Multiplayer Bandit Learning, from Competition to Cooperation

Simina Brânzei SIMINA@PURDUE.EDU
Purdue University

Yuval Peres YUVAL@YUVALPERES.COM
Kent State University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

The stochastic multi-armed bandit model captures the tradeoff between exploration and exploitation. We study the effects of competition and cooperation on this tradeoff. Suppose there are two arms, one predictable and one risky, and two players, Alice and Bob. In every round, each player pulls an arm, receives the resulting reward, and observes the choice of the other player but not their reward. Alice's utility is $\Gamma_A + \lambda \Gamma_B$ (and similarly for Bob), where Γ_A is Alice's total reward and $\lambda \in [-1, 1]$ is a cooperation parameter. At $\lambda = -1$ the players are competing in a zero-sum game, at $\lambda = 1$, their interests are aligned, and at $\lambda = 0$, they are neutral: each player's utility is their own reward. The model is related to the economics literature on strategic experimentation, where usually players observe each other's rewards.

Suppose the predictable arm has success probability p and the risky arm has prior μ. If the discount factor is β, then the value of p where a single player is indifferent between the arms is the Gittins index $g = g(\mu, \beta) > m$, where m is the mean of the risky arm.

Our first result answers, in this setting, a fundamental question posed by Rothschild (1974). We show that competing and neutral players eventually settle on the same arm (even though it may not be the best arm) in every Nash equilibrium, while this can fail for players with aligned interests.

Moreover, we show that *competing players* explore *less* than a single player: there is $p^* \in (m, g)$ so that for all $p > p^*$, the players stay at the predictable arm. However, the players are not myopic: they still explore for some $p > m$. On the other hand, *cooperating players* (with $\lambda = 1$) explore *more* than a single player. We also show that *neutral players* learn from each other, receiving strictly higher total rewards than they would playing alone, for all $p \in (p^*, g)$, where p^* is the threshold above which competing players do not explore.

Keywords: bandit learning, multiplayer learning, Nash equilibrium, zero-sum games, cooperating players, strategic experimentation

1. Introduction

The multi-armed bandit learning problem is a paradigm that captures the tradeoffs between exploration and exploitation (Gittins et al. (2011); Bubeck and Cesa-Bianchi (2012); Slivkins (2019); Lattimore and Szepesvari (2019)). We study the effects of competition and cooperation on exploration in a multiplayer stochastic bandit problem, where multiple players are selecting arms and receiving the corresponding rewards in each round. Examples of scenarios that can be captured by this model include competing firms in a saturated market, which can be seen as a zero-sum game where the utility of a player is the difference between their rewards and the rewards of the opponent. At the other extreme, the interaction of organisms that are (almost) genetically identical, such as ants and bees, can be modeled by a fully cooperative game (Hamilton (1964)) where the utility of a player is the sum of the rewards of all players.

We consider a unifying framework that interpolates between these extremes, and focus on the so-called one-armed bandit problem, where the choices are playing a predictable arm or a risky one. The feedback received by each player is their own reward and the action taken by the other player. If Alice's total reward is Γ_A and Bob's total reward is Γ_B, then Alice's utility is $\Gamma_A + \lambda\Gamma_B$ and similarly for Bob, where $\lambda \in \mathbb{R}$. We study three choices for λ: the zero-sum case ($\lambda = -1$), the fully aligned case ($\lambda = 1$), and the neutral setting ($\lambda = 0$).

We study the long term behavior of the players, and show that both neutral and competing eventually settle on the same arm (even though it may not be the best arm) in every Nash equilibrium (Theorem 1). However, this can fail for cooperating players, where there are Nash equilibria in which one of the players switches infinitely often between the arms.

Rothschild (1974) studied the theory of market pricing using a two-armed bandit model and asked the question of whether players eventually settle on the same arm. Aoyagi (1998, 2011) answered this question positively in the model with imperfect monitoring, where players can see each other's actions but only their own rewards, under an assumption on the distributions. To the best of our knowledge, our paper is the first progress on the Rothschild conjecture since Aoyagi's work.

A key question in the zero sum scenario is to quantify the value of information obtained by exploring (i.e. experimenting with the risky arm). Note that while Alice does not know Bob's rewards, she may try to infer them from his actions, which in turn may lead to Bob trying to change his actions to hide information from her. We find that under optimal play, information is less valuable in the zero-sum game than in the one player setting, which leads to reduced exploration compared to the one player optimum (Theorem 3). However, Theorem 7 shows that information still has positive value. Our model is close to the games of incomplete information analyzed by Aumann and Maschler (1995), where a player may forego some rewards to hide information from their opponent.

In contrast to the zero-sum scenario, we show that exploration is increased in the fully cooperative setting compared to the single player optimum (Theorem 10). We also study the neutral regime and find that in a range of parameters, with probability 1 the players explore in every Nash equilibrium. Moreover, if the equilibrium is also perfect Bayesian, then the players learn from each other: each player gets in expectation strictly higher total reward than they would when playing alone (Theorem 13). A corollary is that with positive probability the players do not follow the same trajectory in any perfect Bayesian equilibrium.

Another natural feedback model is perfect monitoring, where all players see the actions and rewards of each other. This is not interesting in the zero sum case (see Remark 9), but in the neutral and cooperative cases it is close to classic papers in the economics of strategic experimentation. Bolton and Harris (1999), and Cripps et al. (2005) study strategic experimentation with perfect monitoring, while Heidhues et al. (2015) study imperfect monitoring where players can additionally communicate via cheap talk.

In statistical learning, there has been work on multi-player bandits (e.g., Lai et al. (2008); Liu and Zhao (2010)), where multiple cooperating players select arms and then collect the corresponding rewards. If the players collide at an arm, then they receive a reward of zero. A different line of work studied the effects of competition on learning. For example, Mansour et al. (2018) consider a setting where players arrive and depart over time, while Rosenberg et al. (2007) study multi-player learning in the one-armed bandit model, where the move from the risky to the predictable arm is irreversible.

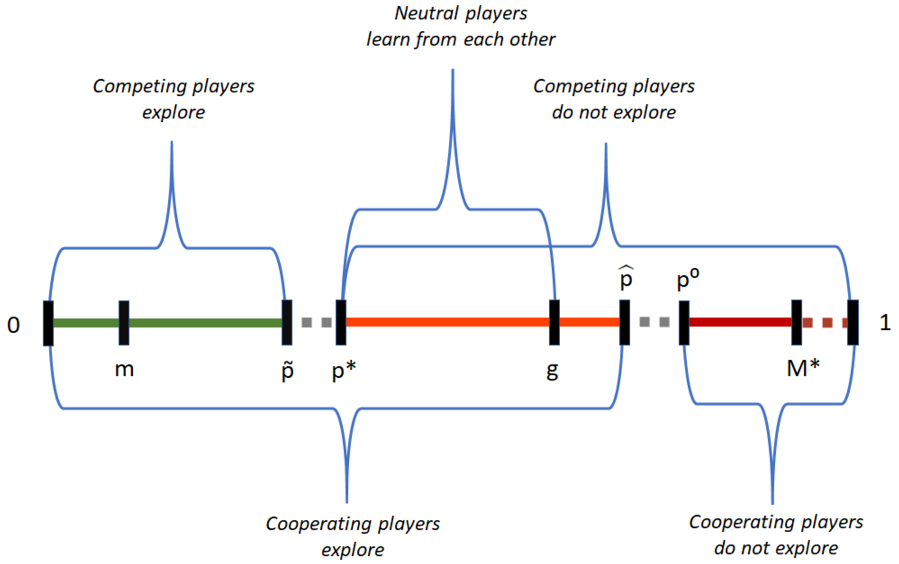

Figure 1: *Different regions in which players explore depending on the success probability p of the left arm as a function of the prior μ and the discount factor β. Here m is the mean of μ, while $g = g(\mu, \beta)$ is the Gittins index of the right arm, $\widetilde{p}$ is the threshold where for all $p < \widetilde{p}$ competing players explore, p^* the threshold where for all $p > p^*$ competing players do not explore, $\widehat{p}$ the threshold where for all $p < \widehat{p}$ cooperating players explore, and p° the threshold above which cooperating players do not explore. M^* is the maximum of the support of μ. Solid intervals are non-empty, dotted intervals may be empty.*

1.1. Model

Suppose there are two players, Alice and Bob, each of which pulls one of K arms in each round. The rewards are drawn from $\{0, 1\}$: for each arm k there is a prior distribution μ_k so that the success probability is picked (by Nature) from μ_k before the game starts. In the finite horizon version, the players play for $T + 1$ rounds indexed from 0 to T. In the discounted version, the game continues forever, but the players have value reduced by a factor of β^t for the reward in round t, where $\beta \in (0, 1)$[1].

Rewards In every round, each player collects the reward from the arm they pulled (even if both chose the same arm). We denote by $\gamma_A(t)$ and $\gamma_B(t)$ the random variable corresponding to the reward received by Alice and Bob, respectively in round $t \geq 0$. The total reward of player i in a finite horizon game is $\Gamma_i = \sum_{t=0}^{T} \gamma_i(t)$, while player i's reward in the discounted game is $\Gamma_i = \sum_{t=0}^{\infty} \gamma_i(t) \cdot \beta^t$. In every round, after selecting the arm to pull, each player observes their own reward and the action taken by the other player, but not their reward.

Utilities The utility of each player is a combination of their own reward and the reward of the other player. More precisely, there is a *cooperation parameter* $\lambda \in [-1, 1]$ so that Alice's utility is $u_A = \Gamma_A + \lambda \cdot \Gamma_B$, while Bob's utility is $u_B = \Gamma_B + \lambda \cdot \Gamma_A$. For $\lambda = -1$, we have a zero-sum

[1] This can alternatively be interpreted as the game stopping with probability $1 - \beta$ in each round.

game while for $\lambda = 1$, the interests of the players are aligned. The case $\lambda = 0$ is the neutral regime where each player's utility is their own total reward.

Arms We focus on the so-called one-armed bandit problem, where there is a predictable left arm, denoted L, with known success probability p and a risky right arm, denoted R, with a prior μ that is not a point mass. A player is said to "explore" if that player selects the right arm at least once. Denote by $m = \int_0^1 x \, d\mu(x)$ the mean of the prior μ, by $w = \int_0^1 (x - m)^2 \, d\mu(x) > 0$ the variance of μ, and by M^* the maximum of the support of μ:

$$M^* = \sup\{x \in [0, 1] : \mu(x, 1] > 0\} \tag{1}$$

Strategies A pure strategy for player i is a function $S_i : \bigcup_{t \in \mathbb{N}} Y_t \times Z_t^i \to \{L, R\}$, where Y_t is the public history (i.e. the sequence of past actions of both players) until the end of round $t - 1$ and Z_t^i is the private history of player i, containing the bits observed by i until the end of round $t - 1$. Thus the pure strategy tells player i which arm to play next given the public and private history. A mixed strategy is a probability distribution over the set of pure strategies. The *expected utility* of a player is computed using the player's beliefs about the private information of the other player.For more details on extensive form games, see Chapter 3 in Maschler et al. (2013).

1.2. Roadmap of the Paper

Section 2 includes more detailed discussion of related work. Section 3 studies the long term behavior of the players, showing that both neutral and competing eventually settle on the same arm in every Nash equilibrium, while this can fail for cooperating players. Section 4 has the analysis for competing players, showing that they explore less than a single player, but are also not completely myopic. Section 5 studies fully cooperating players, showing they explore more than a single player. Section 6 studies neutral players, showing that they learn from each other. Directions for future research are described in Section 7. Background on the Gittins index is given in Section 8.

2. Related Work

This work is related to several streams of research. Bandit learning problems with multiple players have been studied in the collision model, where players are pulling arms independently. The players are cooperating—trying to maximize the sum of rewards—and can agree on a protocol before play, but cannot communicate during the game. Whenever there is a collision at some arm, then no player that selected that arm receives any reward. This is motivated by applications such as cognitive radio networks, where interference at a channel destroys the signal for all the players involved. Several research directions in this setting include designing algorithms that allow the players to maximize the total reward, depending on whether the environment is adversarial (see Alatur et al. (2019), and Bubeck et al. (2020)) or stochastic (e.g., Kalathil (2014), Lugosi and Mehrabian (2018), Bistritz and Leshem (2018)), and on whether they receive feedback about the collision or not (Avner and Mannor (2014), Rosenski et al. (2016), Bonnefoi et al. (2017), Boursier and Perchet (2018). Hillel et al. (2013) study exploration in multi-armed bandits in a setting with pure exploration, where players collaborate to identify an ϵ-optimal arm.

Aoyagi (1998, 2011) studies bandit learning with the same feedback model as ours, when there are multiple neutral players and two risky arms. The main result is that assuming a property on the distributions of the arms, in any Nash equilibrium the players settle eventually on the same arm.

This is related to Aumann's agreement theorem Aumann (1976), which shows in a Bayesian setting that rational players cannot agree to disagree. While this game is not captured by the formal model of the theorem in Aumann (1976), the result is similar conceptually.

Bolton and Harris (1999) study a multiplayer learning problem in in the one-armed bandit model with continuous time and perfect monitoring, where the players can observe each other's past actions and rewards. The main effects observed in symmetric equilibria are a free rider effect and an encouragement effect, where a player may explore more in order to encourage further exploration from others. Cripps et al. (2005) characterize the unique Markovian equilibrium of the game. They also show that asymmetric equilibria are more efficient than symmetric ones, as it is more useful for the players to take turns experimenting.

Heidhues et al. (2015) study the discrete version of this model and establish that in any Nash equilibrium, players stop experimenting once the common belief falls below a single-agent cutoff. They also show that the total number of experiments performed in equilibrium differs from the single-agent optimum by at most one. Heidhues et al. (2015) additionally study a model with imperfect monitoring, where the players can observe each other's actions but only their own rewards. The players can also communicate via cheap talk. One of the main findings is that cheap talk is incentive compatible and the socially optimal symmetric experimentation profile can be supported as a perfect Bayesian equilibrium. Klein and Rady (2011) study the one-armed bandit model with public monitoring, where the correlation across bandits is negative. Rosenberg et al. (2013) study the model with imperfect monitoring, except the decision to switch from the risky arm to the safe one is irreversible. As they write, *"Dropping the assumption that payoffs are publicly observed raises new issues. Player i would like to make inferences about player j's observations on the basis of player j's actions, but cannot do so without knowing how player j's decisions relate to player j's observations, that is, j's strategy"*.

Another line of work in economics studies the interplay between competition and innovation (see, e.g., D'Aspremont and Jacquemin (1988)). Close to our high level direction is the work of Besanko and Wu (2013), which studies the tradeoff between cooperation and competition in R&D via learning in a one-armed bandit model, where the safe arm is the established product and the risky arm is the novel product. Our model is also related to the literature on stochastic games with imperfect monitoring (see, e.g., Abreu et al. (1990)). Rosenberg et al. (2009) study a model of dynamic games with informational externalities and analyze the rate of experimentation, providing conditions under which players eventually reach a consensus.

In evolutionary biology, there is a line of research dedicated to understanding the extremely high levels of cooperation observed in social insects (see, e.g., Anderson (1984) and Boomsma (2007)), including designing mathematical models to explain why eusociality would evolve from natural selection (e.g., Nowak et al. (2010)). Social learning was studied in the context of understanding altruistic behaviors such as sharing of information about food locations (see, e.g., Gruter et al. (2010)).

The theme of incentivizing exploration was studied, for example, by Kremer et al. (2013), and Frazier et al. (2014), where the problem is that a principal wants to explore a set of arms, but the exploration is done by a stream of myopic agents that have their own incentives and may prefer to exploit instead. Mansour et al. (2015) design Bayesian incentive compatible mechanisms for such settings.

Aridor et al. (2019) empirically study the interplay between exploration and competition in a model where multiple firms are competing for the same market of users and each firm commits to a multi-armed bandit algorithm. The objective of each firm is to maximize its market share and the question is when firms are incentivized to adopt better algorithms. Multi-armed bandit problems with strategic arms have been studied theoretically by Braverman et al. (2019), in the setting where each arm receives a reward for being pulled and the goal of the principal is to incentivize the arms to pass on as much of their private rewards as possible to the principle.

Immorlica et al. (2011) study competitive versions of classic search and optimization problems by converting them to zero-sum games. One of the open questions from Immorlica et al. (2011) was whether competition between algorithms improves or degrades expected performance in that framework, which was answered by Dehghani et al. (2016) for the ranking duel and a more general class of dueling games.

3. Long Term Behavior

In this section we study the long term trajectories of the players in a Nash equilibrium for all values of λ. We show that in every Nash equilibrium, competing and neutral players converge to playing the same arm forever from some point on. For $\lambda \notin [-1, 0]$, this can fail. The omitted material of this section can be found in Appendix D.

Theorem 1 *Consider two players, Alice and Bob, playing a one armed bandit problem with discount factor $\beta \in (0, 1)$. The left arm has success probability p and the right arm has prior distribution μ such that $\mu(p) = 0$. Then in any Nash equilibrium, in both the competing ($\lambda = -1$) and neutral ($\lambda = 0$) cases, the players eventually settle on the same arm with probability* 1.

The intuition behind the proof for neutral players is that if both players explore finitely many times, then we are done. Otherwise, there is a player, say Alice, who explores infinitely many times. Then Alice will eventually know which arm is better, so if she continues exploring we must have $\Theta > p$. Thus if Bob sees that Alice keeps exploring, he will eventually realize that $\Theta > p$ and will join her at the right arm. Two obstacles make the proof delicate. The first is that the theorem applies to all Nash equilibria, without assuming subgame perfection, so we need to rule out non-credible threats (which do occur in the case $\lambda = 1$, see Example 1 below). The second obstacle is that Θ might be very close to p, which delays the time at which Alice determines the better arm. This issue did not arise in Aoyagi's work since the distributions were discrete. We overcome this obstacle by careful concentration arguments and Bayesian analysis; see sections D.2 and D.3 in the appendix.

In the zero-sum setting there is an additional difficulty compared to the neutral case: the players might refrain from switching to the optimal arm in order to not trigger an adverse reaction from the opponent. The key to overcoming this difficulty is realizing that it is impossible for both players to benefit from repeatedly pulling the inferior arm. Each player might subjectively believe for some time that they are playing the better arm, but if a player keeps exploring, then their subjective evaluation of the risky arm will converge to the objective reality. The proof can be found in section D.4 in the appendix.

The next example shows that when $\lambda = 1$ there are Nash equilibria where aligned players do not settle on the same arm.

Example 1 (Nash equilibria where players do not converge, $\lambda = 1$) *Suppose Alice and Bob are aligned players in a one-armed bandit problem with discount factor β, where the left arm has success probability p and the right arm has prior distribution μ that is a point mass at $m > p$.*

Then for every discount factor $\beta > 1/2$, there is a Nash equilibrium in which Bob visits both arms infinitely often.

Let $k \in \mathbb{N}$. Bob's strategy S_B is to play left in rounds $0, k, 2k, 3k, \ldots$, and right in the remaining rounds. Alice's strategy S_A is to play right if Bob follows the trajectory above; if Bob ever deviates from S_B, then Alice switches to playing left forever. See section D.1 in Appendix D for a proof that if k is large enough, then this is indeed a Nash equilibrium.

Note this Nash equilibrium is not Pareto optimal: both players could improve by a joint deviation to always playing right. This Nash equilibrium is also not subgame perfect since it has a non-credible threat by Alice (of playing left forever after any deviation by Bob).

4. Competitive Play

In this section we study the zero-sum game, corresponding to $\lambda = -1$. This is an extensive-form game with an initial move by Nature; see Maschler et al. (2013); Karlin and Peres (2017). The game has a value by Sion's minimax theorem Sion (1958); moreover, the value is zero by symmetry.

Recall that in the discounted setting, the **Gittins index** $g = g(\mu, \beta)$ of the right arm is defined as the infimum of the success probabilities p where playing always left is optimal for a single player.

Observation 1 *If selecting L in every round is the only optimal strategy for a single player, then in every optimal strategy for Alice in the zero sum game, she never explores, and similarly for Bob.*

Proof If Alice explores with positive probability and Bob never explores, then her expected net payoff will be negative by the hypothesis, since she is not learning anything from Bob's actions. ■

We also give a simple lower bound on the Gittins index, the proof of which is included for completeness in Appendix A, together with the other omitted proofs of this section. Recall that for an arm with prior μ, we have $g = g(\mu, \beta)$ is the Gittins index of the arm, m is the mean, and w is the variance of μ.

Lemma 2 *Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$.*

Our first theorem shows that when using optimal strategies, competing players explore less than a single player, for both discounted games and finite horizon.

Theorem 3 (Competing players explore less) *Suppose arm L has a known probability p and arm R has i.i.d. rewards with unknown success probability with prior μ (which is not a point mass). Assume that Alice and Bob are playing optimally in the zero sum game with discount factor β. Then there exists a threshold $p^* = p^*(\mu, \beta, -1) < g$, where $g = g(\mu, \beta)$ is the Gittins index of the right arm, such that for all $p > p^*$, with probability 1 the players will not explore arm R. More precisely, define*

$$p^* = p^*(\mu, \beta, \lambda) = \sup\Big\{p : \textit{arm R is explored in some Nash equilibrium}\Big\} \qquad (2)$$

Then $p^*(\mu, \beta, -1) \leq (m\beta + g)/(1 + \beta)$, where m is the mean of μ.

Remark 4 *Note that the upper bound on $p^*(\mu, \beta, -1)$ tends to $(m+g)/2$ as $\beta \to 1$.*

We establish the theorem by showing that a player using an optimal strategy is never the first to explore. However, each player will need as part of their strategy a contingency plan for what to do if the other player deviates from the main path. In fact, the strategy of always playing left is not part of any equilibrium. If Bob always plays left, then Alice can play the one player optimum (of pulling the arm with the highest Gittins index) and win.

Proof [Proof of Theorem 3] Consider the following strategy S_B for Bob: play left until Alice selects the right arm, say in some round k. Then play left again in round $k+1$, and then starting with round $k+2$ copy Alice's move from the previous round. In particular, Bob never plays left first. Fix an arbitrary pure strategy S_A for Alice. If S_A never explores first, then we are done. Otherwise, suppose S_A explores first in round k. Then Alice's total reward $\Gamma_A = \Gamma_A(S_A, S_B)$ has expectation

$$\mathbb{E}(\Gamma_A) = \sum_{t=0}^{k-1} p \cdot \beta^t + \sum_{t=k}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t \,.$$

Bob's total reward $\Gamma_B = \Gamma_B(S_A, S_B)$ has expectation

$$\begin{aligned} \mathbb{E}(\Gamma_B) &= \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+2}^{\infty} \mathbb{E}(\gamma_B(t)) \cdot \beta^t = \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+2}^{\infty} \mathbb{E}(\gamma_A(t-1)) \cdot \beta^t \\ &= \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^{t+1} \end{aligned} \tag{3}$$

Note that $\mathbb{E}(\gamma_A(k)) = m$. Then the difference in rewards is

$$\begin{aligned} \mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) &= \left(\sum_{t=0}^{k-1} p \cdot \beta^t + m \cdot \beta^k + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t\right) - \left(\sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^{t+1}\right) \\ &= (m\beta - p - p\beta)\beta^k + (1-\beta)\beta^k \cdot \left(m + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t\right) \end{aligned} \tag{4}$$

Since Bob is copying Alice, she is not learning anything from his actions, so her total reward from round $k+1$ to ∞ is at most the maximum reward that a single player can obtain, namely $g/(1-\beta)$. Thus we can bound the difference between the players by

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \leq \Big(m\beta - p(1+\beta) + g\Big)\beta^k \,. \tag{5}$$

The right hand side of inequality (5) is negative when

$$m\beta - p(1+\beta) + g < 0 \iff p > \frac{m\beta + g}{1+\beta} \,. \tag{6}$$

Note that $(m\beta + g)/(1+\beta) \in (m, g)$ and the players do not explore arm R for any p above this threshold. Thus $p^* \leq (m\beta + g)/(1+\beta)$ as required. ■

We also show that competition reduces exploration for a large finite horizon.

Theorem 5 (Competing players explore less, finite horizon) *Suppose arm L has a known probability p and arm R has a known distribution μ with mean m. Let T be the horizon and M^* the maximum of the support of μ. We have*

1. *If $p > (m + M^*)/2$, then the players do not explore arm R.*
2. *However, for every $p < M^*$, if T is large enough, then a single player will explore given the same two arms.*

Remark 6 *Part 2 of Theorem 5 is due to Bradt et al. (1956) (see page 1073) and included for comparison. We can define in the finite horizon setting an index $g_T = g(\mu, T)$ as the infimum of p where playing always left is optimal for a single player. Part 2 of Theorem 5 is equivalent to $\lim_{T\to\infty} g_T = M^*$. A similar proof shows that for any prior μ, we have $\lim_{\beta\to 1} g(\mu, \beta) = M^*$.*

In the single player one-armed bandit, obtaining information about the risky arm can compensate for a lower mean reward compared to the predictable arm. In the zero-sum case, the value of acquiring information is less clear since it can be copied by the opponent in the next round. The next theorem shows that such information still has value: competing players do not follow a myopic policy of pulling the arm with the highest mean reward in each round.

Theorem 7 (Competing players are not completely myopic) *In the setting of Theorem 3, there exists a threshold $\widetilde{p} = \widetilde{p}(\mu, \beta, -1) > m$, such that for all $p < \widetilde{p}$, with probability 1 both players will explore arm R in the initial round of any optimal play. More precisely, define*

$$\widetilde{p} = \widetilde{p}(\mu, \beta, \lambda) = \inf\Big\{p : \textit{arm L is not explored in any Nash equilibrium}\Big\} \tag{7}$$

The proof shows that $\widetilde{p}(\mu, \beta, -1) \geq m + \beta w/2$, where w is the variance of μ.

Remark 8 *Note that $\widetilde{p} \leq p^*$ and we conjecture they are in fact equal.*

Remark 9 *In the feedback model with perfect monitoring, where players observe each other's actions and rewards, the myopic strategy of selecting the arm with the highest current mean (in each round) is optimal. The reason is that exploration gives no future (informational) advantage to the exploring player.*

In Appendix E we show improved bounds for a uniform prior and a plot with the bounds.

5. Cooperative Play

In this section we study the scenario of $\lambda = 1$, where the players are cooperating. Cooperating players aim to maximize the sum of their rewards and can agree on their strategies before the game starts. Later the players cannot communicate beyond seeing each other's actions. The omitted proofs of this section can be found in Appendix B.

We show that when the players are cooperating they explore even more than a single player. In particular, they explore even if the known arm has a probability p higher than the Gittins index of the right arm.

Recall the definitions of $m_1 = \frac{1}{m}\int_0^1 x^2 \cdot d\mu$ as posterior mean at the right arm after observing 1 in round zero, and of $w = m \cdot (m_1 - m)$.

Theorem 10 (Cooperating players explore more, discounted) *Consider two cooperating players, Alice and Bob, playing a one armed bandit problem with discount factor $\beta \in (0, 1)$. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Then there exists $\widehat{p} > g = g(\mu, \beta)$, so that for all $p < \widehat{p}$, at least one of the players explores the right arm with positive probability under any optimal strategy pair maximizing their total reward.*

Proof In this proof we write $\Gamma_A = \Gamma_A(p)$ to emphasize the dependence on p. At $p = g$, by definition of the Gittins index, Alice has a strategy starting at the right arm such that $\mathbb{E}(\Gamma_A(g)) = \frac{g}{1-\beta}$. Suppose Bob responds by playing L in rounds zero and one, then from round two onwards copies Alice's previous move. Then by inequality (5), applied at $p = g$, we get $\mathbb{E}\Big(\Gamma_A(g) - \Gamma_B(g)\Big) \leq (m - g)\beta$.

Therefore $\mathbb{E}\Big(\Gamma_A(g) + \Gamma_B(g)\Big) = 2\,\mathbb{E}(\Gamma_A(g)) - \mathbb{E}\Big(\Gamma_A(g) - \Gamma_B(g)\Big) \geq \frac{2g}{1-\beta} + (g - m)\beta$. Applying the same strategies at $p = g + \delta$ and comparing the total rewards to both players staying left gives $\Delta_p := \mathbb{E}\Big(\Gamma_A(g) + \Gamma_B(g)\Big) - 2p/(1-\beta) = (g - m)\beta - 2\delta/(1-\beta)$. By Lemma 2, we infer $\Delta_p \geq \beta^2 w/2 - 2\delta/(1-\beta)$. This is positive provided that $\delta < \beta^2 w(1-\beta)/4$. ■

Recall the index g_T for finite horizon was defined in Remark 6.

Theorem 11 (Cooperating players explore more, finite horizon) *Consider two cooperating players, Alice and Bob, playing a one armed bandit problem. The left arm is known and has probability p, while the right arm has a known distribution μ that is not a point mass. Let T be the horizon and M^* the maximum of the support of μ. Then there exists $\delta = \delta(T) > 0$ so that optimal players explore the right arm for all $p < g_T + \delta$.*

Note that cooperating players do prefer the predictable arm for high enough values of p. Recall that M^* is the maximum of the support of μ.

Proposition 12 *There is a threshold $p^\circ < M^*$ so that cooperating players do not explore for any $p > p^\circ$.*

6. Neutral Play

In this section we study the scenario of $\lambda = 0$, where the players are only interested in their own rewards. The omitted proofs of this section are in Appendix C.

Recall the solution concepts are Nash equilibrium and perfect Bayesian equilibrium. For neutral play the utility of each player is their total reward. The solution concepts will be Nash equilibrium and perfect Bayesian equilibrium. Recall player i's strategy σ_i is a *best response* to player j's strategy σ_j if no strategy σ_i' achieves a higher expected utility against σ_j. A mixed strategy profile (σ_A, σ_B) is a *Bayesian Nash equilibrium* if σ_i is a best response for each player i. For brevity, we refer to such strategy profiles as Nash equilibria. A *Perfect Bayesian Equilibrium* is the version of subgame perfect equilibrium for games with incomplete information. A pair of strategies (σ_A, σ_B) is a perfect Bayesian equilibrium if (i) starting from any information set, subsequent play is optimal, and (ii) beliefs are updated consistently with Bayes' rule on every path of play that occurs with positive probability. Such equilibria are guaranteed to exist in this setting (Fudenberg and Levine (1983)).

Observation 2 *For all $p < g(\mu, \beta)$, in any Nash equilibrium, each player explores with strictly positive probability.*

We will say that players learn from each other under some strategies if the total expected reward of each player is strictly higher than it would be for a single player using an optimal strategy. This can happen if the players infer additional information from each other's actions, beyond the bits that they observe themselves.

Theorem 13 (Neutral players learn from each other, discounted) *Consider two neutral players, Alice and Bob, playing a one armed bandit problem with discount factor $\beta \in (0, 1)$. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Then in any Nash equilibrium:*

1. *For all $p < g(\mu, \beta)$, with probability 1 at least one player explores. Moreover, the probability that no player explores by time t decays exponentially in t.*
2. *Suppose $p \in (p^*, g)$, where p^* is the threshold above which competing players do not explore [2]. If the equilibrium is furthermore perfect Bayesian, then every (neutral) player has expected reward strictly higher than a single player using an optimal strategy.*

Corollary 14 *In each perfect Bayesian equilibrium, with positive probability the players do not have the same trajectory.*

Proof The conclusion follows from part 2 of Theorem 13. If the players always had the same trajectory, then the best total reward achievable would be the one player optimum. ■

In contrast, there exist Nash equilibria where the players have the same trajectory with probability 1. Such play can be supported by using (non-credible) threats in case one of the players deviates from the Nash equilibrium path (see Example 2 in Appendix C).

Observation 3 *For $p \geq g(\mu, \beta)$, there is a perfect Bayesian equilibrium where the players never explore: the pair of strategies in which each player stays left no matter what happens is a perfect Bayesian equilibrium. This is not a Nash equilibrium for $p < g(\mu, \beta)$, since a player can switch to the single player optimum.*

We show players also learn from each other in the finite horizon game. Recall g_T is the index for finite horizon T and M^* is the maximum of the support of μ. Formally, we have:

Theorem 15 (Neutral players learn from each other, finite horizon) *Consider two neutral players, Alice and Bob, playing a one armed bandit problem with horizon T. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Then in any subgame perfect equilibrium:*

1. *For all $p < M^*$, with probability that converges to 1 as T grows, at least one player explores. Moreover, the probability that no player explores by time t decays polynomially in t.*
2. *Suppose $(m + M^*)/2 < p < M^*$. Then every (neutral) player has expected reward strictly higher than a single optimal player.*

[2] For the formal definition of p^*, see Theorem 3.

7. Concluding Remarks and Questions

Several open questions arise from this work, such as understanding whether randomization is required sometimes, whether $p^* = \widetilde{p}$, and whether in the setting with multiple risky arms neutral and competing players settle eventually on the same arm with probability 1 in every Nash equilibrium. More precisely, we propose the following questions:

1. Do competing and neutral players always have optimal pure strategies or is randomization required sometimes?

2. Recall the threshold $p^* = p^*(\mu, \beta, \lambda)$, defined in Theorem 3, is the smallest number so that for all $p > p^*$ and in all Nash equilibria, the players do not explore. Similarly, $\widetilde{p} = \widetilde{p}(\mu, \beta, \lambda)$ from Theorem 7, is the largest number so that for all $p < \widetilde{p}$, the right arm is explored in some Nash equilibrium.

   (a) Observation 3 shows that for neutral players we have $\widetilde{p}(\mu, \beta, 0) \geq g(\mu, \beta)$ and Theorem 10 implies that $\widetilde{p}(\mu, \beta, 1) > g(\mu, \beta)$. Is $\widetilde{p}(\mu, \beta, 0) > g(\mu, \beta)$?

   (b) Is $p^* = \widetilde{p}$ true in general?

   (c) Are p^* and $\widetilde{p}$ monotone in β and λ?

3. Consider the following variant of the model in this paper: each player learns the other player's rewards, but with a delay of k rounds. Can one describe explicitly Nash equilibria in this variant as Bolton and Harris (1999) do in the model with perfect monitoring?

4. Can one induce more exploration in this model by incorporating patent protection? For example, if Alice explores an arm for the first time, then a patent for her could mean that Bob cannot pull that arm for k rounds afterwards, or alternatively, Bob should give Alice a fraction α of his reward whenever he pulls that arm in the k rounds following her first exploration.

5. When there are multiple risky arms, do neutral and competing players eventually settle with probability 1 on the same arm in every Nash equilibrium? For neutral players, this is a conjecture by Rothschild (1974) made explicit by Aoyagi (1998), who solved the case of two arms with finitely supported priors.

6. Consider the setting with multiple risky arms. If there exist optimal pure strategies, can they be obtained from an index analogous to the Gittins index for a single player?

8. Background on Gittins index

The Gittins index was first defined by Bradt et al. (1956) and its importance for the multi-armed bandit problem was shown by Gittins and Jones (1974); see also Gittins (1979); Berry and Fristedt (1985). The description below mainly follows Weber (1992).

In the classic multi-armed bandit problem, there is a gambler who can play any of n one-armed bandit machines. The goal of the gambler is to play in a way that maximizes its expected total discounted reward.

Let $[n] = \{1, \ldots, n\}$ be the set of bandits, each of which is a Markov process. The state of bandit j at time step $t \in \{0, 1, \ldots, \}$ is denoted $x_j(t)$. When playing bandit j, the gambler receives reward $R_j(x_j(t))$ and the state of bandit j changes in a known Markov fashion, while the states of the other bandits remain unchanged.

A policy stipulates which bandit to play next, given the history of play and the rewards obtained so far. Given policy π, let $j(t)$ denote the bandit played at time step t. The goal is to find a policy that maximizes the expected discounted reward, defined as

$$V_\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \beta^t \cdot R_{j(t)}\big(x_j(t)\big) \mid x(0) = x\right] \tag{8}$$

The solution to this optimization problem is described by functions G_j, which are known as Gittins indices. Each function G_j only depends on the current state bandit j. Gittins and Jones (1974) proved that playing bandit j at time t is optimal if and only if

$$G_j(x_j(t)) = \max_{1\le i\le n} G_i\big(x_i(t)\big)\,. \tag{9}$$

The Gittins index has several equivalent definitions. The most convenient for us is the *retirement value* of the arm. Suppose at every step the Gambler may choose to "retire" and receive a payment p in the current step and in all subsequent steps. i Alternatively, he may select arm j and receive the current reward at that arm, while maintaining the option to retire at any point in the future. Given that arm j is currently at state x_j, the Gittins index $G(x_j)$ is the infimum of the values p for which retirement now is preferable.

Bernoulli bandits In the paper, we focus on a special case known as Bernoulli bandits, where the rewards are 1 (successes) or 0 (failures). Arm j has a known prior μ_j^0 on $[0,1]$; its success probability Θ_j (unknown to the player) is drawn from μ_j, and each time arm j is selected, the reward is 1 with probability Θ_j, independently of previous picks. The state of arm j at time t is described by a pair $(s_j(t), f_j(t))$, where $s_j(t)$ and $f_j(t)$ are the number of successes and failures, respectively, obtained at arm j until time t. The posterior distribution of the success probability Θ_j after step t is μ_j^t which has density proportional to $\theta^{s_j(t)}(1-\theta)^{f_j(t)}$ with respect to μ_j^0, i.e., for Borel sets $A \subset [0,1]$,

$$\mu_j^t(A) = \frac{\int_A \theta^{s_j(t)}(1-\theta)^{f_j(t)}\, d\mu_j^0(\theta)}{\int_0^1 \theta^{s_j(t)}(1-\theta)^{f_j(t)}\, d\mu_j^0(\theta)}\,.$$

If the player selects arm j at time $t+1$, then the expected reward from that move (given the history) is the mean of the posterior

$$\int_0^1 \theta d\mu_j^t(\theta)\,.$$

That expected reward is also the transition probability from the state $(s_j(t), f_j(t))$ to the state $(s_j(t)+1, f_j(t))$. The transition probability to $(s_j(t), f_j(t)+1)$ is the complementary probability $\int_0^1 (1-\theta) d\mu_j^t(\theta)\,.$

9. Acknowledgements

We thank Sebastien Bubeck and Mark Sellke for useful discussions and the COLT reviewers for feedback that helped improve the paper.

References

Dilip Abreu, David Pearce, and Ennio Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5):1041–1063, 1990. ISSN 00129682, 14680262.

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

M Anderson. The evolution of eusociality. *Annual Review of Ecology and Systematics*, 15(1): 165–189, 1984.

Masaki Aoyagi. Mutual observability and the convergence of actions in a multi-person two-armed bandit model. *Journal of Economic Theory*, 82:405–424, 1998.

Masaki Aoyagi. Corrigendum: Mutual observability and the convergence of actions in a multi-person two-armed bandit model, 2011.

Guy Aridor, Kevin Liu, Aleksandrs Slivkins, and Zhiwei Steven Wu. The perils of exploration under competition: A computational modeling approach. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019.*, pages 171–172, 2019.

Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976. ISSN 00905364.

Robert J. Aumann and Michael Maschler. *Repeated Games with Incomplete Information*. MIT Press, 1995.

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

Donald A. Berry and Bert Fristedt. *Bandit problems*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1985. ISBN 0-412-24810-7. doi: 10.1007/978-94-015-3711-7. Sequential allocation of experiments.

David Besanko and Jianjun Wu. The impact of market structure and learning on the tradeoff between r and d competition and cooperation. *Journal of Industrial Economics*, 61(1):166–201, 2013.

Ilai Bistritz and Amir Leshem. Distributed multiplayer bandits - a games of thrones approach. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 7222–7232, 2018.

P. Bolton and C. Harris. Strategic experimentation. *Econometrica*, 67(2):349–374, 1999.

Rémi Bonnefoi, Lilian Besson, Christophe Moy, Emilie Kaufmann, and Jacques Palicot. Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings. In *Cognitive Radio Oriented Wireless Networks - 12th International Conference, CROWNCOM 2017, Lisbon, Portugal, September 20-21, 2017, Proceedings*, pages 173–185, 2017.

Jacobus J. Boomsma. Kin selection versus sexual selection: Why the ends do not meet. *Current Biology*, 17(16):R673 – R683, 2007. ISSN 0960-9822.

Etienne Boursier and Vianney Perchet. SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. *CoRR*, abs/1809.08151, 2018.

R. N. Bradt, S. M. Johnson, and S. Karlin. On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27:1060–1074, 1956.

Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 383–416, 2019.

Sebastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Yuanzhi Li, Yuval Peres, and Mark Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *COLT*, 2020.

Martin Cripps, Godfrey Keller, and Sven Rady. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, 2005.

Claude D'Aspremont and Alexis Jacquemin. Cooperative and noncooperative research and development in duopoly with spillovers. *The American Economic Review*, 78(5):1133–1137, 1988.

Sina Dehghani, Mohammad Taghi Hajiaghayi, Hamid Mahini, and Saeed Seddighin. Price of competition and dueling games. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 21:1–21:14, 2016.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pages 5–22, New York, NY, USA, 2014. ACM.

Drew Fudenberg and David Levine. Subgame-perfect equilibria of finite- and infinite-horizon games. *Journal of Economic Theory*, 31(2):251 – 268, 1983. ISSN 0022-0531.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, pages 148–177, 1979.

J.C. Gittins and D.M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, 1974.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed Bandit Allocation Indices*. Wiley, 2011.

Christoph Gruter, Ellouise Leadbeater, and Francis L. W. Ratnieks. Social learning: The importance of copying others. *Current Biology*, 20(16), 2010.

W. D. Hamilton. The genetical evolution of social behaviour. i, ii. *Journal of Theoretical Biology*, 7(1):1–16, 1964.

Paul Heidhues, Sven Rady, and Philipp Strack. Strategic experimentation with private payoffs. *Journal of Economic Theory*, 159:531–551, 2015.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 854–862. Curran Associates, Inc., 2013.

Nicole Immorlica, Adam Tauman Kalai, Brendan Lucier, Ankur Moitra, Andrew Postlewaite, and Moshe Tennenholtz. Dueling algorithms. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 215–224, 2011.

Dileep Kalathil. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4), 2014.

Anna R. Karlin and Yuval Peres. *Game theory, alive*. American Mathematical Society, Providence, RI, 2017. ISBN 978-1-4704-1982-0.

Nicolas Klein and Sven Rady. Negatively correlated bandits. *The Review of Economic Studies*, 78 (2):693–732, 2011.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the ”wisdom of the crowd”. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pages 605–606, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1962-1.

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

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. http://downloads.tor-lattimore.com/banditbook/book.pdf, 2019.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *Trans. Sig. Proc.*, 58(11):5667–5681, 2010.

Gábor Lugosi and Abbas Mehrabian. Multiplayer bandits without observing collision information. *CoRR*, abs/1808.08416, 2018.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, pages 565–582, New York, NY, USA, 2015. ACM.

Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 48:1–48:27, 2018.

Michael Maschler, Eilon Solan, and Shmuel Zamir. *Game Theory*. Cambridge University Press, 2013.

Martin A. Nowak, Corina E. Tarnita, and Edward O. Wilson. The evolution of eusociality. *Nature*, 466(7310):1057–1062, 2010.

Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. Social learning in one-arm bandit problems. *Econometrica*, 75(6):1591–1611, 2007.

Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. Informational externalities and emergence of consensus. *Games Econ. Behav.*, 66(2):979–994, 2009.

Dinah Rosenberg, Antoine Salomon, and Nicolas Vieille. On games of strategic experimentation. *Games and Economic Behavior*, 82:31–51, 2013.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits - a musical chairs approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 155–163, 2016.

Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9 (2):185 – 202, 1974. ISSN 0022-0531.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286, 2019. ISSN 1935-8237.

Richard Weber. On the Gittins Index for Multiarmed Bandits. *The Annals of Applied Probability*, 2 (4):1024 – 1033, 1992.

Appendix A. Competitive Play

In this section we include the omitted proofs for the scenario of $\lambda = -1$, where the players are competing with each other.

Lemma 2 (restated). *Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$.*

Proof Suppose Alice is playing the one-armed bandit game by herself where the right arm has distribution μ that is not a point mass and left arm has known probability $p = g = g(\mu, \beta)$. Consider the following strategy for Alice:

- Round zero: play right.
- Round one: play left if 0 was observed in round zero, and right if 1 was observed.
- Round two onwards: play left.

Then by the definition of g, we have that this strategy for Alice is at most as good as retiring and receiving g forever, and so

$$\frac{g}{1-\beta} \geq m + (1-m) \cdot \frac{g\beta}{1-\beta} + m\left(m_1\beta + \frac{g\beta^2}{1-\beta}\right) \tag{10}$$

Recall that $m \cdot m_1 = m^2 + w$. Using this in (10) and rearranging, we obtain:

$$g \geq m + \frac{\beta w}{1 + m\beta} \geq m + \frac{\beta w}{2} \tag{11}$$

■

Theorem 5 [Competing players explore less, finite horizon] (restated). *Suppose arm L has a known probability p and arm R has a known distribution μ with mean m. Let T be the horizon and M^* the maximum of the support of μ. We have*

1. *If $p > (m + M^*)/2$, then the players do not explore arm R.*
2. *However, for every $p < M^*$, if T is large enough, then a single player will explore given the same two arms.*

Proof For part 1 of the statement, we use the same strategy S_B for Bob as in the proof of Theorem 3: play left until Alice selects the right arm, say in some round k. Then play left again in round $k + 1$, and then starting with round $k + 2$ copy Alice's move from the previous round.

Fix an arbitrary pure strategy S_A for Alice. If S_A never plays right first we are done. If Alice plays right for the first time in round k, then $\gamma_A(k) = m$. Her expected total reward can be written as

$$\mathbb{E}(\Gamma_A) = \left(\sum_{t=0}^{k-1} p \right) + m + \sum_{t=k+1}^{T} \mathbb{E}(\gamma_A(t)) \,.$$

Since from round $k + 2$ Bob is copying Alice's previous move, we have $\mathbb{E}(\gamma_B(t)) = \mathbb{E}(\gamma_A(t - 1))$ for all $t \geq k + 2$. Then Bob's total expected reward is

$$\mathbb{E}(\Gamma_B) = p \cdot (k + 2) + \sum_{t=k+2}^{T} \mathbb{E}(\gamma_B(t)) = p \cdot (k + 2) + \sum_{t=k+1}^{T-1} \mathbb{E}(\gamma_A(t)) \,.$$

If $k = T$, the difference in rewards is

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) = \mathbb{E}(\gamma_A(T)) - \mathbb{E}(\gamma_B(T)) = m - p < 0 \,.$$

If $k \leq T - 1$, by choice of $p > (m + M^*)/2$, the difference in rewards is bounded by

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) = (m - p) + \Big(\mathbb{E}(\gamma_A(T)) - p \Big) \leq m - 2p + \max\{M^*, p\} < 0 \,.$$

Thus Alice loses in expectation.

Part 2 of the statement holds since a single player (say Bob) optimizing his expected total reward has an algorithm with sublinear regret. Formally, the expected mean of the best arm is

$$\xi = p \cdot \mu(0, p) + \int_p^1 x \, d\mu(x) \ .$$

If $p < M^*$, then $\xi > p$. Bob's expected reward when playing optimally can be bounded from below by using a low regret algorithm (see, e.g., Bubeck and Cesa-Bianchi (2012)), which gives

$$\mathbb{E}(\Gamma_B) \geq \xi \cdot (T + 1) - C \cdot \sqrt{T + 1}, \text{ for some constant } C \geq 0$$

We want that

$$\xi \cdot (T + 1) - C \cdot \sqrt{T + 1} > p \cdot (T + 1)$$

This will hold if

$$T > \frac{C^2}{(\xi - p)^2}\,.$$

For such values of T, a regret minimizing algorithm will do better than playing left forever (in expectation). ■

Theorem 7 [Competing players are not completely myopic] (restated). *In the setting of Theorem 3, there exists a threshold $\widetilde{p} = \widetilde{p}(\mu, \beta, -1) > m$, such that for all $p < \widetilde{p}$, with probability 1 both players will explore arm R in the initial round of any optimal play. More precisely, define*

$$\widetilde{p} = \widetilde{p}(\mu, \beta, \lambda) = \inf\Big\{p : \textit{arm L is not explored in any Nash equilibrium}\Big\} \tag{12}$$

Proof Recall that $m = \int_0^1 x\, d\mu(x)$ and $w = \int_0^1 (x-m)^2\, d\mu(x)$. Let S_A and S_B denote strategies for Alice and Bob, respectively. Define

$$v_0^{(p)} = v_0 = \sup_{S_A} \inf_{S_B} \mathbb{E}(\Gamma_A(S_A, S_B) - \Gamma_B(S_A, S_B)).$$

We will show there is a threshold $p^* = p^*(m, w, \beta) > m$ such that for $p < p^*$ we have $v_0^{(p)} > 0$.

Take any mixed strategy of Alice in which she plays R in round zero and consider what happens from round one onwards. First recall that if Alice saw 1 in round zero, then the posterior mean is

$$m_1 = \frac{1}{m} \cdot \int_0^1 x^2 \cdot d\mu(x)\,.$$

If Alice saw 0, then the posterior mean is

$$m_0 = \frac{1}{1-m} \cdot \int_0^1 x(1-x) \cdot d\mu(x)\,.$$

Note that $m \cdot (m_1 - m) = (1-m) \cdot (m - m_0) = w$.

Let $v_t(x, y) = v_t^p(x, y)$ denote the net value for Alice (= maxmin over mixed strategies) from the beginning of round t, when in round 1 she plays x and Bob plays y. Here, since we will let Alice declare her strategy first, we have $y \in \{L, R\}$ (that is, $y = L$ means that Bob plays left in round 1 no matter what; similarly $y = R$ means that he plays right in round 1) and $x \in \{L, R, S\}$, where S is the following Alice round 1 strategy:

- Play left upon seeing 0 in round zero.

- Play right upon seeing 1 in round zero.

Write $\delta = p - m \geq 0$. Then $v_0(x, y) = -\delta + v_1(x, y)$. Let $\tilde{v}_t(x, y)$ be defined in the same way as $v_t(x, y)$ when Bob knows Alice's record in round zero. Clearly, $\tilde{v}_t(x, y) \leq v_t(x, y)$, for all strategies x, y.

By comparing the information available to the players based on the strategies played in rounds 0 and 1, observe:

- $v_2(x, L) \geq 0$ for all $x \in \{L, R, S\}$.

- $v_2(R, y) \geq 0$ for all $y \in \{L, R\}$.

Next we bound Alice's net value at round zero if Bob plays L and Alice plays strategy S:

$$v_0(S, L) = -\delta + \beta m(m_1 - p) + v_2(S, L) \geq -\delta + \beta m(m_1 - m + \delta) = \delta(\beta m - 1) + \beta w\,. \tag{13}$$

If Bob plays L in round one and Alice plays R in round one no matter what bit she observed at round zero, then we have:

$$v_0(R, L) = -\delta + \beta \cdot (-\delta) + v_2(R, L) \geq -2\delta\,. \tag{14}$$

If both players play right in round one we get

$$v_0(R, R) = -\delta + v_2(R, R)\,. \tag{15}$$

We establish a few facts about the net gains of the players.

Lemma 16 $\tilde{v}_2(L, R) = -v_2(R, R)$.

Proof Recall that by definition of $\tilde{v}_2(L, R)$, Bob knows two results from the right arm, while Alice knows only one. For $v_2(R, R)$ this is reversed, since in this case Alice knows two results from the right arm while Bob knows only one. ■

Lemma 17 $v_2(S, R) \geq \tilde{v}_2(S, R) \geq \tilde{v}_2(L, R) = -v_2(R, R)$.

Proof Note the first inequality holds since for $\tilde{v}_2(S, R)$ Alice tells Bob the bit she saw in round zero, while in the case of $v_2(S, R)$ she does not. The second inequality holds since by playing strategy S Alice has at least as much information as when playing L in round one, so she will do at least as well under S as she would do under L from round onwards. ■

By decomposing $v_0(S, R)$ into the payoff obtained in rounds zero, one, and the payoff from round two onwards, we obtain

$$\begin{aligned} v_2(R, R) + v_0(S, R) &\geq -\delta + (1 - m)\beta(p - m_0) \\ &= -\delta + \beta(1 - m)(m - m_0 + \delta) \\ &= \delta\left[\beta(1 - m) - 1\right] + \beta w \end{aligned} \tag{16}$$

Denote by $\frac{R+S}{2}$ the mixed strategy of Alice for round one in which she plays strategy R with probability $1/2$ and strategy S with probability $1/2$. By (15) and (16), we have

$$v_0\left(\frac{R+S}{2}, R\right) \geq \delta\left[\frac{\beta(1-m)}{2} - 1\right] + \frac{\beta w}{2} \geq \frac{\beta w}{2} - \delta \tag{17}$$

By (14) and (13), we obtain

$$v_0\left(\frac{R+S}{2}, L\right) \geq \delta\left[\frac{\beta m}{2} - 1\right] + \frac{\beta w}{2} \geq \frac{\beta w}{2} - \delta \tag{18}$$

Thus we can take $\tilde{p} = m + \beta w/2$, which completes the proof. ■

Appendix B. Cooperative Play

In this section we include the omitted proofs for the scenario of $\lambda = 1$, where the players are cooperating.

Theorem 11 [Cooperating players explore more, finite horizon] (restated). *Consider two cooperating players, Alice and Bob, playing a one armed bandit problem. The left arm is known and has probability p, while the right arm has a known distribution μ that is not a point mass. Let T be the horizon and M^* the maximum of the support of μ. Then there exists $\delta = \delta(T) > 0$ so that optimal players explore the right arm for all $p < g_T + \delta$.*

Proof Consider the same strategy for the players as in Theorem 5: Alice plays the one player optimal, while Bob plays left in rounds zero and one, and then from round two onwards he copies Alice's move from the previous round. The expected total reward of Alice is

$$\mathbb{E}(\Gamma_A) = m + \sum_{t=1}^{T} \mathbb{E}(\gamma_A(t)) \,.$$

Bob's expected reward is

$$\mathbb{E}(\Gamma_B) = 2p + \sum_{t=2}^{T} \mathbb{E}(\gamma_A(t-1)) = 2p + \sum_{t=1}^{T-1} \mathbb{E}(\gamma_A(t)) \,.$$

On the other hand, the expected total reward of the players if they never explore is $2p(T+1)$. The difference between the two strategies is

$$\Delta = \mathbb{E}(\Gamma_A) + \mathbb{E}(\Gamma_B) - 2p(T+1) = 2p + 2 \cdot \sum_{t=0}^{T} \mathbb{E}(\gamma_A(t)) - m - \mathbb{E}(\gamma_A(T)) - 2p(T+1) \,.$$

Given that Alice is playing her single player optimal strategy, we have that for all $p \geq g_T$ her total reward is bounded as follows: $\sum_{t=0}^{T} \mathbb{E}(\gamma_A(t)) \geq g_T \cdot (T+1)$. Then

$$\Delta \geq 2p + 2g_T \cdot (T+1) - m - \mathbb{E}(\gamma_A(T)) - 2p(T+1) = 2g_T \cdot (T+1) - m - \mathbb{E}(\gamma_A(T)) - 2pT \,.$$

Taking $p = g_T + \delta$ for $\delta \geq 0$, the previous inequality is equivalent to

$$\Delta \geq 2g_T - m - \mathbb{E}(\Gamma_A(T)) - 2\delta T \,.$$

For Δ to be strictly positive it suffices that

$$\mathbb{E}(\gamma_A(T)) < 2g_T - m - 2\delta T \tag{19}$$

Recall that $\lim_{T \to \infty} g_T = M^*$ (Remark 6). Moreover, Alice's expected reward in the last round satisfies $\mathbb{E}(\gamma_A(T)) \leq M^*$ when $p < M^*$.

Let $\alpha = (M^* - m)/4$. Then there exists $T_0 = T_0(\alpha)$ so that for all $T \geq T_0$, we have $|g_T - M^*| < \alpha$. For all such T, by choice of α we have that

$$2g_T - m - \mathbb{E}(\gamma_A(T)) > 2(M^* - \alpha) - m - M^* = \frac{M^* - m}{2} > 0$$

Then inequality (19) holds for all $\delta < (M^* - m)/(4T)$ as required. ■

Proposition 12 (restated). *There is a threshold $p^\circ < M^*$ so that cooperating players do not explore for any $p > p^\circ$.*

Proof When $m < p$, the best case scenario for the cooperating players is that one of them (say Alice) plays right in round zero while Bob stays at left, and then from round one onwards both play right and the mean of the right arm is M^* from round one onwards. Then any strategies S_A and S_B of the cooperating players will give at most this total reward, so:

$$\mathbb{E}(\Gamma(S_A)) + \mathbb{E}(\Gamma(S_B)) \leq p + m + \frac{2\beta M^*}{1-\beta}$$

The total reward of both players staying at the left arm is $2p/(1-\beta)$ which is more than any strategies involving exploration when

$$p + m + \frac{2\beta M^*}{1-\beta} < \frac{2p}{1-\beta} \iff p > \frac{(1-\beta)m + 2\beta M^*}{1+\beta}$$

Setting $p^\circ = \left((1-\beta)m + 2\beta M^*\right)/(1+\beta)$ gives the required statement. ■

Appendix C. Neutral Play

In this section we include the omitted proofs for the scenario of $\lambda = 0$, where the players are neutral, each maximizing their own rewards.

Observation 2 (restated). *For all $p < g(\mu, \beta)$, in any Nash equilibrium, each player explores with strictly positive probability.*

Proof Suppose there is a Nash equilibrium with strategies (S_A, S_B), in which one player - say Alice - explores with probability zero. Then $\mathbb{E}(\Gamma_A(S_A, S_B)) = p/(1-\beta)$.

Consider now the modified Alice strategy S'_A of pulling the arm with the highest Gittins index in each round (ignoring any information from Bob). Then since $p < g(\mu, \beta)$, there is $\alpha \in (p, g)$ so that $\mathbb{E}(\Gamma_A(S'_A, S_B)) = \alpha/(1-\beta)$ This is strictly higher than $p/(1-\beta)$, so S'_A is an improving deviation, in contradiction with (S_A, S_B) being an equilibrium. Then Alice explores with strictly positive probability. ■

Theorem 13 [Neutral players learn from each other, discounted] (restated). *Consider two neutral players, Alice and Bob, playing a one armed bandit problem with discount factor $\beta \in (0, 1)$. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Then in any Nash equilibrium:*

1. *For all $p < g(\mu, \beta)$, with probability 1 at least one player explores. Moreover, the probability that no player explores by time t decays exponentially in t.*
2. *Suppose $p \in (p^*, g)$, where p^* is the threshold above which competing players do not explore [3]. If the equilibrium is furthermore perfect Bayesian, then every (neutral) player has expected reward strictly higher than a single player using an optimal strategy.*

[3] For the formal definition of p^*, see Theorem 3.

Proof Let $\alpha \in (p, g)$ denote the expected reward per round of a single player that follows the strategy of pulling the arm with the highest Gittins index in each round.

Part 1: Assume $p < g = g(\mu, \beta)$. Since $p < \alpha$, there exists $\epsilon > 0$ so that $p + \epsilon < \alpha$. Let $\Lambda = \alpha/(1-\beta)$ denote the expected total reward achievable by a single optimal player. Fix a pair of strategies (S_A, S_B) that define a Nash equilibrium. Let $\Phi_k = \mathbb{P}(D_k^c)$, where

$$D_k = \Big\{\text{No player explores in rounds } \ \{0, \dots, k-1\} \ \text{ under strategies } \ (S_A, S_B)\,.\Big\}$$

Since (S_A, S_B) is an equilibrium, we can bound the expected reward of a single player by

$$\Lambda \le (1 - \Phi_k) \left[\sum_{t=0}^{k-1} p \cdot \beta^t + \sum_{t=k}^{\infty} 1 \cdot \beta^t \right]$$

Rearranging the terms gives

$$\Lambda \le \sum_{t=0}^{k-1} \beta^t \Big[(1 - \Phi_k)p + \Phi_k \Big] + \sum_{t=k}^{\infty} \beta^t\,.$$

Expanding the terms in the right hand side, we get a bound on α:

$$\alpha \le (1 - \beta^k) \Big[p + \Phi_k (1-p) \Big] + \beta^k \implies \frac{\alpha - p}{1-p} - \beta^k \le \Phi_k\,.$$

Choose k so that $\beta^k \le (\alpha - p)/2$. Then

$$\Phi_k \ge \frac{\alpha - p}{2} \implies \mathbb{P}(D_k) \le 1 - \frac{\alpha - p}{2}\,.$$

The same argument gives that for every $\ell \ge 0$ we have $\mathbb{P}(D_{k(\ell+1)} \mid D_{k\ell}) \le 1 - (\alpha - p)/2$, so inductively we obtain $\mathbb{P}(D_{k\ell}) \le \Big(1 - (\alpha - p)/2\Big)^{\ell}$. This implies the required bound for D_t:

$$\mathbb{P}(D_t) \le \left(1 - \frac{\alpha - p}{2}\right)^{\lfloor t/k \rfloor}\,.$$

Part 2: Assume that $p \in (p^*, g)$, where p^* is the threshold from equation (2) such that competing players do not explore for any $p > p^*$. In any perfect Bayesian equilibrium, the total expected reward of each player is at least $\alpha/(1-\beta)$, i.e. the single player optimum.

Suppose towards a contradiction that there was a perfect Bayesian equilibrium with strategies (S_A, S_B) where at least one of the players - say Bob - was getting exactly the expected reward of a single optimal player: $\mathbb{E}(\Gamma_B) = \alpha/(1-\beta)$. Note that Alice must explore with positive probability regardless of what Bob does, since the single player optimum does better than never exploring. Consider now the following strategy S_B' for Bob: stay left in every round until Alice explores for the first time. If Alice's exploration occurs in some round k, then from round $k+1$ onwards, let Bob use the optimal strategy that he would play when competing against Alice in the corresponding zero-sum game that starts at round k (where Alice is forced to play right and Bob left in round k, and both play optimally afterwards). By Theorem 3, if Alice does explore the right arm at k, then when $p \in (p^*, g)$ Bob wins against Alice.

Let Γ'_A and Γ'_B be the total rewards of the players under strategies (S_A, S'_B). Since the equilibrium is perfect Bayesian, we have $\mathbb{E}(\Gamma'_A) \geq \alpha/(1-\beta)$. Moreover, since Alice is not learning anything from Bob's strategy S'_B, she can only realize this minimum expected value by using an optimal strategy for a single player, which requires exploring with positive probability.

We classify the trajectories realizable under strategies (S_A, S'_B) in two types, depending on whether Alice explores the right arm or not on that trajectory:

1. Alice does not explore: then Bob always plays left too, so they get the same reward.

2. Alice explores: then Bob's expected reward is strictly greater than Alice's total reward on that trajectory (since he wins in the corresponding zero sum game that starts with Alice's first exploration).

Taking expectation over all possible trajectories and noting that Alice explores with positive probability, we get that Bob's deviation S'_B ensures he gets expected total reward strictly greater than Alice, so $\mathbb{E}(\Gamma'_B) > \mathbb{E}(\Gamma'_A) \geq \alpha/(1-\beta) = \mathbb{E}(\Gamma_B)$. Thus deviation S'_B is profitable, in contradiction with (S_A, S_B) being an equilibrium. Then both players get strictly more than the single player optimum. ■

Example 2 (Perfect Bayesian vs. Nash equilibrium) *There is a Nash equilibrium in which both players follow the same trajectory and play the one player optimal strategy of pulling the arm with the highest Gittins index in each round.*

To see why this is the case, define the strategies of the players as follows:

- *Pull the arm with the highest Gittins index in each round, breaking ties in favor of the left arm if both arms have the same index, as long as the other player did the same in the previous round.*

- *If in some round k one of the players, say Alice, chooses a different action, then Bob switches to playing left forever from round $k+1$ onwards (and similarly if Bob deviates).*

Note that with these strategies, if Alice deviates in some round k, then she cannot learn anything from Bob in rounds $\{k+1, k+2, \ldots\}$. Alice also did not learn anything from Bob in rounds 0 to $k-1$ since they had the same trajectory and in round k since she cannot see Bob's reward in round k and his reward at k does not affect his subsequent actions. Then Alice's expected reward overall cannot be better than the single player optimum, so the deviation is not strictly improving. Thus the strategies form a Nash equilibrium.

Theorem 15 [Neutral players learn from each other, finite horizon] (restated). *Consider two neutral players, Alice and Bob, playing a one armed bandit problem with horizon T. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Then in any subgame perfect equilibrium:*

1. *For all $p < M^*$, with probability that converges to 1 as T grows, at least one player explores. Moreover, the probability that no player explores by time t decays polynomially in t.*

2. *Suppose $(m + M^*)/2 < p < M^*$. Then every (neutral) player has expected reward strictly higher than a single optimal player.*

Proof Recall g_T denotes the index of the right arm. Let $\alpha \cdot (T+1)$ be the expected reward of an optimal single player, where $\alpha \in (p, g_T)$.

Part 1: Let $p < M^*$. We show there exist $\psi = \psi(\mu, p) > 0$ and $C < \infty$ so that the inequality

$$\mathbb{P}\Big(\text{No player explores in rounds } \{0, \ldots, T\} \text{ under } (S_A, S_B)\Big) \leq C \cdot T^{-\psi}$$

holds for all Nash equilibria (S_A, S_B).

First, since $p < M^*$, there exist T_1 and $\alpha > p$ so that for all $T \geq T_1$, the single player optimum over rounds $\{0, \ldots, T\}$ is in expectation at least $\alpha \cdot (T+1)$.

For a sequence $\{T_j\}$ monotonically decreasing (to be specified later) and for $T > T_1$, write $D_j = \Big\{\text{No player explores in rounds } \{0, \ldots, T\} \text{ under } (S_A, S_B)\Big\}$.

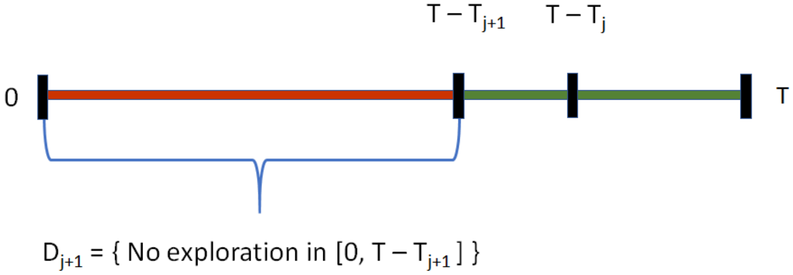

Figure 2: Depiction of the intervals induced by the sequence T_j and the events D_j.

Fix j and let $k = T_{j+1} - T_j$. A depiction of the intervals induced by the sequence T_j is given in Figure 2. Write $\Phi_j = \mathbb{P}(D_j^c \mid D_{j+1})$. Since (S_A, S_B) is a Nash equilibrium,

$$\alpha \cdot (T_j + 1) \leq (1 - \Phi_j) \cdot \Big[p_k + T_j + 1\Big] + \Phi_j \cdot \Big[T_{j+1} + 1\Big]$$
$$\alpha \cdot (k + T_j + 1) \leq T_j + 1 + \Phi_j \cdot (1-p) \cdot k + p \cdot k\,.$$

This implies the inequality:

$$\Phi_j \cdot (1-p) \cdot k \geq (\alpha - p)k - (1-\alpha)(T_j + 1)\,.$$

Choose k such that $(\alpha - p) \cdot k \in \Big[T_j + 1, T_j + 2\Big]$. We obtain

$$0 \leq T_{j+1} - \left(1 + \frac{1}{\alpha - p}\right) T_j \leq \frac{2}{\alpha - p} \tag{20}$$

Then

$$\alpha \cdot T_j \leq k \cdot \Phi_j \leq \Phi_j \cdot \left(\frac{T_j + 2}{\alpha - p}\right) \implies \Phi_j \geq \frac{\alpha(\alpha - p)T_j}{T_j + 2} \geq \frac{\alpha(\alpha - p)}{2}\,.$$

Now suppose $T \in (T_\ell, T_{\ell+1})$. Then

$$\mathbb{P}(D_{\ell-1}) \leq \mathbb{P}(D_{\ell-1} | D_\ell) \leq 1 - \frac{\alpha(\alpha - p)}{2} \implies \mathbb{P}(D_\ell) \leq \left(1 - \frac{\alpha(\alpha - p)}{2}\right)^{\ell - 1} \tag{21}$$

Since $\{T_j\}$ grows exponentially, by (20), the claim follows from (21).

Part 2: The proof is similar to the discounted case, so we give a sketch highlighting the differences. Let (S_A, S_B) be a perfect Bayesian equilibrium where Bob gets exactly the single player optimum: $\mathbb{E}(\Gamma_B) = \alpha \cdot (T+1)$. Since the horizon T is large, Alice explores with positive probability (see part 2 of Theorem 5).

Consider deviation S'_B for Bob: stay left in every round until Alice explores for the first time (e.g. in round k), and from round $k+1$ onwards, use the optimal zero-sum strategy. Let Γ'_A and Γ'_B be the total rewards under (S_A, S'_B). Since the equilibrium is perfect Bayesian, $\mathbb{E}(\Gamma'_A) \geq \alpha \cdot (T+1)$. Alice is not learning from Bob under S'_B, so she must explore with positive probability. We classify the trajectories realizable under (S_A, S'_B) in two types:

1. Alice does not explore: then Bob always plays left too, so they get the same total reward.
2. Alice explores: then by Theorem 5, Bob gets strictly more than Alice for $p > (m + M^*)/2$.

Taking expectation over all possible trajectories and noting that Alice explores with positive probability, it follows that Bob's deviation S'_B ensures he gets expected total reward strictly greater than Alice. Then we get $\mathbb{E}(\Gamma'_B) > \mathbb{E}(\Gamma'_A) \geq \alpha \cdot (T+1) = \mathbb{E}(\Gamma_B)$. Then Bob's deviation S'_B is profitable, in contradiction with the choice of equilibrium strategies (S_A, S_B). ■

Appendix D. Long Term Behavior

We show that in every Nash equilibrium, competing and neutral players converge to playing the same arm forever from some point on. For $\lambda \notin [-1, 0]$, this can fail.

D.1. Nash Equilibria with Oscillations

In this section we construct Nash equilibria with oscillations when $\lambda > 0$ and $\lambda < -1$. The next example shows that for $\lambda > 0$ there cannot be a general theorem that the players settle on the same arm in every Nash equilibrium. The construction shows that for every $\lambda > 0$ and every sufficiently large discount factor $\beta < 1$, there is a Nash equilibrium in which one of the players alternates between the two arms infinitely often.

Proposition 18 (Nash equilibria where players do not converge, $\lambda > 0$) *Suppose Alice and Bob are playing a one-armed bandit problem with discount factor β, where the left arm has success probability p and the right arm has prior distribution μ that is a point mass at $m > p$.*

Then for every $\lambda > 0$, for any discount factor $\beta > 1/(1+\lambda)$, there is a Nash equilibrium in which Bob visits both arms infinitely often.

Proof Let $k \in \mathbb{N}$. Define strategies (S_A, S_B) by:

1. Bob's strategy S_B is to visit the left arm in rounds $0, k, 2k, 3k, \ldots$, and the right arm in the remaining rounds, no matter what Alice does.
2. Alice's strategy S_A is to stay at the right arm if Bob follows the trajectory above. If Bob ever deviates from S_B for the first time in some round ℓ, then Alice plays left forever starting with round $\ell + 1$.

To show the strategies are in Nash equilibrium, consider first the total rewards Γ_A and Γ_B obtained by Alice and Bob, respectively, on the main path under strategy pair (S_A, S_B) starting from round zero:

$$\Gamma_A = \frac{m}{1-\beta} \quad \text{and} \quad \Gamma_B = \frac{m}{1-\beta} - \frac{m-p}{1-\beta^k}$$

It is clear that Alice has no incentive to deviate, since her deviation does not change Bob's behavior and adding any round of playing the left arm only worsens her own total rewards. Consider next any strategy S'_B of Bob that deviates from S_B in some round ℓ. Observe first that the best case deviation for Bob is when he deviates in some round ℓ divisible by k so that instead of playing left in the next round he plays right. Then w.l.o.g. $\ell = 0$. The expected total rewards under (S_A, S'_B) starting from round zero are

$$\Gamma'_A = m + \beta \cdot \frac{p}{1-\beta} \quad \text{and} \quad \Gamma'_B \leq \frac{m}{1-\beta}$$

Then Bob's utility under strategy pairs (S_A, S_B) and (S_A, S'_B), respectively, is

$$u_B = \frac{(1+\lambda) \cdot m}{1-\beta} - \frac{m-p}{1-\beta^k} \quad \text{and} \quad u'_B \leq \frac{m}{1-\beta} + \lambda \left(m + \beta \cdot \frac{p}{1-\beta} \right)$$

If $\beta > 1/(1+\lambda)$ and k is large enough, then $1 - \beta < \beta \cdot \lambda \cdot (1 - \beta^k)$. This implies that $u'_B < u_B$, so (S_A, S_B) is indeed a Nash equilibrium. ■

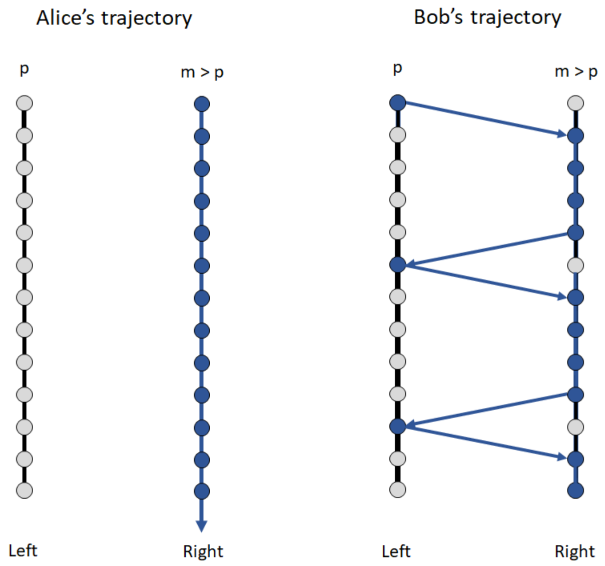

Figure 3: Trajectories of the players on the main line under strategies (S_A, S_B) for $\lambda > 0$ and $k = 5$. The circles represent time units $0, 1, 2, \ldots$. The left arm has success probability p and the right arm has prior distribution which is a point mass at $m > p$.

Note this Nash equilibrium is not Pareto optimal: both players could improve their utility by a joint deviation to always playing right. This Nash equilibrium is also not subgame perfect because it involves a non-credible threat by Alice (of playing left forever after any deviation by Bob).

Remark 19 *In the example discussed above, if $\beta < 1/(1+\lambda)$, then the only Nash equilibrium is where both players always play right.*

Proposition 18 does not preclude the possibility that the players settle on the same arm in every perfect Bayesian equilibrium when $\lambda > 0$.

For every $\lambda < -1$, there are perfect Bayesian equilibria where the players do not settle on the same arm in the long term (for some choice of μ and β), and where a player oscillates infinitely often between the arms.

Proposition 20 (Perfect equilibria where competing players do not converge, $\lambda < -1$) *Suppose Alice and Bob are playing a one-armed bandit problem with discount factor β, where the left arm has success probability p and the right arm has prior distribution μ which is a point mass at $m > p$. For every $\lambda < -1$, if the discount factor satisfies $\beta^2 \cdot |\lambda| > 1$, then there is a perfect Bayesian equilibrium in which Bob visits both arms infinitely often.*

Note in this case there is no uncertainty since both arms are known, so perfect Bayesian equilibria coincide with subgame perfect equilibria.

Proof [Proof of Proposition 20] Suppose Alice fixes a trajectory τ for Bob that requires him to play right in rounds $0, k, 2k, 3k, \ldots$, and to play left in all other rounds. Let S_A be the Alice strategy of playing left as long as Bob follows trajectory τ, while if Bob ever deviates from τ in some round s, then S_A switches to playing right forever from round $s+1$ onwards. Let S_B be the Bob strategy of following trajectory τ, unless Alice ever deviates from playing left in some round s; in that case, S_B switches to playing right forever from round $s+1$ onwards. Whenever the players find themselves off the main line given by this prescription, they just play right forever.

We claim that (S_A, S_B) represent a subgame perfect equilibrium. The expected total rewards of the players under (S_A, S_B) on the main line of play are:

$$\mathbb{E}(\Gamma_A(S_A, S_B)) = \frac{p}{1-\beta} \text{ and } \mathbb{E}(\Gamma_B(S_A, S_B)) = \frac{m-p}{1-\beta^k} + \frac{p}{1-\beta} \tag{22}$$

Then the expected utilities are:

$$\begin{aligned}\mathbb{E}(u_A(S_A, S_B)) &= (1+\lambda) \cdot \frac{p}{1-\beta} + \lambda \cdot \frac{m-p}{1-\beta^k} \\ \mathbb{E}(u_B(S_A, S_B)) &= (1+\lambda) \cdot \frac{p}{1-\beta} + \frac{m-p}{1-\beta^k}\end{aligned} \tag{23}$$

To check the equilibrium property, it will suffice to compare the utility each player i gets when following S_i with the utility obtained when deviating to playing left forever or to playing right forever. Note the response of the other player to any deviation is to switch to playing right forever, thus if there is an improving deviation S_i' for player i, then there is an improving deviation S_i'' in which i plays a fixed arm forever. Moreover, since the right arm is always better, that deviation will be to play right forever.

If Bob switches to playing right from some round on, then the highest gain can be obtained when the switch takes place from round 1 onwards. In this case, Alice will play right forever from round 2. Thus

$$\mathbb{E}(\Gamma_A(S_A, Right)) = p + \beta p + \beta^2 \cdot \frac{m}{1-\beta} \text{ and } \mathbb{E}(\Gamma_B(S_A, Right)) = \frac{m}{1-\beta} \tag{24}$$

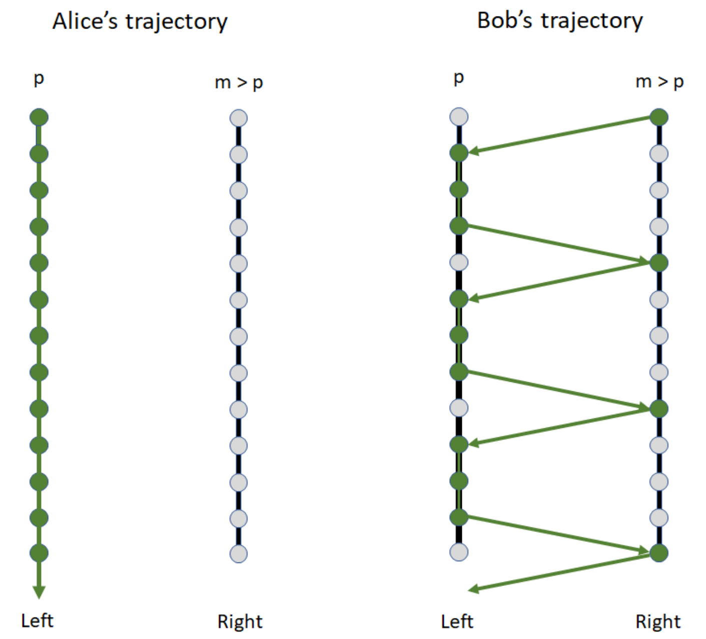

Figure 4: Trajectories of the players on the main line under strategies (S_A, S_B) for $\lambda < -1$ and $k = 4$. The circles represent time units $0, 1, 2, \ldots$.

Bob's expected utility is

$$\mathbb{E}(u_B(S_A, Right)) = (1+\lambda)\cdot\frac{m}{1-\beta} + \lambda\cdot[(p-m)(1+\beta)] \tag{25}$$

From equations (25) and (23), we get that $\mathbb{E}(u_B(S_A, Right)) < \mathbb{E}(u_B(S_A, S_B))$ whenever

$$1+\beta+\ldots+\beta^{k-1} > \frac{1+\lambda(1+\beta)}{1+\lambda} = \frac{|\lambda|\cdot(1+\beta)-1}{|\lambda|-1} \tag{26}$$

Such a k exists whenever $|\lambda|\beta > 1$.

For Alice the best time to deviate is just before S_B tells Bob to play right (in round $k-1$).

$$\mathbb{E}(u'_A(k-1,\infty)) = \frac{m}{1-\beta} + \lambda\cdot\left[p+\beta\cdot\frac{m}{1-\beta}\right] \tag{27}$$

On the other hand, under strategies (S_A, S_B), Alice's utility from round i on is

$$\mathbb{E}(u_A(k-1,\infty)) = \frac{p}{1-\beta} + \lambda p + \lambda\beta\cdot\left(\frac{m-p}{1-\beta^k} + \frac{p}{1-\beta}\right) \tag{28}$$

Comparing Alice's utility before and after the deviation, we obtain

$$\begin{aligned} \mathbb{E}(u'_A(k-1,\infty)) &< \mathbb{E}(u_A(k-1,\infty)) \\ \frac{m-p}{1-\beta} + \frac{\lambda\beta(m-p)}{1-\beta} &< \frac{\lambda\beta(m-p)}{1-\beta^k} \iff \\ 1+\lambda\beta &< \lambda\beta \cdot \frac{1-\beta}{1-\beta^k} = \frac{\lambda\beta}{1+\beta+\ldots+\beta^{k-1}} \iff \\ 1+\beta+\ldots+\beta^{k-1} &> \frac{|\lambda|\beta}{|\lambda|\beta-1} \end{aligned} \tag{29}$$

This holds whenever $|\lambda|\beta^2 > 1$ and k is large enough. ■

D.2. A Concentration Lemma

Lemma 21 *Suppose Alice is playing a one armed bandit problem, where the left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Let R_∞ denote Alice's total number of pulls of the right arm, R_n the number of pulls by time n, and w_n the number of successes in the first $\min\{n, R_\infty\}$ draws by Alice.*

Then for each $\epsilon > 0$ and $k \in \mathbb{N}$, we have

$$\mathbb{P}\Big(\big(|w_k - k\Theta| > k\epsilon\big) \cap \big(R_\infty \geq k\big)\Big) \leq 2e^{-2k\epsilon^2} \tag{30}$$

Proof Recall we are working in the probability space where Θ is picked according to μ and every toss of the risky arm is a Bernoulli variable with parameter Θ.

The goal is to show the next inequality holds for every k:

$$\mathbb{P}\Big(\big(|w_k - k\Theta| > k\epsilon\big) \cap \big(R_\infty \geq k\big) \mid \Theta\Big) \leq 2e^{-2k\epsilon^2} \tag{31}$$

Fix a value of k. It will be enough to show that the following inequality holds for every n:

$$\mathbb{P}\Big(\big(|w_k - k\Theta| > k\epsilon\big) \cap \big(R_n \geq k\big) \mid \Theta\Big) \leq 2e^{-2k\epsilon^2} \tag{32}$$

We define an auxiliary variable $\widetilde{w}_k$ to represent the number of successes in the first k pulls, but where the pulls are padded in case their number is too low. That is, if $R_n \geq k$, then $\widetilde{w}_k = w_k$. Otherwise, if $R_n < k$, then add another $k - R_n$ pulls after time n and let $\widetilde{w}_k$ denote the number of successes in the first k pulls defined this way. Note the containment

$$\Big(\big(|w_k - k\Theta| > k\epsilon\big) \cap \big(R_n \geq k\big)\Big) \subseteq \Big(|\widetilde{w}_k - k \cdot \Theta| > k\epsilon\Big).$$

Using Hoeffding's inequality applied to the k pulls of the risky arm, we have

$$\mathbb{P}\Big(|\widetilde{w}_k - k \cdot \Theta| > k\epsilon \mid \Theta\Big) \leq 2e^{-2k\epsilon^2} \tag{33}$$

This implies inequality (32), and so (31) also holds. Taking expectation in (31) implies the required inequality (30). ■

D.3. Neutral Players in the Long Term

Theorem 22 (Neutral players eventually settle on the same arm) *Consider two neutral players, Alice and Bob, playing a one armed bandit problem with discount factor $\beta \in (0,1)$. The left arm has success probability p and the right arm has prior distribution μ such that $\mu(p) = 0$. Then in any Nash equilibrium, with probability 1 the players eventually settle on the same arm.*

Proof [Proof of Theorem 22] Let (S_A, S_B) be an arbitrary pair of strategies in Nash equilibrium.

For each round n, let H_n^i be the history from player i's point of view up to round n; this contains the actions of both players until the end of round $n-1$, the rewards of player i in the first $n-1$ rounds, and the action prescribed to player i by strategy S_i for round n. Let N be a stopping time for player i.

Let $\widehat{\Theta}_k^i$ be the empirical mean of the success probability at the right arm as observed by player i after k explorations, defined if $R_\infty^i \geq k$, where R_∞^i is player i's total number of pulls of the right arm.

Let $\epsilon \in (0, 1/2)$. Pick $\delta \in (0, \epsilon)$ such that $\mu([p - 3\delta, p + 3\delta]) < \epsilon$. Let $\delta_1 \in (0, \delta)$ be such that $\mu([p - 3\delta_1, p + 3\delta_1]) \leq \epsilon^2\delta/2$.

We start by showing that if Alice explores at least k times, for k large enough, then she will settle afterwards on the better of the two arms with high probability. Then we will study Bob's behavior and show that he will also pick the better of the two arms if at least one of the players explores k times.

We bound Alice's expected value for pulling the right arm in the future by considering three cases, depending on the value of $\widehat{\Theta}_k^A$.

Case 1: $\widehat{\Theta}_k^A \leq p - 2\delta$. The goal will be to show that the following event has small probability

$$\widetilde{D}_1 = \{(R_\infty^A \geq k+1) \cap (\widehat{\Theta}_k^A \leq p - 2\delta)\}.$$

Let N be the (random) time of the $k+1$-st exploration by Alice, if it exists, and otherwise $N = \infty$. Define the random variable

$$\Psi_k^A = \mathbb{P}\Big((|w_k - k\Theta| > k\delta) \cap (R_\infty^A \geq k+1) \mid H_N^A\Big) \tag{34}$$

Using Lemma 21, we get $\mathbb{E}(\Psi_k^A) \leq 2e^{-2k\delta^2}$. From this we deduce by Markov's inequality that

$$\mathbb{P}(\Psi_k^A \geq e^{-k\delta^2}) \leq 2e^{-k\delta^2} \tag{35}$$

We define the event $D_1 = \{\Psi_k^A \leq e^{-k\delta^2}\} \cap \widetilde{D}_1$. By Lemma 21, we have

$$\mathbb{P}(\widetilde{D}_1 \setminus D_1) \leq 2e^{-k\delta^2} \tag{36}$$

If $|\widehat{\Theta}_k^A - \Theta| \leq \delta$ and D_1 holds, then $\Theta \leq p - \delta$. Let $\Gamma_A(N, \infty) = \sum_{j=0}^{\infty} \beta^j \cdot \gamma_A(N+j)$ denote Alice's total normalized reward starting from round N under strategy S_A. Then Alice's total reward satisfies:

$$\mathbb{1}_{D_1}\, \mathbb{E}\Big[\Gamma_A(N,\infty) \mid H_N^A, \Theta\Big] \leq \mathbb{1}_{D_1}\Big[\mathbb{1}_{|\Theta - \widehat{\Theta}_k^A| \leq \delta}\left(p - \delta + \frac{p\beta}{1-\beta}\right) + \mathbb{1}_{|\Theta - \widehat{\Theta}_k^A| > \delta}\frac{1}{1-\beta}\Big] \tag{37}$$

Taking expectation over Θ gives (since $D_1 \subseteq \{R^A_\infty \geq k+1\}$):

$$\begin{aligned}\mathbb{1}_{D_1}\,\mathbb{E}\Big[\Gamma_A(N,\infty) \mid H^A_N\Big] &\leq \mathbb{1}_{D_1}\Big[p - \delta + \frac{p\beta}{1-\beta} + \mathbb{P}\Big(\{R^A_\infty \geq k+1\} \cap |\Theta - \widehat{\Theta}^A_k| > \delta \mid H^A_N\Big) \cdot \frac{1}{1-\beta}\Big] \\ &= \mathbb{1}_{D_1}\Big[\frac{p}{1-\beta} - \delta + \Psi^A_k \cdot \frac{1}{1-\beta}\Big] \\ &\leq \mathbb{1}_{D_1}\Big[\frac{p}{1-\beta} - \delta + \frac{e^{-k\delta^2}}{1-\beta}\Big] \qquad (38)\end{aligned}$$

Consider an alternative strategy S'_A as follows: play S_A until time N when it is about to do its $k+1$-st exploration. If $\widehat{\Theta}^A_k \leq p - 2\delta$ and $\Psi^A_k \leq e^{-k\delta^2}$, then play left forever, otherwise continue with S_A. Note that S'_A differs from S_A only on the event D_1.

Using inequality (38), on the event D_1 we have that for all $k > \delta^{-2} \cdot \log 1/(\delta(1-\beta))$:

$$\mathbb{E}\Big[\Gamma_A(N,\infty) - \Gamma'_A(N,\infty) \mid H^A_N\Big] \leq \Big(\frac{e^{-k\delta^2}}{1-\beta} - \delta\Big) < 0 \qquad (39)$$

Since the original strategies (S_A, S_B) are in equilibrium, we have

$$\begin{aligned}0 \leq \mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma'_A) &= \mathbb{E}\Big[\beta^N \Gamma_A(N,\infty) - \beta^N \Gamma'_A(N,\infty)\Big] \\ &= \mathbb{E}\Big[\beta^N \mathbb{1}_{D_1}\,\mathbb{E}\Big(\Gamma_A(N,\infty) - \Gamma_{A'}(N,\infty) \mid H^A_N\Big)\Big] \qquad (40)\end{aligned}$$

This forces $\mathbb{P}(D_1) = 0$ by (39). By inequality (36), we get that $\mathbb{P}(\widetilde{D}_1) \leq 2e^{-k\delta^2}$.

Case 2: $\widehat{\Theta}^A_k \geq p + 2\delta$. Define M as the first round after the k-th exploration at which Alice does not explore. If there is no such time, then $M = \infty$. The goal will be to bound the probability of the event

$$\widetilde{D}_2 = \{(M < \infty) \cap (\widehat{\Theta}^A_k \geq p + 2\delta)\} \qquad (41)$$

Recall H^A_M is the history from Alice's point of view until the beginning of round M and Φ^A_k as the random variable

$$\Phi^A_k = \mathbb{P}\Big((|w_k - k\Theta| > k\epsilon) \cap (M < \infty) \mid H^A_M\Big) \qquad (42)$$

Using Lemma 21, we get $\mathbb{E}(\Phi^A_k) \leq 2e^{-2k\epsilon^2}$. From this we deduce by Markov's inequality that

$$\mathbb{P}(\Phi^A_k \geq e^{-k\epsilon^2}) \leq 2e^{-k\epsilon^2} \qquad (43)$$

Define the event $D_2 = \{\Phi^A_k \leq e^{-k\delta^2}\} \cap \widetilde{D}_2$. By Lemma 21, we have $\mathbb{P}(\widetilde{D}_2 \setminus D_2) \leq 2e^{-k\delta^2}$.

Consider Alice's strategy S_A. If M is finite, then at round M Alice is playing left by definition. We will upper bound Alice's total reward from round M onwards and compare it to the expected reward obtained by playing right in all rounds $t \geq M$. The following inequality is immediate: $\max\{\Theta, p\} \leq \Theta + p \cdot \mathbb{1}_{\Theta < p}$. By taking expectation and conditioning on the history, on the event D_2

we get $\mathbb{E}(\mathbb{1}_{\Theta<p} \mid H_M^A) \leq e^{-k\delta^2}$. Then on the event D_2, Alice's expected total reward from round M onwards under strategy S_A satisfies

$$\begin{aligned}\mathbb{E}\Big[\Gamma_A(M,\infty) \mid H_M^A\Big] &\leq p + \frac{\beta}{1-\beta}\cdot\mathbb{E}(\Theta \mid H_M^A) + \frac{\beta\cdot p}{1-\beta}\cdot\mathbb{P}\Big((\Theta<p)\cap(M<\infty) \mid H_M^A\Big)\\ &\leq p + \frac{\beta}{1-\beta}\cdot\mathbb{E}(\Theta \mid H_M^A) + \frac{\beta\cdot p}{1-\beta}\cdot\Phi_k^A \qquad (44)\end{aligned}$$

On the other hand, on the event D_2, by playing right forever starting with round M, Alice's expected reward is Γ'_A, which has expectation:

$$\mathbb{E}\Big[\Gamma'_A(M,\infty) \mid H_M^A\Big] = \frac{\mathbb{E}(\Theta \mid H_M^A)}{1-\beta} \qquad (45)$$

If $\widehat{\Theta}_k^A - \Theta \leq \delta$ and the event D_2 holds, then $\Theta \geq p+\delta$. On the event D_2 we have

$$\mathbb{E}\Big[\Theta \mid H_M^A\Big] \geq p + \delta - e^{-k\delta^2} \qquad (46)$$

By combining inequalities (44), (45), and (46), we get that on the event D_2 the following inequalities hold for all k large enough so that $e^{-k\delta^2} < \delta(1-\beta)$.

$$\begin{aligned}\mathbb{E}\Big[\Gamma_A(M,\infty) \mid H_M^A\Big] - \mathbb{E}\Big[\Gamma'_A(M,\infty) \mid H_M^A\Big] &\leq p + \frac{\beta\cdot p}{1-\beta}\cdot\Phi_k^A + \frac{\beta}{1-\beta}\cdot\mathbb{E}(\Theta \mid H_M^A) - \frac{\mathbb{E}(\Theta \mid H_M^A)}{1-\beta}\\ &\leq p + \frac{\beta p}{1-\beta}\cdot e^{-k\delta^2} - \mathbb{E}(\Theta \mid H_M^A)\\ &\leq p + \frac{\beta}{1-\beta}\cdot e^{-k\delta^2} - p - \delta + e^{-k\delta^2}\\ &= \frac{e^{-k\delta^2}}{1-\beta} - \delta < 0 \qquad (47)\end{aligned}$$

Consider the alternative strategy S'_A defined as follows: play S_A until time M when it is about to play left. If $\Phi_k^A \leq e^{-k\delta^2}$ and $\widehat{\Theta}_k^A \geq p + 2\delta$, then play right forever. Otherwise, continue with S_A. Similarly to Case 1, since (S_A, S_B) is an equilibrium, we obtain that $\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma'_A) \geq 0$. Moreover, the strategies S_A and S'_A have different rewards only on the event D_2, so

$$\begin{aligned}0 \leq \mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma'_A) &= \mathbb{E}\Big[\Gamma_A(M,\infty) - \Gamma'_A(M,\infty)\Big]\\ &= \mathbb{E}\Big[\beta^M \mathbb{1}_{D_2}\,\mathbb{E}\Big(\Gamma_A(M,\infty) - \Gamma'_A(M,\infty) \mid H_M^A\Big)\Big] \qquad (48)\end{aligned}$$

By (47), we get $\mathbb{P}(D_2) = 0$. Recall $\mathbb{P}(\widetilde{D}_2 \setminus D_2) \leq 2e^{-k\delta^2}$. Then $\mathbb{P}(\widetilde{D}_2) \leq 2e^{-k\delta^2}$.

Case 3: $\Theta \in [p-3\delta, p+3\delta]$. By choice of δ, we have $\mu([p-3\delta,p+3\delta]) < \epsilon$.

Case 4: $|\Theta - \widehat{\Theta}_k^A| > \delta$ and $R_\infty^A \geq k$. By Lemma 21, this happens with probability at most $2\epsilon^2$.

Let τ_k^A be the round in which Alice explores for the k-th time (if $R_\infty^A < k$, then $\tau_k^A = \infty$). From cases $(1-4)$, we obtain that for k additionally satisfying the inequality $e^{-k\delta^2} < \epsilon$ we have

$$\mathbb{P}\Big(\exists\, t > \tau_k^A \;:\; \text{Alice pulls the worse arm at time } t\Big) \leq 7\epsilon \qquad (49)$$

We consider now Bob's behavior and show that if Alice explores at least $k+1$ times, then Bob will explore at all later times with high probability. Let $Q > \tau_{k+1}^A$ be the first round after τ_{k+1}^A where S_B plays left; if there is no such round, then $Q = \infty$.

Let $\widetilde{H}_Q$ be Alice's public history (i.e. containing the sequence of arms she played, but not her rewards) running from round 0 until the end of round $Q-1$. If $Q = \infty$, then $\widetilde{H}_Q$ is the whole history. Note that $\widetilde{H}_Q$ is observable by Bob.

Let $\mu_{\widetilde{H}_Q}$ denote Bob's posterior distribution for Θ (at the end of round $Q-1$) given $\widetilde{H}_Q$:

$$\mu_{\widetilde{H}_Q}(a,b) = \mathbb{P}\Big(\Theta \in (a,b) \mid \widetilde{H}_Q\Big) \cdot \mathbb{1}_{Q<\infty} \tag{50}$$

For $Q < \infty$, the history $\widetilde{H}_Q$ is said to be "good" if $\mu_{\widetilde{H}_Q}\big(p, p+2\delta\big) < \sqrt{\epsilon}$ and $\mu_{\widetilde{H}_Q}\big(0,p\big) < \epsilon\delta$. Define strategy S'_B as follows: play S_B until the beginning of round Q, when S_B plays left. If the history is good, then play right forever. Otherwise, continue with S_B. The goal will be to compare the total reward Γ_B from strategy S_B with the reward Γ'_B from S'_B.

Taking expectation over the history in (50) gives $\mathbb{E}\big[\mu_{\widetilde{H}_Q}\big(a,b\big)\big] = \mathbb{P}\big(\{\Theta \in (a,b)\} \cap \{Q < \infty\}\big)$. We additionally require that

$$\frac{6e^{-k\delta_1^2}}{1-\beta} \le \epsilon^2\delta/2\,.$$

By choice of δ, δ_1, and k, we get

- $\mathbb{E}\big[\mu_{\widetilde{H}_Q}\big(p,p+2\delta\big)\big] \le \mu\big(p,p+2\delta\big) \le \epsilon$
- $\mathbb{E}\big[\mu_{\widetilde{H}_Q}\big(0,p\big)\big] \le \mathbb{P}\big(\{\Theta \in [0,p]\} \cap \{R_\infty^A \ge k+1\}\big) \le \mu(p-3\delta_1,p) + \dfrac{6e^{-k\delta_1^2}}{1-\beta} \le \epsilon^2\delta \quad (51)$

Applying Markov's inequality in (51) gives: $\mathbb{P}\big(\mu_{\widetilde{H}_Q}\big(p,p+2\delta\big) \ge \sqrt{\epsilon},\ Q < \infty\big) \le \sqrt{\epsilon}$ and $\mathbb{P}\big(\mu_{\widetilde{H}_Q}\big(0,p\big) \ge \epsilon\delta,\ Q < \infty\big) \le \epsilon$. Combining these implies:

$$\mathbb{P}\big((Q < \infty) \cap (\widetilde{H}_Q \text{ is bad})\big) \le 2\sqrt{\epsilon}\,. \tag{52}$$

We claim that $\mathbb{P}\big(Q < \infty \text{ and } \widetilde{H}_Q \text{ is good}\big) = 0$. On the event that $\widetilde{H}_Q$ is good, we have

$$\begin{aligned}
\mathbb{E}\Big[\Gamma_B(Q,\infty) \mid \widetilde{H}_Q\Big] &\le p + \frac{\beta}{1-\beta}\,\mathbb{E}\Big[\max(\Theta,p) \mid \widetilde{H}_Q\Big] \\
&\le p + \frac{\beta}{1-\beta}\Big(\mathbb{E}\big[\Theta \mid \widetilde{H}_Q\big] + p\cdot\mathbb{P}(\Theta < p \mid \widetilde{H}_Q)\Big) \\
&\le p + \frac{\beta}{1-\beta}\,\mathbb{E}\big[\Theta \mid \widetilde{H}_Q\big] + \epsilon\delta
\end{aligned} \tag{53}$$

On the other hand, the reward from strategy S'_B on the event that $\widetilde{H}_Q$ is good is

$$\mathbb{E}\Big[\Gamma'_B(Q,\infty) \mid \widetilde{H}_Q\Big] = \frac{\mathbb{E}\big[\Theta \mid \widetilde{H}_Q\big]}{1-\beta}\,. \tag{54}$$

Using inequalities (53) and (54), the difference between the rewards under S_B and S'_B from round Q onwards given a good history $\widetilde{H}_Q$ satisfies the inequality:

$$\Delta(\widetilde{H}_Q) = \mathbb{E}\Big[\Gamma_B(Q,\infty) - \Gamma'_B(Q,\infty) \mid \widetilde{H}_Q\Big] \leq p - \mathbb{E}\Big[\Theta \mid \widetilde{H}_Q\Big] + \epsilon\delta \tag{55}$$

For any good history $\widetilde{H}_Q$, we can bound $\mathbb{E}[\Theta \mid \widetilde{H}_Q]$ as follows:

$$\begin{aligned}\mathbb{E}\Big[\Theta \mid \widetilde{H}_Q\Big] &\geq (p+2\delta)\cdot \mu_{\widetilde{H}_Q}\big(p+2\delta,1\big) + p\cdot \mu_{\widetilde{H}_Q}\big(p,p+2\delta\big)\\ &= p\cdot \mu_{\widetilde{H}_Q}\big(p,1\big) + 2\delta\cdot \mu_{\widetilde{H}_Q}\big(p+2\delta,1\big)\end{aligned} \tag{56}$$

Then on the event that $\widetilde{H}_Q$ is good: $\mathbb{E}[\Theta \mid \widetilde{H}_Q] \geq p(1-\epsilon\delta)+\delta \geq p+\delta/2$. Replacing in (55) implies that

$$\Delta(\widetilde{H}_Q) = \mathbb{E}\Big[\Gamma_B(Q,\infty) - \Gamma'_B(Q,\infty) \mid \widetilde{H}_Q\Big] < -\delta/3 \tag{57}$$

for all good histories $\widetilde{H}_Q$ when $\epsilon < 1/2$. Then the difference between the rewards under strategies S_B and S'_B satisfies

$$0 \leq \mathbb{E}(\Gamma_B - \Gamma'_B) = \mathbb{E}\Big[\beta^Q \cdot \Delta(\widetilde{H}_Q)\cdot \mathbb{1}_{\{\widetilde{H}_Q \text{ is good}\}}\Big]\,. \tag{58}$$

This forces $\mathbb{P}\big((\widetilde{H}_Q \text{ is good}) \cap (Q<\infty)\big) = 0$. Combining (52) and (58) implies

$$\mathbb{P}(Q<\infty) \leq 2\sqrt{\epsilon} \tag{59}$$

By (59), the probability that Alice explores infinitely often and Bob plays left infinitely often is bounded by $2\sqrt{\epsilon}$ for every ϵ, so the probability is zero. Similarly with roles reversed.

Now we can conclude the proof of convergence to the same arm. If both players explore finitely many times, then they converge to the left arm. Otherwise, one of the players - say Alice - explores infinitely often. This implies that Bob plays right for all but finitely many rounds. Reversing the roles, Alice also plays right for all but finitely many rounds. ■

We give an example showing why the Ψ_k^A condition is needed. In this example, the empirical mean Θ_k^A is significantly below p, yet the posterior mean is larger than p.

Example 3 *Let $p=0.7$ and μ be the uniform distribution on $[0,1/4]\cup[3/4,1]$. Suppose that in some history Alice obtains after k observations a value of $\widehat{\Theta}_k^A = 0.65 < p$. Then since μ has zero mass in $[1/4,3/4]$, it is much more likely that $\Theta \in [3/4,1]$ rather than $\Theta \in [0,1/4]$. In this case, Ψ_k^A is large and the strategy S'_A does not tell Alice to stop exploring, even though $\widehat{\Theta}_k^A < p-2\delta$ for small $\delta>0$.*

D.4. Competing Players in the Long Term

Theorem 23 (Competing players eventually settle on the same arm) *Consider two competing players, Alice and Bob, playing a one armed bandit problem with discount factor $\beta \in (0,1)$. The left arm has success probability p and the right arm has prior distribution μ with $\mu(p)=0$. Then in any Nash equilibrium, with probability 1 the players eventually settle on the same arm.*

In this setting there is an additional difficulty compared to the neutral case: players might refrain from switching to the optimal arm in order to not trigger an adverse reaction from the opponent. The key to overcoming this difficulty is realizing that it is impossible for both players to benefit from repeatedly pulling the inferior arm. Each player might subjectively believe for some time that they are playing the better arm, but if a player keeps exploring, then their subjective evaluation of the risky arm will converge to the objective reality.

Proof [Proof of Theorem 23] Let (S_A, S_B) be an optimal strategy pair. Note the strategies may be randomized.

As in the neutral case, for each n, let H_n^i be the history from player i's point of view up to round n. Let N be a stopping time for player i; this means that for every fixed n, the event $N = n$ is determined by H_n^i. Then H_N^i is a history for player i up to round N.

Recall $\widehat{\Theta}_k^i$ is the fraction of successes in the first k explorations by player i and R_∞^i is the total number of explorations by player i.

Define the events $\Lambda_f = \{$Alice explores infinitely often and Bob explores finitely many times$\}$ and $\Lambda_\infty = \{$Both players explore infinitely often$\}$.

Choose the following parameters:

$$\epsilon \in \left(0, \frac{1-\beta}{2}\right) \text{ and } \delta \in (0, \epsilon) \text{ such that } \mu([p - 10\delta, p + 10\delta]) < \epsilon \text{ and } \mu(p - 8\delta) = 0 \tag{60}$$

Case 1: $\Theta < p$. We consider two sub-cases, depending on whether both players explore infinitely often or only one player does (note that on the event that both players explore finitely often the conclusion holds immediately).

Case 1.a: $\Theta < p$, Alice explores infinitely often and Bob finitely many times (i.e., Λ_f holds). Define the event $S_k = \{$Bob explores at or after Alice's k-th exploration$\}$. Let

$$\epsilon_1 = \delta(1-\beta)/4 \,.$$

Let N_ℓ be the time of Alice's ℓ-th exploration if it exists, and otherwise $N_\ell = \infty$. Let $\mathcal{B}_1(k)$ be the collection of histories $H_{N_k}^A$ such that $\mathbb{P}(\Lambda_f \cap S_k \mid H_{N_k}^A) > \epsilon_1$ and $\mathcal{B}_2(k)$ the collection of histories $H_{N_k}^A$ for which $\mathbb{P}(\Lambda_f^c \mid H_{N_k}^A) > \epsilon_1$. Applying the Levy zero-one law gives

$$\lim_{k \to \infty} \mathbb{P}\Big(\Lambda_f \cap \big(H_{N_k}^A \in \mathcal{B}_2(k)\big)\Big) = 0 \,. \tag{61}$$

We also have $\bigcap_{k=0}^\infty (\Lambda_f \cap S_k) = \emptyset$ by the definition of Λ_f and S_k. Together with (61), this implies that for large enough k the next two conditions are satisfied:

$$\mathbb{P}(\Lambda_f \cap S_k) \leq \delta\epsilon_1 \text{ and } \mathbb{P}\Big(\Lambda_f \cap \big(H_{N_k}^A \in \mathcal{B}_2(k)\big)\Big) \leq \epsilon_1 \,. \tag{62}$$

Fix k so that (62) is satisfied for all $k' \geq k$. Letting $k' = k + 1$ and taking expectation over the history gives

$$\mathbb{P}(\Lambda_f \cap S_{k+1}) = \mathbb{E}(\mathbb{P}(\Lambda_f \cap S_{k+1}) \mid H_{N_{k+1}}^A) \leq \delta\epsilon_1 \,.$$

The collection of *bad* histories is defined as $\mathcal{B} = \mathcal{B}_1(k+1) \cup \mathcal{B}_2(k+1)$. Note that if the history $H^A_{N_{k+1}}$ is good, then

$$\mathbb{P}(S_{k+1} \mid H^A_{N_{k+1}}) = \mathbb{P}(\Lambda_f^c \mid H^A_{N_{k+1}}) + \mathbb{P}(\Lambda_f \cap S_k \mid H^A_{N_{k+1}}) \leq 2\epsilon_1 \tag{63}$$

By Markov's inequality, we obtain:

$$\begin{aligned}\mathbb{P}(H^A_{N_{k+1}} \in \mathcal{B}_1(k+1)) &= \mathbb{P}\Big(\mathbb{P}(\Lambda_f \cap S_{k+1} \mid H^A_{N_{k+1}}) > \epsilon_1\Big) \\ &\leq \frac{\mathbb{E}(\mathbb{P}(\Lambda_f \cap S_{k+1} \mid H^A_{N_{k+1}}))}{\epsilon_1} \leq \frac{\delta\epsilon_1}{\epsilon_1} = \delta \end{aligned} \tag{64}$$

By (64) and (62), we have

$$\mathbb{P}(\Lambda_f \cap (H^A_{N_{k+1}} \text{ is bad})) \leq \epsilon_1 + \delta \tag{65}$$

Then we can define, similarly to Theorem 22, the event $\widetilde{D}_1$ and the random variable Ψ^A_k (both from Alice's point of view):

$$\begin{aligned}\widetilde{D}_1 &= \{(R^A_\infty \geq k+1) \cap (\widehat{\Theta}^A_k \leq p - 2\delta)\} \\ \Psi^A_k &= \mathbb{P}\Big((|w_k - k\Theta| > k\delta) \cap (R^A_\infty \geq k+1) \mid H^A_{N_{k+1}}\Big)\end{aligned}$$

Define the event

$$D_1 = \{\Psi^A_k \leq e^{-k\delta^2}\} \cap \widetilde{D}_1 \cap \{H^A_{N_{k+1}} \text{ is good}\}\,.$$

Using inequality (38) in Theorem 22, we get that Alice's total reward from round N_{k+1} on is bounded as follows:

$$\mathbb{1}_{D_1}\,\mathbb{E}\Big[\Gamma_A(N_{k+1},\infty) \mid H^A_{N_{k+1}}\Big] \leq \mathbb{1}_{D_1}\Big[\frac{p}{1-\beta} - \delta + \frac{e^{-k\delta^2}}{1-\beta}\Big] \tag{66}$$

Bob's expected total reward given Alice's information is bounded, using (63), by

$$\mathbb{1}_{D_1}\,\mathbb{E}\Big[\Gamma_B(N_{k+1},\infty) \mid H^A_{N_{k+1}}\Big] \geq \mathbb{1}_{D_1}\Big[\frac{p}{1-\beta}\cdot(1-2\epsilon_1)\Big] \tag{67}$$

Combining inequalities (66) and (67) implies that Alice's expected utility satisfies

$$\begin{aligned}\mathbb{1}_{D_1}\,\mathbb{E}\Big[u_A(N_{k+1},\infty) \mid H^A_{N_{k+1}}\Big] &= \mathbb{1}_{D_1}\,\mathbb{E}\Big[\Gamma_A(N_{k+1},\infty) - \Gamma_B(N_{k+1},\infty) \mid H^A_{N_{k+1}}\Big] \\ &\leq \mathbb{1}_{D_1}\Big[\frac{p}{1-\beta} - \delta + \frac{e^{-k\delta^2}}{1-\beta} - \frac{p}{1-\beta}\cdot(1-\delta(1-\beta)/2)\Big] \\ &= \mathbb{1}_{D_1}\Big[\frac{p\delta}{2} - \delta + \frac{e^{-k\delta^2}}{1-\beta}\Big]\end{aligned} \tag{68}$$

We require k to be large enough so that

$$\frac{e^{-k\delta^2}}{1-\beta} \leq \delta\epsilon \tag{69}$$

Consider instead the following strategy S'_A for Alice: play S_A until time N_{k+1} when it is about to do its $k+1$-st exploration. If $\widehat{\Theta}^A_k \leq p - 2\delta$ and $\Psi^A_k \leq e^{-k\delta^2}$ and $\{H^A_{N_{k+1}}$ is good$\}$, then play left forever. (In other words, Alice deviates from S_A exactly on the event D_1.) Otherwise, continue with S_A. Then Alice's expected utility from round N_{k+1} on under strategy pair (S'_A, S_B) satisfies

$$\mathbb{1}_{D_1}\, \mathbb{E}\Big[u'_A(N_{k+1}, \infty) \mid H^A_{N_{k+1}}\Big] \geq \mathbb{1}_{D_1}\Big[\frac{p}{1-\beta} - \Big(\frac{p}{1-\beta} + \frac{e^{-k\delta^2}}{1-\beta}\Big)\Big] = \mathbb{1}_{D_1}\Big[\frac{-e^{-k\delta^2}}{1-\beta}\Big] \tag{70}$$

From inequalities (68), (69) (70), we get that on the event D_1, for $\epsilon < 1/4$, the inequality $u'_A(N_{k+1}, \infty) > u_A(N_{k+1}, \infty)$ holds. Since the original strategy pair (S_A, S_B) is optimal, we get that $\mathbb{P}(D_1) = 0$.

Now consider the event $\big(\Lambda_f \cap (\Theta < p)\big) \setminus D_1$ and note that it is contained in the union of the following events:

- $\Theta \in [p - 3\delta, p]$. Note that $\mu([p - 3\delta, p]) < \epsilon$.
- $\Lambda_f \cap (|\widehat{\Theta}^A_k - \Theta| > \delta)$. By Lemma 4, the probability of this event is bounded by $2e^{-2k\delta^2} \leq 2\delta\epsilon$.
- $\Psi^A_k > e^{-k\delta^2}$. By inequality (35), the probability of this event is at most $2e^{-k\delta^2} \leq 2\delta\epsilon$.
- $\Lambda_f \cap (H^A_{N_{k+1}}$ is bad$)$. By (65), the probability of this event is at most $\epsilon_1 + \delta$.

Since $\mathbb{P}(D_1) = 0$, we conclude that

$$\mathbb{P}\Big(\Lambda_f \cap (\Theta < p)\Big) = \mathbb{P}\Big(\big(\Lambda_f \cap (\Theta < p)\big) \setminus D_1\Big) \leq 7\epsilon$$

Thus taking $\epsilon \to 0$ gives $\mathbb{P}(\Lambda_f \cap (\Theta < p)) = 0$, so case (1.a) occurs with probability zero.

Case 1.b: $\Theta < p$, Both players explore infinitely often (i.e., Λ_∞ holds).

For each random time $N \in \mathbb{N}$ at which player i has explored at least k times, define

$$\Psi^i_k = \mathbb{P}\Big((|\widehat{\Theta}^i_k - \Theta| > \delta) \cap (R^i_\infty \geq k+1) \mid H^i_N\Big)$$

For each player i and each $\ell \in \mathbb{N}$, let τ^i_ℓ be the (random) time of the ℓ-th exploration by that player. Let N be the first time strictly greater than $\max\{\tau^A_k, \tau^B_k\}$ at which Alice explores. If there is no such time, then $N = \infty$.

Define the event

$$D = \{\widehat{\Theta}^A_k < p - 8\delta\} \cap \{\Psi^A_k \leq \delta\} \cap \{\max(\tau^A_k, \tau^B_k) < N < \infty\}.$$

Let $u^A_N(S_A, S_B) = \Gamma_A(N, \infty) - \Gamma_B(N, \infty)$ be Alice's normalized expected utility from round N onwards under (S_A, S_B) and define $u^A_N(Left, S_B)$ similarly when she plays left forever from round N on the event D. Since strategy pair (S_A, S_B) is optimal, Alice does at least as well by playing S_A as she would do by playing left from round N onwards on the event D:

$$\begin{aligned}\mathbb{E}\Big[u^A_N(S_A, S_B)\mathbb{1}_D \mid H^A_N\Big] &\geq \mathbb{E}\Big[u^A_N(Left, S_B)\mathbb{1}_D \mid H^A_N\Big] \\ &\geq \Big(p - \mathbb{E}\Big[\gamma_B(N) \mid H^A_N\Big] - \frac{\beta \cdot e^{-k\delta^2}}{1-\beta}\Big)\mathbb{1}_D, \end{aligned} \tag{71}$$

since Bob's action at time N is not influenced by Alice's action at time N. On the other hand, Alice's utility can be bounded from above by

$$\mathbb{E}\Big[u_N^A(S_A, S_B)\mathbb{1}_D \mid H_N^A\Big] \leq \mathbb{E}\Big[\gamma_A(N)\mathbb{1}_D \mid H_N^A\Big] - \mathbb{E}\Big[\gamma_B(N)\mathbb{1}_D \mid H_N^A\Big] + \mathbb{E}\Big[u_{N+1}^A(S_A, S_B)\mathbb{1}_D \mid H_N^A\Big]$$

Alice's expected reward in round N on the event D is at most

$$\begin{aligned}\mathbb{E}\Big[\gamma_A(N)\mathbb{1}_D \mid H_N^A\Big] &\leq \mathbb{E}\Big[\Theta \cdot \mathbb{1}_D \mid H_N^A\Big] \\ &= \mathbb{E}\Big[\Theta \cdot \mathbb{1}_D \mathbb{1}_{|\widehat{\Theta}_k^A - \Theta| > \delta} \mid H_N^A\Big] + \mathbb{E}\Big[\Theta \cdot \mathbb{1}_D \mathbb{1}_{|\widehat{\Theta}_k^A - \Theta| \leq \delta} \mid H_N^A\Big] \\ &\leq \Psi_k^A \cdot \mathbb{1}_D + \mathbb{E}\big[(p - 7\delta)\mathbb{1}_D \mid H_N^A\big] \\ &\leq (p - 6\delta)\mathbb{1}_D \end{aligned} \tag{72}$$

Then we can bound Alice's normalized expected utility from round N on by

$$\mathbb{E}\Big[u_N^A(S_A, S_B)\mathbb{1}_D \mid H_N^A\Big] \leq (p - 6\delta)\mathbb{1}_D - \mathbb{E}\Big[\gamma_B(N)\mathbb{1}_D \mid H_N^A\Big] + \mathbb{E}\Big[u_{N+1}^A(S_A, S_B)\mathbb{1}_D \mid H_N^A\Big] \tag{73}$$

Let $\eta \in (0, \delta)$ so that $\mu([p - 8\delta - 2\eta, p - 8\delta + 2\eta]) < \delta(1 - \beta)\epsilon$. Select k large enough so that

$$\frac{e^{-k\eta^2}}{1 - \beta} \leq \delta\epsilon \tag{74}$$

Comparing inequality (73) to (71) yields

$$\mathbb{E}\Big[u_{N+1}^A(S_A, S_B)\mathbb{1}_D \mid H_N^A\Big] \geq \Big(6\delta - \frac{\beta \cdot e^{-k\delta}}{1 - \beta}\Big)\mathbb{1}_D \geq 5\delta \cdot \mathbb{1}_D \tag{75}$$

Taking expectation over the history gives

$$\mathbb{E}\Big[u_{N+1}^A(S_A, S_B)\mathbb{1}_D\Big] \geq 5\delta \cdot \mathbb{P}(D) \tag{76}$$

Define the event

$$D^* = \Big\{\Theta < p - 8\delta\Big\} \cap \Big\{\max\left(\tau_k^A, \tau_k^B\right) < N < \infty\Big\}.$$

We argue that the event D^* is well approximated by the event D by showing their symmetric difference is small. By the choice of k in (74), we have

$$\mathbb{P}(\Psi_k^A \geq \delta) \leq 2\delta(1 - \beta)\epsilon \text{ and } \mathbb{P}(|\widehat{\Theta}_k^A - \Theta| > \eta) \leq 2\delta(1 - \beta)\epsilon\,.$$

The symmetric difference is included in the union of the following events:

$$D^* \triangle D \subset \Big\{\Theta \in \big[p - 8\delta - 2\eta, p - 8\delta + 2\eta\big]\Big\} \cup \Big\{|\widehat{\Theta}_k^A - \Theta| > \eta\Big\} \cup \Big\{\Psi_k^A \geq \delta\Big\}$$

By choice of δ, η, and k, we obtain

$$\mathbb{P}(D^* \triangle D) = \mathbb{E}\big(|\mathbb{1}_D - \mathbb{1}_{D^*}|\big) \leq 5\delta(1 - \beta)\epsilon$$

This implies that

$$\mathbb{E}\Big[u^A_{N+1}(S_A,S_B)|\mathbb{1}_{D^*}-\mathbb{1}_D|\Big]\leq \frac{1}{1-\beta}\cdot 5\delta(1-\beta)\epsilon = 5\delta\epsilon \tag{77}$$

Combining (77) with (76) implies

$$\mathbb{E}\Big[u^A_{N+1}(S_A,S_B)\mathbb{1}_{D^*}\Big]\geq 5\delta\cdot\mathbb{P}(D)-5\delta\epsilon \tag{78}$$

Define an event similar to D for Bob:

$$D' = \Big\{\big(\widehat{\Theta}^B_k < p-8\delta\big)\cap\big(\Psi^B_k\leq\delta\big)\cap\big(\max\{\tau^A_k,\tau^B_k\}<N\big)\Big\}$$

Then Bob's normalized expected utility from round N onwards given his history satisfies

$$\mathbb{E}\Big[u^B_{N+1}(S_A,S_B)\cdot\mathbb{1}_{D'}\mid H^B_N\Big]\geq\mathbb{E}\Big[u^B_{N+1}(S_A,Left)\cdot\mathbb{1}_{D'}\mid H^B_N\Big] \tag{79}$$

$$\geq -\frac{e^{-k\delta^2}}{1-\beta}\mathbb{1}_{D'}\geq -\delta\epsilon \tag{80}$$

A similar argument to the proof of (77) gives

$$\mathbb{E}\Big[u^B_{N+1}(S_A,S_B)\mathbb{1}_{D*}\Big]\geq -\delta\epsilon-5\delta\epsilon \tag{81}$$

Since the game is zero-sum, adding (78) and (81) gives

$$0=\mathbb{E}\Big[u^A_{N+1}(S_A,S_B)\mathbb{1}_{D*}\Big]+\mathbb{E}\Big[u^B_{N+1}(S_A,S_B)\mathbb{1}_{D*}\Big]\geq 5\delta\cdot\mathbb{P}(D)-11\delta\epsilon$$

This implies that $\mathbb{P}(D)\leq 3\epsilon$.

Note the containment

$$\Big\{\big(\Theta<p-9\delta\big)\cap\big(\max\{\tau^A_k,\tau^B_k\}<N<\infty\big)\Big\}\subset D\cup\Big\{\big(|\Theta-\widehat{\Theta}^A_k|\geq\delta\big)\cup\big(\Psi^A_k>\delta\big)\Big\}$$

Since $\mathbb{P}(D)\leq 3\epsilon$ and $\mathbb{P}(|\Theta-\widehat{\Theta}^A_k|\geq\delta)\leq\epsilon$ and $\mathbb{P}(\Psi^A_k>\delta)\leq\epsilon$, we get

$$\mathbb{P}\Big(\big(\Theta<p-9\delta\big)\cap\big(\max\{\tau^A_k,\tau^B_k\}<N<\infty\big)\Big)\leq 5\epsilon\,.$$

By the choice of δ, since $\big(R^A_\infty=R^B_\infty=\infty\big)\subseteq\big(\max\{\tau^A_k,\tau^B_k\}<N<\infty\big)$, this implies that

$$\mathbb{P}\Big(\big(\Theta<p\big)\cap\big(R^A_\infty=R^B_\infty=\infty\big)\Big)\leq 6\epsilon$$

Letting $\epsilon\to 0$ implies that case (1.b) occurs with probability zero.

Case 2. $\Theta>p$. We define M as a stopping time with the following property: M is the first time where $M>\tau^A_k$ and $M>\tau^B_k$ and Bob plays left at time M under S_B. If there is no such time, then $M=\infty$. Let S'_B be the following Bob strategy: play S_B until time $M-1$. From round M on, play right.

Similarly to the neutral players case, let $\widetilde{H}_M$ be Alice's public history (i.e. containing the sequence of arms she played but not her rewards) running from round 0 until the end of round $M-1$. If $M=\infty$, then $\widetilde{H}_M$ is the whole history. Note that $\widetilde{H}_M$ is observable by both players at the beginning of round $M+1$. For $M<\infty$, define

$$\widetilde{H}_M \text{ is good } \iff \mu_{\widetilde{H}_M}\big(p, p+2\delta\big) < \sqrt{\epsilon} \text{ and } \mu_{\widetilde{H}_M}\big(0,p\big) < \epsilon\delta\,.$$

Let $D = \{(\widetilde{H}_M \text{ is good}) \cap (M<\infty)\}$.

By combining cases $(1.a)$ and $(1.b)$ for competing players, we get that

$$\mathbb{P}\Big((\Theta < p) \cap (R^A_\infty = \infty)\Big) = 0$$

Thus, for sufficiently large k, we get

$$\mathbb{P}\Big((\Theta < p) \cap (R^A_\infty \geq k+1)\Big) < \epsilon^2\delta\,. \tag{82}$$

We follow the formulas from (50) to (52) of the analysis for neutral players, with the only change that in (51) we use the bound from (82). Thus we obtain the same bound as in (52):

$$\mathbb{P}\big((M<\infty)\cap(\widetilde{H}_M \text{ is bad})\big) \leq 2\sqrt{\epsilon}\,. \tag{83}$$

Define S_B' as the following Bob strategy: play S_B until the end of round $M-1$. At round M, if D holds, then play right forever. Otherwise, continue with S_B. On the event D, since $\mu_{\widetilde{H}_M}(0,p) < \epsilon\delta$, by playing right forever from round $M+1$ onwards Bob guarantees a minimal utility of at least

$$\mathbb{E}\Big[\big(u^B_{M+1}(S_A, S_B')\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] \geq -\frac{\epsilon\delta}{1-\beta}\cdot \mathbb{1}_D$$

Let $\gamma_i(M)$ be the reward of player i in round M under strategies (S_A,S_B) and $\gamma_i'(M)$ the reward of player i in round M under strategies (S_A,S_B'). Note that $\gamma_A(M)=\gamma_A'(M)$. Then

$$\begin{aligned}\mathbb{E}\Big[\big(u^B_M(S_A,S_B')\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] &= \mathbb{E}\Big[\big(\gamma_B'(M)-\gamma_A'(M)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] + \mathbb{E}\Big[\big(u^B_{M+1}(S_A,S_B')\big)\mathbb{1}_D \mid \widetilde{H}_M\Big]\\ &\geq \left(p+\frac{\delta}{2}\right)\mathbb{1}_D - \mathbb{E}\Big[\big(\gamma_A(M)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] - \frac{\epsilon\delta}{1-\beta}\cdot\mathbb{1}_D \end{aligned}\tag{84}$$

For strategy pair (S_A,S_B) we get

$$\begin{aligned}\mathbb{E}\Big[\big(u^B_M(S_A,S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] &= \mathbb{E}\Big[\big(\gamma_B(M)-\gamma_A(M)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] + \mathbb{E}\Big[\big(u^B_{M+1}(S_A,S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big]\\ &= p\cdot\mathbb{1}_D - \mathbb{E}\Big[\big(\gamma_A(M)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] + \mathbb{E}\Big[\big(u^B_{M+1}(S_A,S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big]\end{aligned}\tag{85}$$

Since $\mathbb{E}\Big[\big(u^B_M(S_A,S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] \geq \mathbb{E}\Big[\big(u^B_M(S_A,S_B')\big)\mathbb{1}_D \mid \widetilde{H}_M\Big]$, by combining the previous inequalities we obtain

$$\begin{aligned}&p\cdot\mathbb{1}_D - \mathbb{E}\big[\big(\gamma_A(M)\big)\mathbb{1}_D \mid \widetilde{H}_M\big] + \mathbb{E}\big[\big(u^B_{M+1}(S_A,S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\big] \geq\\ &\left(p+\frac{\delta}{2}\right)\mathbb{1}_D - \mathbb{E}\Big[\big(\gamma_A(M)\big)\mathbb{1}_D \mid \widetilde{H}_M\Big] - \epsilon\delta \implies\\ &\mathbb{E}\big[\big(u^B_{M+1}(S_A,S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\big] \geq \frac{\delta}{2}\cdot\mathbb{1}_D\end{aligned}\tag{86}$$

On the other hand, Alice can ensure a utility of at least $-\epsilon\delta/(1-\beta)$ from round $M+1$ onwards by always playing right, and so

$$\mathbb{E}\big[\big(u^A_{M+1}(S_A, S_B)\big)\mathbb{1}_D \mid \widetilde{H}_M\big] \geq -\frac{\epsilon\delta}{1-\beta} \cdot \mathbb{1}_D \tag{87}$$

Since the game is zero-sum, we obtain that

$$0 = \mathbb{E}\big[u^B_{M+1}(S_A, S_B)\big] + \mathbb{E}\big[u^A_{M+1}(S_A, S_B)\big] \geq \Big(\frac{\delta}{2} - \frac{\epsilon\delta}{1-\beta}\Big) \cdot \mathbb{P}(D) \tag{88}$$

By choice of ϵ in (60) we have $\delta/2 - \epsilon\delta/(1-\beta) > 0$, so $\mathbb{P}(D) = 0$. Combined with inequality (83), we get that $\mathbb{P}(M < \infty) < 2\sqrt{\epsilon}$. This implies

$$\mathbb{P}\Big((R^A_\infty = R^B_\infty = \infty) \cap \{\text{Bob plays left infinitely often}\}\Big) < 2\sqrt{\epsilon}$$

By letting $\epsilon \to 0$, we conclude the latter probability must be zero. By symmetry, we obtain that if Bob and Alice explore infinitely often, then they both settle on the right arm from some point on with probability 1. ■

Appendix E. Competitive Play – Improved Bounds for a Uniform Prior

In this section we give improved bounds for the thresholds in the case of an arm with a uniform prior. In particular, we show that both players will explore the risky arm for all $p < 5/9$ and will not explore for all $p > 2 - \sqrt{2}$.

Proposition 24 *Let arm L have a known probability p and arm R with a uniform prior that is common knowledge. Then both players will explore arm R with positive probability for all $p < 5/9$ as the discount factor $\beta \to 1$.*

Proof To prove this, we consider the scenario where in round zero Bob is forced to play L and Alice is forced to play R, while the players can play optimally afterwards. Then we show that Alice wins for all $p < 5/9$ as $\beta \to 1$ regardless of Bob's strategy, which will imply that in any equilibrium the players will both play the right arm in round zero.

Consider now the scenario where they are forced to play as described above and define the following strategy for Alice:

1. If the bit observed in round zero is 0, then play L in round 1. Then
   - If Bob played R in round 1, then play L in round 2 and from round 3 onwards copy Bob's arm from the previous round.
   - If Bob played L in round 1, then play optimally from round 2 onwards.
2. Else, if the bit observed in round zero is 1, then play R again in round 1. Play optimally from round 2 onwards.

We show that Alice wins in expectation regardless of Bob's counterstrategy. For round zero we have $\mathbb{E}(\gamma_A(0)) = 1/2$ and $\mathbb{E}(\gamma_B(0)) = p$. Alice's expected payoff in round 1 depends on whether she got a 0 or a 1 in round zero: $\mathbb{E}(\gamma_A(1)) = 1/2 \cdot p + 1/2 \cdot 2/3 = 1/2 \cdot p + 1/3$.

To analyze Alice's payoff in round two and Bob's maximum expected reward overall we consider two cases, depending on Bob's move in round one. We assume that Bob plays optimally from round two onwards. Note also that Bob's decision for what arm to play in round one cannot depend on Alice's bit, so it is in fact determined at round zero.

Case 1: Bob plays R in round one. Denote by $\gamma_A(t|1,R)$ Alice's reward in round t given that she saw a 1 in round zero and that Bob's strategy is to play R in round one, and similarly for $\gamma_A(t|0,R)$. Then Alice's expected total reward is:

$$\mathbb{E}(\Gamma_A) = \mathbb{E}(\gamma_A(0)) + \mathbb{E}(\gamma_A(1)) \cdot \beta + \mathbb{E}(\gamma_A(2)) \cdot \beta^2 + \sum_{t=3}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t = \frac{1}{2} + \beta\left(\frac{p}{2} + \frac{1}{3}\right)$$
$$+ \beta^2 \left(\frac{p}{2} + \frac{\mathbb{E}(\gamma_A(2|1,R))}{2}\right) + \frac{1}{2}\sum_{t=3}^{\infty} \mathbb{E}(\gamma_A(t|1,R)) \cdot \beta^t + \frac{1}{2}\sum_{t=3}^{\infty} \mathbb{E}(\gamma_B(t-1|0,R)) \cdot \beta^t . \tag{89}$$

Bob's expected total reward is:

$$\mathbb{E}(\Gamma_B) = p + \frac{1}{2} \cdot \beta + \frac{1}{2}\sum_{t=2}^{\infty} \mathbb{E}(\gamma_B(t|0,R)) \cdot \beta^t + \frac{1}{2}\sum_{t=2}^{\infty} \mathbb{E}(\gamma_B(t|1,R)) \cdot \beta^t \tag{90}$$

In the case where Alice observes a 1 in round zero, then Alice has an advantage from round two onwards, so $\mathbb{E}(\gamma_A(t|1,R)) \geq \mathbb{E}(\gamma_B(t|1,R))$ for all $t \geq 2$. Using this fact and simplifying the expressions, we obtain that Alice's net gain in the zero sum game is

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \geq \frac{1}{2} + \beta\left(\frac{p}{2} + \frac{1}{3}\right) + \frac{p\beta^2}{2} - p - \frac{\beta}{2} - \frac{1-\beta}{2} \cdot \sum_{t=2}^{\infty} \mathbb{E}(\gamma_B(t|0,R)) \cdot \beta^t$$

To bound $\mathbb{E}(\gamma_B(t|0,R))$ we consider two scenarios, depending on whether Bob saw a zero or a one in round 1. Moreover, by round two Bob knows the bit seen by Alice in round zero since her behavior is different depending on that bit. Let X_1 and X_2 be two random variables with densities $3(1-x)^2$ and $6x(1-x)$ for all $x \in [0,1]$ respectively. Then

$$\mathbb{E}(\gamma_B(t|0,R)) \leq \frac{2}{3} \cdot \mathbb{E}(\max(p, X_1)) + \frac{1}{3} \cdot \mathbb{E}(\max(p, X_2)) = 1/3 - 1/3 \cdot p^3$$

Then as $\beta \to 1$, the difference in rewards is $\lim_{\beta\to 1}\left[\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B)\right] \geq 1/6 + p^3/6 - p^2/2$. Then Alice wins in expectation for all $p \leq 0.65$.

Case 2: Bob plays L in round one. Note we will write again $\gamma_A(t|b,L)$ to denote Alice's reward in round t given that Alice saw the bit b in round zero and Bob plays left in round one. Bob's expected total reward is $\mathbb{E}(\Gamma_B)) = p + p \cdot \beta + \sum_{t=2}^{\infty} \mathbb{E}(\gamma_B(t)) \cdot \beta^t$, while Alice's is

$$\mathbb{E}(\Gamma_A) = \frac{1}{2} + \beta\left(\frac{p}{2} + \frac{1}{3}\right) + \frac{1}{2} \cdot \sum_{t=2}^{\infty} \mathbb{E}(\gamma_A(t|1,L)) \cdot \beta^t + \frac{1}{2} \cdot \sum_{t=2}^{\infty} \mathbb{E}(\gamma_B(t-1|0,L)) \cdot \beta^t.$$

Again Alice has an informational advantage so she can at least equalize from round two onwards regardless of the value of the bit observed in round zero, and so $\mathbb{E}(\gamma_A(t|b,L)) \geq \mathbb{E}(\gamma_B(t|b,L))$ for all $t \geq 2$ and $b \in \{0,1\}$. The difference in expected payoffs can be bounded in this case by $\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \geq 1/2 + \beta\,(p/2 + 1/3) - p(1+\beta)$. Taking $\beta \to 1$ gives $\lim_{\beta\to 1}\left[\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B)\right] \geq 5/6 - 3p/2$. Then Alice wins for all $p < 5/9$ when Bob plays left.

Taking into account both cases implies the players explore for all $p < 5/9$ as required. ■

Proposition 25 *Let arm L have a known probability $p > 2 - \sqrt{2}$ and arm R with a uniform prior, both of which are common knowledge. Then with probability 1 the players will not explore arm R in any equilibrium.*

Proof To analyze this, we study the game where round zero is fixed such that Alice is at the right arm and Bob is at the left arm.

Consider the following strategy for Bob: stay at arm L in rounds zero and one, then in each round t, where $t \geq 2$, copy Alice's move from round $t-1$. Suppose Alice plays optimally given Bob's strategy. Then Alice's expected total is: $\mathbb{E}(\Gamma_A) = 1/2 + \sum_{t=1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t$. For Bob, note that $\mathbb{E}(\gamma_B(t)) = \mathbb{E}(\gamma_A(t-1))$ for all $t \geq 2$. Then his expected total reward is

$$\mathbb{E}(\Gamma_B) = p \cdot (1+\beta) + \sum_{t=2}^{\infty} \mathbb{E}(\gamma_A(t-1)) \cdot \beta^t = p \cdot (1+\beta) + \sum_{t=1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^{t+1}$$

Bob's expected net gain is $\mathbb{E}(\Gamma_B) - \mathbb{E}(\Gamma_A) = p \cdot (1+\beta) - 1/2 - (1-\beta) \cdot \sum_{t=1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t$. Let X be a random variable with a uniform prior. Then Alice's expected value at round t can be bounded by the expected maximum of p and X, so

$$\mathbb{E}(\gamma_A(t)) \leq \mathbb{E}(\max(p, X)) = \int_0^p p\,dx + \int_p^1 x\,dx = \frac{1}{2} + \frac{p^2}{2} \tag{91}$$

Using the bound on $\mathbb{E}(\gamma_A(t))$ above gives

$$\mathbb{E}(\Gamma_B) - \mathbb{E}(\Gamma_A) \geq p \cdot (1+\beta) - \frac{1}{2} - (1-\beta) \cdot \sum_{t=1}^{\infty} \left(\frac{1}{2} + \frac{p^2}{2}\right) \cdot \beta^t$$

Taking $\beta \to 1$ gives $\lim_{\beta\to 1}\left[\mathbb{E}(\Gamma_B) - \mathbb{E}(\Gamma_A)\right] \geq 2p - 1/2 - \left(1/2 + 1/2 \cdot p^2\right) = 2p - 1 - 1/2 \cdot p^2$. Then for all $p > 2 - \sqrt{2}$ Bob's net gain is strictly positive, thus Alice is at a disadvantage by playing the right arm in round zero. It follows that Alice is better off by playing the left arm in round zero, and so the players will never explore arm R in any equilibrium. ■

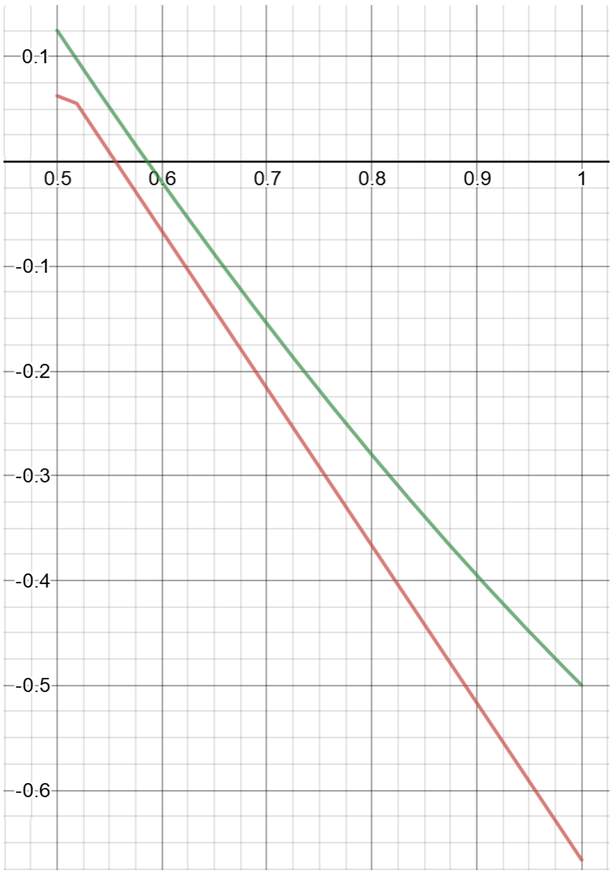

Figure 5: Bounds as a function of $p \in [0.5, 1]$ where the right arm has a uniform prior and $\beta \to 1$: the red line shows the lower bound on Alice's net gain given by the function $\mathrm{lb}(p) = \min\{1/6 + p^3/6 - p^2/2, 5/6 - 3p/2\}$ (Proposition 24) when in round zero she starts at the right arm and Bob starts at left, after which they both play optimally. The green line shows the corresponding upper bound on Alice's net gain given by $\mathrm{ub}(p) = p^2/2 + 1 - 2p$ (Proposition 25).