Multiplayer Bandit Learning

Simina Branzei
Purdue University

Toulouse Workshop on Learning in Games
July 2024

One player bandit learning

Will start with one decision maker that has to pick between different actions.

Example – gold mining

Alice has n gold mines and a gold-mining machine.

Example – gold mining

Alice has n gold mines and a gold-mining machine.

Each day she must assign the machine to one of the mines.

Example – gold mining

Alice has n gold mines and a gold-mining machine.

Each day she must assign the machine to one of the mines. When the machine is assigned to mine i, there is a probability p_i that it extracts a proportion q_i of the gold left in the mine and a probability $1 - p_i$ that it extracts no gold and breaks down permanently.

Example – gold mining

Alice has n gold mines and a gold-mining machine.

Each day she must assign the machine to one of the mines. When the machine is assigned to mine i, there is a probability p_i that it extracts a proportion q_i of the gold left in the mine and a probability $1 - p_i$ that it extracts no gold and breaks down permanently.

Question: to what sequence of mines should the machine be assigned before it breaks down?

Example – search

An object is hidden in one of n boxes.

Example – search

An object is hidden in one of n boxes. The probability that a search in box i finds the object (assuming it is in box i) is q_i.

Example – search

An object is hidden in one of n boxes. The probability that a search in box i finds the object (assuming it is in box i) is q_i. The probability that the object is in box i is p_i and changes by Bayes' Theorem as successive boxes are searched.

Example – search

An object is hidden in one of n boxes. The probability that a search in box i finds the object (assuming it is in box i) is q_i. The probability that the object is in box i is p_i and changes by Bayes' Theorem as successive boxes are searched. The cost of a single search of box i is c_i.

Example – search

An object is hidden in one of n boxes. The probability that a search in box i finds the object (assuming it is in box i) is q_i. The probability that the object is in box i is p_i and changes by Bayes' Theorem as successive boxes are searched. The cost of a single search of box i is c_i.

Question: in what sequence should the boxes be searched to minimize the expected cost of finding the object?

Example – medical treatments

There are n treatments (arms) that must be pulled (allocated to patients) in some order. Each pull results in a success or a failure.

Example – medical treatments

There are n treatments (arms) that must be pulled (allocated to patients) in some order. Each pull results in a success or a failure.

The sequence of successes and failures from pulling arm i forms a Bernoulli process with unknown success probability Θ_i.

Example – medical treatments

There are n treatments (arms) that must be pulled (allocated to patients) in some order. Each pull results in a success or a failure.

The sequence of successes and failures from pulling arm i forms a Bernoulli process with unknown success probability Θ_i.

- A **success** at the t-th pull yields reward 1, while a **failure** yields reward zero.

Example – medical treatments

There are n treatments (arms) that must be pulled (allocated to patients) in some order. Each pull results in a success or a failure.

The sequence of successes and failures from pulling arm i forms a Bernoulli process with unknown success probability Θ_i.

- A **success** at the t-th pull yields reward 1, while a **failure** yields reward zero.
- At time $t = 0$, each Θ_i has a beta prior distribution; the distributions are independent for different arms.

Example – medical treatments

There are n treatments (arms) that must be pulled (allocated to patients) in some order. Each pull results in a success or a failure.

The sequence of successes and failures from pulling arm i forms a Bernoulli process with unknown success probability Θ_i.

- A **success** at the t-th pull yields reward 1, while a **failure** yields reward zero.
- At time $t = 0$, each Θ_i has a beta prior distribution; the distributions are independent for different arms.

The prior distributions are converted by Bayes' theorem to successive posterior distributions as arms are pulled (Note the posterior distributions are beta distributions too).

Example – medical treatments

There are n treatments (arms) that must be pulled (allocated to patients) in some order. Each pull results in a success or a failure.

The sequence of successes and failures from pulling arm i forms a Bernoulli process with unknown success probability Θ_i.

- A **success** at the t-th pull yields reward 1, while a **failure** yields reward zero.
- At time $t = 0$, each Θ_i has a beta prior distribution; the distributions are independent for different arms.

The prior distributions are converted by Bayes' theorem to successive posterior distributions as arms are pulled (Note the posterior distributions are beta distributions too).

Question: in what order should the arms be pulled to maximize total expected (discounted) reward from an infinite sequence of pulls?

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

E.g., if the current distribution for Θ_i is $Beta(\alpha_i, \beta_i)$, for $i = 1, \ldots, n$, pull next arm j such that

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \max_{i=1,\ldots,n} \frac{\alpha_i}{\alpha_i + \beta_i}.$$

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

E.g., if the current distribution for Θ_i is $Beta(\alpha_i, \beta_i)$, for $i = 1, \ldots, n$, pull next arm j such that

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \max_{i=1,\ldots,n} \frac{\alpha_i}{\alpha_i + \beta_i}.$$

This policy is intuitive, but not always optimal.

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

E.g., if the current distribution for Θ_i is $Beta(\alpha_i, \beta_i)$, for $i = 1, \ldots, n$, pull next arm j such that

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \max_{i=1,\ldots,n} \frac{\alpha_i}{\alpha_i + \beta_i}.$$

This policy is intuitive, but not always optimal.

Why? Suppose $n = 2$ and $\frac{\alpha_1}{\alpha_1+\beta_1} = \frac{\alpha_2}{\alpha_2+\beta_2}$.

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

E.g., if the current distribution for Θ_i is $Beta(\alpha_i, \beta_i)$, for $i = 1, \dots, n$, pull next arm j such that

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \max_{i=1,\dots,n} \frac{\alpha_i}{\alpha_i + \beta_i}.$$

This policy is intuitive, but not always optimal.

Why? Suppose $n = 2$ and $\frac{\alpha_1}{\alpha_1+\beta_1} = \frac{\alpha_2}{\alpha_2+\beta_2}$. Then the policy suggests that pulling either arm in the next step is optimal.

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

E.g., if the current distribution for Θ_i is $Beta(\alpha_i, \beta_i)$, for $i = 1, \dots, n$, pull next arm j such that

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \max_{i=1,\dots,n} \frac{\alpha_i}{\alpha_i + \beta_i}.$$

This policy is intuitive, but not always optimal.

Why? Suppose $n = 2$ and $\frac{\alpha_1}{\alpha_1+\beta_1} = \frac{\alpha_2}{\alpha_2+\beta_2}$. Then the policy suggests that pulling either arm in the next step is optimal.

But suppose $\alpha_2 + \beta_2 \gg \alpha_1 + \beta_1$. Then the variance at arm 1 is much greater than at arm 2: even a few pulls can change $\frac{\alpha_1}{\alpha_1+\beta_1}$ by a lot.

Example – medical treatments

In this case the solution is obvious: at each step, pull the arm with highest expected value of Θ_i currently.

E.g., if the current distribution for Θ_i is $Beta(\alpha_i, \beta_i)$, for $i = 1, \ldots, n$, pull next arm j such that

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \max_{i=1,\ldots,n} \frac{\alpha_i}{\alpha_i + \beta_i}.$$

This policy is intuitive, but not always optimal.

Why? Suppose $n = 2$ and $\frac{\alpha_1}{\alpha_1+\beta_1} = \frac{\alpha_2}{\alpha_2+\beta_2}$. Then the policy suggests that pulling either arm in the next step is optimal.

But suppose $\alpha_2 + \beta_2 \gg \alpha_1 + \beta_1$. Then the variance at arm 1 is much greater than at arm 2: even a few pulls can change $\frac{\alpha_1}{\alpha_1+\beta_1}$ by a lot.

Thus an optimal strategy should pull arm 1 next, since the immediate expected rewards from the two arms are the same, but there is more info to be gained from pulling arm 1.

Multi-armed bandit model with one player

A gambler/player can play any of n one-armed bandit machines.

Multi-armed bandit model with one player

A gambler/player can play any of n one-armed bandit machines. The set of bandits is $[n] = \{1, \dots, n\}$; each bandit is a Markov process.

Multi-armed bandit model with one player

A gambler/player can play any of n one-armed bandit machines. The set of bandits is $[n] = \{1, \ldots, n\}$; each bandit is a Markov process.

The goal of the gambler is to maximize its expected total discounted reward.

Multi-armed bandit model with one player

A gambler/player can play any of n one-armed bandit machines. The set of bandits is $[n] = \{1, \dots, n\}$; each bandit is a Markov process.

The goal of the gambler is to maximize its expected total discounted reward.

The state of bandit j at time $t \in \{0,1, \dots\}$ is denoted $x_j(t)$.

Multi-armed bandit model with one player

A gambler/player can play any of n one-armed bandit machines. The set of bandits is $[n] = \{1, \dots, n\}$; each bandit is a Markov process.

The goal of the gambler is to maximize its expected total discounted reward.

The state of bandit j at time $t \in \{0,1, \dots\}$ is denoted $x_j(t)$.

When playing bandit j, the player 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.

Multi-armed bandit model with one player

A policy states 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 t.

Multi-armed bandit model with one player

A policy states 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 t.

The goal is to find a policy that maximizes the expected discounted reward:

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

where β is a discount factor and $x(0) = \left(x_1(0), \dots, x_n(0)\right)$ is the vector of initial states.

Multi-armed bandit model with one player

A policy states 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 t.

The goal is to find a policy that maximizes the expected discounted reward:

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

where β is a discount factor and $x(0) = \left(x_1(0), \ldots, x_n(0)\right)$ is the vector of initial states.

Question: what is the optimal policy?

The Gittins index

The optimal policy is described by functions G_j, which are known as Gittins indices. Every function G_j only depends on the state of bandit j.

Gittins and Jones ('74) showed that playing bandit j at time t is optimal if and if

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

The Gittins index

The optimal policy is described by functions G_j, which are known as Gittins indices. Every function G_j only depends on the state of bandit j.

Gittins and Jones ('74) showed that playing bandit j at time t is optimal if and if

$$G_j\left(x_j(t)\right) = \max_{i=1,\dots,n} G_i\left(x_i(t)\right).$$

That is, at each time step, compute the Gittins index of each arm and pull the arm with the highest index at that point in time. This strategy is deterministic and represents independence of irrelevant alternatives.

The Gittins index

The optimal policy is described by functions G_j, which are known as Gittins indices. Every function G_j only depends on the state of bandit j.

Gittins and Jones ('74) showed that playing bandit j at time t is optimal if and if

$$G_j\left(x_j(t)\right) = \max_{i=1,\ldots,n} G_i\left(x_i(t)\right).$$

Interpretation as retirement value:

The Gittins index

The optimal policy is described by functions G_j, which are known as Gittins indices. Every function G_j only depends on the state of bandit j.

Gittins and Jones ('74) showed that playing bandit j at time t is optimal if and if

$$G_j\left(x_j(t)\right) = \max_{i=1,\ldots,n} G_i\left(x_i(t)\right).$$

Interpretation as retirement value: Suppose at every step the gambler can

- retire and receive a payment p every round from now onwards; or

The Gittins index

The optimal policy is described by functions G_j, which are known as Gittins indices. Every function G_j only depends on the state of bandit j.

Gittins and Jones ('74) showed that playing bandit j at time t is optimal if and if

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

Interpretation as retirement value: Suppose at every step the gambler can

- retire and receive a payment p every round from now onwards; or
- pull arm j, receive the current reward at arm j, while keeping the option to retire at any point in the future.

The Gittins index

The optimal policy is described by functions G_j, which are known as Gittins indices. Every function G_j only depends on the state of bandit j.

Gittins and Jones ('74) showed that playing bandit j at time t is optimal if and if

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

Interpretation as retirement value: Suppose at every step the gambler can

- retire and receive a payment p every round from now onwards; or
- pull arm j, receive the current reward at arm j, while keeping the option to retire at any point in the future.

Given that arm j is currently at state x_j, the Gittins index $\mathrm{G}(x_j)$ is the infimum of the values p for which retirement now is preferable.

Special case: Bernoulli bandits

In the setting of Bernoulli bandits, the rewards are 1 (success) or 0 (failure).

Special case: Bernoulli bandits

In the setting of Bernoulli bandits, the rewards are 1 (success) or 0 (failure).

Arm j has known prior μ_j^0 on $[0,1]$; the success probability Θ_j of the arm is unknown to the player and drawn from μ_j^0.

Special case: Bernoulli bandits

In the setting of Bernoulli bandits, the rewards are 1 (success) or 0 (failure).

Arm j has known prior μ_j^0 on $[0,1]$; the success probability Θ_j of the arm is unknown to the player and drawn from μ_j^0.

The state of arm j at time t is described by a pair $\left(s_j(t), f_j(t)\right)$, where $s_j(t)$ and $f_j(t)$ are the number of successes and failures, respectively, obtained at arm j until time t.

Special case: Bernoulli bandits

In the setting of Bernoulli bandits, the rewards are 1 (success) or 0 (failure).

Arm j has known prior μ_j^0 on $[0,1]$; the success probability Θ_j of the arm is unknown to the player and drawn from μ_j^0.

The state of arm j at time t is described by a pair $\left(s_j(t), f_j(t)\right)$, where $s_j(t)$ and $f_j(t)$ are the number of successes and failures, respectively, obtained at arm j until time t.

Bayesian updating is used to obtain posterior distribution μ_j^t of the success probability Θ_j after t steps; i.e. for any Borel set $A \subseteq [0,1]$, its density is given by

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

Special case: Bernoulli bandits

In the setting of Bernoulli bandits, the rewards are 1 (success) or 0 (failure).

Arm j has known prior μ_j^0 on $[0,1]$; the success probability Θ_j of the arm is unknown to the player and drawn from μ_j^0.

The state of arm j at time t is described by a pair $\left(s_j(t), f_j(t)\right)$, where $s_j(t)$ and $f_j(t)$ are the number of successes and failures, respectively, obtained at arm j until time t.

Note: If player pulls arm j at time $t+1$, its expected reward given the history is

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

This expected reward is also the transition probability from state $(s_j(t), f_j(t))$ to state $(s_j(t)+1, f_j(t))$.

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, $m_1 = \frac{1}{m} \int_0^1 x^2 \, d\mu$ the posterior mean at the right arm after observing 1 in round zero, and $w = \int (x-m)^2 \, d\mu = m \cdot (m_1 - m)$ is the variance of μ.

Recall for discount factor β, 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.

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 = $g(\mu, \beta)$.

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu,\beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

- Round zero: play right.

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

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

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

- 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.

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

- 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.

By definition of g, this Alice strategy is at most as good as retiring and receiving g forever.

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

- 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.

By definition of g, this Alice strategy is at most as good as retiring and receiving g forever. Thus

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

Gittins index is more than the mean of the arm

Lemma: Consider one arm with prior μ. Then $g(\mu, \beta) \geq m + \beta w/2$, where m is the mean, m_1 is the posterior mean at the right arm after observing 1 in round zero, and $w = m \cdot (m_1 - m)$ is the variance of μ.

Proof sketch. Suppose the gambler (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 Alice strategy:

- 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.

By definition of g, this Alice strategy is at most as good as retiring and receiving g forever. Thus

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

Using $w = m \cdot (m_1 - m)$ and rearranging $\bigstar$, we get $g(\mu, \beta) \geq m + \beta w/2$.

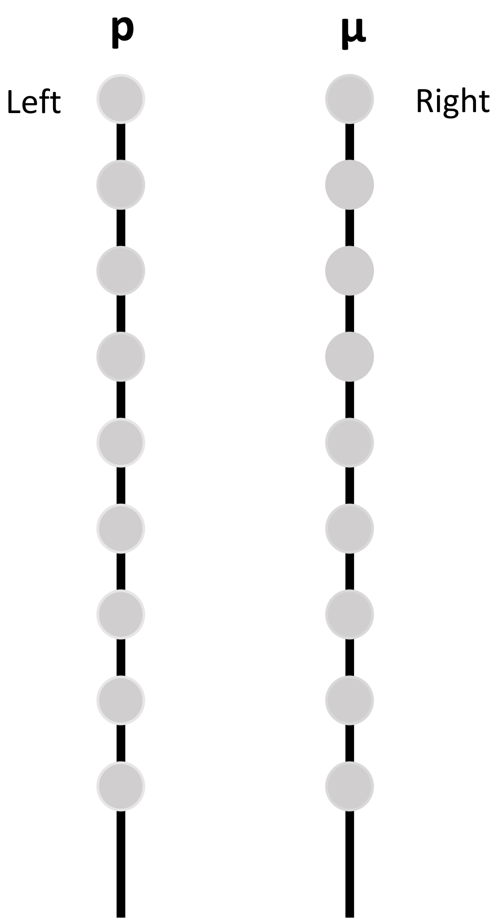
Multiplayer Model

Will summarize a framework on multiplayer bandit learning and results from “Multiplayer bandit learning, from competition to cooperation” (joint with Y. Peres, appeared in COLT ’21).

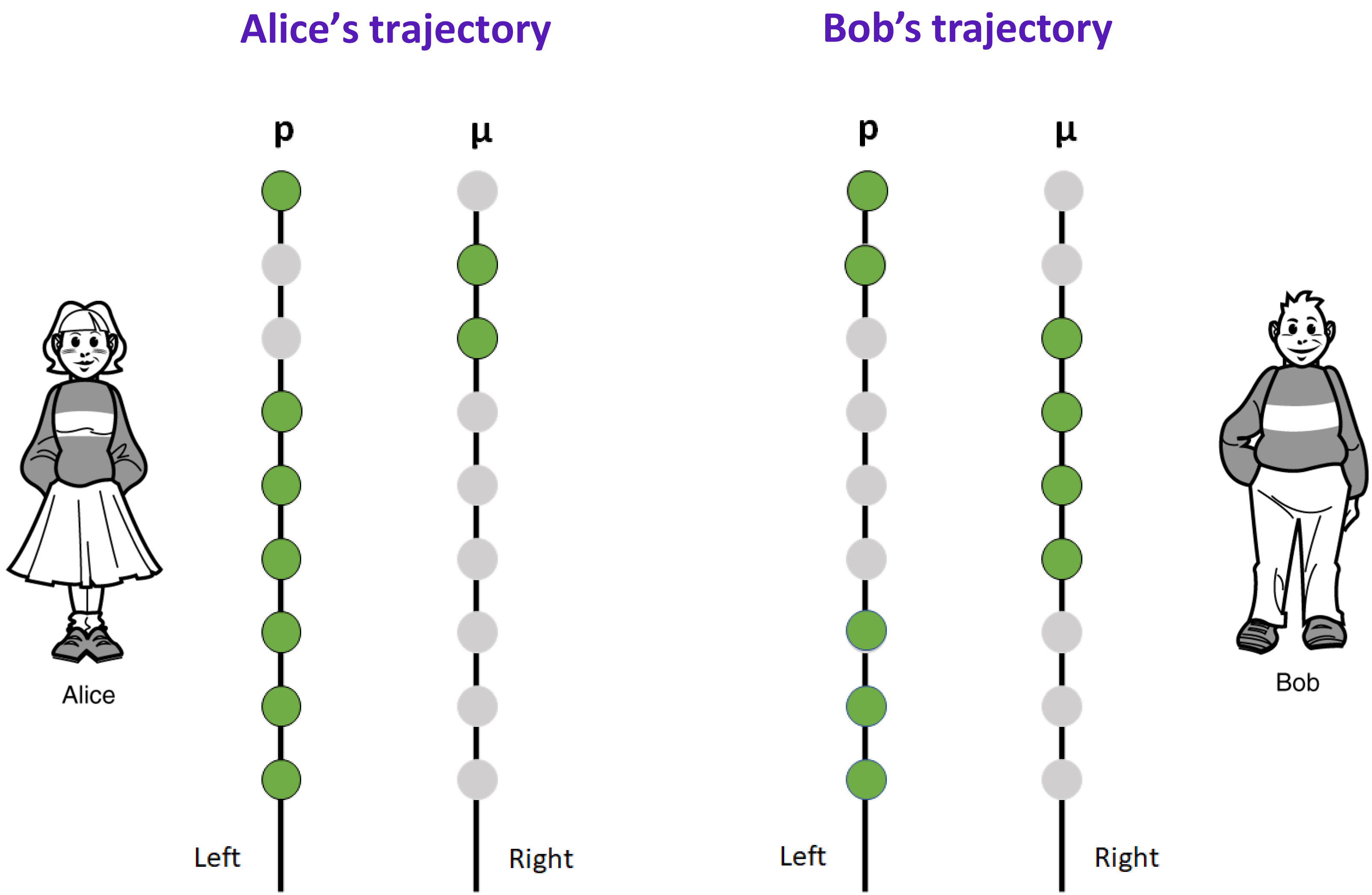
Multiplayer Model

- 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**.

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
Bob's trajectory
p
μ
p
μ
Alice
Bob
Left
Right
Left
Right

Utilities

$\gamma_A(t)$ is the random variable corresponding to Alice's **reward** in round t; similarly $\gamma_B(t)$ for Bob

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

- $\Gamma_A = \sum_{t=0}^{\infty} \gamma_A(t) \cdot \beta^t$ and $\Gamma_B = \sum_{t=0}^{\infty} \gamma_B(t) \cdot \beta^t$ are Alice and Bob's discounted rewards, respectively
- β is the discount factor: the game stops with probability $1 - \beta$ in each round (or one dollar today is worth β dollars tomorrow).

Utilities

$\gamma_A(t)$ is the random variable corresponding to Alice's **reward** in round t; similarly $\gamma_B(t)$ for Bob

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

- $\Gamma_A = \sum_{t=0}^{\infty} \gamma_A(t) \cdot \beta^t$ and $\Gamma_B = \sum_{t=0}^{\infty} \gamma_B(t) \cdot \beta^t$ are Alice and Bob's discounted rewards, respectively
- β is the discount factor: the game stops with probability $1 - \beta$ in each round (or one dollar today is worth β dollars tomorrow).

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

Will focus on discounted game; similar statements hold for finite horizon.

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., animals competing for food or phone companies competing for users in a saturated market)

Neutral setting

- $\boldsymbol{\lambda = 0}$: Each player's utility is their own rewards.
- So Alice's utility is Γ_A and Bob's utility is Γ_B.

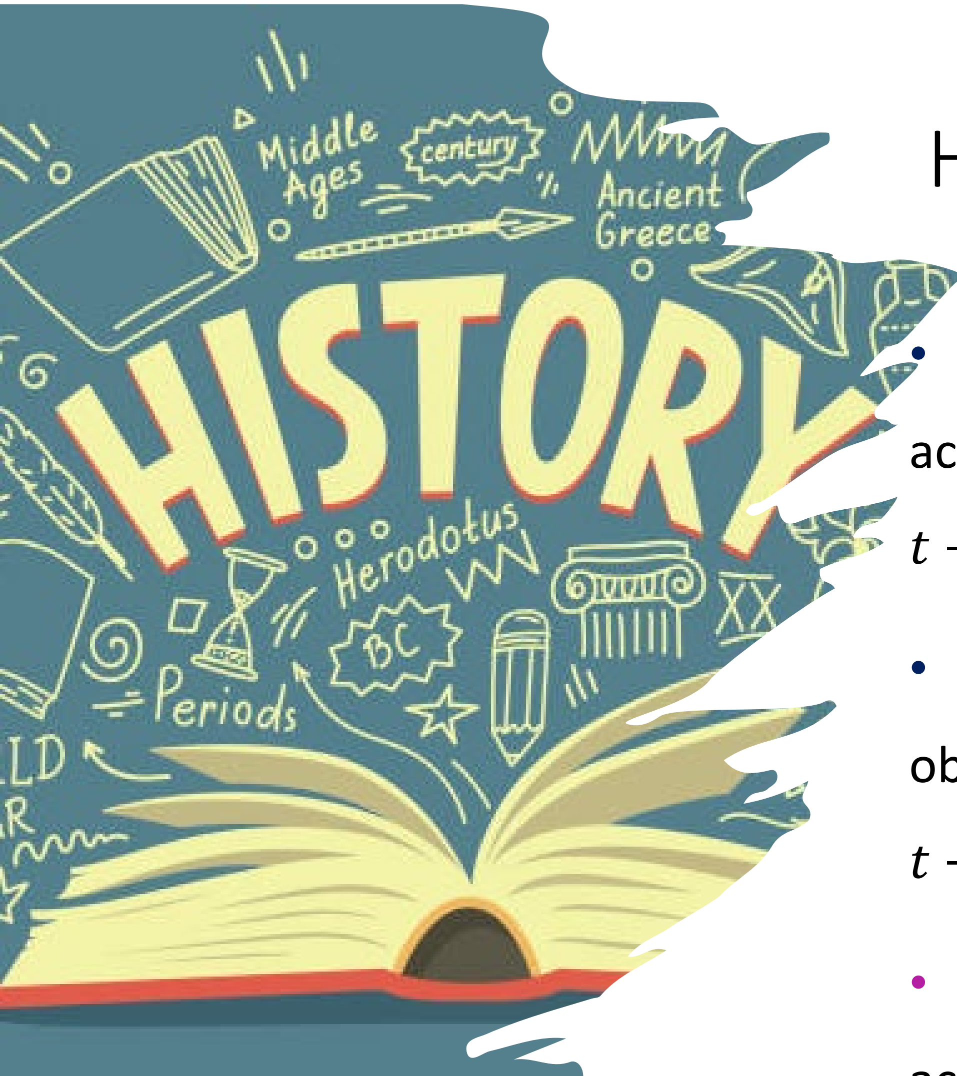
Cooperative setting

- $\boldsymbol{\lambda = 1}$: Both Alice and Bob have utility $\Gamma_A + \Gamma_B$

players are aligned, maximize total rewards collected (e.g. genetically identical organisms)

Partly cooperative setting

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



History and pure strategies

- **Public history at time t:** sequence of past actions of both players until the end of round $t-1$.
- **Private history of a player i at time t:** bits observed by player i until the end of round $t-1$.
- **Pure strategy:** map that tells a player what action to play at each point given the public and private history

Randomized Strategies

Mixed Strategy: probability distribution over pure strategies

Equivalent to **behavioral strategies**:

- Given by map that tells at each node, what probability mixture to play over the actions available at that node

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

Multiplayer bandits 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.

Multiplayer bandits 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]**)**

Multiplayer bandits 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)

Multiplayer bandits 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

Multiplayer bandits 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 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.

Multiplayer bandits literature

Survey on multiplayer bandits:

Boursier-Perchet 2024 summarize these various models and results.

Related literature in biology

Evolutionary biology

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


Competitive setting

Competitive setting

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

How do competing players behave?

Different hypotheses possible: they play as one player would (pulling the arm with the highest Gittins index in each round), or they both play the same arm in every round, or on the contrary they randomize…

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 the safe arm has known probability p and the risky arm 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

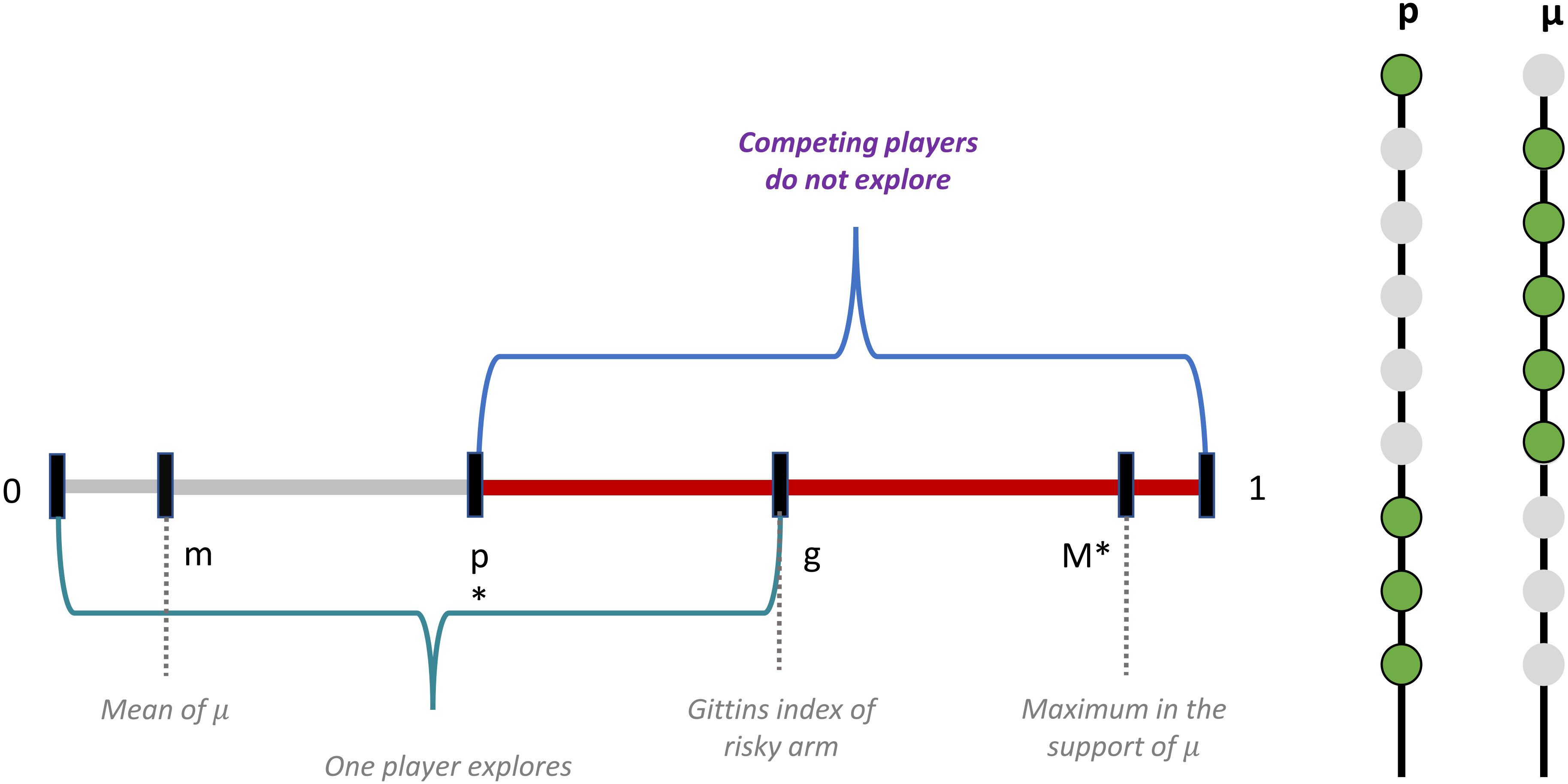
Zero-sum game has a value by Sion's minimax theorem.

Theorem 1 (Competing players explore less). Suppose the safe arm has known probability p and the risky arm 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 risky arm, such that **for all $\boldsymbol{p > p^*}$**, with probability 1 the players ***will not explore the risky arm***.

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

Competing players explore less



Competitive setting

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

Theorem 1 (Competing players explore less). Suppose the safe arm has known probability p and the risky arm 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 risky arm, such that **for all $\boldsymbol{p > p^*}$**, with probability 1 the players ***will not explore the risky arm***.

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

Competitive setting

Proof of Theorem 1 (Competing players explore less).

Recall Lemma: $g(\mu, \beta) \geq m + \frac{\beta w}{2}$, where $m = \int_0^1 x \, d\mu(x)$ is the mean of the risky arm, β is the discount factor, and $w = \int_0^1 (x - m)^2 d\mu(x) > 0$ is the variance of μ.

Competitive setting

Proof of Theorem 1 (Competing players explore less).

Recall Lemma: $g(\mu, \beta) \geq m + \frac{\beta w}{2}$, where $m = \int_0^1 x \, d\mu(x)$ is the mean of the risky arm, β is the discount factor, and $w = \int_0^1 (x-m)^2 d\mu(x) > 0$ is the variance of μ.

If $p > g$, it's easy to see neither player explores. So we can assume $p \leq g$.

Competitive setting

Proof of Theorem 1 (Competing players explore less).

Recall Lemma: $g(\mu,\beta) \geq m + \frac{\beta w}{2}$, where $m = \int_0^1 x \, d\mu(x)$ is the mean of the risky arm, β is the discount factor, and $w = \int_0^1 (x-m)^2 d\mu(x) > 0$ is the variance of μ.

If $p > g$, it's easy to see neither player explores. So we can assume $p \leq g$.

Consider the following Bob strategy S_B:

- *Play the safe arm until Alice plays the risky one, say in some round k.*
- *Then play the safe arm again in round k+1 and starting with round k + 2 copy Alice's move from the previous round.*

(In particular, Bob never plays the risky arm first.)

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and 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 "explores" (i.e. plays the risky arm) first, then: done.
- Else, suppose S_A explores first in round k:

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and 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 "explores" (i.e. plays the risky arm) first, then: done.
- Else, suppose S_A explores first in round k:

Alice's total reward has expectation

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

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and 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 "explores" (i.e. plays the risky arm) first, then: done.
- Else, suppose S_A explores first in round k:

Alice's total reward has expectation

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

Bob's total reward has expectation:

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

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

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and starting with round k+ 2 copy Alice's move from the previous round.*

If Alice's strategy S_A explores first in round k:

- Alice's total reward has expectation $\mathrm{E}[\Gamma_A] = \sum_{t=0}^{k-1} p \cdot \beta^t + \sum_{t=k}^{\infty} E[\gamma_A(t)] \cdot \beta^t$.
- Bob's total reward has expectation $\mathrm{E}[\Gamma_B] = \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+1}^{\infty} E[\gamma_A(t)] \cdot \beta^{t+1}$.

Since $E[\gamma_A(k)] = m$, the difference in rewards is:

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

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and starting with round k+ 2 copy Alice's move from the previous round.*

If Alice's strategy S_A explores first in round k:

- Alice's total reward has expectation $\mathrm{E}[\Gamma_A] = \sum_{t=0}^{k-1} p \cdot \beta^t + \sum_{t=k}^{\infty} E[\gamma_A(t)] \cdot \beta^t$.
- Bob's total reward has expectation $\mathrm{E}[\Gamma_B] = \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+1}^{\infty} E[\gamma_A(t)] \cdot \beta^{t+1}$.

Since $E[\gamma_A(k)] = m$, the difference in rewards is:

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

$$= (m\beta - p - p\beta)\beta^k + (1-\beta) \cdot \left(m\beta^k + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t\right).$$

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and starting with round k+ 2 copy Alice's move from the previous round.*

If Alice's strategy S_A explores first in round k:

- Alice's total reward has expectation $\mathrm{E}[\Gamma_A] = \sum_{t=0}^{k-1} p \cdot \beta^t + \sum_{t=k}^{\infty} E[\gamma_A(t)] \cdot \beta^t$.
- Bob's total reward has expectation $\mathrm{E}[\Gamma_B] = \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+1}^{\infty} E[\gamma_A(t)] \cdot \beta^{t+1}$.

Since $E[\gamma_A(k)] = m$, the difference in rewards is:

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) = (m\beta - p - p\beta)\beta^k + (1-\beta) \cdot \left(m\beta^k + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t \right).$$

Bob is copying Alice ⇒ she is not learning anything from him. So from round $k+1$ onwards her reward is at most the one player optimum $\frac{g}{1-\beta} \cdot \beta^k$.

Proof of Theorem 1 (Competing players explore less).

Bob's strategy: *Play the safe arm until Alice plays the risky one, say in some round k. Then play the safe arm again in round k+1 and starting with round k+ 2 copy Alice's move from the previous round.*

If Alice's strategy S_A explores first in round k:

- Alice's total reward has expectation $\mathrm{E}[\Gamma_A] = \sum_{t=0}^{k-1} p \cdot \beta^t + \sum_{t=k}^{\infty} E[\gamma_A(t)] \cdot \beta^t$.
- Bob's total reward has expectation $\mathrm{E}[\Gamma_B] = \sum_{t=0}^{k+1} p \cdot \beta^t + \sum_{t=k+1}^{\infty} E[\gamma_A(t)] \cdot \beta^{t+1}$.

Since $E[\gamma_A(k)] = m$, the difference in rewards is:

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) = (m\beta - p - p\beta)\beta^k + (1-\beta) \cdot \left(m\beta^k + \sum_{t=k+1}^{\infty} \mathbb{E}(\gamma_A(t)) \cdot \beta^t \right).$$

Bob is copying Alice ⇒ she is not learning anything from him. So from round $k+1$ onwards her reward is at most the one player optimum $\frac{g}{1-\beta} \cdot \beta^k$. Then the difference in rewards is at most

$$\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \leq \left(m\beta - p(1+\beta) + g \right)\beta^k .$$

Proof of Theorem 1 (Competing players explore less).

Recall lemma: $g(\mu, \beta) \geq m + \frac{\beta w}{2}$.

Recall difference in rewards is at most: $\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \leq \Big(m\beta - p(1+\beta) + g\Big)\beta^k$.

Proof of Theorem 1 (Competing players explore less).

Recall lemma: $g(\mu, \beta) \geq m + \frac{\beta w}{2}$.

Recall difference in rewards is at most: $\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \leq \left(m\beta - p(1+\beta) + g\right)\beta^k$.

The right hand side of the inequality above is negative when

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

Proof of Theorem 1 (Competing players explore less).

Recall lemma: $g(\mu, \beta) \geq m + \frac{\beta w}{2}$.

Recall difference in rewards is at most: $\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \leq \Big(m\beta - p(1+\beta) + g\Big)\beta^k$.

The right hand side of the inequality above is negative when

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

Note that $\frac{m\beta+g}{1+\beta} \in (m, g)$.

Proof of Theorem 1 (Competing players explore less).

Recall lemma: $g(\mu, \beta) \geq m + \frac{\beta w}{2}$.

Recall difference in rewards is at most: $\mathbb{E}(\Gamma_A) - \mathbb{E}(\Gamma_B) \leq \Big(m\beta - p(1+\beta) + g\Big)\beta^k$.

The right hand side of the inequality above is negative when

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

Note that $\frac{m\beta+g}{1+\beta} \in (m, g)$.

Thus the players do not explore the risky arm for any p above this threshold.

This also implies $p^* \leq \frac{m\beta+g}{1+\beta}$.

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).

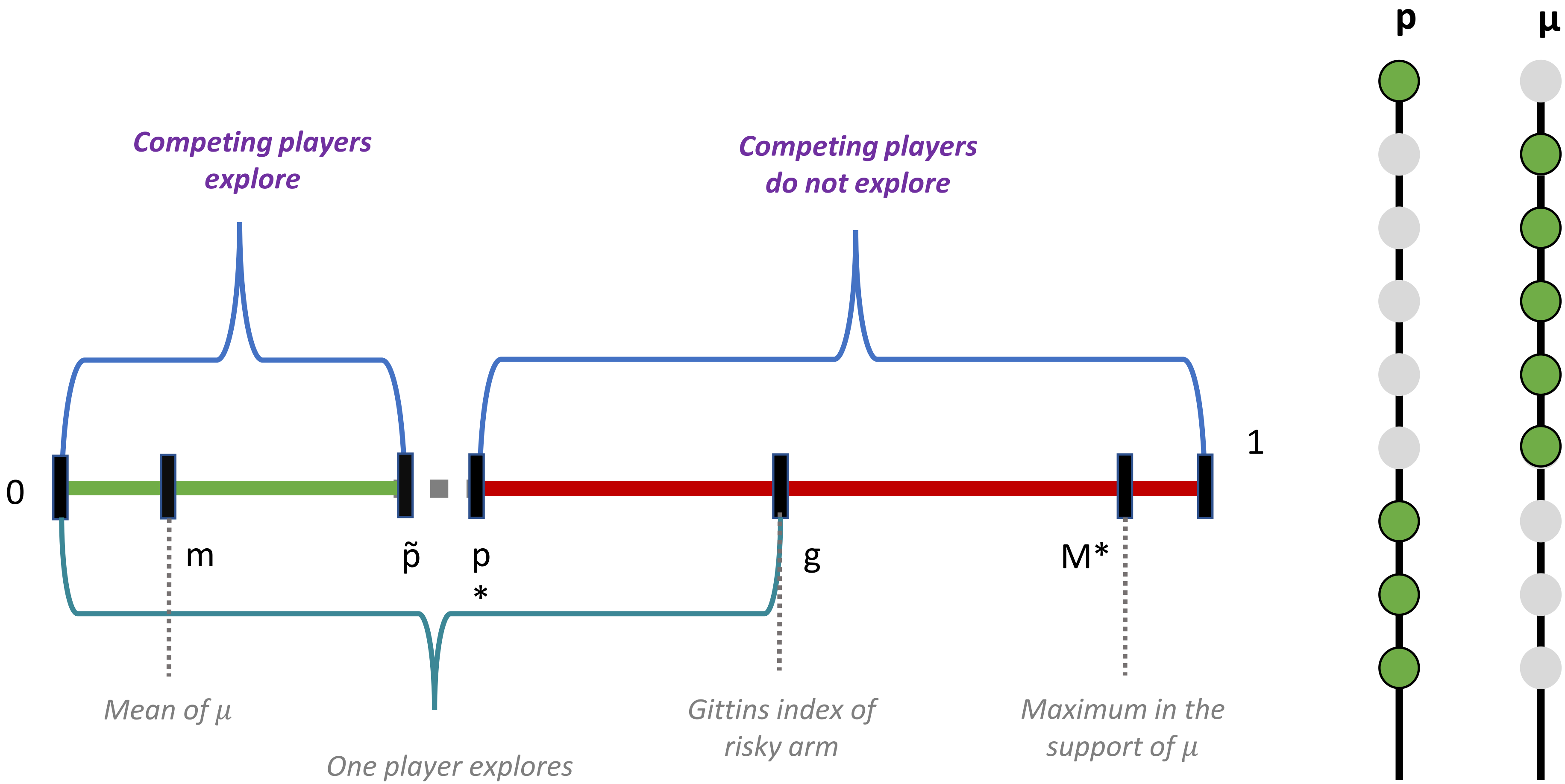
Competitive setting

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 the risky arm*** in the initial round of optimal play.

More precisely, $\tilde{p} \geq m + \frac{\beta w}{4}$, 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 safe arm has success probability p and the risky arm has prior distribution μ that is not a point mass.

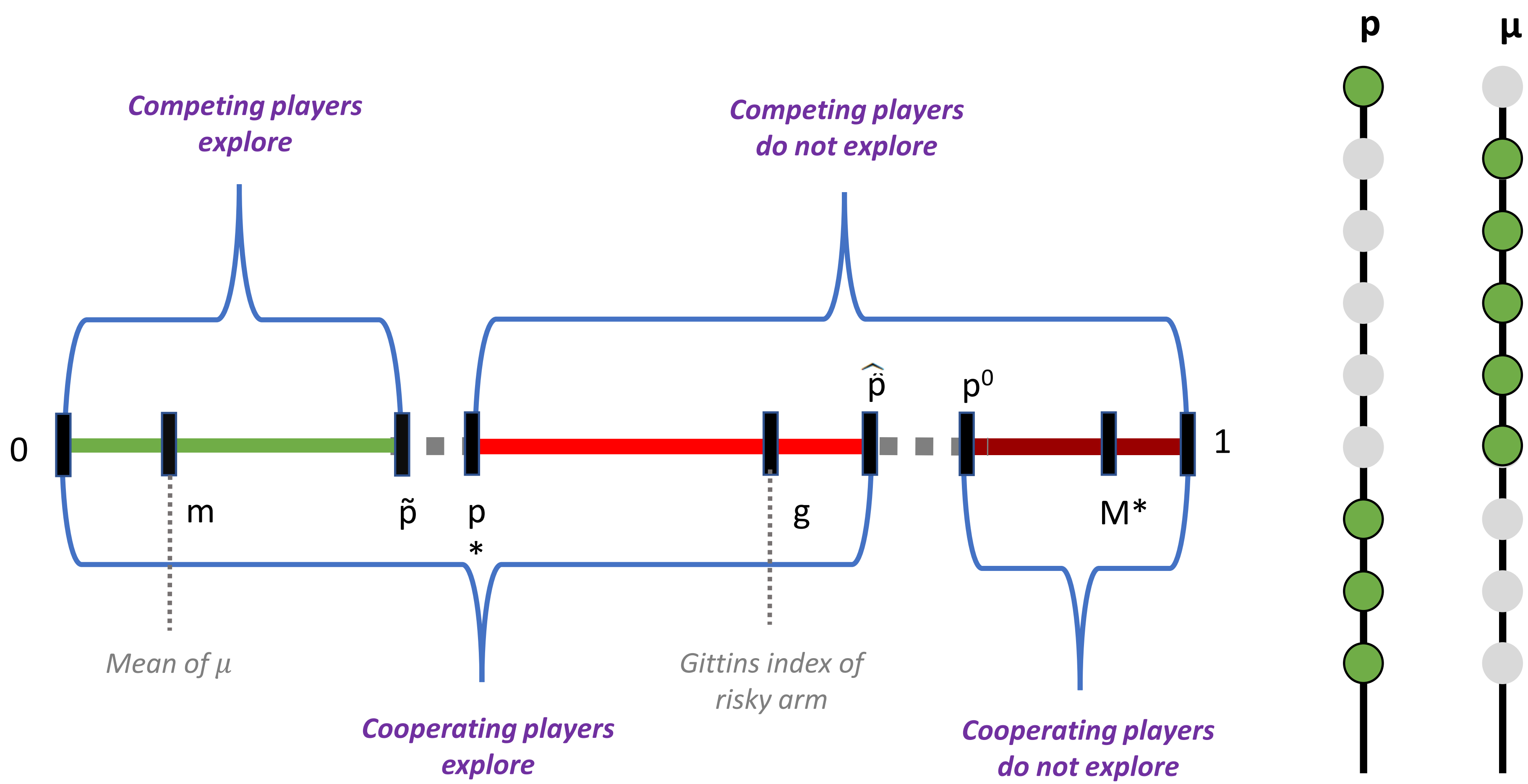
Cooperative setting

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 safe arm has success probability p and the risky 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 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 $\sigma_i{}'$ 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 safe arm has success probability p and the risky arm has prior distribution μ that is not a point mass. Then ***in any Nash equilibrium***:

Neutral setting

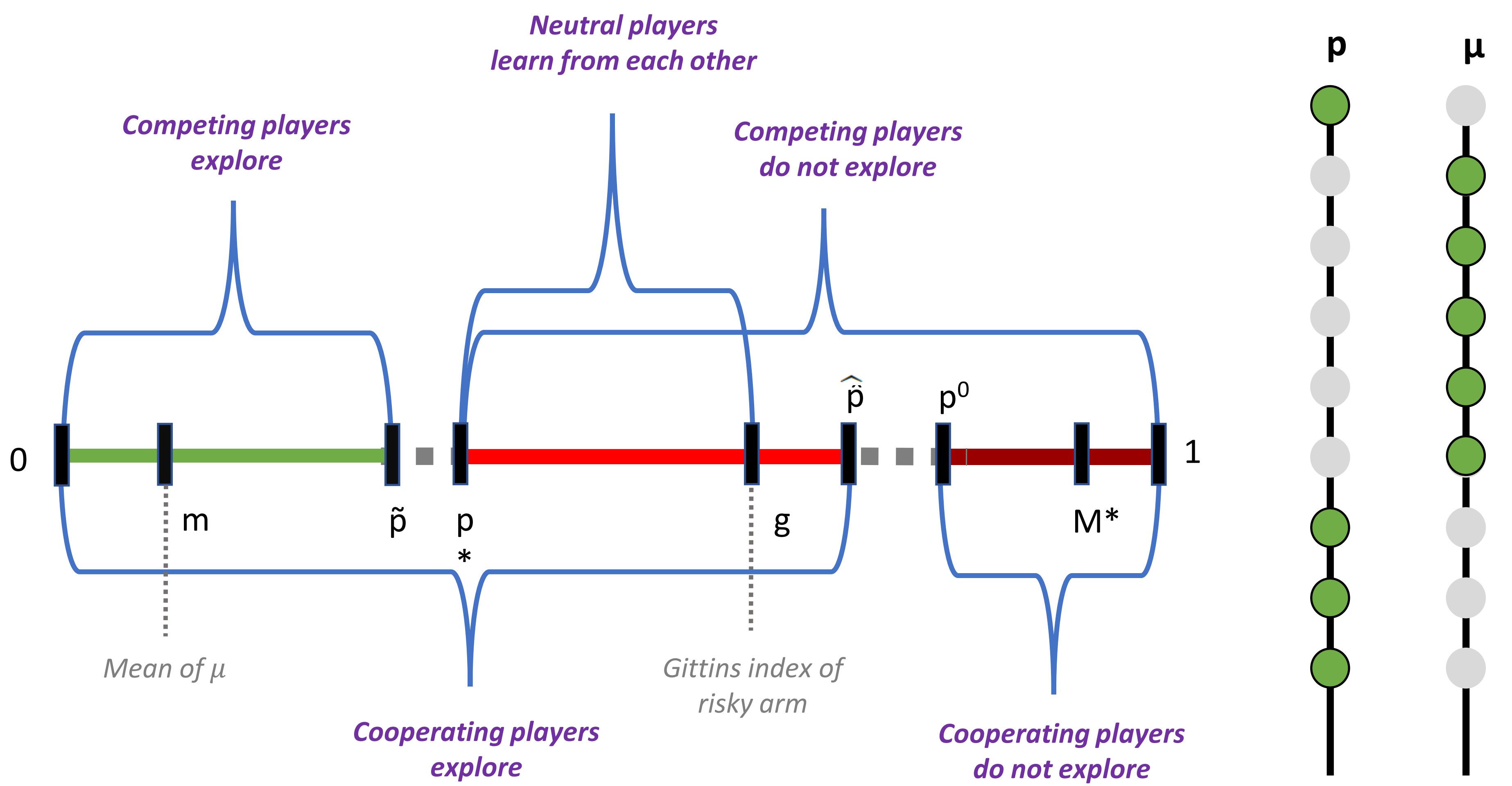
Theorem 4 (Neutral players learn from each other). Let Alice and Bob be neutral players in a one armed bandit problem with discount factor β. The safe arm has success probability p and the risky arm has prior distribution μ that is not a point mass. Then ***in any Nash equilibrium***:

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 safe arm has success probability p and the risky 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
m
p̃
p
*
g
p̂
p0
M*
1
Mean of μ
Gittins index of
risky arm
Cooperating players
explore
Cooperating players
do not explore
p
μ

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

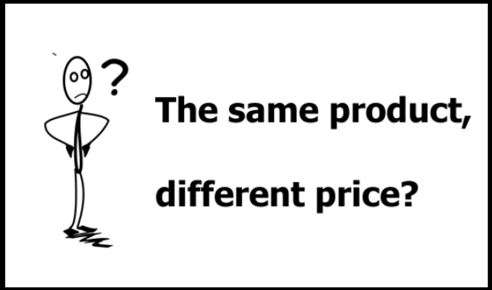
- **Rothschild [1974]** studies a single-person two-armed bandit, and shows that the player ends up with the wrong arm with positive probability. Rothschild conjectures that two players observing each other's actions may settle on different arms.

- **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 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 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 λ = 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.

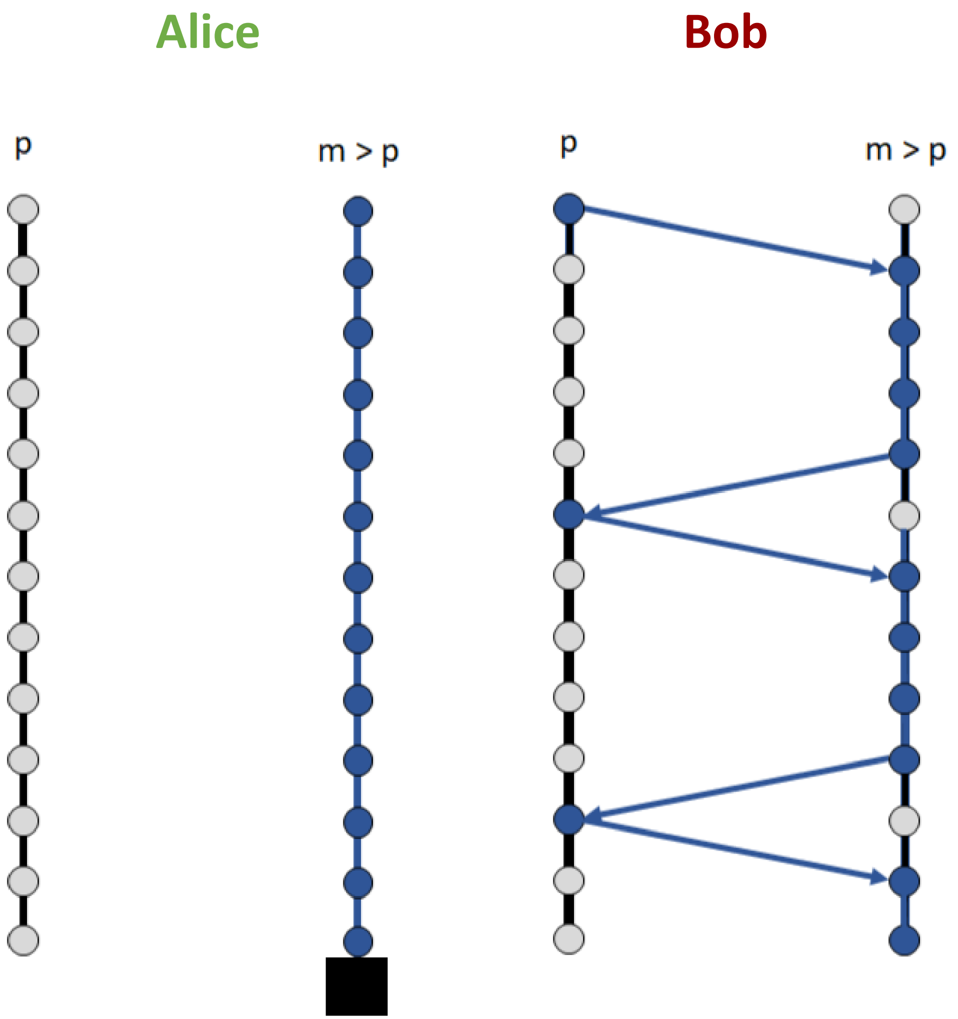
Example (Nash equilibria where players do not converge, $\lambda = 1$). Suppose Alice and Bob are aligned players in a one-armed bandit problem with discount factor β, where the left arm has success probability p and the right arm has prior distribution μ that is a point mass at $m > p$.

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

Long term behavior

Proof sketch (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 μ(p) = 0.

Then in any Nash equilibrium, in both the competing (λ=−1) and neutral (λ= 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.

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 $\Theta > p$.

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.

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?

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 com settle with probability 1 on the same arm in every Nash players, this is Rotschild's conjecture.

Computational issues – finite memory for players?

THANKS