Multi-Player Bandits without Communication

Mark Sellke

Based on collaborations with Sébastien Bubeck, Thomas Budzinski, Allen Liu

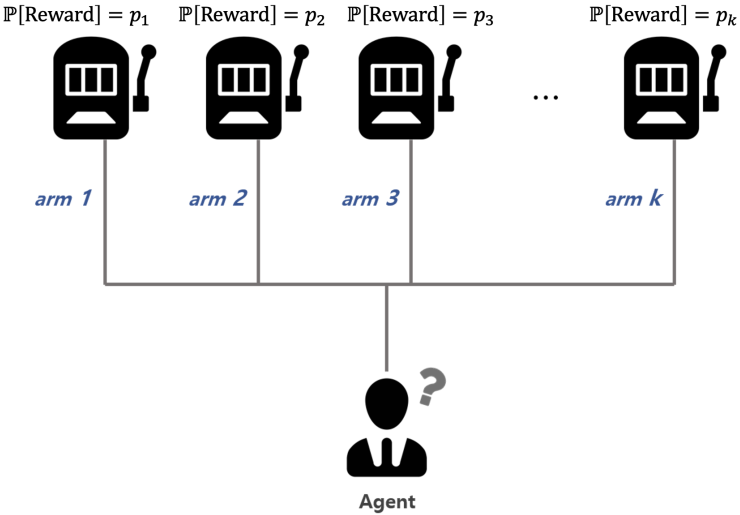


Plan for Today

1. Intro to multi-player (stochastic) bandits.
2. The power of (explicit or implicit) communication.
3. $T^{1/2}$ regret with no collisions.
4. Pareto optimal instance dependence with no communication.

Classical Bandits

Classic (stochastic) bandit problem: learn the best of K actions **online**.



Ordinary Bandits

K actions $a_1, \ldots, a_K$. Unknown reward probabilities $\mathbf{p} = (p_1, \ldots, p_K) \in [0,1]$.

Each time $t \in [T]$, play action a_{i_t}. Receive (and observe) reward

$$\mathrm{rew}_{i_t} \sim Ber(p_{i_t}) \in \{0,1\}.$$

Minimize expected regret

$$R_T(\mathbf{p}) = \mathbb{E}\left[T \cdot \max_i p_i \; - \; \sum_{i=1}^{T} \mathrm{rew}_{i_t}\right].$$

Ordinary Bandits

K actions $a_1, \ldots, a_K$. Unknown reward probabilities $\mathbf{p} = (p_1, \ldots, p_K) \in [0,1]$.

Each time $t \in [T]$, play action a_{i_t}. Receive (and observe) reward

$$\mathrm{rew}_{i_t} \sim Ber(p_{i_t}) \in \{0,1\}.$$

Minimize expected regret

$$R_T(\mathbf{p}) = \mathbb{E}\left[T \cdot \max_i p_i \; - \sum_{i=1}^{T} \mathrm{rew}_{i_t}\right].$$

Both **minimax** and **gap-dependent** regret are important. Gold standards:

Ordinary Bandits

K actions $a_1, \ldots, a_K$. Unknown reward probabilities $\mathbf{p} = (p_1, \ldots, p_K) \in [0,1]$.

Each time $t \in [T]$, play action a_{i_t}. Receive (and observe) reward

$$\mathrm{rew}_{i_t} \sim Ber(p_{i_t}) \in \{0,1\}.$$

Minimize expected regret

$$R_T(\mathbf{p}) = \mathbb{E}\left[T \cdot \max_i p_i \; - \sum_{i=1}^{T} \mathrm{rew}_{i_t} \right].$$

Both **minimax** and **gap-dependent** regret are important. Gold standards:

1. Minimax regret

$$R_T = \max_{\mathbf{p}} R_T(\mathbf{p}) \lesssim \sqrt{T}.$$

K actions $a_1, \ldots, a_K$. Unknown reward probabilities $\mathbf{p} = (p_1, \ldots, p_K) \in [0,1]$.

Each time $t \in [T]$, play action a_{i_t}. Receive (and observe) reward

$$\mathrm{rew}_{i_t} \sim Ber(p_{i_t}) \in \{0,1\}.$$

Minimize expected regret

$$R_T(\mathbf{p}) = \mathbb{E}\left[T \cdot \max_i p_i \; - \; \sum_{i=1}^{T} \mathrm{rew}_{i_t}\right].$$

Both **minimax** and **gap-dependent** regret are important. Gold standards:

1. Minimax regret
$$R_T = \max_{\mathbf{p}} R_T(\mathbf{p}) \lesssim \sqrt{T}.$$

2. Gap-dependent regret (with $\Delta = p_1^* - p_2^*$ the gap between best and 2nd best):
$$R_{T,\Delta} = \max_{\Delta(\mathbf{p}) \geq \Delta} R_T(\mathbf{p}) \lesssim \frac{\log(T)}{\Delta}.$$

Multi-player (Cooperative) Bandits

Consider $m > 1$ players $X, Y, \dots$. We assume colliding on the same action $i_t^X = i_t^Y$ is very bad and yields zero reward.

Multi-player (Cooperative) Bandits

Consider $m > 1$ players $X, Y, \ldots$. We assume colliding on the same action $i_t^X = i_t^Y$ is very bad and yields zero reward.

We consider **full cooperation**. No ~~incentives~~ or ~~zero sum~~. Just maximize total reward.

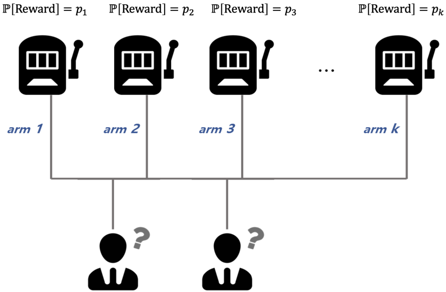
Multi-player (Cooperative) Bandits

Consider $m > 1$ players $X, Y, \ldots$. We assume colliding on the same action $i_t^X = i_t^Y$ is very bad and yields zero reward.

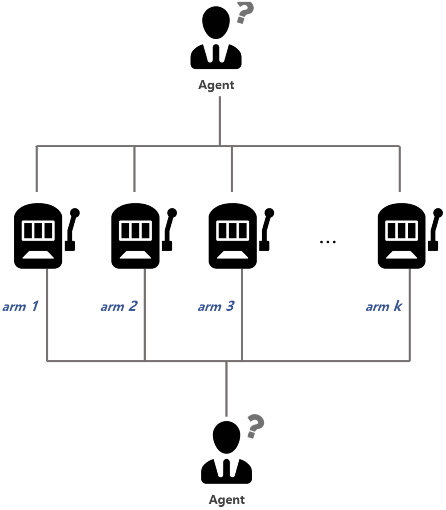
We consider **full cooperation**. No ~~incentives~~ or ~~zero sum~~. Just maximize total reward.

Proposed for wireless radio – learn good signal frequencies without interference. [Lai-Jiang-Poor 08, Liu-Zhao 10, Anandkumar-Michael-Tang-Swami 11].

With communication between players this is **semibandit**. E.g. online shortest path.



Catch: the players **cannot** communicate. We want a **distributed** algorithm.



Many, many possibilities for algorithms. A central difficulty is to smoothly break ties between top arms.

The Power of Communication

Many, many possibilities for algorithms. A central difficulty is to smoothly break ties between top arms.

If players can detect collisions, this helps a lot!

The Power of Communication

Many, many possibilities for algorithms. A central difficulty is to smoothly break ties between top arms.

If players can detect collisions, this helps a lot!

[Alatur-Levy-Krause 20]: $T^{2/3}$ regret. Communication via (detectable) collisions.

- On each block of length $T^{1/3}$, every player stays on a fixed action.
- Every $T^{1/3}$ time-steps, players synchronize information using $O(\log T)$ collisions.
- Reduces to semibandit game with $T^{2/3}$ "rounds" of true length $T^{1/3}$.

The Power of Communication

Many, many possibilities for algorithms. A central difficulty is to smoothly break ties between top arms.

If players can detect collisions, this helps a lot!

[Alatur-Levy-Krause 20]: $T^{2/3}$ regret. Communication via (detectable) collisions.

- On each block of length $T^{1/3}$, every player stays on a fixed action.
- Every $T^{1/3}$ time-steps, players synchronize information using $O(\log T)$ collisions.
- Reduces to semibandit game with $T^{2/3}$ "rounds" of true length $T^{1/3}$.

This is not **really** distributed... Communicating via collisions is super powerful.

In fact $\tilde{O}(T^{1/2})$ regret is possible even with adversarial rewards [Bubeck-Li-Peres-S. 20].

The Power of Communication

Many, many possibilities for algorithms. A central difficulty is to smoothly break ties between top arms.

If players can detect collisions, this helps a lot!

[Alatur-Levy-Krause 20]: $T^{2/3}$ regret. Communication via (detectable) collisions.

- On each block of length $T^{1/3}$, every player stays on a fixed action.
- Every $T^{1/3}$ time-steps, players synchronize information using $O(\log T)$ collisions.
- Reduces to semibandit game with $T^{2/3}$ "rounds" of true length $T^{1/3}$.

This is not **really** distributed... Communicating via collisions is super powerful.

In fact $\tilde{O}(T^{1/2})$ regret is possible even with adversarial rewards [Bubeck-Li-Peres-S. 20].

Takeaway: to get distributed algorithms, need to set the problem up carefully.

Precise Setup

Fix $\mathbf{p} = (p_1, p_2, \ldots, p_K) \in [0,1]^K$. Generate KmT independent Bernoulli reward variables $\mathrm{rew}_t^X(i)$ for $(t, i, X) \in [T] \times [K] \times [M]$:

$$\mathbb{P}\left[\mathrm{rew}_t^X(i) = 1\right] = p_i \quad \text{and} \quad \mathbb{P}\left[\mathrm{rew}_t^X(i) = 0\right] = 1 - p_i.$$

Precise Setup

Fix $\mathbf{p} = (p_1, p_2, \dots, p_K) \in [0,1]^K$. Generate KmT independent Bernoulli reward variables $\mathrm{rew}_t^X(i)$ for $(t, i, X) \in [T] \times [K] \times [M]$:

$$\mathbb{P}\left[\mathrm{rew}_t^X(i) = 1\right] = p_i \quad \text{and} \quad \mathbb{P}\left[\mathrm{rew}_t^X(i) = 0\right] = 1 - p_i.$$

At time t, each player $(P_X)_{X \in [m]}$ picks arm i_t^X, and receives the reward:

$$\mathrm{rew}_t(X) = \mathrm{rew}_t^X(i_t^X) \cdot \mathbb{1}_{i_t^X \neq i_t^Y \ \forall Y \neq X}.$$

Players **do not observe** whether collisions occur.

Fix $\mathbf{p} = (p_1, p_2, \ldots, p_K) \in [0,1]^K$. Generate KmT independent Bernoulli reward variables $\mathrm{rew}_t^X(i)$ for $(t, i, X) \in [T] \times [K] \times [M]$:

$$\mathbb{P}\left[\mathrm{rew}_t^X(i) = 1\right] = p_i \quad \text{and} \quad \mathbb{P}\left[\mathrm{rew}_t^X(i) = 0\right] = 1 - p_i.$$

At time t, each player $(P_X)_{X\in[m]}$ picks arm i_t^X, and receives the reward:

$$\mathrm{rew}_t(X) = \mathrm{rew}_t^X(i_t^X) \cdot \mathbb{1}_{i_t^X \neq i_t^Y \ \forall Y \neq X}.$$

Players **do not observe** whether collisions occur.
$\mathbf{p}^* = \sum_{j=1}^m p_j^*$, the sum of the best m arms, is the regret benchmark:

$$R_T = \mathbb{E}\left[T\mathbf{p}^* - \left(\sum_{t=1}^T \sum_{X=1}^m \mathrm{rew}_t(X)\right)\right].$$

Fix $\mathbf{p} = (p_1, p_2, \ldots, p_K) \in [0,1]^K$. Generate KmT independent Bernoulli reward variables $\mathrm{rew}_t^X(i)$ for $(t, i, X) \in [T] \times [K] \times [M]$:

$$\mathbb{P}\left[\mathrm{rew}_t^X(i) = 1\right] = p_i \quad \text{and} \quad \mathbb{P}\left[\mathrm{rew}_t^X(i) = 0\right] = 1 - p_i.$$

At time t, each player $(P_X)_{X\in[m]}$ picks arm i_t^X, and receives the reward:

$$\mathrm{rew}_t(X) = \mathrm{rew}_t^X(i_t^X) \cdot \mathbb{1}_{i_t^X \neq i_t^Y \ \forall Y \neq X}.$$

Players **do not observe** whether collisions occur.
$\mathbf{p}^* = \sum_{j=1}^m p_j^*$, the sum of the best m arms, is the regret benchmark:

$$R_T = \mathbb{E}\left[T\mathbf{p}^* - \left(\sum_{t=1}^{T}\sum_{X=1}^{m} \mathrm{rew}_t(X)\right)\right].$$

More specification needed!? Collisions may be **weakly detectable** or **undetectable**.

Precise Setup

Fix $\mathbf{p} = (p_1, p_2, \ldots, p_K) \in [0,1]^K$. Generate KmT independent Bernoulli reward variables $\mathrm{rew}_t^X(i)$ for $(t, i, X) \in [T] \times [K] \times [M]$:

$$\mathbb{P}\left[\mathrm{rew}_t^X(i) = 1\right] = p_i \quad \text{and} \quad \mathbb{P}\left[\mathrm{rew}_t^X(i) = 0\right] = 1 - p_i.$$

At time t, each player $(P_X)_{X \in [m]}$ picks arm i_t^X, and receives the reward:

$$\mathrm{rew}_t(X) = \mathrm{rew}_t^X(i_t^X) \cdot \mathbb{1}_{i_t^X \neq i_t^Y \ \forall Y \neq X}.$$

Players **do not observe** whether collisions occur.
$\mathbf{p}^* = \sum_{j=1}^m p_j^*$, the sum of the best m arms, is the regret benchmark:

$$R_T = \mathbb{E}\left[T\mathbf{p}^* - \left(\sum_{t=1}^{T}\sum_{X=1}^{m} \mathrm{rew}_t(X)\right)\right].$$

More specification needed!? Collisions may be **weakly detectable** or **undetectable**.

- A Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)

Precise Setup

Fix $\mathbf{p} = (p_1, p_2, \ldots, p_K) \in [0,1]^K$. Generate KmT independent Bernoulli reward variables $\mathrm{rew}_t^X(i)$ for $(t, i, X) \in [T] \times [K] \times [M]$:

$$\mathbb{P}\left[\mathrm{rew}_t^X(i) = 1\right] = p_i \quad \text{and} \quad \mathbb{P}\left[\mathrm{rew}_t^X(i) = 0\right] = 1 - p_i.$$

At time t, each player $(P_X)_{X \in [m]}$ picks arm i_t^X, and receives the reward:

$$\mathrm{rew}_t(X) = \mathrm{rew}_t^X(i_t^X) \cdot \mathbb{1}_{i_t^X \neq i_t^Y \ \forall Y \neq X}.$$

Players **do not observe** whether collisions occur.
$\mathbf{p}^* = \sum_{j=1}^m p_j^*$, the sum of the best m arms, is the regret benchmark:

$$R_T = \mathbb{E}\left[T\mathbf{p}^* - \left(\sum_{t=1}^T \sum_{X=1}^m \mathrm{rew}_t(X)\right)\right].$$

More specification needed!? Collisions may be **weakly detectable** or **undetectable**.

- A Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)
- B Observe reward $\mathrm{rew}_t^X(i_t^X)$. (Collisions affect rewards but **NOT** feedback.)

Weakly Detectable vs Undetectable Collisions

Collisions may be **weakly detectable** or **undetectable**.

- **A** Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)
- **B** Observe reward $\mathrm{rew}_t^X(i_t^X)$. (Collisions affect rewards but **NOT** feedback.)

Weakly Detectable vs Undetectable Collisions

Collisions may be **weakly detectable** or **undetectable**.

- A. Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)
- B. Observe reward $\mathrm{rew}_t^X(i_t^X)$. (Collisions affect rewards but **NOT** feedback.)

Questions:

- Are these models equivalent?
- How do they compare to settings with "arbitrary" collision behavior?

Weakly Detectable vs Undetectable Collisions

Collisions may be **weakly detectable** or **undetectable**.

- A. Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)
- B. Observe reward $\mathrm{rew}_t^X(i_t^X)$. (Collisions affect rewards but **NOT** feedback.)

Questions:

- Are these models equivalent?
- How do they compare to settings with "arbitrary" collision behavior?

Answers:

- $\tilde{O}(T^{1/2})$ regret is possible in both. But $\tilde{O}(1/\Delta)$ requires weak detectability.

Weakly Detectable vs Undetectable Collisions

Collisions may be **weakly detectable** or **undetectable**.

- A. Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)
- B. Observe reward $\mathrm{rew}_t^X(i_t^X)$. (Collisions affect rewards but **NOT** feedback.)

Questions:

- Are these models equivalent?
- How do they compare to settings with "arbitrary" collision behavior?

Answers:

- $\tilde{O}(T^{1/2})$ regret is possible in both. But $\tilde{O}(1/\Delta)$ requires weak detectability.
- With undetectable collisions, **Pareto optimal** gap dependence ranges from $\tilde{O}(T^{1/2})$ to $\tilde{O}(\Delta^{-2})$. These are attained by **collision-free** algorithms.

Weakly Detectable vs Undetectable Collisions

Collisions may be **weakly detectable** or **undetectable**.

- A. Observe reward $\mathrm{rew}_t(X)$. (Collisions affect rewards **AND** feedback.)
- B. Observe reward $\mathrm{rew}_t^X(i_t^X)$. (Collisions affect rewards but **NOT** feedback.)

Questions:

- Are these models equivalent?
- How do they compare to settings with "arbitrary" collision behavior?

Answers:

- $\tilde{O}(T^{1/2})$ regret is possible in both. But $\tilde{O}(1/\Delta)$ requires weak detectability.
- With undetectable collisions, **Pareto optimal** gap dependence ranges from $\tilde{O}(T^{1/2})$ to $\tilde{O}(\Delta^{-2})$. These are attained by **collision-free** algorithms.
- Corollary: $\tilde{O}(\sqrt{T})$ is the minimax regret in **any feedback model**. For gap-dependence, <u>weakly detectable is easiest</u> and <u>undetectable is hardest</u>.

Optimal Algorithms with Weakly Detectable Collisions

Idea of [Huang-Combes-Trinh 21] (see also [Pacchiano-Bartlett-Jordan 21]).
First step: players agree on a "decent" arm with $p_j \geq \max_i p_i/2$.

Optimal Algorithms with Weakly Detectable Collisions

Idea of [Huang-Combes-Trinh 21] (see also [Pacchiano-Bartlett-Jordan 21]).
First step: players agree on a "decent" arm with $p_j \geq \max_i p_i/2$.

Now, a_j will be used for communication.

- If I want to communicate, I play a_j for $\tilde{O}(1/p_j)$ timesteps.
- If I observe no reward, I "found" someone else to communicate with.
- We now exchange bits via $\tilde{O}(1/p_j)$-length blocks of zero reward on a_j.

Optimal Algorithms with Weakly Detectable Collisions

Idea of [Huang-Combes-Trinh 21] (see also [Pacchiano-Bartlett-Jordan 21]).
First step: players agree on a "decent" arm with $p_j \geq \max_i p_i/2$.

Now, a_j will be used for communication.

- If I want to communicate, I play a_j for $\tilde{O}(1/p_j)$ timesteps.
- If I observe no reward, I "found" someone else to communicate with.
- We now exchange bits via $\tilde{O}(1/p_j)$-length blocks of zero reward on a_j.

Because $p_j \geq \max_i p_i/2$, the regret is only $O(p_j)$ per timestep. Hence communication cost is just $\tilde{O}(1)$ per bit, regardless of p_j.

Idea of [Huang-Combes-Trinh 21] (see also [Pacchiano-Bartlett-Jordan 21]).
First step: players agree on a "decent" arm with $p_j \geq \max_i p_i/2$.

Now, a_j will be used for communication.

- If I want to communicate, I play a_j for $\tilde{O}(1/p_j)$ timesteps.
- If I observe no reward, I "found" someone else to communicate with.
- We now exchange bits via $\tilde{O}(1/p_j)$-length blocks of zero reward on a_j.

Because $p_j \geq \max_i p_i/2$, the regret is only $O(p_j)$ per timestep. Hence communication cost is just $\tilde{O}(1)$ per bit, regardless of p_j.

This is surprising and cool! But it is **still not really decentralized**. Rather, it shows that weak detection still allows players to find each other and communicate.

Optimal Algorithms with Weakly Detectable Collisions

Idea of [Huang-Combes-Trinh 21] (see also [Pacchiano-Bartlett-Jordan 21]).
First step: players agree on a "decent" arm with $p_j \geq \max_i p_i/2$.

Now, a_j will be used for communication.

- If I want to communicate, I play a_j for $\tilde{O}(1/p_j)$ timesteps.
- If I observe no reward, I "found" someone else to communicate with.
- We now exchange bits via $\tilde{O}(1/p_j)$-length blocks of zero reward on a_j.

Because $p_j \geq \max_i p_i/2$, the regret is only $O(p_j)$ per timestep. Hence communication cost is just $\tilde{O}(1)$ per bit, regardless of p_j.

This is surprising and cool! But it is **still not really decentralized**. Rather, it shows that weak detection still allows players to find each other and communicate.

Explicit communication protocols are brittle. What if the effect of collisions varies unpredictably or is just extremely negative?

$T^{1/2}$ Minimax Regret with No Collisions

With undetectable collisions, communication is truly impossible.

$T^{1/2}$ Minimax Regret with No Collisions

With undetectable collisions, communication is truly impossible.

In fact is possible to have no collisions at all. Such algorithms **work under any collision behavior**.

$T^{1/2}$ Minimax Regret with No Collisions

With undetectable collisions, communication is truly impossible.

In fact is possible to have no collisions at all. Such algorithms **work under any collision behavior**.

Theorem (Bubeck-Budzinski 20 and Bubeck-Budzinski-S. 21)

There is a algorithm with no collisions and $\widetilde{O}(\sqrt{T})$ regret. More precisely,

$$\max_{\mathbf{p}} \mathbb{E}[R_T] = O\left(mK^{11/2}\sqrt{T\log T}\right),$$

$$\mathbb{P}\left(\textit{there is ever a collision}\right) = O(T^{-2}).$$

$T^{1/2}$ Minimax Regret with No Collisions

With undetectable collisions, communication is truly impossible.

In fact is possible to have no collisions at all. Such algorithms **work under any collision behavior**.

Theorem (Bubeck-Budzinski 20 and Bubeck-Budzinski-S. 21)

There is a algorithm with no collisions and $\widetilde{O}(\sqrt{T})$ regret. More precisely,

$$\max_{\mathbf{p}} \mathbb{E}[R_T] = O\left(mK^{11/2}\sqrt{T \log T}\right),$$

$$\mathbb{P}\left(\textit{there is ever a collision}\right) = O(T^{-2}).$$

The log is real: $\Theta(\sqrt{T \log T})$ is optimal even with full feedback [Bubeck-Budzinski 20].

Definition of full feedback: all $K \times m \times T$ rewards are independent. I.e. Player X and Y's observations of arm 1 are independent.

For illustration, work in the plane $P = \{p_1 + p_2 + p_3 = \text{constant}\}$ with full feedback.

Undetectability means Player Y's decisions do not influence Player X at all.

Hence the protocol consists of pre-specified functions

$$(f_1^X, f_1^Y, \ldots, f_T^X, f_T^Y) : P \to \{1, 2, 3\}.$$

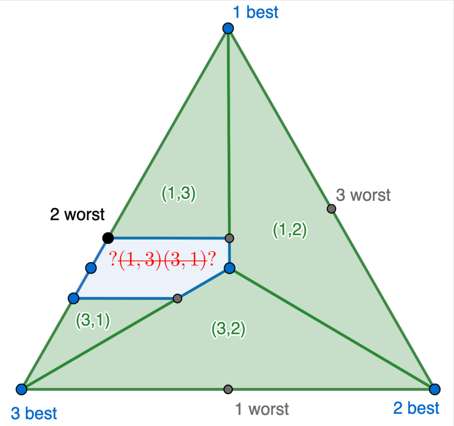
Topological Obstruction for 2 players and 3 arms

For illustration, work in the plane $P = \{p_1 + p_2 + p_3 = \text{constant}\}$ with full feedback.

Undetectability means Player Y's decisions do not influence Player X at all.

Hence the protocol consists of pre-specified functions

$$(f_1^X, f_1^Y, \ldots, f_T^X, f_T^Y) : P \to \{1, 2, 3\}.$$



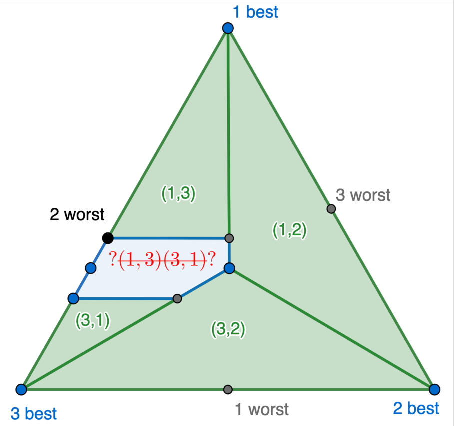
Topological Obstruction for 2 players and 3 arms

For illustration, work in the plane $P = \{p_1 + p_2 + p_3 = \text{constant}\}$ with full feedback.

Undetectability means Player Y's decisions do not influence Player X at all.

Hence the protocol consists of pre-specified functions

$$(f_1^X, f_1^Y, \ldots, f_T^X, f_T^Y) : P \to \{1, 2, 3\}.$$



Full feedback ensures the estimates $\hat{\mathbf{p}}_t^X, \hat{\mathbf{p}}_t^Y$ are within $\tilde{O}(t^{-1/2})$ of each other.

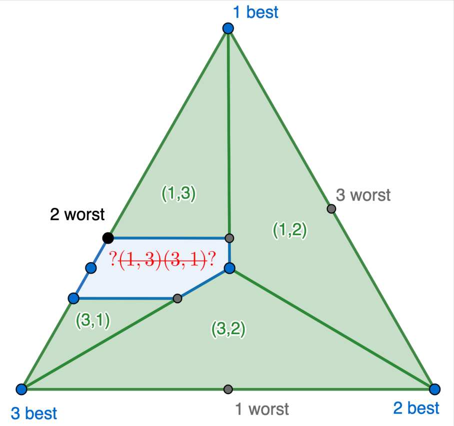
Topological Obstruction for 2 players and 3 arms

For illustration, work in the plane $P = \{p_1 + p_2 + p_3 = \text{constant}\}$ with full feedback.

Undetectability means Player Y's decisions do not influence Player X at all.

Hence the protocol consists of pre-specified functions

$$(f_1^X, f_1^Y, \ldots, f_T^X, f_T^Y) : P \to \{1, 2, 3\}.$$

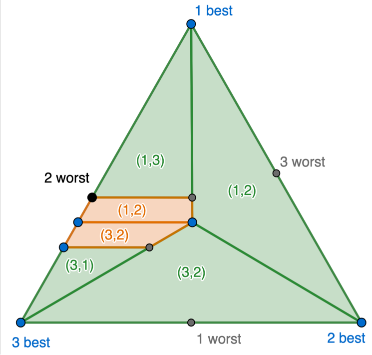


Full feedback ensures the estimates $\hat{\mathbf{p}}_t^X, \hat{\mathbf{p}}_t^Y$ are within $\tilde{O}(t^{-1/2})$ of each other.

Topological obstruction: cannot always play the top 2 arms without colliding for some $\mathbf{p}$.

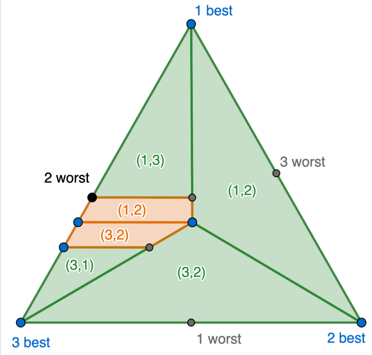
Collision-Free Solution for 2 players and 3 arms

Idea of [Bubeck-Budzinski 20]: separate $(1,3)$ and $(3,1)$ with a padding layer.



Collision-Free Solution for 2 players and 3 arms

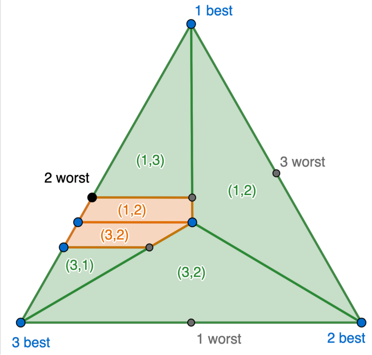
Idea of [Bubeck-Budzinski 20]: separate $(1,3)$ and $(3,1)$ with a padding layer.



Padding width $\sqrt{\frac{\log T}{t}}$.

Estimates $\mathbf{p}_t^X, \mathbf{p}_t^Y$ close $\implies$ land in adjacent regions.

Idea of [Bubeck-Budzinski 20]: separate $(1,3)$ and $(3,1)$ with a padding layer.



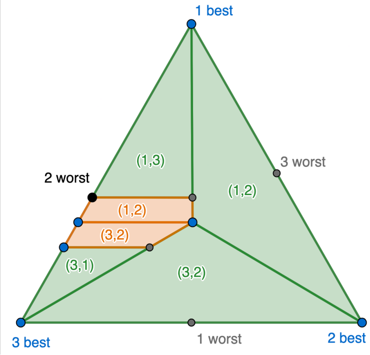
Padding width $\sqrt{\frac{\log T}{t}}$.

Estimates $\mathbf{p}_t^X, \mathbf{p}_t^Y$ close $\implies$ land in adjacent regions.

With suitable padding labels, this ensures no collisions.

Collision-Free Solution for 2 players and 3 arms

Idea of [Bubeck-Budzinski 20]: separate $(1,3)$ and $(3,1)$ with a padding layer.



Padding width $\sqrt{\frac{\log T}{t}}$.

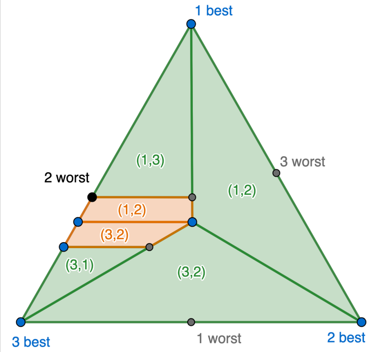
Estimates $\mathbf{p}_t^X, \mathbf{p}_t^Y$ close $\implies$ land in adjacent regions.

With suitable padding labels, this ensures no collisions.

Random angle $\implies$ $\widetilde{O}(\sqrt{T})$ extra regret for any $\mathbf{p}$.

Collision-Free Solution for 2 players and 3 arms

Idea of [Bubeck-Budzinski 20]: separate $(1,3)$ and $(3,1)$ with a padding layer.



Padding width $\sqrt{\frac{\log T}{t}}$.

Estimates $\mathbf{p}_t^X, \mathbf{p}_t^Y$ close $\implies$ land in adjacent regions.

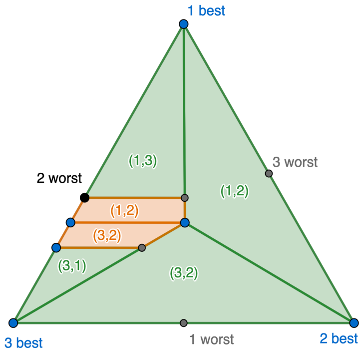
With suitable padding labels, this ensures no collisions.

Random angle $\implies$ $\widetilde{O}(\sqrt{T})$ extra regret for any $\mathbf{p}$.

Outside padding, just greedy. Hence $\widetilde{O}(\sqrt{T})$ regret. Bandit feedback is harder.

Collision-Free Solution for 2 players and 3 arms

Idea of [Bubeck-Budzinski 20]: separate $(1,3)$ and $(3,1)$ with a padding layer.



Padding width $\sqrt{\frac{\log T}{t}}$.

Estimates $\mathbf{p}_t^X, \mathbf{p}_t^Y$ close $\implies$ land in adjacent regions.

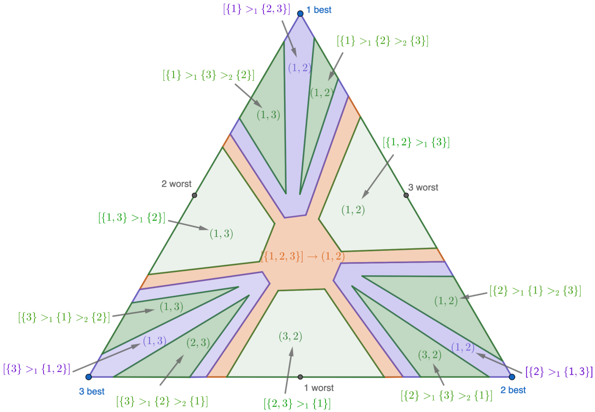
With suitable padding labels, this ensures no collisions.

Random angle $\implies$ $\widetilde{O}(\sqrt{T})$ extra regret for any $\mathbf{p}$.

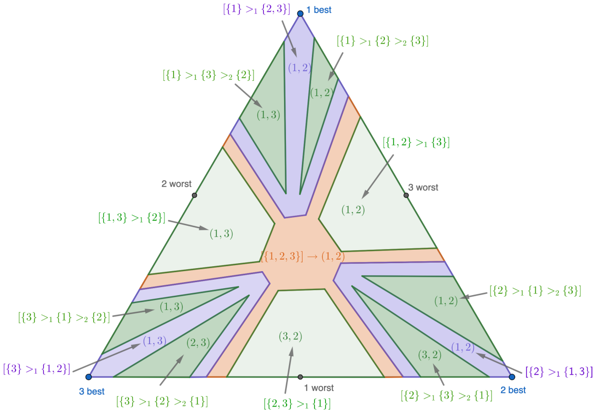
Outside padding, just greedy. Hence $\widetilde{O}(\sqrt{T})$ regret. Bandit feedback is harder.

Larger (K, m): need to generalize this picture.

General partition in the case $(K, m) = (3, 2)$:



General partition in the case $(K, m) = (3, 2)$:



Regions form a tree, defined by arm inequalities added **in order**.

Example region: $\{1, 3, 5\} >_2 \{4, 8\} >_3 \{2, 6\} >_1 \{7, 9, 10\}$.

Computing With the Partition

Never compute the full partition tree. (More than $K!$ regions...)

Luckily, computing the correct region for any estimate $\hat{\mathbf{p}}_t^X \in [0,1]^K$ is efficient.

Never compute the full partition tree. (More than $K!$ regions...)

Luckily, computing the correct region for any estimate $\hat{\mathbf{p}}_t^X \in [0,1]^K$ is efficient.

Repeatedly add new inequalities to separate arms that *might* be in top m.
Once top m and bottom $K-m$ are determined, stop. E.g. for $(K,m)=(10,5)$:

$$
\begin{aligned}
&\{1,2,3,4,5,6,7,8,9,10\} \\
\to\ &\{1,2,3,4,5,6,8\} >_1 \{7,9,10\} \\
\to\ &\{1,3,5\} >_2 \{2,4,6,8\} >_1 \{7,9,10\} \\
\to\ &\{1,3,5\} >_2 \{4,8\} >_3 \{2,6\} >_1 \{7,9,10\}.
\end{aligned}
$$

Computing With the Partition

Never compute the full partition tree. (More than $K!$ regions...)

Luckily, computing the correct region for any estimate $\hat{\mathbf{p}}_t^X \in [0,1]^K$ is efficient.

Repeatedly add new inequalities to separate arms that *might* be in top m.
Once top m and bottom $K-m$ are determined, stop. E.g. for $(K,m)=(10,5)$:

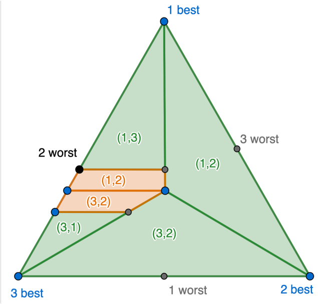
$$\begin{aligned} &\{1,2,3,4,5,6,7,8,9,10\} \\ \to\ &\{1,2,3,4,5,6,8\} >_1 \{7,9,10\} \\ \to\ &\{1,3,5\} >_2 \{2,4,6,8\} >_1 \{7,9,10\} \\ \to\ &\{1,3,5\} >_2 \{4,8\} >_3 \{2,6\} >_1 \{7,9,10\}. \end{aligned}$$

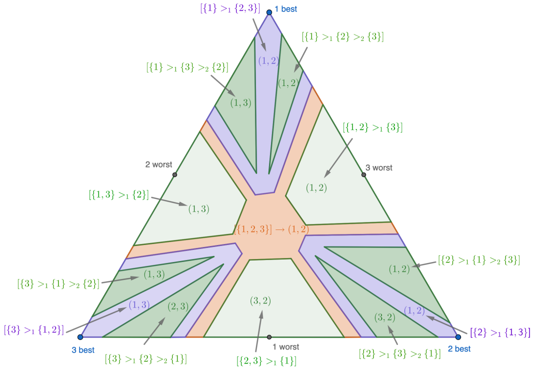
Generalization of padding layers using random threshold $\tau > 0$:

- If margin for new inequality is **above** τ, add it.
- If margin is **well below** τ, try next potential inequality.
- If margin is **barely below** τ, stop early (enter padding).

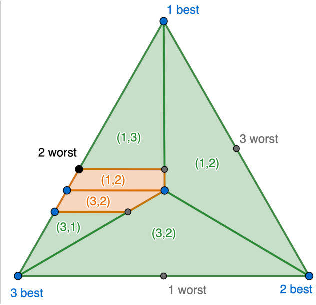
Large Gaps Still Incur Regret $T^{1/2}$

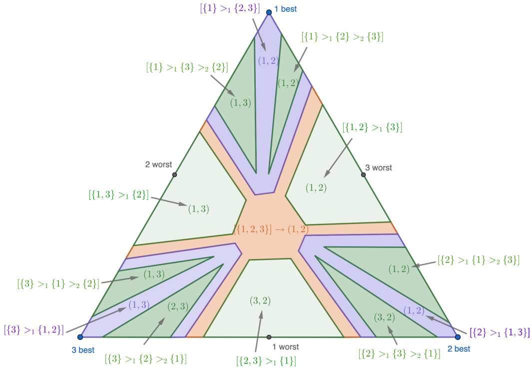
In both constructions, the $\tilde{O}(T^{1/2})$ extra regret from **padding** applies for all gaps Δ.





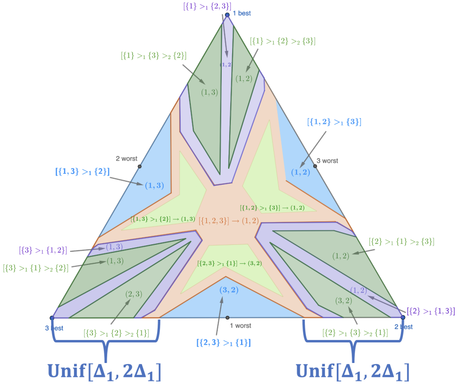
In both constructions, the $\tilde{O}(T^{1/2})$ extra regret from **padding** applies for all gaps Δ.





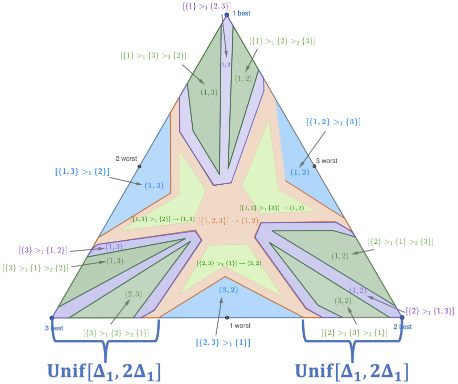
How to improve things for large gaps? Push the padding somewhere else!

Idea: designate those $\mathbf{p}$ with $\Delta(\mathbf{p}) \gtrsim \Delta_1$ as safe zones with no padding.



A Modified Gap-Dependent Algorithm

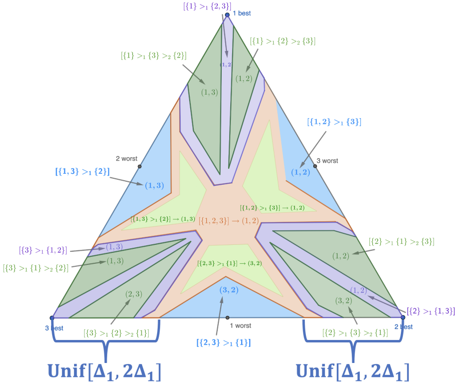
Idea: designate those $\mathbf{p}$ with $\Delta(\mathbf{p}) \gtrsim \Delta_1$ as **safe zones** with no padding.



$\Delta(\mathbf{p}) \gg \Delta_1$: zero regret once $t \gg \widetilde{O}\left(\frac{1}{\Delta_1^2}\right)$.

A Modified Gap-Dependent Algorithm

Idea: designate those $\mathbf{p}$ with $\Delta(\mathbf{p}) \gtrsim \Delta_1$ as **safe zones** with no padding.

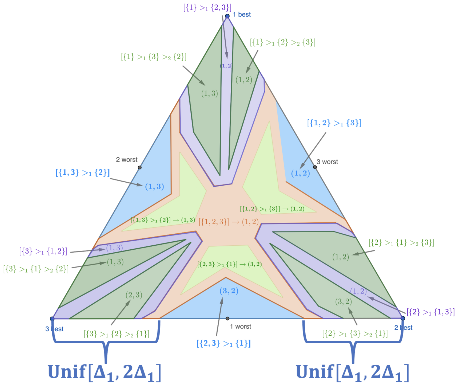


$\Delta(\mathbf{p}) \gg \Delta_1$: zero regret once $t \gg \widetilde{O}\left(\frac{1}{\Delta_1^2}\right)$.

$\Delta(\mathbf{p}) \ll \Delta_1$: padding cost increases.

$$\sqrt{T} \to \frac{\sqrt{T}}{\Delta_1}.$$

Idea: designate those $\mathbf{p}$ with $\Delta(\mathbf{p}) \gtrsim \Delta_1$ as **safe zones** with no padding.



$\Delta(\mathbf{p}) \gg \Delta_1$: zero regret once $t \gg \widetilde{O}\left(\frac{1}{\Delta_1^2}\right)$.

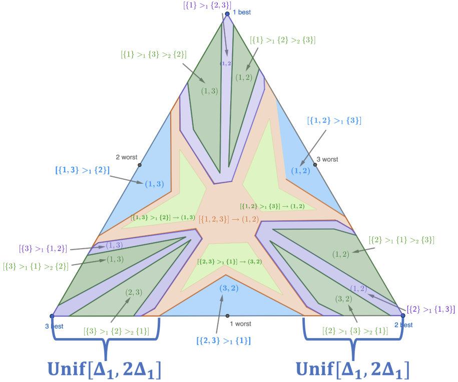
$\Delta(\mathbf{p}) \ll \Delta_1$: padding cost increases.

$$\sqrt{T} \to \frac{\sqrt{T}}{\Delta_1}.$$

Why? Padding around the safe zone is "less random".

A Modified Gap-Dependent Algorithm

Idea: designate those $\mathbf{p}$ with $\Delta(\mathbf{p}) \gtrsim \Delta_1$ as safe zones with no padding.



$\Delta(\mathbf{p}) \gg \Delta_1$: zero regret once $t \gg \widetilde{O}\left(\frac{1}{\Delta_1^2}\right)$.

$\Delta(\mathbf{p}) \ll \Delta_1$: padding cost increases.

$$\sqrt{T} \to \frac{\sqrt{T}}{\Delta_1}.$$

Why? Padding around the safe zone is "less random".

Better performance for large gaps. Worse for small gaps.

$$R_{T,\Delta} \leq \begin{cases} \widetilde{O}\left(1/\Delta_1\right), & \Delta \geq \Delta_1 \\ \widetilde{O}\left(\sqrt{T}/\Delta_1\right), & \Delta \leq \Delta_1. \end{cases}$$

Pareto Optimal Gap Dependence

More generally, use a sequence $1 \geq \Delta_1 \geq \cdots \geq \Delta_J \geq T^{-1/2}$. Use Δ_j once $t \gg \Delta_j^{-2}$.

Theorem (Liu-S. 22)

The Pareto-optimal regret guarantees with undetectable collisions are:

$$R_{T,\Delta} \leq \widetilde{O}\left(\frac{1}{\Delta_i \cdot \Delta_{i+1}}\right), \quad \Delta \in [\Delta_i, \Delta_{i+1}].$$

(Up to $\mathrm{poly}(K, \log(T))$ *factors.)*

Pareto Optimal Gap Dependence

More generally, use a sequence $1 \geq \Delta_1 \geq \cdots \geq \Delta_J \geq T^{-1/2}$. Use Δ_j once $t \gg \Delta_j^{-2}$.

Theorem (Liu-S. 22)

The Pareto-optimal regret guarantees with undetectable collisions are:

$$R_{T,\Delta} \leq \widetilde{O}\left(\frac{1}{\Delta_i \cdot \Delta_{i+1}}\right), \quad \Delta \in [\Delta_i, \Delta_{i+1}].$$

(Up to $\mathrm{poly}(K, \log(T))$ *factors.)*

Example: with bounded ratios $\frac{\Delta_i}{\Delta_{i+1}} = O(1)$, regret is $R_{T,\Delta} = \tilde{O}(\Delta^{-2})$.

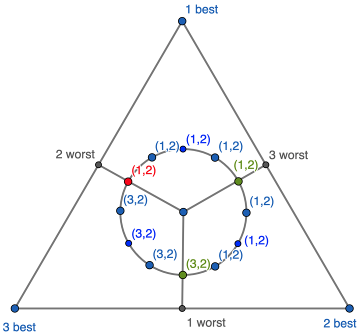
Several consequences of Pareto optimality. For example:

Corollary (Liu-S. 22)

Suppose $R_T \leq T^{0.51}$*. Then* $R_{T,\Delta} \gtrsim T^{1/2}$ *for all* $\Delta \lesssim T^{-0.01}$*.*

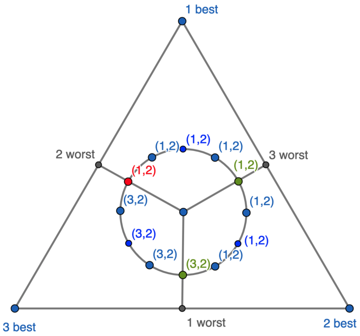
A One-Step Lower Bound

Assume $(K, m) = (3, 2)$. Consider $\sqrt{T}$ points equally spaced on a constant-size circle.



A One-Step Lower Bound

Assume $(K, m) = (3, 2)$. Consider $\sqrt{T}$ points equally spaced on a constant-size circle.

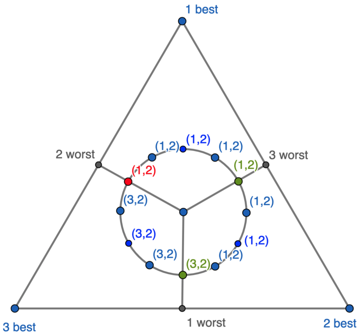


Topological obstruction: for any labelling, there is a **FAIL** with constant regret. Meaning either:

1. There is a collision, OR
2. The worst two actions are played.

A One-Step Lower Bound

Assume $(K, m) = (3, 2)$. Consider $\sqrt{T}$ points equally spaced on a constant-size circle.



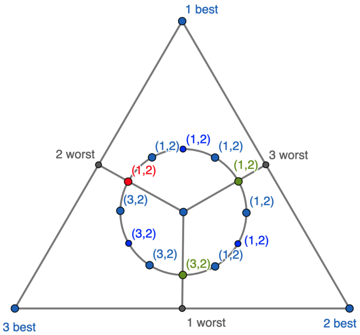
Topological obstruction: for any labelling, there is a **FAIL** with constant regret. Meaning either:

1. There is a collision, OR
2. The worst two actions are played.

The estimates $\hat{\mathbf{p}}_t^X, \hat{\mathbf{p}}_t^Y$ are basically adjacent points...

A One-Step Lower Bound

Assume $(K, m) = (3, 2)$. Consider $\sqrt{T}$ points equally spaced on a constant-size circle.



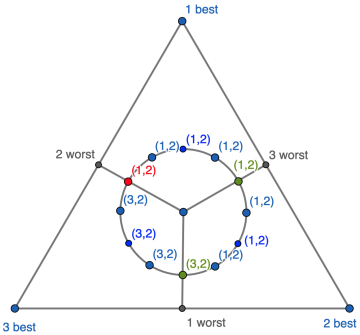
Topological obstruction: for any labelling, there is a **FAIL** with constant regret. Meaning either:

1. There is a collision, OR
2. The worst two actions are played.

The estimates $\hat{\mathbf{p}}_t^X, \hat{\mathbf{p}}_t^Y$ are basically adjacent points...

Each **FAIL** point has some gap Δ.

Assume $(K, m) = (3, 2)$. Consider $\sqrt{T}$ points equally spaced on a constant-size circle.



Topological obstruction: for any labelling, there is a **FAIL** with constant regret. Meaning either:

1. There is a collision, OR
2. The worst two actions are played.

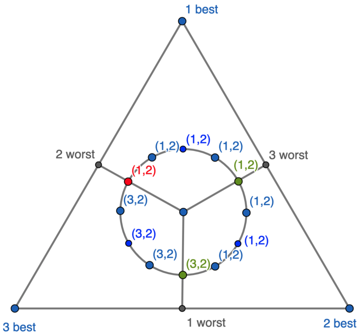
The estimates $\hat{\mathbf{p}}_t^X, \hat{\mathbf{p}}_t^Y$ are basically adjacent points...

Each **FAIL** point has some gap Δ.

By dyadic pigeonhole, there **exists** a gap Δ_J with $\widetilde{\Omega}(T)$ **FAILs** across $1 \leq t \leq T$.

A One-Step Lower Bound

Assume $(K, m) = (3, 2)$. Consider $\sqrt{T}$ points equally spaced on a constant-size circle.



Topological obstruction: for any labelling, there is a **FAIL** with constant regret. Meaning either:

1. There is a collision, OR
2. The worst two actions are played.

The estimates $\hat{\mathbf{p}}_t^X, \hat{\mathbf{p}}_t^Y$ are basically adjacent points...

Each **FAIL** point has some gap Δ.

By dyadic pigeonhole, there **exists** a gap Δ_J with $\widetilde{\Omega}(T)$ **FAILs** across $1 \leq t \leq T$.

There are $\approx \Delta_J\sqrt{T}$ points on the circle with gap $\approx \Delta_J$ to absorb the **FAILs**. Hence

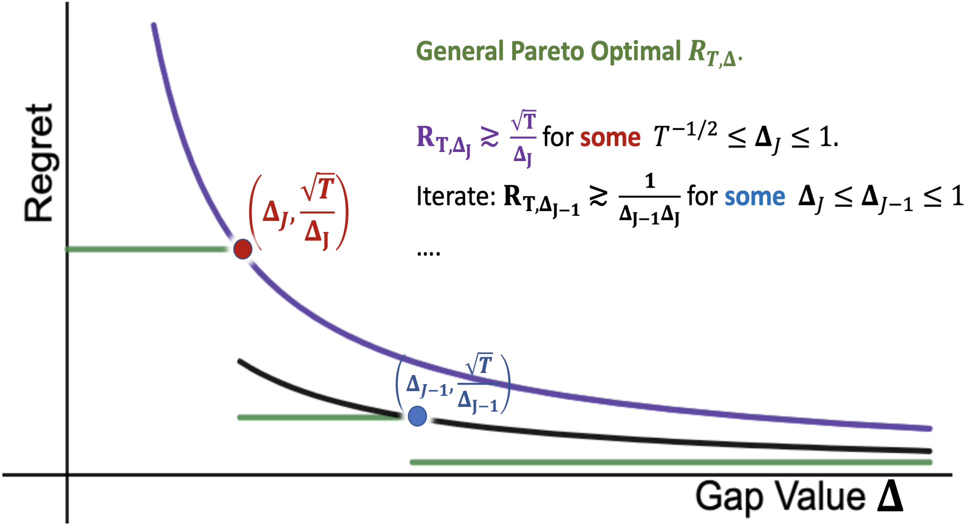
$$R_{T,\Delta_J} \gtrsim \frac{T}{\Delta_J\sqrt{T}} = \frac{\sqrt{T}}{\Delta_J}.$$

General Pareto Optimal $R_{T,\Delta}$.

$\mathbf{R_{T,\Delta_J}} \gtrsim \frac{\sqrt{T}}{\Delta_J}$ for **some** $T^{-1/2} \le \mathbf{\Delta}_J \le 1.$

Iterate: $\mathbf{R_{T,\Delta_{J-1}}} \gtrsim \frac{1}{\Delta_{J-1}\Delta_J}$ for **some** $\mathbf{\Delta}_J \le \mathbf{\Delta}_{J-1} \le 1.$

....



Summary

For multiplayer bandit, collision details matter!

Summary

For multiplayer bandit, collision details matter!

- **Weak detectability**: implicit communication allows "best case" regret

$$\mathbb{E}[R_T] \lesssim \min(\sqrt{T}, \log(T)/\Delta).$$

Summary

For multiplayer bandit, collision details matter!

- **Weak detectability**: implicit communication allows "best case" regret
$$\mathbb{E}[R_T] \lesssim \min(\sqrt{T}, \log(T)/\Delta).$$
- **Undetectable** collisions yields maximal regret. Pareto optima include $\sqrt{T}$ and Δ^{-2}. Achieved with **no collisions at all**.

Summary

For multiplayer bandit, collision details matter!

- **Weak detectability**: implicit communication allows "best case" regret
$$\mathbb{E}[R_T] \lesssim \min(\sqrt{T}, \log(T)/\Delta).$$
- **Undetectable** collisions yields maximal regret. Pareto optima include $\sqrt{T}$ and Δ^{-2}. Achieved with **no collisions at all**.

Many directions aren't well understood.

Summary

For multiplayer bandit, collision details matter!

- **Weak detectability**: implicit communication allows "best case" regret

$$\mathbb{E}[R_T] \lesssim \min(\sqrt{T}, \log(T)/\Delta).$$

- **Undetectable** collisions yields maximal regret. Pareto optima include $\sqrt{T}$ and Δ^{-2}. Achieved with **no collisions at all**.

Many directions aren't well understood.

- Non-stochastic rewards? Network structure?

Summary

For multiplayer bandit, collision details matter!

- **Weak detectability**: implicit communication allows "best case" regret

$$\mathbb{E}[R_T] \lesssim \min(\sqrt{T}, \log(T)/\Delta).$$

- **Undetectable** collisions yields maximal regret. Pareto optima include $\sqrt{T}$ and Δ^{-2}. Achieved with **no collisions at all**.

Many directions aren't well understood.

- Non-stochastic rewards? Network structure?
- Implications for complex RL problems?

For multiplayer bandit, collision details matter!

- **Weak detectability**: implicit communication allows "best case" regret

$$\mathbb{E}[R_T] \lesssim \min(\sqrt{T}, \log(T)/\Delta).$$

- **Undetectable** collisions yields maximal regret. Pareto optima include $\sqrt{T}$ and Δ^{-2}. Achieved with **no collisions at all**.

Many directions aren't well understood.

- Non-stochastic rewards? Network structure?
- Implications for complex RL problems?