Multiplayer Bandit Learning, from Competition to Cooperation

Simina Branzei

Purdue University

Mathematics of Online Decision Making Workshop, Simons Institute

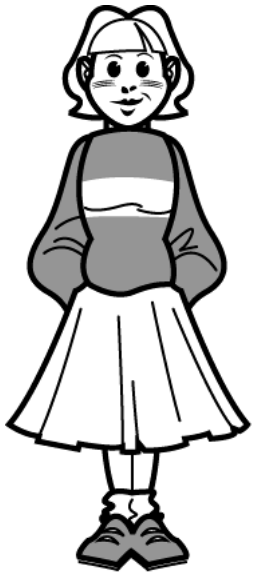
October 29, 2020

Joint work with Yuval Peres

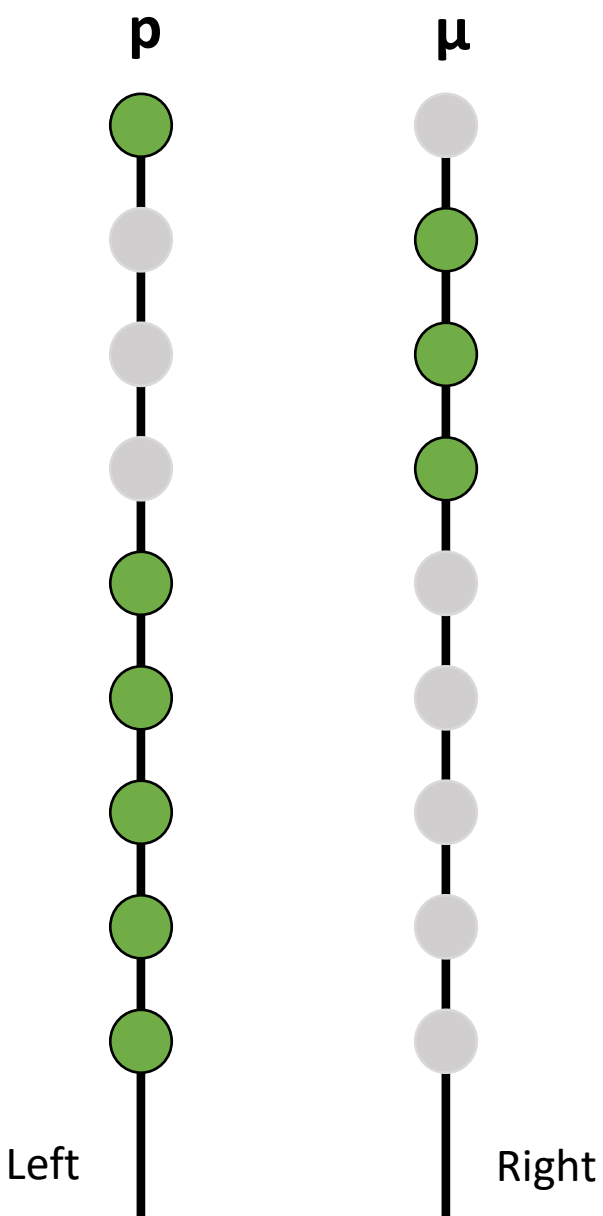
Multiplayer Model

- Alice and Bob play in a multi-armed bandit problem.
- One arm is safe (known probability p), the other is volatile (unknown probability of success θ with prior μ).
- In every round, each player
  - pulls an arm
  - gets the reward (0 or 1) from the arm they pulled, and
  - observes **the action of the other player but not their reward**.

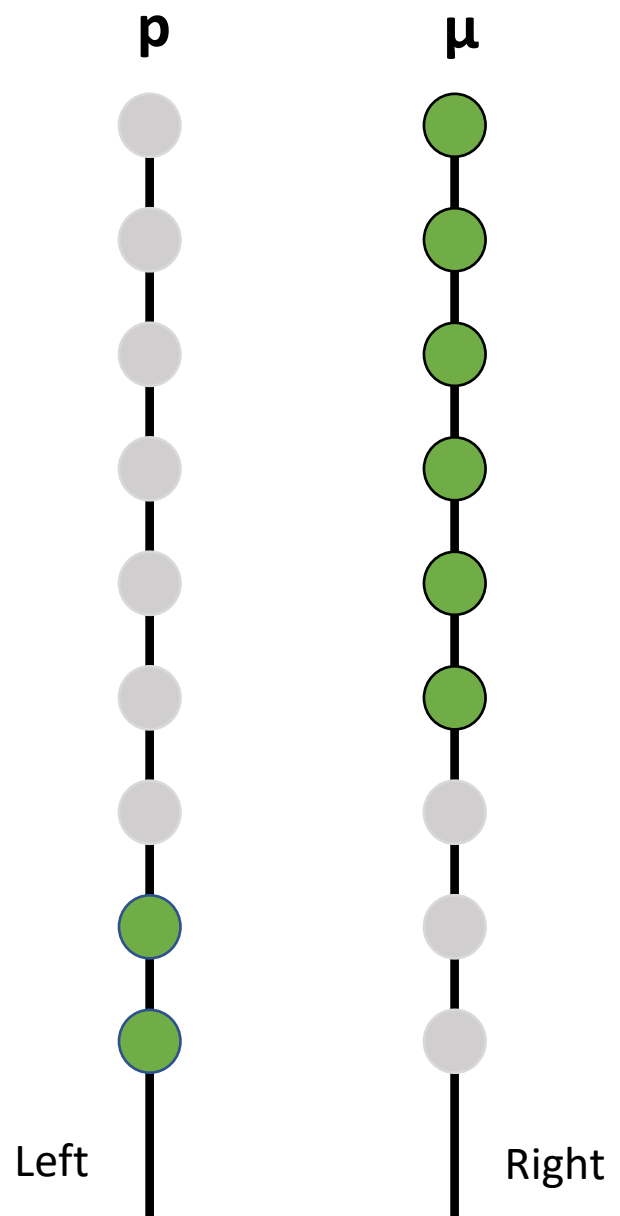
Alice's trajectory



Alice



Bob's trajectory



Bob

Utilities

Alice's utility is: $\boldsymbol{\Gamma_A + \lambda \cdot \Gamma_B}$, and similarly for Bob, where

- $\boldsymbol{\Gamma_A = \sum_{t=0}^{\infty} \gamma_A(t) \cdot \beta^t}$ and $\boldsymbol{\Gamma_B = \sum_{t=0}^{\infty} \gamma_B(t) \cdot \beta^t}$ are Alice and Bob's discounted rewards, respectively
- $\boldsymbol{\beta}$ is the discount factor

Definition for finite horizon is similar: $\boldsymbol{\Gamma_A = \sum_{t=0}^{T} \gamma_A(t)}$, i.e. sum of rewards

Competitive setting: zero sum game

$\boldsymbol{\lambda = -1}$: Alice's utility is: $\Gamma_A - \Gamma_B$ and Bob's is $\Gamma_B - \Gamma_A$ (E.g., competing phone companies in a saturated market)

Neutral setting

$\boldsymbol{\lambda = 0}$: Alice's utility is Γ_A and Bob's is Γ_B $\Rightarrow$ each player cares about its own rewards (i.e. purely selfish)

Cooperative setting

$\boldsymbol{\lambda = 1}$: Both Alice and Bob have utility $\Gamma_A + \Gamma_B \Rightarrow$ players are aligned, maximize total rewards collected (e.g. genetically identical organisms)

Partly cooperative setting

$\boldsymbol{\lambda = \frac{1}{2}}$: Alice has utility $\Gamma_A + \frac{1}{2} \cdot \Gamma_B \Rightarrow$ players are partly aligned (e.g. siblings – share ½ of the genes)

Strategies

Strategy: map from history to the next action to play; may be randomized.

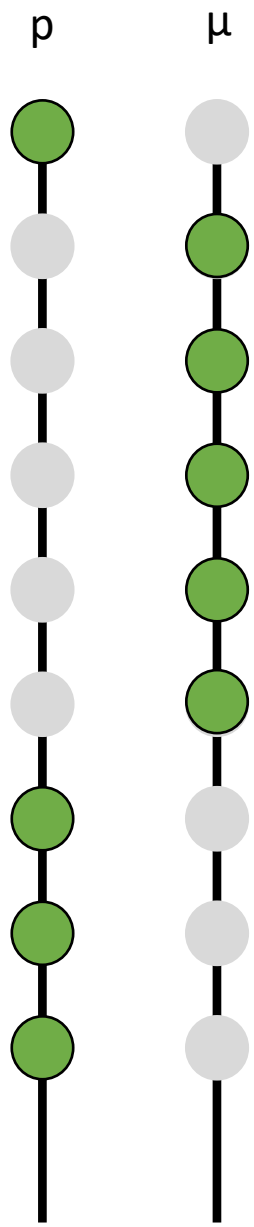
- Alice's history = actions of both players + Alice's rewards; same for Bob.

Expected utility: computed using the player's beliefs about the private information of the other player

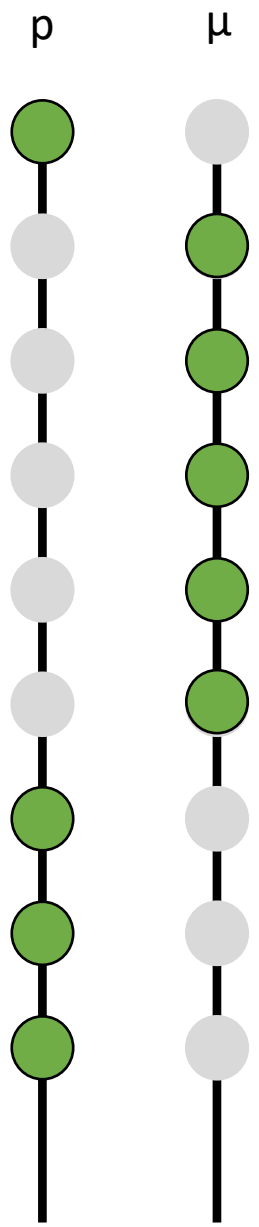
Gittins Index

Gittins index g=g(μ,β) of the risky arm is defined as the infimum of the success probabilities p where playing always a safe arm with probability p is optimal for a single player.

- It's a **retirement value** – what payment p the player would accept every period instead of exploring more the risky arm
- Note g > m, where m is the mean of μ and μ is not a point mass.



Gittins Index



Gittins index g=g(μ,β) of the risky arm is defined as the infimum of the success probabilities p where playing always a safe arm with probability p is optimal for a single player.

- It's a **retirement value** – what payment p the player would accept every period instead of exploring more the risky arm
- Note g > m, where m is the mean of μ and μ is not a point mass.

Optimal strategy of one player facing k risky arms: in each round, play the arm with the highest Gittins index at that point.

Related literature

Multiplayer learning in the collision model

- players are pulling arms independently.
- 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.

Related literature

Multiplayer learning in the collision model

- players are pulling arms independently.
- 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.
- **Adversarial setting:** Alatur et al (2019), Bubeck et al (2019); **stochastic setting**: Kalathil et al (14), Lugosi and Mehrabian (18), Bistritz and and Leshem (18)
- **May receive input about collision or not** (Avner and Mannor [AM14], Rosenski, Shamir, and Szlak [RSS16], Bonnefoi et al [BBM+17], Boursier and Perchet [BP18]**)**

Related literature

Multiplayer bandit learning in the same feedback model

- Aoyagi (98, 11) – with two risky arms where priors have discrete support
- Rosenberg et al (13) – same model but decision to switch to the safe arm is irreversible

Interplay between competition and innovation modeled with bandit learning in R&D (D'Aspremont and Jackquemi (88), Besanko and Wu (13)

Related literature

Multiplayer bandit learning, same setting except feedback is immediate (everyone can observe all the past actions and all past rewards)

- Bolton and Harris (99) – free rider effect and encouragement effect: a player may explore more in order to encourage further exploration from others
- Cripps, Keller, and Rady (05) - characterize the unique Markovian equilibrium of the game
- Heidhues, Rady, and Strack (15) - 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

Related literature

Incentivizing exploration

- Kremer et al (13), Frazier et al (14), Mansour et al (15) - principal wants to explore a set of arms, but exploration is done by stream of myopic agents
- Aridor et al (19) - empirically study the interplay between exploration and competition in a model where multiple firms are competing for the same market of usersand each firm commits to a multi-armed bandit algorithm
- Braverman et al (19) - each arm receivesa 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

Related literature

Evolutionary biology

- How cooperation evolved in insects (ants, bees) – Hamilton (64), Anderson (84), Boomsma (07)

Competitive setting

Competitive setting

Zero-sum game has a value by Sion’s minimax theorem.

Competitive setting

Zero-sum game has a value by Sion's minimax theorem.

How do competing players behave?

Competitive setting

Zero-sum game has a value by Sion's minimax theorem.

Theorem 1 (Competing players explore less).

Competitive setting

Zero-sum game has a value by Sion's minimax theorem.

Theorem 1 (Competing players explore less). Suppose arm L has 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 Alice and Bob are playing optimally in the zero sum game with discount factor β.

Competitive setting

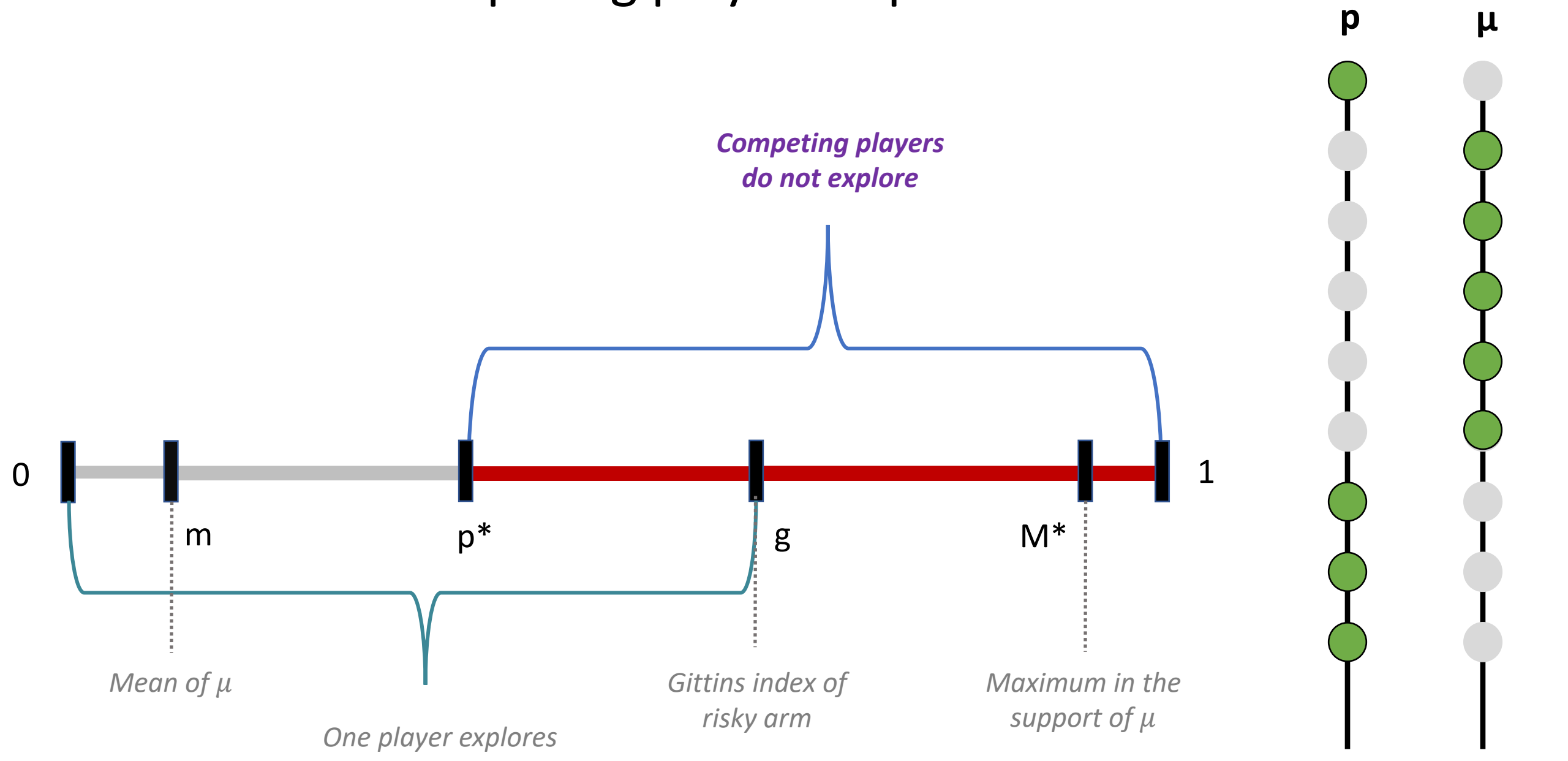
Zero-sum game has a value by Sion's minimax theorem.

Theorem 1 (Competing players explore less). Suppose arm L has 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 Alice and Bob are playing optimally in the zero sum game with discount factor β.

Then there exists a threshold $\boldsymbol{p^* < g}$, where $g = g(\mu, \beta)$ is the Gittins index of the right arm, such that **for all $\boldsymbol{p > p^*}$**, with probability 1 the players ***will not explore arm R***.

More precisely, $p^* \leq \frac{m \cdot \beta + g}{1+\beta}$, where m is the mean of μ.

Competing players explore less



Competitive setting

Theorem 1 shows information is less valuable in the zero sum setting. Does it have any value?

Competitive setting

Theorem 1 shows information is less valuable in the zero sum setting. Does it have any value?

Theorem 2 (Competing players are not completely myopic).

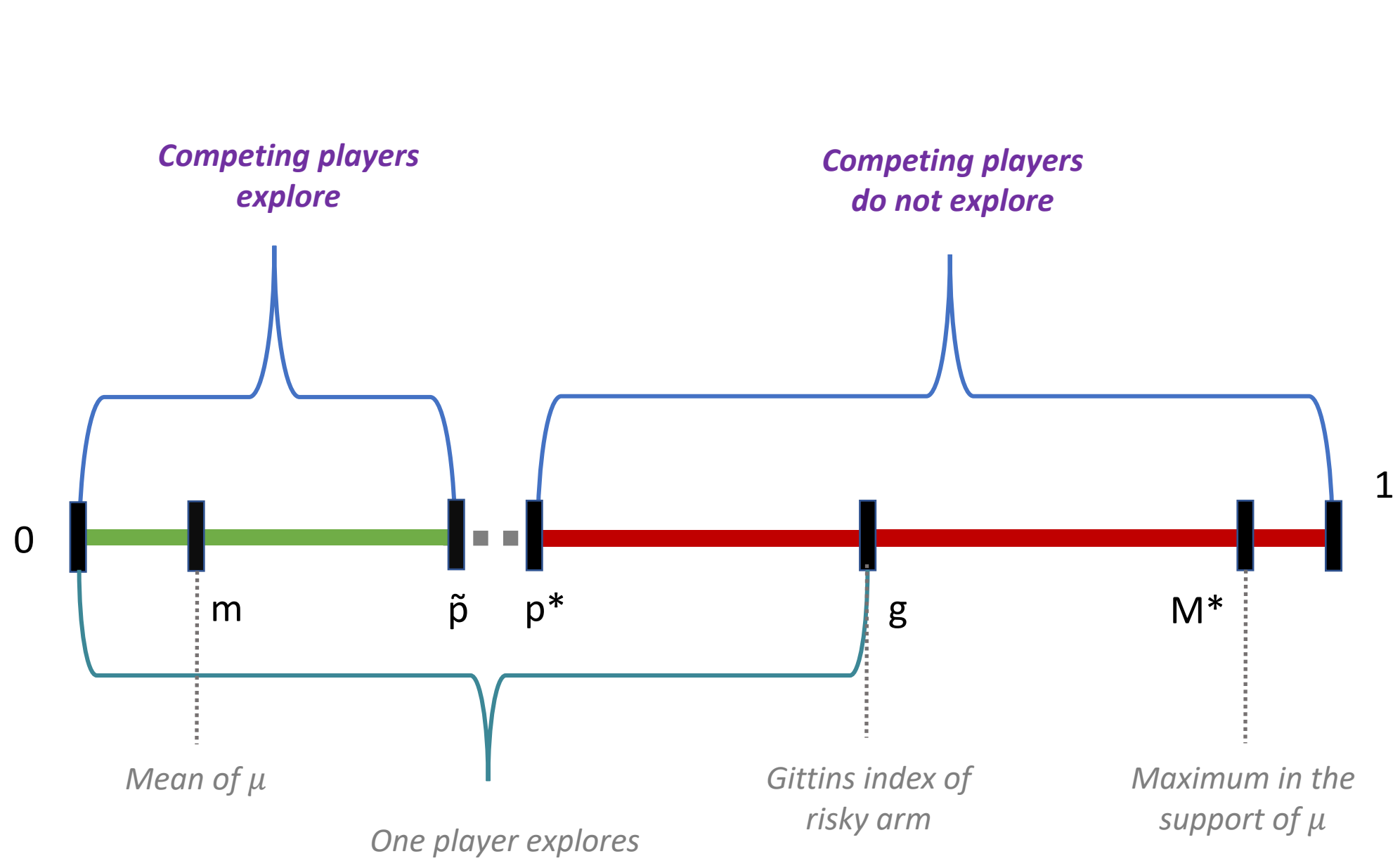
Competitive setting

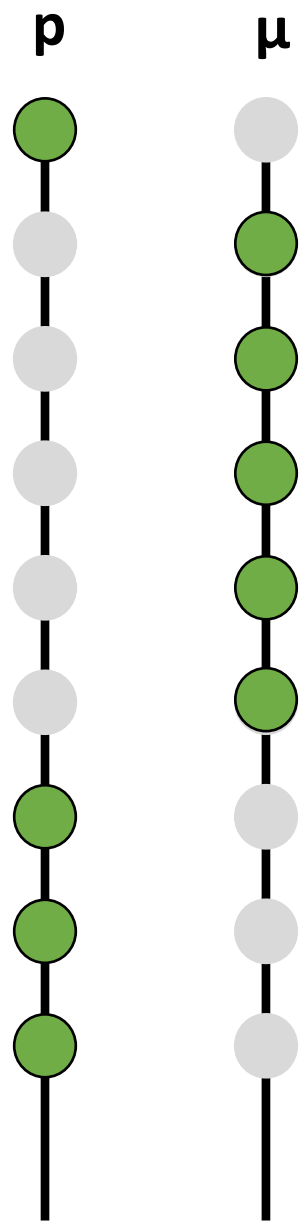
Theorem 1 shows information is less valuable in the zero sum setting. Does it have any value?

Theorem 2 (Competing players are not completely myopic). In the same setting of Theorem 1, there exists a threshold $\tilde{p} > m$, such that ***for all*** $p < \tilde{p}$, with probability 1 both players ***will explore arm R*** in the initial round of optimal play.

More precisely, $\tilde{p} \geq m + \frac{\beta w}{2}$, where m is the mean of μ and w its variance.

Competing players are not completely myopic





Cooperative setting

Cooperative setting

Players aim to maximize the sum of their rewards; can agree on their strategies before play

Cooperative setting

Players aim to maximize the sum of their rewards; can agree on their strategies before play

Theorem 3 (Cooperating players explore more).

Cooperative setting

Players aim to maximize the sum of their rewards; can agree on their strategies before play

Theorem 3 (Cooperating players explore more). Suppose Alice and Bob are players with aligned interests playing a one armed bandit problem with discount factor β. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass.

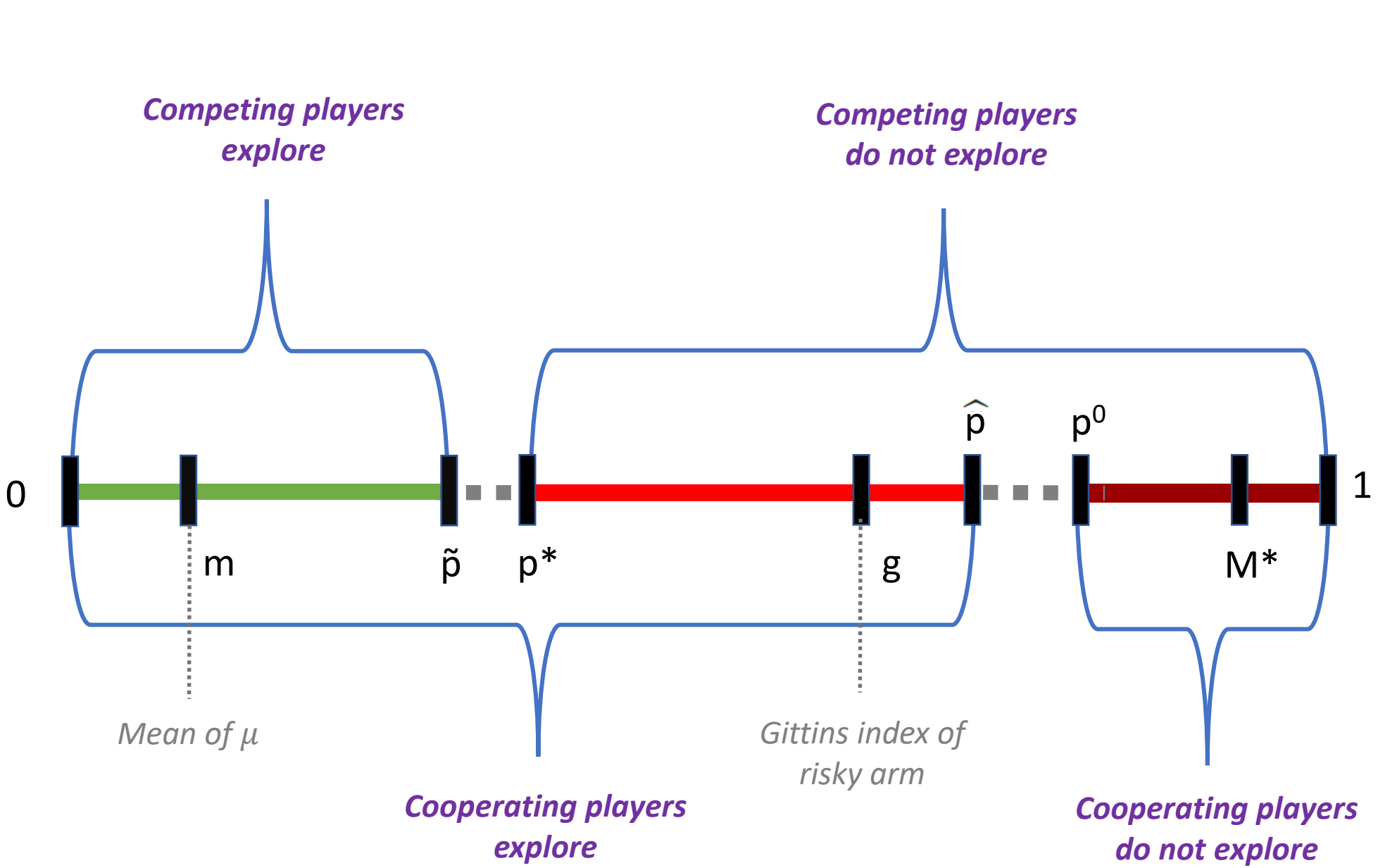
Cooperative setting

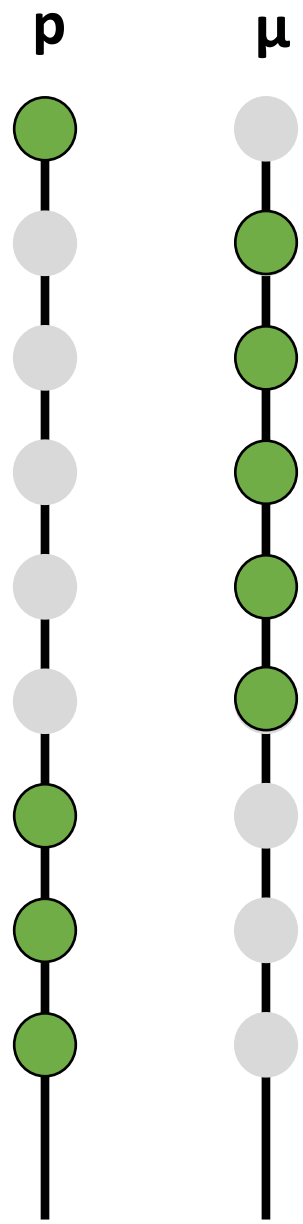
Players aim to maximize the sum of their rewards; can agree on their strategies before play

Theorem 3 (Cooperating players explore more). Suppose Alice and Bob are players with aligned interests playing a one armed bandit problem with discount factor β. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass.

Then there exists $\tilde{p} > g = g(\mu, \beta)$, so that ***for all*** $p < \hat{p}$, at least one of the players explores the risky (right) arm with positive probability under any optimal strategy pair maximizing their total reward.

Cooperating players explore more





Neutral setting

Neutral setting

Utility of each player is their own reward (selfish)

Solution concepts: Nash equilibrium and perfect Bayesian equilibrium.

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_i, σ_j) is a **Bayesian Nash equilibrium** if σ_i is a best response for each player i.

Neutral setting

A **Perfect Bayesian Equilibrium** is the version of subgame perfect equilibrium for games with incomplete information. A pair of strategies (σ_i, σ_j) is a perfect Bayesian equilibrium if

- starting from any information set, subsequent play is optimal, and
- beliefs are updated consistently with Bayes' rule on every path of play that occurs with positive probability.

Note: Such equilibria are guaranteed to exist in this setting; unlike Nash equilibria, there cannot be ***non-credible threats***.

Neutral setting

Does each neutral player play the one player optimum strategy? (i.e. pull the arm with highest Gittins index in each round)

Neutral setting

Theorem 4 (Neutral players learn from each other).

Neutral setting

Theorem 4 (Neutral players learn from each other). Let Alice and Bob be neutral players in a one armed bandit problem with discount factor β. 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***:

Neutral setting

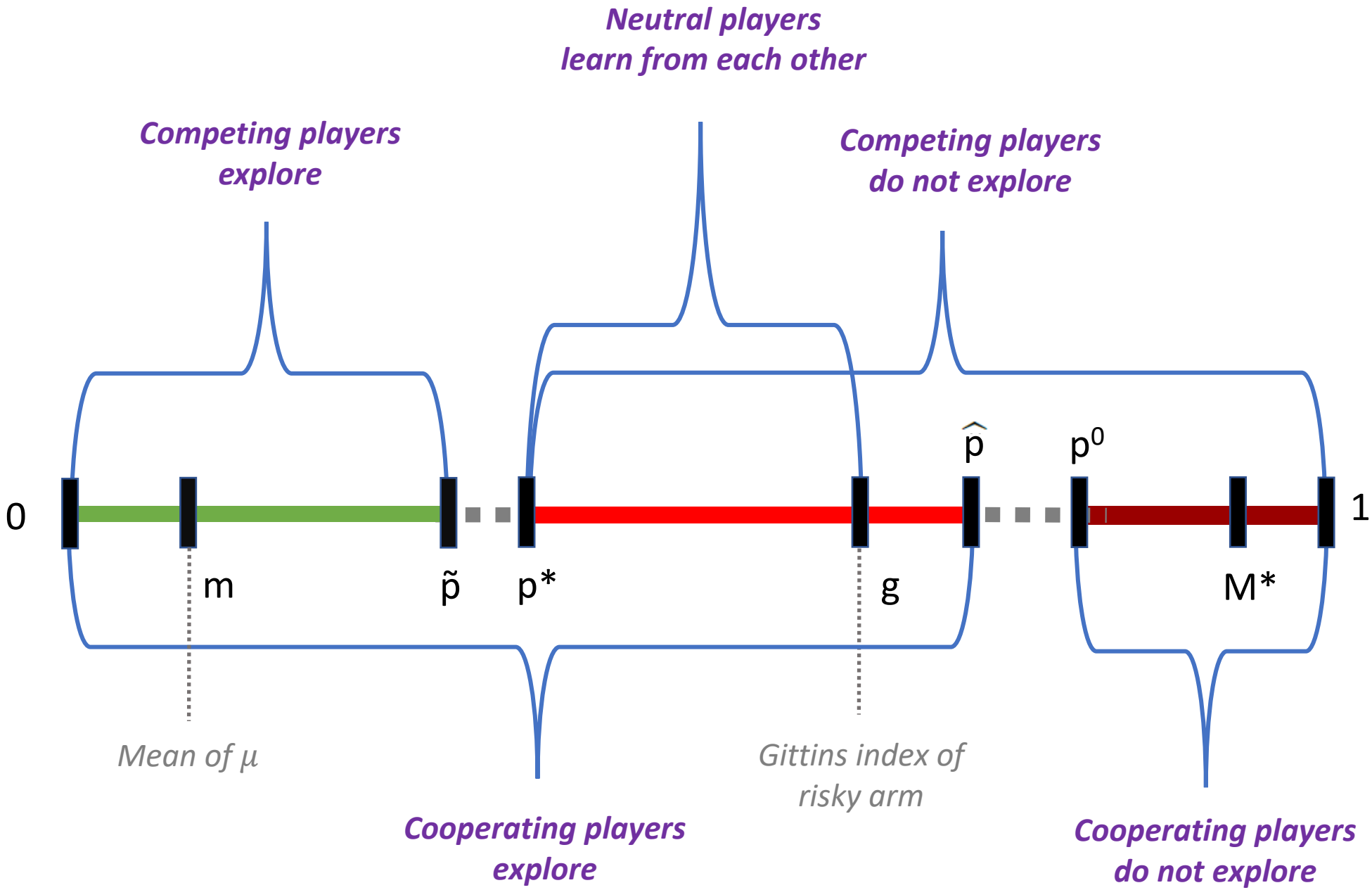
Theorem 4 (Neutral players learn from each other). Let Alice and Bob be neutral players in a one armed bandit problem with discount factor β. The left arm has success probability p and the right arm has prior distribution μ that is not a point mass. Then ***in any Nash equilibrium***:

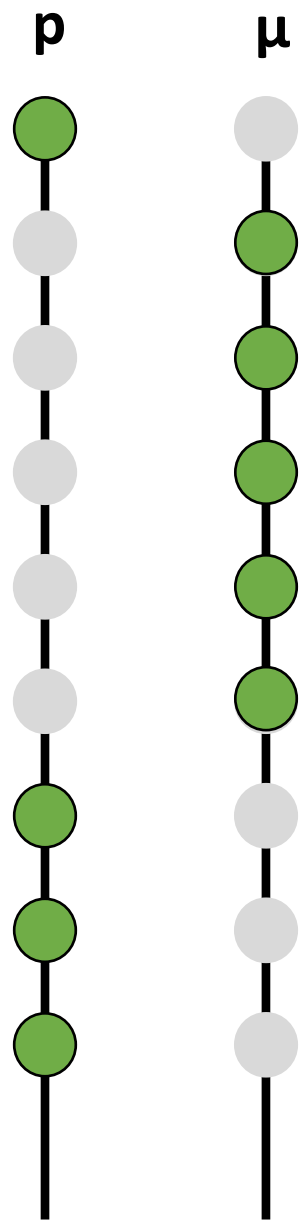
1. For all $p < g(\mu, \beta)$, with probability 1 at least one player explores. Moreover, the probability that no player explores by time t decays exponentially in t.

Neutral setting

Theorem 4 (Neutral players learn from each other). Let Alice and Bob be neutral players in a one armed bandit problem with discount factor β. 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***. If the equilibrium is furthermore perfect Bayesian, then ***every (neutral) player has expected reward strictly higher than a single player using an optimal strategy***.


Neutral players
learn from each other
Competing players
explore
Competing players
do not explore
0
1
m
p̃
p*
g
p̂
p⁰
M*
Mean of μ
Gittins index of
risky arm
Cooperating players
explore
Cooperating players
do not explore


p
μ


SHORT TERM
LONG TERM

Long term behavior

What do strategies look like in the long term?

The Rothschild conjecture

- **Rothschild [1974]** studies a single-person two-armed bandit, and shows that the player ends up with the wrong arm with positive probability. Rothschild conjectures that two players observing each other's actions may settle on different arms.

The Rothschild conjecture

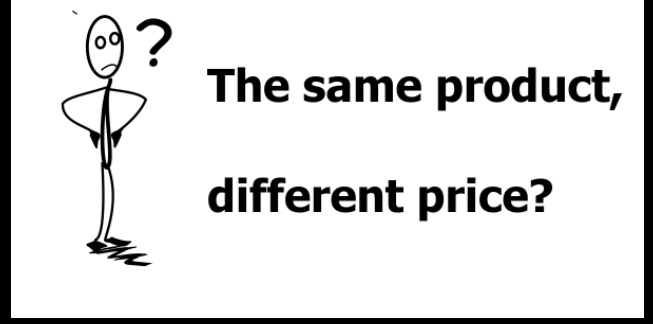
- **Rothschild [1974]** studies a single-person two-armed bandit, and shows that the player ends up with the wrong arm with positive probability. Rothschild conjectures that two players observing each other's actions may settle on different arms.

- **High level reasoning:** When a single player plays a two-armed bandit, he settles on the wrong arm with positive probability because he will give up the right arm if he happens to have bad draws on that arm. Even if there are two players, therefore, they may settle on different arms both thinking it is the other player who is playing the wrong arm after having had bad draws on the right arm. ***[Discussion in Ayoyagi '98]***

The Rothschild conjecture

- **Rothschild [1974]** studies a single-person two-armed bandit, and shows that the player ends up with the wrong arm with positive probability. Rothschild conjectures that two players observing each other's actions may settle on different arms.
- **High level reasoning:** When a single player plays a two-armed bandit, he settles on the wrong arm with positive probability because he will give up the right arm if he happens to have bad draws on that arm. Even if there are two players, therefore, they may settle on different arms both thinking it is the other player who is playing the wrong arm after having had bad draws on the right arm. ***[Discussion in Ayoyagi '98]***
- Ayoyagi [98, 01] proves convergence in discrete case.

The Rothschild conjecture



Rothschild writes

- ``... One could well ask whether they (stores) would be content charging the prices that they think are best while observing that other stores presumably rational are charging different prices. I do not think this is a particularly compelling point.
- Unless store A has access to store B's books, the mere fact that store B is charging a price different from A's and not going bankrupt is not conclusive evidence that A is doing the wrong thing. Who is to say A's experience is not a better guide to the true state of affairs than B's?''

Aumann’s agreement theorem (1976): rational players with common knowledge of each other's beliefs cannot agree to disagree.

Aumann’s agreement theorem (1976): rational players with common knowledge of each other's beliefs cannot agree to disagree.

But the bandit setting has elements not found in the setting of Aumann’s theorem: players keep getting different information.

Long term behavior

- When $\lambda = 1$ there are Nash equilibria where (aligned) players do not settle on the same arm; one player alternates infinitely often between the two arms.

Long term behavior

- When $\lambda = 1$ there are Nash equilibria where (aligned) players do not settle on the same arm; one player alternates infinitely often between the two arms.

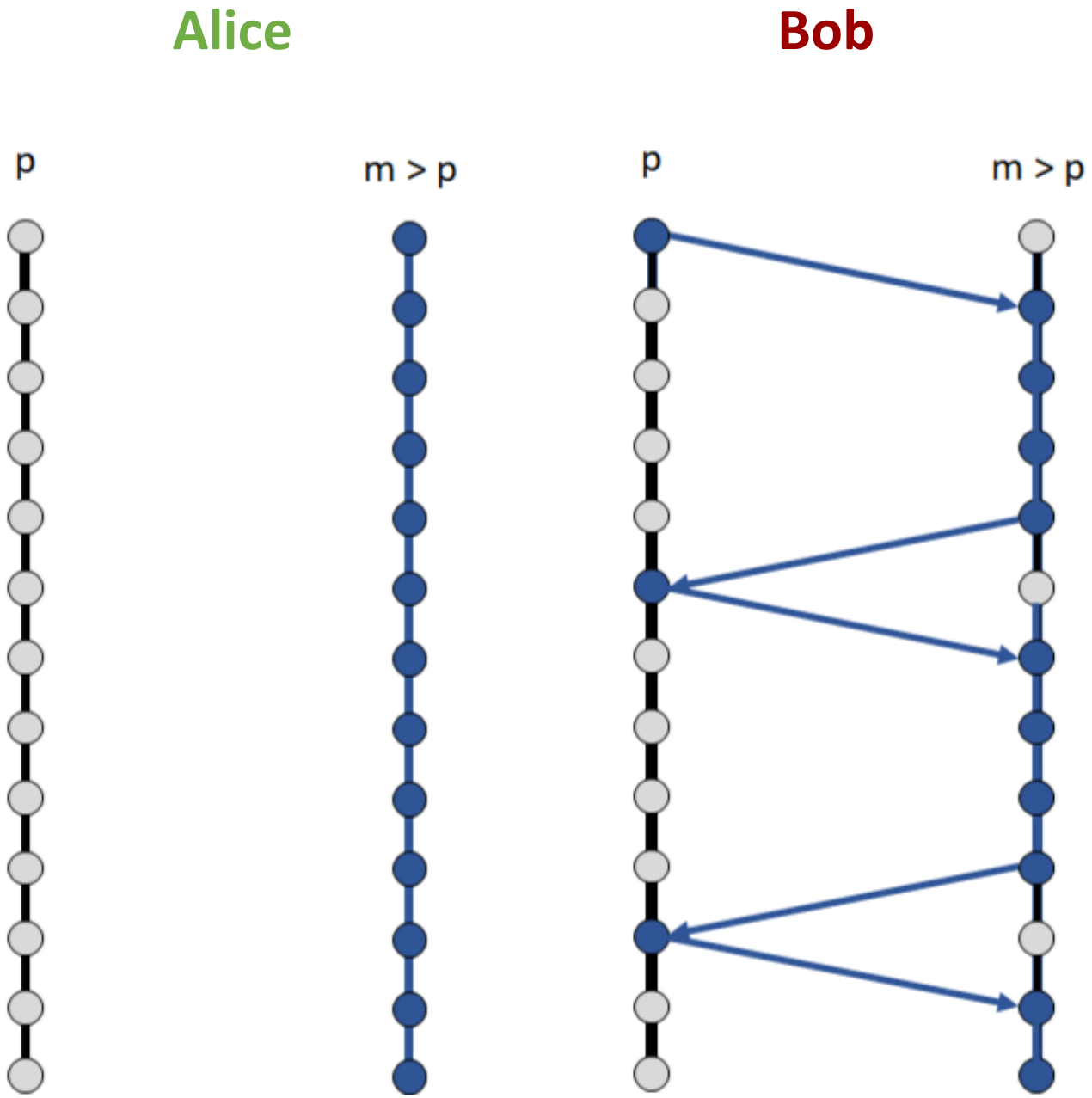
Example (Nash equilibria where players do not converge, $\lambda = 1$). Suppose Alice and Bob are aligned players in a one-armed bandit problem with discount factor β, where the left arm has success probability p and the right arm has prior distribution μ that is a point mass at $m > p$.

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

Long term behavior

Nash equilibria where aligned players do not converge, λ= 1: Let $k \in N$.

- Bob's strategy S_B : play left in rounds $0, k, 2k, 3k, \ldots$ and right in the remaining rounds.
- Alice's strategy S_A : play right if Bob follows the trajectory above; if Bob ever deviates from S_B, then Alice switches to playing left forever.



Long term behavior

Theorem 5 (Competing and neutral players settle on the same arm).

Suppose Alice and Bob are playing a one-armed bandit game, where 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.

Long term behavior

Theorem 5 (Competing and neutral players settle on the same arm).

Intuition: 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, then $\Theta > p$.

So if Bob sees that Alice keeps exploring, he will eventually realize that $\Theta > p$ and will join her at the right arm.

Challenge: Θ might be very close to p, which delays the time at which Alice determines the better arm.

Discussion and open questions

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

Discussion and open questions

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

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?

Discussion and open questions

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

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?

Is $\tilde{p} = p^*$? (Monotonicity)

Discussion and open questions

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

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?

Is $\tilde{p} = p^*$? (Monotonicity)

Patent protection: each player learns the other player's rewards, but with a delay of k rounds, or is given a "patent" – the other player cannot explore for k rounds after its first exploration

Discussion and open questions

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

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?

Is $\tilde{p} = p^*$? (Monotonicity)

Patent protection: each player learns the other player's rewards, but with a delay of k rounds, or is given a "patent" – the other player cannot explore for k rounds after its first exploration

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 Rotschild's conjecture.

Discussion and open questions

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

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?

Is $\tilde{p} = p^*$? (Monotonicity)

Patent protection: each player learns the other player's rewards, but with a delay of k rounds, or is given a "patent" – the other player cannot explore for k rounds after its first exploration

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 Rotschild's conjecture.

Computational issues – finite memory for players?


THANKS

Competitive setting

Theorem 1 (Competing players explore less). Suppose arm L has 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 Alice and Bob are playing optimally in the zero sum game with discount factor β.

Then there exists a threshold $\boldsymbol{p^* < g}$, where $g = g(\mu, \beta)$ is the Gittins index of the right arm, such that **for all $\boldsymbol{p > p^*}$**, with probability 1 the players ***will not explore arm R***.

More precisely, $p^* \leq \frac{m \cdot \beta + g}{1+\beta}$, where m is the mean of μ.

Proof sketch for Theorem 1 (Competing players explore less).

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 we can calculate Alice and Bob's expected reward and bound them (since Bob is copying Alice, she is not learning from his actions), so from round k+1 on her maximum reward is what a single player can do.

Using inequality on Gittins index $g(\mu, \beta) \geq m + \frac{\beta w}{2}$

we obtain the conclusion.