBUBECK@MICROSOFT.COM
Microsoft Research

Thomas Budzinski BUDZINSKI@MATH.UBC.CA
University of British Columbia

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider two agents playing simultaneously the same stochastic three-armed bandit problem. The two agents are cooperating but they cannot communicate. Under the assumption that shared randomness is available, we propose a strategy with no collisions at all between the players (with very high probability), and with near-optimal regret $O(\sqrt{T\log(T)})$. We also argue that the extra logarithmic term $\sqrt{\log(T)}$ should be necessary by proving a lower bound for a full information variant of the problem.

1. Introduction

We consider the (cooperative) multi-player version of the classical stochastic multi-armed bandit problem. We focus on the case of two players, Alice and Bob, and three actions. The problem can be defined as follows. The environment[1] is described by the mean losses $\mathbf{p} = (p_1, p_2, p_3) \in [0,1]^3$ for the three actions. The parameter $\mathbf{p}$ is unknown to the players. Denote $(\ell_t(i))_{1\leq i\leq 3, 1\leq t\leq T}$ for a sequence of independent random variables such that $\mathbb{P}(\ell_t(i) = 1) = p_i$ and $\mathbb{P}(\ell_t(i) = 0) = 1 - p_i$. At each time step $t = 1, \dots, T$, Alice and Bob choose independently two actions $i_t^A \in \{1,2,3\}$ and $i_t^B \in \{1,2,3\}$. If they collide, i.e. $i_t^A = i_t^B$, then they both suffer the maximal loss of 1. Otherwise they respectively suffer the losses $\ell_t(i_t^A)$ and $\ell_t(i_t^B)$. As is usual in bandit scenarios, each player receives only its own loss as feedback (in particular when a player receives a loss of 1, they don't know if they have collided or if it came from the loss ℓ_t). The goal of the players is to minimize their (combined) cumulative losses. To evaluate the performance of Alice and Bob we measure the regret R_T, defined as the (expected) difference between their cumulative losses and the best they could have done if they knew $\mathbf{p}$, namely $T \cdot \mathbf{p}^*$ where $\mathbf{p}^* = \min(p_1 + p_2, p_1 + p_3, p_2 + p_3)$. That is:

$$R_T = \sum_{t=1}^{T} \left(2 \cdot \mathbb{1}_{i_t^A = i_t^B} + \mathbb{1}_{i_t^A \neq i_t^B} (p_{i_t^A} + p_{i_t^B}) - \mathbf{p}^* \right). \tag{1}$$

1.1. Main result and related works

The above problem is motivated by cognitive radio applications, where players correspond to devices trying to communicate with a cell tower, and the actions correspond to different channels.

1. We focus on $\{0,1\}$-valued losses. Note that it is easy to reduce $[0,1]$-valued losses to $\{0,1\}$.

The model was first introduced roughly at the same time in Lai et al. (2008); Liu and Zhao (2010); Anandkumar et al. (2011), and has been extensively studied since then (Avner and Mannor, 2014; Rosenski et al., 2016; Bonnefoi et al., 2017; Lugosi and Mehrabian, 2018; Boursier and Perchet, 2018; Alatur et al., 2019; Bubeck et al., 2019). Despite all this attention, at the moment the state of the art regret bound is $\widetilde{O}(T^{3/4})$. The latter regret was obtained for two players in Bubeck et al. (2019) (in fact it holds in the more general non-stochastic case), and it can also be recovered from the bounds in Lugosi and Mehrabian (2018); Boursier and Perchet (2018) as we explain in the end of Section 2. On the other hand no non-trivial lower bound is known (i.e. only $\Omega(\sqrt{T})$ is known). A near-optimal regret of $\widetilde{O}(\sqrt{T})$ has been obtained under various extra assumptions such as revealed collisions, or assuming that players can abstain from playing, or assuming that the mean losses are bounded away from 1 (Lugosi and Mehrabian, 2018; Boursier and Perchet, 2018; Bubeck et al., 2019).

Our main contribution is the first $\widetilde{O}(\sqrt{T})$ algorithm for this problem, in the case where there are 3 arms:

Theorem 1 *There exists a randomized strategy (with shared randomness) for Alice and Bob such that, for any* $\mathbf{p} \in [0,1]^3$*, we simultaneously have*

$$\mathbb{E}[R_T] \leq 2^{20}\sqrt{T\log(T)}$$

and

$$\mathbb{P}\left(\forall t \in [T], i_t^A \neq i_t^B\right) \geq 1 - \frac{1}{T}\,, \tag{2}$$

where the expectation and the probability are with respect to both the loss sequence and the randomness in Alice and Bob's strategies[2].

The property (2) is an important part of our result, and it points to a fundamental difference between our approach and all previous works on cooperative multi-player multi-armed bandits. Indeed, all previous works have proposed strategies that use collisions as a form of implicit communication between the players, since Alice can affect Bob's feedback by trying to force collisions. For example, assume as in Lugosi and Mehrabian (2018); Boursier and Perchet (2018) that the mean-losses are bounded from above by $1-\mu$, i.e., $\|\mathbf{p}\|_\infty \leq 1-\mu$. Then if Bob plays an action for $\Omega(1/\mu)$ rounds and does not observe a single 0 loss, he knows that with high probability Alice must have been playing that action too, effectively making communication possible. Leveraging this implicit communication device, Lugosi and Mehrabian (2018); Boursier and Perchet (2018) obtain a strategy with regret $\widetilde{O}(\sqrt{T}+1/\mu)$ (we explain at the end of Section 2 how to use this result to obtain an algorithm with $\widetilde{O}(T^{3/4})$ regret without any assumption). In Bubeck et al. (2019) another $\widetilde{O}(T^{3/4})$ strategy is proposed. It is epoch-based, with Alice playing a fixed action in an epoch, and Bob playing a sleeping-bandit strategy where arms awaken as losses with value 0 are observed (i.e., an arm is awake for Bob when he can guarantee that Alice is not there for this epoch). Thus we see that both methods heavily rely on collisions for implicit communication. The approach presented in this paper is fundamentally different, in that with very high probability the two players *do not collide at all*. Thus we achieve one of the key properties required by the underlying cognitive radio

2. By our method, we can actually obtain a slightly stronger version where, with probability at least $1-1/T$ with respect to the i.i.d. loss sequence, we have both the expected regret bound and almost surely no collision (with respect to the players' randomness).

application, namely that the two agents *do not communicate in any way* once the game has started. We note however that for the strategy presented here it is crucial that Alice and Bob have shared randomness. However, in the arxiv version of the present work Bubeck and Budzinski (2020), we present a different algorithm achieving regret $O(\sqrt{T \log T})$ without shared randomness, but also without the no-collision property[3].

1.2. A toy problem

In order to motivate our new strategy, it will be useful to first consider a different model which contains the essence of the difficulty of *coordination without communication*, but without the usual *exploration or exploitation* dilemma. The first modification that we propose is to assume that, even under collisions, a "real" loss is revealed. Precisely, if both players play the same action i at round t, then we assume that they both observe independent samples from $\mathrm{Ber}(p_i)$ (rather than observing 1 in the original model). This modification completely removes the possibility for implicit communication, since Alice's feedback is now completely unaffected by the presence of Bob (and vice versa). Concretely we denote $(\ell_t^X(i))_{1 \leq i \leq 3, 1 \leq t \leq T, X \in \{A,B\}}$ for a sequence of independent random variables such that $\mathbb{P}(\ell_t^X(i) = 1) = p_i$ and $\mathbb{P}(\ell_t^X(i) = 0) = 1 - p_i$. When player $X \in \{A, B\}$ plays action i, they observe the loss $\ell_t^X(i)$ (irrespective of the other player's action). Note that in this model we still assume that the players suffer a loss of 1 if they collide, they simply don't observe their actual suffered loss (to put it differently, we are still concerned with the regret (1)). The problem now looks significantly more difficult for the players[4], and it is not clear a priori that any non-trivial guarantee can be obtained. In fact it is non-trivial even with *full information*: that is at the end of round t, player $X \in \{A, B\}$ observes $(\ell_t^X(1), \ell_t^X(2), \ell_t^X(3))$. For this modified model we assume such a full information feedback. The reason why we have chosen to have two different, independent loss sequences ℓ^A and ℓ^B is that if we had $\ell^A = \ell^B$, then A and B would have exactly the same information, in which case it is very easy to avoid collisions.

Our first task will be to give a strategy with regret $O(\sqrt{T \log(T)})$ for the full-information toy model, which we do in Section 3. The extension to the bandit scenario is then done in Section 4. An interesting property of the toy model is that it is amenable to lower bound arguments, since we avoid the difficulty created by implicit communication. In particular, in Appendix E we prove the first non-trivial lower bound for multi-player online learning, by showing that the extra factor $\sqrt{\log(T)}$ is necessary:

Theorem 2 *There exists a universal constant $c > 0$ and a distribution over $\mathbf{p}$ such that, for any strategy in the full-information toy model, one has:*

$$\mathbb{E}_{\mathbf{p}} R_T \geq c\sqrt{T \log(T)}\,.$$

Unfortunately, there does not seem to be a direct way to transfer this lower bound to the original bandit problem.

3. Note that Bubeck et al. (2019) show that for the adaptive adversary model in non-stochastic bandit, the shared randomness assumption is necessary to get sublinear regret. Observe also that from a minimax perspective, the shared randomness assumption is most natural as it is needed to even *define* a minimax strategy. Finally we note that we do propose a derandomization approach (Appendix C) for a toy variant of the problem, see below for more details.
4. It is not strictly speaking more difficult, since always receiving the feedback $\ell_t^X(i_t)$ means that the players have a slightly more accurate estimate of $\mathbf{p}$.

2. Difficulties of coordination without communication

Whether we consider the toy model, or strategies for the bandit scenario that do not exploit the extra 1's due to collisions, we face the same question: how can two agents with imperfect information coordinate without communicating? In this section we illustrate some of the difficulties of *coordination without communication*. We focus on the most basic bandit strategy, namely explore then exploit. We show how to appropriately modify it to obtain $T^{4/5}$ regret for the bandit scenario, using shared randomness. All the discussion applies similarly to the full-information toy model, and as we note at the end of the section it gives $T^{3/4}$ regret in that case.

2.1. Explore then exploit

Consider the following protocol:

1. Alice and Bob first explore in a round-robin way for $\Theta(T^b)$ rounds, where $b \in (0,1)$ is a fixed parameter. Denote $q^A(i)$ for the average loss observed by Alice on action i (and similarly $q^B(i)$ for Bob).

2. Using these estimates, the players can order the arms in terms of expected performances. Denote (A_1, A_2, A_3) (respectively (B_1, B_2, B_3)) for the order Alice (respectively Bob) obtains, in ascending order of average empirical loss (i.e., $q^A(A_1) \leq q^A(A_2) \leq q^A(A_3)$).

3. For the remaining rounds they want to exploit. Alice and Bob could have agreed that Alice will play the best action, and Bob the second best, thus for the remaining of the game Alice plays A_1 and Bob plays B_2.

The problem with this naive implementation of explore/exploit is clear: there could be ambiguity on which action is the best, for example if $p_1 = p_2 \ll p_3$, in which case both A_1 and B_2 are independent and uniform in $\{1,2\}$. Thus in this case there is a constant probability of collision, resulting in a linear regret. A natural fix is for Alice to build a set of "potential top action" $\mathcal{A}$ and for Bob to build a set of "potential second best action" $\mathcal{B}$. To decide whether an action is "potentially the top action" we fix an "ambiguity threshold" τ, and now replace step 3 above with:

3' If $q^A(A_1) \leq q^A(A_2) - \tau$ (A_1 is "clearly" the best) then let $\mathcal{A} = \{A_1\}$ (in the same case for B let $\mathcal{B} = B_2$), if not but $q^A(A_2) \leq q^A(A_3) - \tau$ (A_3 is "clearly" worse than A_1 and A_2) then let $\mathcal{A} = \{A_1, A_2\}$ (in the same case for B let $\mathcal{B} = \{B_1, B_2\}$), and if neither then let $\mathcal{A} = \{1,2,3\}$ (same for B). To avoid collisions it makes sense for Alice to play $\min(\mathcal{A})$ and for Bob to play $\max(\mathcal{B})$.

Unfortunately this is just pushing the problem to a different configuration of $\mathbf{p}$. Indeed consider for example $p_3 \gg p_1 > p_2 = p_1 - \tau$. With a constant probability Alice could end up with $\mathcal{A} = \{2\}$ and Bob with $\mathcal{B} = \{1,2\}$, in which case we have again a collision, and hence we get linear regret.

2.2. The root of the problem

Geometrically, the issues above come from the boundary regions of the "decision map" $\sigma : ([0,1]^3)^2 \to \{1,2,3\}^2$ from empirical estimates of the mean-losses to actions to be played in the exploitation phase. All our results will come from careful considerations of these boundaries. Moreover, most of the difficulties already arise for our proposed full-information toy model, hence the focus on the

toy model first. We also note that the geometric considerations are much easier with two players and three actions, which is why we focus on this case in this paper. The "high-dimensional" version of the strategy proposed in Section 3 probably requires different tools.

Before going into the geometric considerations, we can illustrate one of our insights in the simple case of the explore/exploit strategy above. Namely we propose to make the decision boundaries *random*. For the explore/exploit strategy this means taking the ambiguity threshold τ to be random. Say we take it random at scale T^{-a} for some parameter $a \in [0,1]$. More precisely let $\tau = U/T^a$ with U a uniform random variable in $[0,1]$. In particular, since we don't distinguish differences below the scale T^{-a}, we might suffer a regret of T^{1-a}. On the other hand, the only risk of collision is if Alice and Bob disagree on whether some gap $\Delta = |p_i - p_j|$ is smaller than τ or not. Since the fluctuations of the empirical means are of order $T^{-b/2}$, we have that a collision might happen if $|\tau - \Delta| = \widetilde{O}(T^{-b/2})$. To put it differently, with high probability (over the observed losses during the exploration phase), collisions happen only if

$$|U - T^a \Delta| = \widetilde{O}(T^{a-b/2})\,.$$

Because we have taken U uniform on $[0,1]$, the above event has probability (over the realization of U) at most $\widetilde{O}(T^{a-b/2})$. Thus finally we get a regret of order:

$$T \cdot T^{a-b/2} + T \cdot T^{-a} + T^b\,,$$

which is optimized at $b = 4/5$ and $a = 1/5$, resulting in a $\widetilde{O}(T^{4/5})$ regret.

2.3. Minor variants

We note that the same argument applies to the full-information toy model, where we are effectively taking $b = 1$, resulting in a $\widetilde{O}(T^{3/4})$ regret. Furthermore the same technique can be used to estimate μ in Lugosi and Mehrabian (2018); Boursier and Perchet (2018), improving upon the above $T^{4/5}$ to give $T^{3/4}$ for the bandit case.

3. Toy model upper bound

We prove here the following theorem:

Theorem 3 *There exists a deterministic strategy for Alice and Bob in the full-information toy model such that with probability at least $1 - 1/T$, one has both:*

$$R_T \leq 320\sqrt{T\log(T)}\,, \tag{3}$$

and $\forall t \in [T], i_t^A \neq i_t^B$.

For $2 \leq t \leq T$, $i \in \{1,2,3\}$ and $X \in \{A,B\}$, we write

$$q_t^X(i) = \frac{1}{t-1}\sum_{s=1}^{t-1} \ell_s^X(i),$$

with the convention $q_1^X(i) = 0$. In other words q_t^X is the estimate of the vector $\mathbf{p}$ by player X at time t. Our strategy is based on a subtle partition of the cube $[0,1]^3$. Precisely we build a map

$\sigma_t : [0,1]^3 \to \{1,2,3\} \times \{1,2,3\}$, with $\sigma_t = \left(\sigma_t^A, \sigma_t^B\right)$, such that Alice plays $i_t^A = \sigma_t^A(q_t^A)$ and Bob plays $i_t^B = \sigma_t^B(q_t^B)$. An interesting aspect of Theorem 3 compared to Theorem 1 is that we do not require shared randomness for the full-information toy model. However it will be easier for us to first describe a shared randomness strategy, and then explain how to remove that assumption. More precisely, we first build a *random partition* σ, and we prove Theorem 3 with (3) holding in expectation over this random partition. We explain how to derandomize in Section C with a *dynamic* partition.

We denote $w_t = 16\sqrt{\frac{\log(T)}{t}}$, and we fix the event

$$\Omega = \left\{\forall t \in [T], i \in \{1,2,3\}, X \in \{A,B\}, |q_t^X(i) - p_i| < \frac{w_t}{4}\right\}. \tag{4}$$

Applying Hoeffding's inequality and an union bound, one obtains

$$\mathbb{P}(\Omega) \geq 1 - \frac{1}{T}.$$

For the remainder of the section, we fix loss sequences for which Ω holds true. All probabilities will be taken with respect to the randomness of Alice and Bob. We note in particular that under Ω we have $\|q_t^X - p\|_\infty \leq \frac{w_t}{4}$ for $X \in \{A,B\}$, so we get

$$\|q_t^A - q_t^B\|_2 < w_t. \tag{5}$$

3.1. A random partition of the cube

3.1.1. CYLINDRICAL COORDINATES

To describe our partition, it will be more convenient to use cylindrical coordinates around the axis $\mathcal{D} = \{\mathbf{p} | p_1 = p_2 = p_3\}$. More precisely, for $\mathbf{p} = (p_1, p_2, p_3)$ we write

$$m_{\mathbf{p}} = \frac{p_1 + p_2 + p_3}{3},$$

$$r_{\mathbf{p}} = d(\mathbf{p}, \mathcal{D}) = \sqrt{(p_1 - m_{\mathbf{p}})^2 + (p_2 - m_{\mathbf{p}})^2 + (p_3 - m_{\mathbf{p}})^2},$$

and $\theta_{\mathbf{p}} \in [0, 2\pi)$ for the angle between the line from $\mathbf{p}$ to its orthogonal projection $(m_{\mathbf{p}}, m_{\mathbf{p}}, m_{\mathbf{p}})$ on the axis $\mathcal{D}$ and the half-line $\{(m_{\mathbf{p}} - t, m_{\mathbf{p}} + 2t, m_{\mathbf{p}} - t) \,|\, t \geq 0\}$ (this angle is contained in the plane orthogonal to $\mathcal{D}$ passing through $\mathbf{p}$). We write $\mathbf{p} = (p_1, p_2, p_3) = [m_{\mathbf{p}}, r_{\mathbf{p}}, \theta_{\mathbf{p}}]$.

An equivalent way to describe these cylindrical coordinates is as follows. Let us denote $\mathbf{a} = \frac{1}{\sqrt{3}}(1,1,1)$ (the main axis direction), $\mathbf{b} = \sqrt{\frac{2}{3}}\left(-\frac{1}{2}, 1, -\frac{1}{2}\right)$ (the direction of the half-line mentioned above), and $\mathbf{c} = \sqrt{\frac{2}{3}}\left(\frac{\sqrt{3}}{2}, 0, -\frac{\sqrt{3}}{2}\right)$ (the direction so that $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ forms an orthonormal basis). We have:

$$\begin{aligned}\mathbf{p} &= \langle \mathbf{p}, \mathbf{a}\rangle \mathbf{a} + r_{\mathbf{p}} \cos(\theta_{\mathbf{p}})\mathbf{b} + r_{\mathbf{p}} \sin(\theta_{\mathbf{p}})\mathbf{c} \\ &= \begin{pmatrix} m_{\mathbf{p}} \\ m_{\mathbf{p}} \\ m_{\mathbf{p}} \end{pmatrix} + \sqrt{\frac{2}{3}} \cdot r_{\mathbf{p}} \cdot \begin{pmatrix} \cos\left(\theta_{\mathbf{p}} + \frac{2\pi}{3}\right) \\ \cos(\theta_{\mathbf{p}}) \\ \cos\left(\theta_{\mathbf{p}} - \frac{2\pi}{3}\right) \end{pmatrix},\end{aligned}$$

where the last equality comes from standard trigonometric identities.

The basic partitioning of interest is into the three regions corresponding to different top two actions, namely $p_3 \geq p_1, p_2$ (players should play arms 1 and 2), $p_1 \geq p_2, p_3$, and $p_2 \geq p_1, p_3$. In cylindrical coordinates these regions are described respectively by $\theta \in \left[\frac{\pi}{3}, \pi\right]$, $\theta \in \left[\pi, \frac{5\pi}{3}\right]$, and $\theta \in \left[\frac{5\pi}{3}, 2\pi\right] \cup \left[0, \frac{\pi}{3}\right]$.

3.1.2. TOPOLOGICAL DIFFICULTY

Intuitively, the "topological" difficulty of the problem is that, as θ varies continuously, the players will face a decision boundary with a collision. For example, say that in the region around $\theta = 0$ (namely $\theta \in \left[\frac{5\pi}{3}, 2\pi\right] \cup \left[0, \frac{\pi}{3}\right]$) we play $(i_t^A, i_t^B) = (3, 1)$. As θ increases we enter the region where we should stop playing action 3 and start playing action 2, and thus it is natural to play $(i_t^A, i_t^B) = (2, 1)$ in the region $\theta \in \left[\frac{\pi}{3}, \pi\right]$ (i.e., only Alice is trying to figure out whether she plays action 2 or 3, while Bob stays constant on action 1). On the other hand, as we decrease θ and enter the region $\theta \in \left[\pi, \frac{5\pi}{3}\right]$, we want to play $(i_t^A, i_t^B) = (3, 2)$ (i.e., it is now Bob who tries to figure out whether to play action 2 or 1). The problem with this construction is that at $\theta = \pi$ we go from configuration $(2, 1)$ to configuration $(3, 2)$, thus at this value of θ there is a constant chance of collisions! The same occurs if $(i_t^A, i_t^B) = (3, 1)$. This observation is the core of our lower bound proof in Section E.

To fix this issue, we propose to replace this fixed interface between $(2, 1)$ and $(3, 2)$ by a random cut in the region $\theta \in \left[\frac{\pi}{3}, \pi\right]$, where we will move from $(2, 1)$ to $(1, 2)$ (and thus at $\theta = \pi$ we move from $(1, 2)$ to $(3, 2)$ and there is no risk of collision). We explain this construction next (see also Figure 2).

3.1.3. RANDOM INTERFACE

Let Θ be a uniform random variable in $\left[\frac{\pi}{3}, \pi\right]$ (this is the only randomness needed by the players). We write $\mathcal{P} = \{[m, r, \theta] | \theta = \Theta\}$, which is a (random) half-plane containing the axis $\mathcal{D}$ (this will be our "random cut", to be padded appropriately to move from $(2, 1)$ to $(1, 2)$). More precisely, we recall that $w_t = 16\sqrt{\frac{\log T}{t}}$, and define the following regions:

- $A_t = \{\mathbf{p} = [m, r, \theta] | \frac{\pi}{3} \leq \theta < \Theta \text{ and } d(\mathbf{p}, \mathcal{P}) \geq w_t\}$,
- $B'_t = \{\mathbf{p} = [m, r, \theta] | \frac{\pi}{3} \leq \theta < \Theta \text{ and } d(\mathbf{p}, \mathcal{P}) < w_t\}$,
- $C'_t = \{\mathbf{p} = [m, r, \theta] | \Theta \leq \theta < \pi \text{ and } d(\mathbf{p}, \mathcal{P}) < w_t\} \setminus \mathcal{D}$,
- $D_t = \{\mathbf{p} = [m, r, \theta] | \Theta \leq \theta < \pi \text{ and } d(\mathbf{p}, \mathcal{P}) \geq w_t\}$,
- $B''_t = \{\mathbf{p} = [m, r, \theta] | 0 \leq \theta < \frac{\pi}{3} \text{ or } \frac{5\pi}{3} \leq \theta < 2\pi\}$,
- $C''_t = \{\mathbf{p} = [m, r, \theta] | \pi \leq \theta < \frac{5\pi}{3}\} \setminus \mathcal{D}$.

We finally write $B_t = B'_t \cup B''_t$ and $C_t = C'_t \cup C''_t$. Note that the large or strict inequalities and the convention $\mathcal{D} \not\subset C_t$ were chosen so that (A_t, B_t, C_t, D_t) is a partition of the cube $[0, 1]^3$, but these choices do not really matter.

We illustrate on Figure 2 the restriction of this partition to the plane of equation $p_1+p_2+p_3 = \frac{3}{2}$. Note that the definition of A_t, B_t, C_t, D_t does not depend on the coordinate m. This implies that the full partition is just obtained from Figure 2 by adding one dimension orthogonally to the plane. More precisely, a point of $[0,1]^3$ belongs to a region of the partition if and only if its orthogonal projection on the plane of Figure 2 belongs to that region. Note that B_t'' corresponds exactly to the region where the best two arms are 1 and 3, and C_t'' to the region where the best two arms are 2 and 3.

3.1.4. COLORING THE PARTITION

We now define the map $\sigma_t : [0,1]^3 \to \{1,2,3\} \times \{1,2,3\}$ that the players use to select an action. It will be constant over the regions A_t, B_t, C_t, D_t. Precisely, as on Figure 2:

$$\sigma_t(\mathbf{q}) := \begin{cases} (2,1) & \text{if } \mathbf{q} \in A_t, \\ (3,1) & \text{if } \mathbf{q} \in B_t, \\ (3,2) & \text{if } \mathbf{q} \in C_t, \\ (1,2) & \text{if } \mathbf{q} \in D_t. \end{cases}$$

We denote by σ_t^A and σ_t^B the two coordinates of σ_t. For example, for $\mathbf{q} \in A_t$, we have $\sigma_t^A(\mathbf{q}) = 2$ and $\sigma_t^B(\mathbf{q}) = 1$. As explained above, the strategy is to set $i_t^A = \sigma_t^A(\mathbf{q}_t^A)$ and $i_t^B = \sigma_t^B(\mathbf{q}_t^B)$.

Roughly speaking, the reasons why this strategy works are as follows:

- By (5) q_t^A and q_t^B are never too far away from each other, so they are either in the same region or in two neighbour regions of the partition, and the strategy ensures that there is no collision.
- Under the event Ω of (4), the players almost play the best two arms except in the region $B_t' \cup C_t'$. If $\mathbf{p}$ is close to the axis $\mathcal{D}$, this is not suboptimal by a lot. If $\mathbf{p}$ is far away from $\mathcal{D}$, then $\mathbb{P}(\mathbf{p} \in B_t' \cup C_t')$ is small since Θ is randomized.

3.2. Regret analysis

We give here the proof of Theorem 3, with (3) holding in expectation over Θ (which is the only source of randomness in the players' strategy).

3.2.1. NO COLLISION PROPERTY

First observe that the coloring σ_t is such that there are no collisions for neighboring regions, i.e., if $U, V \in \{A_t, B_t, C_t, D_t\}$ are neighboring regions then $\sigma_t^A(U) \neq \sigma_t^B(V)$ and $\sigma_t^B(U) \neq \sigma_t^A(V)$. Next we note that two non-neighboring regions are well-separated.

Lemma 4 *In the partition (A_t, B_t, C_t, D_t), the distance between any two non-neighboring regions is at least w_t.*

Proof The pairs of non-neighboring regions are (A_t, D_t), (A_t, C_t) and (B_t, D_t). Any of these pairs has its two elements on different sides of the set $\left\{\theta = \Theta \text{ or } \theta = \frac{5\pi}{3}\right\}$. Moreover, simple geometric considerations show that A_t and D_t are both at distance w_t from that set. Thus all these distances are at least w_t. ∎

Finally recall that on Ω the observations of Alice and Bob are close to each other (see (5)), so we can conclude that Alice and Bob never collide when Ω holds true.

3.2.2. CONTROLLING THE REGRET FROM SUBOPTIMAL DECISIONS

We denote by $B(x,r)$ the ball of radius r around x for the Euclidean distance. Given that there are no collisions on Ω, we have:

$$\begin{aligned} R_T = \sum_{t=1}^{T} (p_{i_t^A} + p_{i_t^B} - \mathbf{p}^*) &= \sum_{t=1}^{T} (p_{\sigma_t^A(q_t^A)} + p_{\sigma_t^B(q_t^B)} - \mathbf{p}^*) \\ &\leq \sum_{t=1}^{T} \max_{\mathbf{q},\mathbf{q}' \in B(\mathbf{p},w_t/2)} (p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q}')} - \mathbf{p}^*) \\ &\leq 2 \sum_{t=1}^{T} \max_{\mathbf{q} \in B(\mathbf{p},w_t/2)} (p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q})} - \mathbf{p}^*), \end{aligned} \tag{6}$$

where the second line uses that under Ω we have $q_t^A, q_t^B \in B(\mathbf{p}, w_t/2)$, and the last line uses the bound

$$p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q}')} - \mathbf{p}^* \leq \left(p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q})} - \mathbf{p}^*\right) + \left(p_{\sigma_t^A(\mathbf{q}')} + p_{\sigma_t^B(\mathbf{q}')} - \mathbf{p}^*\right).$$

To control the last quantity of (6), let us first assume that $d(\mathbf{p}, \mathcal{P}) > 2w_t$. Then we know that for any $\mathbf{q} \in B(\mathbf{p}, w_t/2)$, one has $\mathbf{q} \notin B'_t \cup C'_t$. By construction, $q_{\sigma_t^A(\mathbf{q})} + q_{\sigma_t^B(\mathbf{q})} = \mathbf{q}^*$ for any $\mathbf{q} \notin B'_t \cup C'_t$. Moreover the map $\mathbf{q} \mapsto \mathbf{q}^*$ is 2-Lipschitz so we get that $p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q})} \leq w_t + q_{\sigma_t^A(\mathbf{q})} + q_{\sigma_t^B(\mathbf{q})} = w_t + \mathbf{q}^* \leq 2w_t + \mathbf{p}^*$. In other words, so far we have proved that on Ω we have:

$$R_T \leq 4 \sum_{t=1}^{T} w_t + 2 \sum_{t=1}^{T} \mathbb{1}_{d(\mathbf{p},\mathcal{P}) \leq 2w_t} \max_{\mathbf{q} \in B(\mathbf{p},w_t/2)} (p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q})} - \mathbf{p}^*).$$

Note that

$$p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q})} - \mathbf{p}^* \leq \max_{i \neq j} |p_i - p_j| \leq r_{\mathbf{p}}.$$

Thus we get with the two above displays:

$$\mathbb{E}_\Theta R_T \leq 4 \sum_{t=1}^{T} w_t + 2 \sum_{t=1}^{T} r_{\mathbf{p}} \mathbb{P}_\Theta(d(\mathbf{p}, \mathcal{P}) \leq 2w_t). \tag{7}$$

The proof is now concluded with the following lemma, which implies $\mathbb{E}_\Theta R_T \leq 10 \sum_{t=1}^{T} w_t \leq 320\sqrt{T \log T}$. The proof of this lemma is postponed to Appendix B.

Lemma 5 *For every t and $\mathbf{p}$, we have*

$$\mathbb{P}\left(d(\mathbf{p}, \mathcal{P}) \leq 2w_t\right) \leq 3 \frac{w_t}{r_{\mathbf{p}}}. \tag{8}$$

4. Bandit upper bound

We prove here Theorem 1. The extra difficulty introduced by the bandit setting compared to the full-information toy model is that, in addition to coordinating for exploitation (which is the key point of the toy model), the players also have to coordinate their *exploration* of the arms. Moreover,

there needs to be a *smooth* transition between exploration and exploitation, so that there are also no collisions if one player stops exploring before the other. To do so we introduce extra padding around the decision boundaries of the partition built in the previous section, and we give a carefully choreographed dynamic coloring of this new partition. An explicit algorithm is fully described below by combining the definition (9), the partition constructed in Section 4.1 (and represented on Figure 3) and the table on Figure 1.

We denote $w_t = 2^{15}\sqrt{\frac{\log(T)}{t}}$. For $1 \leq t \leq T$, $i \in \{1,2,3\}$ and $X \in \{A,B\}$, we denote by $n_t^X(i)$ the number of times from 1 to $t-1$ where player X has played arm i. We also write

$$q_t^X(i) = \frac{1}{n_t^X(i)} \sum_{\substack{s=1 \\ i_s^X = i}}^{t-1} \max\left(\ell_t(i), \mathbb{1}_{i_t^A = i_t^B}\right), \tag{9}$$

with the convention $q_t^X(i) = 0$ if $n_t^X(i) = 0$. Then $\mathbf{q}_t^X = \left(q_t^X(1), q_t^X(2), q_t^X(3)\right)$ is an estimate at time t, according to player X, of $\mathbf{p}$. Note that this estimator is biased due to the potential collisions. This issue will be handled below (Lemma 6).

We will prove the absence of collisions by induction on t, which means that we need to show that our estimators at time t are not too bad if there has been no collision before. For this reason, we define the following event:

$$\Omega = \Big\{\forall t \in [T], i \in \{1,2,3\}, X \in \{A,B\}, \text{ if there has been no collision at times } 1,\dots,t-1, \text{ then } \left|q_t^X(i) - p_i\right| < \frac{w_{4n_t^X(i)+5}}{32}\Big\}.$$

If there has been no collision before time t, we have $q_t^X(i) = \frac{1}{n_t^X(i)} \sum_{i=1, i_t^X = i}^{t-1} \ell_t(i)$. Note that Ω depends on the $n_t^X(i)$, and therefore on the strategies used by the players. However, for any strategy, if we fix an arm i and list the values $\ell_t(i)$ observed by a player $X \in \{A,B\}$, then these values are i.i.d. Bernoulli with parameter p_i. Therefore, the Hoeffding inequality and a union bound show that $\mathbb{P}(\Omega) \geq 1 - \frac{1}{T}$ for any deterministic strategy of A and B, and thereforee also for a random one. We will later prove the following result, which implies that the assumption of no collisions in Ω can be removed.

Lemma 6 *On the event Ω, our proposed bandit strategy satisfies $i_t^A \neq i_t^B$ for all $t \in [T]$.*

Like in the full-information toy model, in the remainder of this section we fix loss sequences such that Ω holds true, and all probabilities are with respect to the random interface defined by Θ (see below).

4.1. The bandit partition

We recall that $w_t = 2^{15}\sqrt{\frac{\log(T)}{t}}$. We denote by $\mathcal{P}$ the half-plane $\{\theta = \Theta\}$ and by $\mathcal{Q}_1$ (resp. $\mathcal{Q}_2$, $\mathcal{Q}_3$) the half-plane $\{\theta = \frac{\pi}{3}\}$ (resp. $\{\theta = \pi\}$, $\{\theta = \frac{5\pi}{3}\}$). We now define the following sets, that we will refer to as *regions*:

- $E_t = \left\{\mathbf{p} \middle| \frac{\pi}{3} \leq \theta_{\mathbf{p}} < \Theta \text{ and } d(\mathbf{p}, \mathcal{Q}_1) \geq \frac{w_t}{2} \text{ and } d(\mathbf{p}, \mathcal{P}) \geq \frac{3w_t}{2}\right\}$,

- $G_t = \left\{\mathbf{p} | \theta_{\mathbf{p}} \in \left[0, \frac{\pi}{3}\right) \cup \left[\frac{5\pi}{3}, 2\pi\right) \text{ and } d(\mathbf{p}, \mathcal{Q}_1) \geq \frac{w_t}{2} \text{ and } d(\mathbf{p}, \mathcal{Q}_3) \geq \frac{w_t}{2}\right\}$,
- $H_t = \left\{\mathbf{p} | d(\mathbf{p}, \mathcal{P} \cup \mathcal{Q}_3) < \frac{w_t}{2}\right\}$,
- $I_t = \left\{\mathbf{p} | \theta_{\mathbf{p}} \in \left[\pi, \frac{5\pi}{3}\right) \text{ and } d(\mathbf{p}, \mathcal{Q}_2) \geq \frac{w_t}{2} \text{ and } d(\mathbf{p}, \mathcal{Q}_3) \geq \frac{w_t}{2}\right\}$,
- $K_t = \left\{\mathbf{p} | \Theta \leq \theta_{\mathbf{p}} < \pi \text{ and } d(\mathbf{p}, \mathcal{Q}_2) \geq \frac{w_t}{2} \text{ and } d(\mathbf{p}, \mathcal{P}) \geq \frac{3w_t}{2}\right\}$,
- $F_t = \left\{\mathbf{p} | \theta_{\mathbf{p}} \in [0, \Theta) \cup \left[\frac{5\pi}{3}, 2\pi\right)\right\} \setminus (E_t \cup G_t \cup H_t)$,
- $J_t = \left\{\mathbf{p} | \theta_{\mathbf{p}} \in \left[\Theta, \frac{5\pi}{3}\right)\right\} \setminus (H_t \cup I_t \cup K_t)$.

As in the full information case, we have represented on Figure 3 the restriction of this partition to the plane $\left\{p_1 + p_2 + p_3 = \frac{3}{2}\right\}$. Here again, since that plane is orthogonal to the half-planes $\mathcal{P}$, $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$, the full partition is obtained by extending Figure 3 orthogonally to its plane.

4.2. Dynamic coloring

The strategy is now the following: for every $0 \leq t < \frac{T}{4}$, player A will decide according to the region of $\mathbf{q}^A_{4t+1}$ where he plays at times $4t+1$, $4t+2$, $4t+3$, $4t+4$, and similarly for player B according to the region of $\mathbf{q}^B_{4t+1}$. More precisely, player A will play according to the table below. The way to read this table is as follows: if $2/1$ is written at the intersection of the row "$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in E_{4t+1}$" and the column "$4t+2$", this means that if $\mathbf{q}^A_{4t+1} \in E_{4t+1}$, then player A plays arm 2 at time $4t+2$. If $\mathbf{q}^B_{4t+1} \in E_{4t+1}$, then player B plays arm 1 at time $4t+2$.

	$4t+1$	$4t+2$	$4t+3$	$4t+4$
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in E_{4t+1}$	2 / 1	2 / 1	1 / 2	1 / 2
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in F_{4t+1}$	2 / 1	3 / 1	1 / 2	1 / 3
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in G_{4t+1}$	3 / 1	3 / 1	1 / 3	1 / 3
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in H_{4t+1}$	3 / 1	3 / 2	1 / 3	2 / 3
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in I_{4t+1}$	3 / 2	3 / 2	2 / 3	2 / 3
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in J_{4t+1}$	3 / 2	1 / 2	2 / 3	2 / 1
$\mathbf{q}^A_{4t+1}/\mathbf{q}^B_{4t+1} \in K_{4t+1}$	1 / 2	1 / 2	2 / 1	2 / 1

Figure 1: The table describing the arms played by the players at time $4t+1, \ldots, 4t+4$ according to $\mathbf{q}^A_{4t+1}$ and $\mathbf{q}^B_{4t+1}$.

Although it might seem quite complicated, this table is actually a natural adaptation of the full information strategy, where we have "smoothened" the boundaries between regions. Let us first focus on the first two columns: the regions E, G, I and K then correspond to the regions A, B, C, D of the full information strategy. The difference here is that, if for example we are in the region where p_2 and p_3 are close but much larger than p_1, it is necessary to explore both arms 2 and 3 during a long time to find which is the best one. This is the role of region F, and regions H and J play a similar role.

Moreover, the last two columns are the same as the first two, where the roles of A and B have been exchanged. This is necessary to make sure that each of the players has information about all the arms. Of course, such a problem did not exist in the full information case.

4.3. Exploration phase and no collision property

The regions F_t, H_t and J_t can be considered as "exploration" regions, since they are regions where both players play the three arms. It is immediate from the definition of the regions that E_t, G_t, I_t and K_t are increasing in t, which means that $F_t \cup H_t \cup J_t$ is decreasing in t. Therefore, it is natural to expect that $\mathbf{q}_t^A$ will be in $F_t \cup H_t \cup J_t$ in the beginning ("exploration phase"), and in the complementary after some time ("exploitation phase"). We make this intuition precise in the proof of the next lemma.

Lemma 7 *Under Ω, for every $1 \leq t \leq T$, if there has been no collision before time t, then either $\mathbf{p}$, $\mathbf{q}_t^A$ and $\mathbf{q}_t^B$ are in the same region, or $\mathbf{q}_t^A$, $\mathbf{q}_t^B$ belong to the ball of radius $\frac{w_t}{4}$ around $\mathbf{p}$.*

Proof For $X \in \{A, B\}$, we denote by τ^X the first time t such that $q_t^X \notin F_t \cup H_t \cup J_t$, with the convention $\tau^X = +\infty$ if $q_t^X \in F_t \cup H_t \cup J_t$ for all $t \in [T]$. In particular, for any $s < \frac{\tau^X - 5}{4}$ we have $\mathbf{q}_{4s+1}^X \in F_t \cup H_t \cup J_t$, which means that each arm appears at least once among $i_{4s+1}^X, i_{4s+2}^X, i_{4s+3}^X, i_{4s+4}^X$. Therefore, we must have $n_t^X(i) \geq \frac{\min(t,\tau^X)-5}{4}$ for every arm i. Using the event Ω, this implies $\left|q_t^X(i) - p_i\right| < \frac{w_{\min(t,\tau^X)}}{32}$ for all i, and thus

$$d(\mathbf{q}_t^X, \mathbf{p}) < \frac{w_{\min(t,\tau^X)}}{16}\,. \tag{10}$$

In particular, since any point at distance $\leq \frac{w_t}{2}$ from $\mathcal{Q}_1 \cup \mathcal{Q}_2 \cup \mathcal{Q}_3 \cup \mathcal{P}$ is in $F_t \cup H_t \cup J_t$, we have $d(\mathbf{q}_{\tau^X}^X, \mathcal{Q}_1 \cup \mathcal{Q}_2 \cup \mathcal{Q}_3 \cup \mathcal{P}) \geq \frac{w_t}{2}$. Hence $\mathbf{p}$ must be at distance at least $\frac{7}{16} w_{\tau^X}$ from $\mathcal{Q}_1 \cup \mathcal{Q}_2 \cup \mathcal{Q}_3 \cup \mathcal{P}$ (it is also immediate if $\tau_X = +\infty$). Next observe that for $t \geq 16\tau^X$ one has $\frac{7}{16} w_{\tau^X} \geq \frac{3}{2} w_t + \frac{1}{16} w_{\tau^X}$. Since $F_t \cup H_t \cup J_t$ lie entirely at distance at most $\frac{3}{2} w_t$ from $\mathcal{Q}_1 \cup \mathcal{Q}_2 \cup \mathcal{Q}_3 \cup \mathcal{P}$, we deduce that $\mathbf{p}$ is at distance $\frac{w_{\tau^X}}{16}$ from $F_t \cup H_t \cup J_t$, so the ball of center $\mathbf{p}$ and radius $\frac{w_{\tau^X}}{16}$ is contained in the region of $\mathbf{p}$ (which may be E_t, G_t, I_t or K_t). By (10), this implies that $\mathbf{q}_t^X$ is in the same region as $\mathbf{p}$.

On the other hand, for $t \leq 16\tau^X$, (10) gives

$$d(\mathbf{q}_t^X, \mathbf{p}) < \frac{w_{t/16}}{16} = \frac{1}{4} w_t\,,$$

which concludes the proof. ■

We now prove the no collision property. Note that this will allow us to use the event Ω without having to assume that there has been no collision so far.

Proof [Proof of Lemma 6.] As explained earlier, we assume Ω and prove by induction on t the absence of collisions until t. Assume there was no collision at times $1, \ldots, t-1$. By Lemma 7, we know that for every t, either $\mathbf{q}_t^A$ and $\mathbf{q}_t^B$ lie in the same region, or $d(\mathbf{q}_t^A, \mathbf{q}_t^B) < \frac{w_t}{2}$. In the first case, there is no collision.

In the second case, we call two regions *colliding* if it is possible, when A plays according to the first one and B according to the second, that $i_t^A = i_t^B$. By looking at the table of Figure 1, we can list the pairs of colliding pairs: (E_t, H_t), (E_t, I_t), (E_t, J_t), (E_t, K_t), (F_t, I_t), (F_t, J_t), (F_t, K_t), (G_t, J_t), (G_t, K_t) and (H_t, K_t). By the definitions of the regions, the distance between any two colliding regions is always at least w_t (this is very similar to Lemma 4 in the full information case, so we omit the detailed proof). Therefore, no collision can happen if $d(\mathbf{q}_t^A, \mathbf{q}_t^B) < \frac{w_t}{2}$, which proves the lemma. ■

Given Lemma 7 and the absence of collisions, the proof of Theorem 1 is now very similar to the full information case, and is done in Appendix D.

References

P. Alatur, K. Y. Levy, and A. Krause. Multi-player bandits: The adversarial case. *arXiv preprint arXiv:1902.08036*, 2019.

A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

O. Avner and S. Mannor. Concurrent bandits and cognitive radio networks. In *ECML/PKDD*, 2014.

R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot. 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, 2017.

E. Boursier and V. Perchet. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. *arXiv preprint arXiv:1809.08151*, 2018.

S. Bubeck and T. Budzinski. Coordination without communication: optimal regret in two players multi-armed bandits. *arXiv:2002.07596*, 2020.

S. Bubeck, Y. Li, Y. Peres, and M. Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. abs/1904.12233, 2019.

L. Lai, H. Jiang, and H. V. Poor. 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, 2008.

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

G. Lugosi and A. Mehrabian. Multiplayer bandits without observing collision information. *arXiv preprint arXiv:1808.08416*, 2018.

J. Rosenski, O. Shamir, and L. Szlak. Multi-player bandits - a musical chairs approach. In *ICML*, 2016.

Appendix A. Figures

Here, we provide a more visual description of the two partitions of the cube that we used, by drawing their restriction to the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$. Figure 2 is the partition used for the toy model in Section 3, and Figure 3 is the one used for bandits in Section 4.

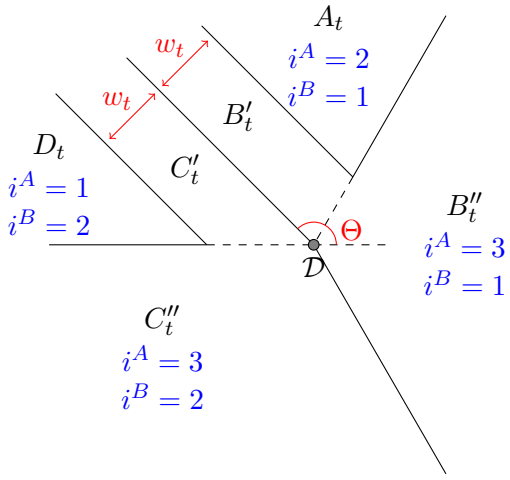

Figure 2: The restriction of our partition of the cube to the plane $\{m_{\mathbf{p}} = \frac{1}{2}\}$. We recall that $B_t = B'_t \cup B''_t$ and $C_t = C'_t \cup C''_t$. The full partition is obtained from here by extending each region orthogonally to that plane. In blue, the arms played by each player in each region.

Appendix B. Proof of Lemma 5

Proof We first note that, since the half-plane $\mathcal{P}$ is orthogonal to the plane $\{m_{\mathbf{p}} = \frac{1}{2}\}$ of Figure 2, both sides of (8) are unchanged if we replace $\mathbf{p}$ by its projection on $\{m_{\mathbf{p}} = \frac{1}{2}\}$, so we can assume $\mathbf{p} \in \mathcal{P}$. Moreover, the distance between $\mathbf{p}$ and $\mathcal{P}$ is equal to the distance in $\{m_{\mathbf{p}} = \frac{1}{2}\}$ between $\mathbf{p}$ and the half-line $\mathcal{P} \cap \{m_{\mathbf{p}} = \frac{1}{2}\}$.

We also note the result is obviously true if $r_{\mathbf{p}} > 2w_t$ (the right-hand side of (8) is larger than 1), so we can assume $r_{\mathbf{p}} \leq 2w_t$. Then we have

$$d(\mathbf{p}, \mathcal{P}) = r_{\mathbf{p}} \sin \alpha \,,$$

where α is the angle between the line from the point $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$ to $\mathbf{p}$ and the half-line $\{\theta = \Theta\}$, in the plane of Figure 2. We have $\alpha = |\Theta - \theta_{\mathbf{p}}|$, so the event of (8) is equivalent to

$$\theta_{\mathbf{p}} - \arcsin \frac{2w_t}{r_{\mathbf{p}}} \leq \Theta \leq \theta_{\mathbf{p}} + \arcsin \frac{2w_t}{r_{\mathbf{p}}} \,.$$

This has probability $\frac{3}{2\pi} \times 2 \arcsin \frac{2w_t}{r_{\mathbf{p}}} \leq 3\frac{w_t}{r_{\mathbf{p}}}$, which concludes the proof of the lemma. ■

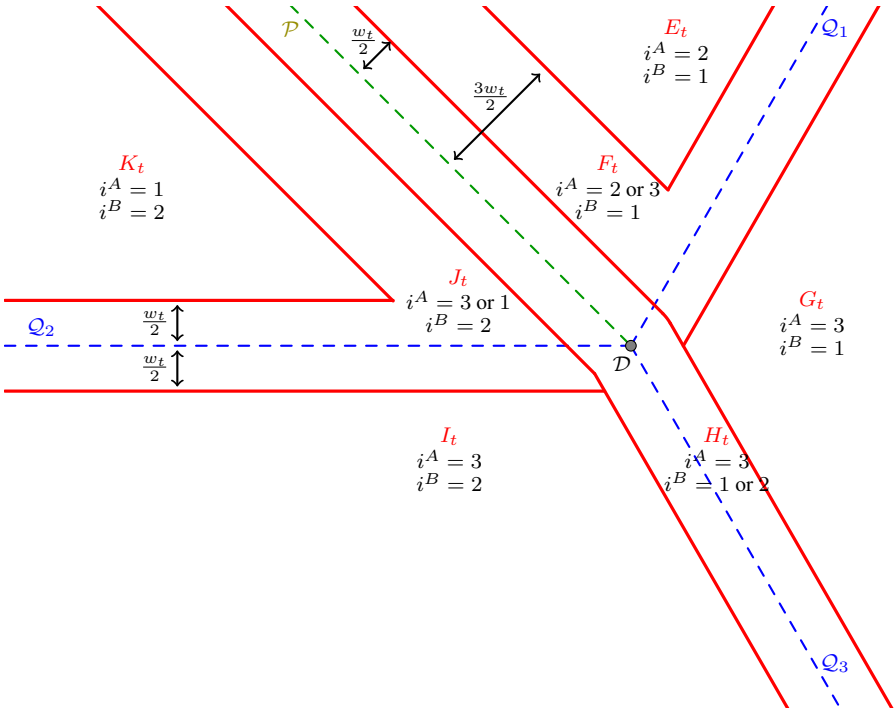

Figure 3: The intersection of our bandit partition with the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$. Below the names of the regions are the arms played by the players in the first two columns of the table, i.e. for $t \equiv 1$ or 2 modulo 4 (for $t \equiv 3$ or 0 modulo 4, the roles of players A and B are exchanged).

Appendix C. Derandomization in the full information case via a dynamic interface

In the proof of Theorem 3, the only place where we used the randomness in Θ is Lemma 5. To derandomize the algorithm, we can replace the random angle Θ by a deterministic, time-dependent angle $(\theta_t)_{t\in[T]}$, with $\frac{\pi}{3} \leq \theta_t \leq \pi$. In this setting, all the proof is the same until (7), which becomes

$$R_T \leq 4\sum_{t=1}^{T} w_t + 2r_{\mathbf{p}} \sum_{t=1}^{T} \mathbb{1}_{d(\mathbf{p},\mathcal{P}_t)\leq 2w_t},$$

where $\mathcal{P}_t = \{[m, r, \theta] | \theta = \theta_t\}$. For the same reason as in the proof of Lemma 5, if $d(\mathbf{p}, \mathcal{P}_t) \leq 2w_t$, then $|\theta_{\mathbf{p}} - \theta_t| \leq \arcsin \frac{2w_t}{r_{\mathbf{p}}} \leq \pi \frac{w_t}{r_{\mathbf{p}}}$. Therefore, to obtain the analog of Lemma 5, it is enough to find (θ_t) such that, for any r and θ, the number of t such that $|\theta - \theta_t| \leq \pi \frac{w_t}{r}$ is at most $\frac{3}{r}\sum_{t=1}^{T} w_t$.

One way to do so is the following: for every t, let k be such that $2^k \leq t < 2^{k+1}$, and take

$$\theta_t = \frac{\pi}{3} + \frac{2\pi}{3}\frac{t - 2^k}{2^k}.$$

In that case, for every fixed k, r and θ, using that w_t is decreasing in t, we have

$$\sum_{t=2^k}^{2^{k+1}-1} \mathbb{1}_{|\theta_t-\theta|\leq\frac{\pi w_t}{r}} \leq \sum_{t=2^k}^{2^{k+1}-1} \mathbb{1}_{|\theta_t-\theta|\leq\frac{\pi w_{2^k}}{r}} \leq 1+\frac{\pi w_{2^k}/r}{2\pi/(3\times 2^k)} = 1+\frac{3}{2}\times 2^k\frac{w_{2^k}}{r} \leq 3\sum_{t=2^k}^{2^{k+1}-1}\frac{w_t}{r},$$

and summing over k yields the result.

Appendix D. End of the proof of Theorem 1

For every t, we write $\underline{t} = 4\lfloor\frac{t-1}{4}\rfloor+1$, so that i_t^A is chosen according to the region of $\mathbf{q}_{\underline{t}}^A$. We denote by $\sigma_t = \left(\sigma_t^A, \sigma_t^B\right)$ the map prescribed by the table of Figure 1, so that $i_t^X = \sigma_t^X(\mathbf{q}_{\underline{t}}^X)$. Using the fact that we have no collisions, we have

$$R_T = \sum_{t=1}^{T}\left(p_{\sigma_t^A(\mathbf{q}_{\underline{t}}^A)} + p_{\sigma_t^B(\mathbf{q}_{\underline{t}}^B)} - \mathbf{p}^*\right).$$

Just like in the full information case (Section 3.2) we decompose the sum into two terms, based on whether $d(\mathbf{p},\mathcal{P}) > 2w_{\underline{t}}$ or not. The case where $d(\mathbf{p},\mathcal{P}) \leq 2w_{\underline{t}}$ is dealt exactly as in the full information case, and gives a term $6\sum_{t=1}^T w_{\underline{t}}$ in expectation over Θ. Now for the other term, we assume that $d(\mathbf{p},\mathcal{P}) > 2w_{\underline{t}}$ and we write, thanks to the dichotomy given by Lemma 7,

$$\begin{aligned} p_{\sigma_t^A(\mathbf{q}_{\underline{t}}^A)} + p_{\sigma_t^B(\mathbf{q}_{\underline{t}}^B)} - \mathbf{p}^* &\leq p_{\sigma_t^A(\mathbf{p})} + p_{\sigma_t^B(\mathbf{p})} - \mathbf{p}^* + 2\max_{\mathbf{q}\in B(\mathbf{p},w_{\underline{t}}/4)}\left(p_{\sigma_t^A(\mathbf{q})} + p_{\sigma_t^B(\mathbf{q})} - \mathbf{p}^*\right) \\ &\leq 3\max_{\mathbf{q}\in B(\mathbf{p},w_{\underline{t}}/4)}\left(q_{\sigma_t^A(\mathbf{q})} + q_{\sigma_t^B(\mathbf{q})} - \mathbf{q}^*\right) + 3w_{\underline{t}}\,, \end{aligned}$$

where the second inequality uses that $\mathbf{q}\mapsto\mathbf{q}^*$ is 2-Lipschitz. Finally it only remains to observe that the construction of the bandit partition is such that for any $\mathbf{q}$ with $d(\mathbf{q},\mathcal{P}) \geq \frac{3w_t}{2}$ one has

$$q_{\sigma_t^A(\mathbf{q})} + q_{\sigma_t^B(\mathbf{q})} - \mathbf{q}^* \leq w_{\underline{t}}\,.$$

Thus we have proved that, $R_T\mathbb{1}_{d(\mathbf{p},\mathcal{P})>2w_{\underline{t}}} \leq 6\sum_{t=1}^T w_{\underline{t}}$, and $\mathbb{E}_\Theta[R_T\mathbb{1}_{d(\mathbf{p},\mathcal{P}))\leq 2w_{\underline{t}}}] \leq 6\sum_{t=1}^T w_{\underline{t}}$. The expected regret is therefore bounded by $12\sum_{t=1}^T w_{\underline{t}} = O(\sqrt{T\log T})$, which concludes the proof of Theorem 1.

Appendix E. Toy model lower bound

We prove here Theorem 2. The goal is essentially to exploit the topological obstruction we alluded to in Section 3.1.2. This topological obstruction is basically Lemma 13.

E.1. The hard instance

We first describe the law of (p_1,p_2,p_3). Let $\varepsilon > 0$ be small (it is actually enough to have $\varepsilon < 1/4$). Let I be the following union of intervals:

$$\begin{aligned} I = &\left[\frac{\pi}{3} - 2T^{-1/2+2\varepsilon}, \frac{\pi}{3} + 2T^{-1/2+2\varepsilon}\right] \cup \left[\frac{3\pi}{3} - 2T^{-1/2+2\varepsilon}, \frac{3\pi}{3} + 2T^{-1/2+2\varepsilon}\right] \\ &\cup \left[\frac{5\pi}{3} - 2T^{-1/2+2\varepsilon}, \frac{5\pi}{3} + 2T^{-1/2+2\varepsilon}\right], \end{aligned}$$

with total measure $12T^{-1/2+2\varepsilon}$. We assume $T^{-1/2+2\varepsilon} < \frac{\pi}{6}$ so that the definition makes sense. Let Θ be a random variable on $[0, 2\pi]$ with distribution

$$\frac{1}{4\pi}\mathrm{d}\theta + \frac{\mathbb{1}_{\theta\in I}}{24T^{-1/2+2\varepsilon}}\mathrm{d}\theta. \tag{11}$$

In other words Θ is picked uniformly in $[0, 2\pi]$ with probability $\frac{1}{2}$ and uniformly in I with probability $\frac{1}{2}$.

Finally using the cylindrical coordinates $\mathbf{p} = [m_{\mathbf{p}}, r_{\mathbf{p}}, \theta_{\mathbf{p}}]$ from Section 3.1.1 we set $m_{\mathbf{p}} = 1/2$, $r_{\mathbf{p}} = \sqrt{\frac{3}{2}}T^{-\varepsilon}$, and $\theta_{\mathbf{p}} = \Theta$. We also denote by $(p_1(\Theta), p_2(\Theta), p_3(\Theta))$ the Cartesian coordinates of $\mathbf{p}$, and write $p^*(\Theta)$ for the sum of the two smallest coordinates.

In particular (p_1, p_2, p_3) is picked on a circle. Moreover, the "reinforcement" near $\frac{\pi}{3}$, π and $\frac{5\pi}{3}$ of the law of Θ implies that the law of (p_1, p_2, p_3) is reinforced at the places where two p_i are almost equal, and much larger than the third.

E.2. Proof skeleton

From now on, we assume that A and B follow a fixed, deterministic strategy. We concentrate on the quantity:

$$r_t(\theta) = \mathbb{E}\left[2 \cdot \mathbb{1}_{i_t^A = i_t^B} + \mathbb{1}_{i_t^A \neq i_t^B}(p_{i_t^A} + p_{i_t^B}) - \mathbf{p}^*\Big)\big|\Theta = \theta\right].$$

It is easy to see (and the standard route for bandit lower bounds) that it is sufficient to prove that, for every $1 \leq t \leq T$, we have

$$\mathbb{E}\left[r_t(\Theta)\right] \geq c\sqrt{\frac{\log T}{T}}. \tag{12}$$

Therefore, we fix such a t until the end of the proof. Key quantities of interest will be the following functions, defined for $i \in \{1, 2, 3\}$ and $X \in \{A, B\}$:

$$f_i^X(\theta) = \mathbb{P}\left(i_t^X = i|\Theta = \theta\right).$$

Even if this depends on t, since t is fixed until the end of the proof, we drop the t in the notation. Since the loss vectors observed by A and B are independent conditionally on Θ, we can write

$$r_t(\theta) = \sum_{i=1}^{3} f_i^A(\theta) f_i^B(\theta)(2 - p^*(\theta)) + \sum_{i\neq j} f_i^A(\theta) f_j^B(\theta)(p_i(\theta) + p_j(\theta) - p^*(\theta)) \geq 0. \tag{13}$$

The proof will now proceed by analyzing properties of the functions f_i^X, in particular the various constraints they must satisfy for the players to hope for a small regret.

E.3. Constraints on the functions f_i^X

In our proof, the fact that the players cannot have a very precise estimate of Θ will be encoded by the fact that the functions f_i^A, f_i^B are smooth enough, so that the players cannot change drastically their choices when θ varies a little. Therefore, the first step is to prove an estimate on the regularity of the functions f_i^A, f_i^B.

Lemma 8 *The functions f_i^A and f_i^B are analytic. Moreover, let $\delta > 0$. Then there is a constant $c > 0$ (depending on δ but not on t or T) such that, for every θ, θ', we have*

$$f_i^A(\theta') \geq \left(f_i^A(\theta) - \delta\right) \exp\left(-c - cT^{1-2\varepsilon}|\theta' - \theta|^2\right),$$

and the same is true for f_i^B.

Proof Both functions are polynomials in (p_1, p_2, p_3), so they are analytic in δ.

For the second point, we start by defining a "truncation" of the functions f_i^A. If E is an event, we write

$$f_i^A(\theta, E) = \mathbb{P}\left(i_t^A = i \text{ and } E \text{ occurs} | \Theta = \theta\right).$$

We fix a constant C, and denote by $E_C(\theta)$ the event that $\left|\sum_{s=1}^{t-1} \ell_s^A(i) - (t-1)p_i(\theta)\right| \leq C\sqrt{T}$ for every $j \in \{1,2,3\}$. By the central limit theorem, if C is chosen large enough (independently of θ, t and T), we have $\mathbb{P}\left(E_C(\theta)\right) \geq 1 - \delta$, so

$$f_i^A\left(\theta, E_C(\theta)\right) \geq f_i^A(\theta) - \delta.$$

On the other hand, we obviously have $f_i^A(\theta') \geq f_i^A\left(\theta', E_C(\theta)\right)$, so it is enough to prove

$$f_i^A\left(\theta', E_C(\theta)\right) \geq f_i^A\left(\theta, E_C(\theta)\right) \exp\left(-c - cT^{1-2\varepsilon}|\theta' - \theta|^2\right). \tag{14}$$

For this, let $\ell = (\ell_s(i))_{1 \leq s \leq t-1, 1 \leq i \leq 3} \in \left(\{0,1\}^3\right)^{t-1}$ be a possible value of the loss vectors observed by A until time $t-1$. For $j \in \{1,2,3\}$, we write $S(j) = \sum_{s=1}^{t-1} \ell_s(j)$. Then we have

$$\log \frac{\mathbb{P}\left(A \text{ observes } \ell | \Theta = \theta'\right)}{\mathbb{P}\left(A \text{ observes } \ell | \Theta = \theta\right)} = \sum_{j=1}^{3} \left(S(j) \log \frac{p_j(\theta')}{p_j(\theta)} + (t - 1 - S(j)) \log \frac{1 - p_j(\theta')}{1 - p_j(\theta)} \right).$$

The ratio $\frac{p_j(\theta')}{p_j(\theta)}$ is going to 1 as $T \to +\infty$, uniformly in θ, so we can use the inequality $\log(1+x) \geq x - x^2$ to bound the above quantity from below by

$$\sum_{j=1}^{3} \left(p_j(\theta') - p_j(\theta)\right) \left(\frac{S(j)}{p_j(\theta)} - \frac{t - 1 - S(j)}{1 - p_j(\theta)} \right) - \sum_{j=1}^{3} |p_j(\theta') - p_j(\theta)|^2 \left(\frac{S(j)}{p_j(\theta)^2} + \frac{t - 1 - S(j)}{(1 - p_j(\theta))^2} \right). \tag{15}$$

The second term is bounded from below by

$$-\sum_{j=1}^{3} \left|p_j(\theta') - p_j(\theta)\right|^2 \times \frac{2t}{1/16} \geq -96T^{1-2\varepsilon}|\theta' - \theta|^2,$$

by using $\frac{1}{4} \leq p_j(\theta) \leq \frac{3}{4}$, and then $\left|\frac{\mathrm{d}p_j(\theta)}{\mathrm{d}\theta}\right| \leq T^{-\varepsilon}$ and $t \leq T$.

On the other hand, since we work on the event $E_C(\theta)$, we have $|S(j) - (t-1)p_j(\theta)| \leq C\sqrt{T}$, so both $\frac{S(j)}{p_j(\theta)}$ and $\frac{t-1-S(j)}{1-p_j(\theta)}$ are close to $t-1$. More precisely, we can bound the absolute value of the first sum of (15) by

$$\sum_{j=1}^{3} \left|p_j(\theta') - p_j(\theta)\right| \times 2\frac{C\sqrt{T}}{1/4} \leq 24CT^{1/2-\varepsilon}|\theta' - \theta| \leq 12C\left(1 + T^{1-2\varepsilon}|\theta' - \theta|^2\right).$$

By combining our estimates on (15), we obtain, for every ℓ compatible with $E_C(\theta)$:

$$\mathbb{P}\left(A \text{ observes } \ell | \Theta = \theta'\right) \geq \mathbb{P}\left(A \text{ observes } \ell | \Theta = \theta\right) \exp\left(-c - cT^{1-2\varepsilon}|\theta' - \theta|^2\right),$$

with $c = 12C + 96$. This proves (14) and the lemma. ■

The next lemma expresses the risk of collision: if $f_i^A(\theta)$ and $f_i^B(\theta')$ are both large for θ' close to θ, then there is a risk that both players pull the arm i and a large loss occurs. In all the rest of the paper, we will write $x \succeq y$ if x is larger than y times an absolute constant, which does not depend on t or T or θ, but which may vary from line to line.

Lemma 9 *There is an absolute constant η such that the following holds. Assume that there is an arm i and θ, θ' with $|\theta' - \theta| \leq \eta T^{\varepsilon}\sqrt{\frac{\log T}{T}}$, such that*

$$f_i^A(\theta) \geq \frac{1}{10} \text{ and } f_i^B(\theta') \geq \frac{1}{10}.$$

Then $r_t(\theta) \succeq T^{-\varepsilon/2}$ and $\mathbb{E}\left[r_t(\Theta)\right] \succeq T^{-\frac{1}{2}+\frac{\varepsilon}{2}}$.

Proof Since every term in (13) is nonnegative, if T is large enough so that all the p_i are at most $\frac{3}{4}$, we can write

$$\begin{aligned} r_t(\theta) &\geq f_i^A(\theta) f_i^B(\theta)\left(2 - p^*(\theta)\right) \\ &\geq \frac{1}{2} f_i^A(\theta)\left(f_i^B(\theta') - \frac{1}{20}\right) \exp\left(-c - c|\theta' - \theta|^2 T^{1-2\varepsilon}\right) \\ &\succeq T^{-c\eta^2} \\ &\succeq T^{-\varepsilon/2}, \end{aligned}$$

provided η was chosen small enough compared to ε. The second inequality uses Lemma 8 with $\delta = \frac{1}{20}$. For the second point of the lemma, assume without loss of generality $\theta < \theta'$. For every θ'' in the interval

$$\left[\theta - T^{\varepsilon - 1/2}, \theta' + T^{\varepsilon - 1/2}\right], \tag{16}$$

we have $|\theta'' - \theta|, |\theta'' - \theta'| \leq 2\eta T^{\varepsilon}\sqrt{\frac{\log T}{T}}$ (provided T is large enough), so Lemma 8 gives

$$f_i^A(\theta'') \succeq T^{-4\eta^2 c} \succeq T^{-\varepsilon/4} \text{ and } f_i^B(\theta'') \succeq T^{-4\eta^2 c} \succeq T^{-\varepsilon/4}$$

provided η is small enough. Hence $r_t(\theta'') \succeq T^{-\varepsilon/2}$. Moreover, we know from (13) that $r_t(\Theta) \geq 0$, so

$$\mathbb{E}\left[r_t(\Theta)\right] \succeq T^{-\varepsilon/2}\mathbb{P}\left(\theta - T^{\varepsilon-1/2} \leq \Theta \leq \theta' + T^{\varepsilon-1/2}\right) \geq T^{-\varepsilon/2} \times \frac{2T^{\varepsilon-1/2}}{4\pi} \succeq T^{-1/2+\varepsilon/2},$$

where in the end we used the law of Θ (11). ■

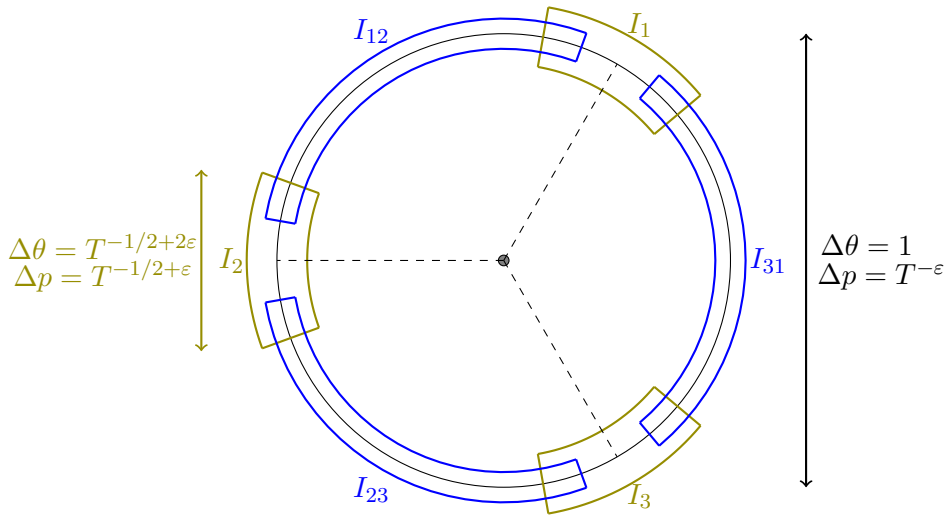

Figure 4: The sets I_i and $I_{i_1 i_2}$.

Remark 10 *This is the only place in the proof where it was necessary that the fluctuations of* (p_1, p_2, p_3) *are of order* $T^{-\varepsilon}$ *instead of* 1. *If the fluctuations were constant, the interval of* (16) *would have size* $T^{-1/2}$ *instead of* $T^{\varepsilon-1/2}$.

We now define several regions on the unit circle. Our goal will then be to show in a quantitative way that the players must make certain choices on each of these regions (Lemmas 11 and 12). More precisely, we write:

- $I_1 = \left[\frac{\pi}{3} - 2T^{-1/2+2\varepsilon}, \frac{\pi}{3} + 2T^{-1/2+2\varepsilon}\right]$,
- $I_2 = \left[\pi - 2T^{-1/2+2\varepsilon}, \pi + 2T^{-1/2+2\varepsilon}\right]$,
- $I_3 = \left[\frac{5\pi}{3} - 2T^{-1/2+2\varepsilon}, \frac{5\pi}{3} + 2T^{-1/2+2\varepsilon}\right]$,
- $I_{12} = \left[\frac{\pi}{3} + T^{-1/2+2\varepsilon}, \pi - T^{-1/2+2\varepsilon}\right]$,
- $I_{23} = \left[\pi + T^{-1/2+2\varepsilon}, \frac{5\pi}{3} - T^{-1/2+2\varepsilon}\right]$,
- $I_{31} = \left[\frac{5\pi}{3} + T^{-1/2+2\varepsilon}, 2\pi\right] \cup \left[0, \frac{\pi}{3} - T^{-1/2+2\varepsilon}\right]$.

See also Figure 4 to see what these intervals look like. Basically, I_i is the region where the arm i is way better than the two others but the two others are close to each other. $I_{i_1 i_2}$ is the region where the arms i_1 and i_2 are significantly better than the last one. Note also that $I_1 \cup I_2 \cup I_3$ is precisely the set I of (11) where the distribution of Θ is "reinforced".

The next lemma means that in the interval I_i, it is absolutely necessary that one of the players picks the arm i.

Lemma 11 *Let* i_1, i_2, i_3 *be any permutation of the indices* $1, 2, 3$. *Assume that there is* $\theta \in I_{i_1}$ *such that*

$$f^A_{i_2}(\theta) f^B_{i_3}(\theta) \geq \frac{1}{100}.$$

Then $r_t(\theta) \succeq T^{-\varepsilon}$ *and* $\mathbb{E}[r_t(\Theta)] \succeq T^{-2\varepsilon}$.

Proof Without loss of generality, assume $i_1 = 1, i_2 = 2, i_3 = 3$. Since each term in (13) is nonnegative, we have

$$r_t(\theta) \geq f_2^A(\theta) f_3^B(\theta) \left(p_2(\theta) + p_3(\theta) - p^*(\theta)\right) \geq \frac{1}{100} \left(\max(p_2(\theta), p_3(\theta)) - p_1(\theta)\right) \succeq T^{-\varepsilon},$$

by the definition of I_1.

Similarly, for every θ' with $|\theta' - \theta| \leq T^{\varepsilon - 1/2}$, by Lemma 8, we have $r_t(\theta') \succeq r_t(\theta) \succeq T^{-\varepsilon}$. Therefore:

$$\mathbb{E}[r_t(\Theta)] \succeq T^{-\varepsilon} \mathbb{P}\left(|\Theta - \theta| \leq T^{\varepsilon - 1/2}\right) \succeq T^{-\varepsilon} \frac{T^{\varepsilon - 1/2}}{T^{2\varepsilon - 1/2}},$$

where the last inequality follows from the law of Θ (11), and more precisely the fact that it is "reinforced" on $I_1 \cup I_2 \cup I_3$. ■

After Lemmas 9 and 11, we now state a third constraint on the strategy of the players. This one states that a suboptimal choice cannot be made on a too large region, and in particular not on the whole region $I_1 \cap I_{12}$.

Lemma 12 *Let* i_1, i_2, i_3 *be any permutation of the indices* $1, 2, 3$.

- *Let* $\theta \in I_{i_1} \cap I_{i_1 i_2}$. *If*

$$f_{i_1}^A(\theta) f_{i_3}^B(\theta) \geq \frac{1}{100} \text{ or } f_{i_3}^A(\theta) f_{i_1}^B(\theta) \geq \frac{1}{100},$$

  then $r_t(\theta) \succeq T^{-1/2+\varepsilon}$.

- *If*

$$f_{i_1}^A(\theta) f_{i_3}^B(\theta) + f_{i_3}^A(\theta) f_{i_1}^B(\theta) \geq \frac{2}{100}$$

  for all $\theta \in I_{i_1} \cap I_{i_1 i_2}$, *then* $\mathbb{E}[r_t(\Theta)] \succeq T^{-1/2+\varepsilon}$.

Proof Without loss of generality, assume $i_1 = 1, i_2 = 2, i_3 = 3$, so that $p_1(\theta) < p_2(\theta) < p_3(\theta)$ on $I_1 \cap I_{12}$. For the first point, by (13), we have

$$r_t(\theta) \geq f_1^A(\theta) f_3^B(\theta) \left(p_3(\theta) - p_2(\theta)\right) \succeq T^{-1/2+\varepsilon},$$

where the last inequality follows from the definition of I_{12}.

This implies that under the assumptions of the second point, we have $r_t(\theta) \succeq T^{-1/2+\varepsilon}$ for all $\theta \in I_1 \cap I_{12}$, so

$$\mathbb{E}[r_t(\Theta)] \succeq T^{-1/2+\varepsilon} \mathbb{P}\left(\Theta \in I_{12} \cap I_1\right) \succeq T^{-1/2+\varepsilon}$$

by (11). ■

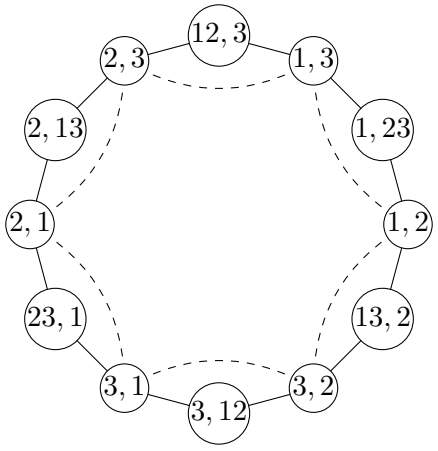

Figure 5: The collision graph: the vertices are the possible values of $E(\theta)$. The pairs of vertices linked by a full edge correspond to regions that may be neighbour. Note that $(1,2)$ and $(1,3)$ are not linked by a full edge, because at the boundary we would have $f_2^B(\theta) = f_3^B(\theta) = \frac{1}{10}$ but $f_1^B(\theta) < \frac{1}{10}$, which is not possible since $f_1^B + f_2^B + f_3^B = 1$. The vertices not linked by any edge correspond to regions which must be separated by at least $\eta T^{\varepsilon}\sqrt{\frac{\log T}{T}}$ to avoid the risk of a collision.

E.4. Proof of Theorem 2

We recall that $1 \leq t \leq T$ is fixed. As noted earlier, it is sufficient to check $\mathbb{E}\left[r_t(\Theta)\right] \succeq \sqrt{\frac{\log T}{T}}$. For each θ, let $a(\theta)$ (resp. $b(\theta)$) be the set of arms i such that $f_i^A(\theta)$ (resp. $f_i^B(\theta)$) is at least $\frac{1}{10}$.

It follows from Lemma 9 that if $\mathbb{E}\left[r_t(\Theta)\right] \preceq \sqrt{\frac{\log T}{T}}$, then $a(\theta) \cap b(\theta) = \emptyset$ and clearly $a(\theta)$ and $b(\theta)$ are nonempty, so only the following situations can occur:

- $a(\theta)$ and $b(\theta)$ are disjoint singletons;
- $a(\theta)$ is a singleton and $b(\theta)$ its complement;
- $b(\theta)$ is a singleton and $a(\theta)$ its complement.

We denote by $E(\theta)$ the pair $(a(\theta), b(\theta))$. We will write $E(\theta)$ in a compact form. For example, if $a(\theta) = \{1,3\}$ and $b(\theta) = \{2\}$, we will write $E(\theta) = (13,2)$. The 12 possible values of $E(\theta)$ split the circle on which θ lives into regions. Since the functions f_i^A and f_i^B are analytic by Lemma 8, these regions are finite unions of intervals. Moreover, Lemma 9 shows that $a(\theta) \cap b(\theta') = \emptyset$ if $|\theta' - \theta| \leq \eta T^{\varepsilon}\sqrt{\frac{\log T}{T}}$, so certain regions may not touch each other. More precisely, the graph of possible adjacence of these regions is summed up on Figure 5.

Moreover, if $\mathbb{E}\left[r_t(\Theta)\right] \preceq \sqrt{\frac{\log T}{T}}$, then Lemmas 11 and 12 imply respectively the following.

1. For $i \in \{1,2,3\}$ and $\theta \in I_i$, we have $i \in a(\theta) \cup b(\theta)$;

2. for any permutation i_1, i_2, i_3 of the indices $1, 2, 3$, there is $\theta_{i_1 i_2} \in I_{i_1} \cap I_{i_1 i_2}$ such that $\{i_1, i_3\}$ is not included in $a(\theta_{i_1 i_2}) \cup b(\theta_{i_1 i_2})$. Since i_1 is always in the union by the previous item, it means that $E(\theta_{i_1 i_2})$ has to be (i_1, i_2) or (i_2, i_1).

Lemma 13 *There is a permutation i_1, i_2, i_3 of the indices $1, 2, 3$ such that:*

$$E(\theta_{i_1 i_2}) = (i_1, i_2) \text{ but } E(\theta_{i_2 i_1}) = (i_2, i_1), \text{ or } E(\theta_{i_1 i_2}) = (i_2, i_1) \text{ but } E(\theta_{i_2 i_1}) = (i_1, i_2).$$

Proof Suppose this is not the case, and assume without loss of generality that $E(\theta_{12}) = (1, 2)$. By Item 1 above, we know that for every $\theta \in I_1$, the arm 1 must be in exactly one of the two sets $a(\theta)$ and $b(\theta)$. Since I_1 is connected, it is always in the same set, so $1 \in a(\theta)$. In particular, since $\theta_{13} \in I_1$, we have $1 \in a(\theta_{13})$, so $E(\theta_{13}) = (1, 3)$.

But by our assumption that we are on a counter-example to Lemma 13, it follows that $E(\theta_{31}) = (1, 3)$. By the same argument using Item 1, this implies $E(\theta_{32}) = (2, 3)$, so by our assumption $E(\theta_{23}) = (2, 3)$. Hence $E(\theta_{21}) = (2, 1)$ by Item 1 and finally $E(\theta_{12}) = (2, 1)$ by our assumption. This is a contradiction. ■

We are now in position to conclude the proof of Theorem 2. We consider a counter-example to (12). By Lemma 13, without loss of generality, we can assume $E(\theta_{12}) = (1, 2)$ and $E(\theta_{21}) = (2, 1)$, where $\theta_{12} \in I_1 \cap I_{12}$ and $\theta_{21} \in I_2 \cap I_{12}$, so $\theta_{12} < \theta_{21}$. We define

$$\widehat{\theta} = \inf\{\theta \in [\theta_{12}, \theta_{21}] | E(\theta) = (2, 1)\},$$

$$\widetilde{\theta} = \sup\{\theta \in [\theta_{12}, \widehat{\theta}] | E(\theta) = (1, 2)\}.$$

We note that by definition of I_{12}, we have

$$\frac{\pi}{3} + T^{-1/2+2\varepsilon} \leq \widetilde{\theta} < \widehat{\theta} \leq \pi - T^{-1/2+2\varepsilon},$$

with $\widehat{\theta} - \widetilde{\theta} \geq \eta T^{\varepsilon} \sqrt{\frac{\log T}{T}}$ to avoid collisions (see Figure 5). By definition, for $\widetilde{\theta} < \theta < \widehat{\theta}$, we have $E(\theta) \neq (1, 2), (2, 1)$, so $3 \in a(\theta) \cup b(\theta)$. But note that on Figure 5, the vertices $(1, 2)$ and $(2, 1)$ disconnect the graph into two parts: the "top" part, where $3 \in b(\theta)$, and the "bottom" part, where $3 \in a(\theta)$. It follows that either $3 \in a(\theta)$ for all $\widetilde{\theta} < \theta < \widehat{\theta}$, or $3 \in b(\theta)$ for all such θ. Without loss of generality, assume that we are in the first case.

To finish the proof, we distinguish three cases according to the values of $\widetilde{\theta}$ and $\widehat{\theta}$ in the interval I_{12}.

- Case 1: $\frac{\pi}{3} + \frac{\pi}{100} \leq \widetilde{\theta} < \widehat{\theta}$.
  In this case, note that by the graph of Figure 5, the region where $E(\theta) = (3, 2)$ must be separated from $\widehat{\theta}$ by at least $\eta T^{\varepsilon} \sqrt{\frac{\log T}{T}}$. Hence, there is an interval J of length at least $\eta T^{\varepsilon} \sqrt{\frac{\log T}{T}}$ where $3 \in a(\theta)$ and $1 \in b(\theta)$. For any θ in this interval, we have

$$\begin{aligned} r_t(\theta) &\geq f_3^A(\theta) f_1^B(\theta) \left(p_1(\theta) + p_3(\theta) - p^*(\theta)\right) \\ &\geq \frac{1}{100} \left(p_3(\theta) - p_2(\theta)\right) \\ &\succeq T^{-\varepsilon}, \end{aligned}$$

where the second inequality follows from the definitions of $a(\theta)$ and $b(\theta)$, and the last one from $\theta > \frac{\pi}{3} + \frac{\pi}{100}$. From the law of Θ, it follows that

$$\mathbb{E}\left[r_t(\Theta)\right] \succeq T^{-\varepsilon}\mathbb{P}\left(\Theta \in J\right) \geq T^{-\varepsilon} \times \frac{1}{4\pi}\eta T^{\varepsilon}\sqrt{\frac{\log T}{T}} \succeq \sqrt{\frac{\log T}{T}}.$$

- Case 2: $\widetilde{\theta} < \widehat{\theta} \leq \pi - \frac{\pi}{100}$.
  This case is similar to the first one where we exchange the roles of the arms 1 and 2: there is an interval $J' \subset \left[\frac{\pi}{3}, \pi - \frac{\pi}{100}\right]$ with length at least $\eta T^{\varepsilon}\sqrt{\frac{\log T}{T}}$ where $3 \in a(\theta)$ and $2 \in b(\theta)$. On this interval, we have

  $$r_t(\theta) \geq f_3^A(\theta) f_2^B(\theta)\left(p_3(\theta) - p_1(\theta)\right) \succeq T^{-\varepsilon},$$

  so we get $\mathbb{E}\left[r_t(\Theta)\right] \succeq \sqrt{\frac{\log T}{T}}$.

- Case 3: $\widetilde{\theta} < \frac{\pi}{3} + \frac{\pi}{100} < \pi - \frac{\pi}{100} < \widehat{\theta}$.
  In this case, we have $3 \in a(\theta)$ on the full interval $\left[\frac{\pi}{3} + \frac{\pi}{100}, \pi - \frac{\pi}{100}\right]$, so for any θ in that interval we have

  $$r_t(\theta) \geq f_3^A(\theta)\left(p_3(\theta) - \max\left(p_1(\theta), p_2(\theta)\right)\right) \succeq T^{-\varepsilon}.$$

  Since this interval is macroscopic, the variable Θ lands in it with probability $\succeq 1$, so $\mathbb{E}\left[r_t(\Theta)\right] \succeq T^{-\varepsilon}$, which concludes the proof.

Remark 14 *Separating different cases was necessary in the end: for example, if the interval $[\widetilde{\theta}, \widehat{\theta}]$ is very close to $\frac{\pi}{3}$, then the arm 2 is barely better than 3, so we lose almost nothing on the interval where $E(\theta) = (3, 1)$. However, we lose a lot when $E(\theta) = (3, 2)$.*