New Algorithms for Multiplayer Bandits

Abbas Mehrabian

McGill University
IVADO Fellow

23 September 2019

Co-authors: Etienne Boursier, Emilie Kaufmann, Gabor Lugosi, Vianney Perchet

The multi-armed bandit problem

The multi-armed bandit model

1. A multi-round single player game, a finite set of actions.
2. In each round the player chooses one of the actions and receives a (stochastic) reward.
3. The rewards of each action come from some unknown distribution.

The multi-armed bandit problem

The multi-armed bandit model

1. A multi-round single player game, a finite set of actions.
2. In each round the player chooses one of the actions and receives a (stochastic) reward.
3. The rewards of each action come from some unknown distribution.

Oracle's strategy. In all rounds, choose the action with the largest expected reward.

Regret of a learning algorithm: difference between algorithm's total reward and the oracle's total reward.

The multi-armed bandit problem

known results

T rounds, K arms, $\Delta =$ gap between best arm and second-best arm

Theorem (Lai and Robbins 1985, Auer, Cesa-Bianchi, Fischer 1998)

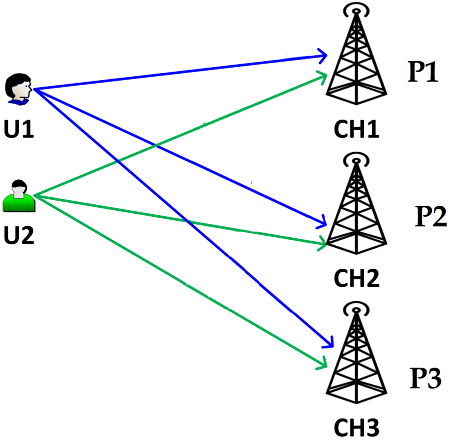
If each single reward $\in [0, 1]$, there is an algorithm with regret $K \log T/\Delta$, and this is tight.

Per round suboptimality $\rightarrow \frac{\log T}{T} \times \frac{K}{\Delta}$

Upper confidence bound (UCB) algorithm.

Multiplayer multi-armed bandits

Opportunistic spectrum access in cognitive radios



Rules of the game

1. The players pull arms simultaneously. If more than one players pull some arm, they all get zero reward.
2. Two feedback models: visible collisions versus invisible collisions
3. Players cannot talk during the game, and do not see each other's actions.
4. Rewards $\in [0, 1]$.
5. Time horizon, number of players/arms are known.
6. Number of arms $\geq$ number of players

Rules of the game

1. The players pull arms simultaneously. If more than one players pull some arm, they all get zero reward.
2. Two feedback models: visible collisions versus invisible collisions
3. Players cannot talk during the game, and do not see each other's actions.
4. Rewards $\in [0, 1]$.
5. Time horizon, number of players/arms are known.
6. Number of arms $\geq$ number of players

$$\begin{aligned}\text{Regret} = &\text{ Expected total system reward obtainable by oracle} \\ &- \text{Expected total system reward obtained by algorithm}\end{aligned}$$

I: Invisible Collisions

Multiplayer multi-armed bandits

Our algorithm for invisible collisions

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Theorem (Lugosi, M 2018)

In the harder setup that players do not observe collisions, there exists a polynomial-time algorithm with regret $\lesssim (KM/\Delta^2)\log T$.

Multiplayer multi-armed bandits

Our algorithm for invisible collisions

M players, K arms, Δ = gap between arm M and $M+1$

Theorem (Lugosi, M 2018)

In the harder setup that players do not observe collisions, there exists a polynomial-time algorithm with regret $\lesssim (KM/\Delta^2)\log T$.

Two main phases:

1. Determine the M best arms.
2. Occupy one of these arms.

Phase 2: occupy one of the best M arms

Musical chairs subroutine

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Musical chairs (MC) subroutine [Rosenski, Shamir, Szlak'16]

1. Pull one of the M best arms randomly.
2. If positive reward received, pull the same arm in subsequent rounds.
3. Otherwise, go to 1.

Phase 2: occupy one of the best M arms

Musical chairs subroutine

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Musical chairs (MC) subroutine [Rosenski, Shamir, Szlak'16]

1. Pull one of the M best arms randomly.
2. If positive reward received, pull the same arm in subsequent rounds.
3. Otherwise, go to 1.

Lemma. Number of rounds to stabilize $\leq 4M \log(M/\delta)/\Delta$ with probability $1-\delta$.

Phase 2: occupy one of the best M arms

Musical chairs subroutine

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Musical chairs (MC) subroutine [Rosenski, Shamir, Szlak'16]

1. Pull one of the M best arms randomly.
2. If positive reward received, pull the same arm in subsequent rounds.
3. Otherwise, go to 1.

Lemma. Number of rounds to stabilize $\leq 4M \log(M/\delta)/\Delta$ with probability $1-\delta$.

Phase 1: find the best M arms

The single-player case

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Hoeffding's inequality. If $X_1, \dots, X_n \sim X \in [0,1]$, then

$$\mathbf{Pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X > t\right|\right] < 2\exp(-2nt^2).$$

Corollary. Arm i has been pulled n times. Can build a confidence interval of width $\sqrt{\log(1/\delta)/n}$.

Phase 1: find the best M arms

The single-player case

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Hoeffding's inequality. If $X_1, \ldots, X_n \sim X \in [0,1]$, then

$$\mathbf{Pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X > t\right|\right] < 2\exp(-2nt^2).$$

Corollary. Arm i has been pulled n times. Can build a confidence interval of width $\sqrt{\log(1/\delta)/n}$.

Algorithm. Pull arms in a round-robin manner, until M of the confidence intervals lie strictly above the other intervals. Number of rounds until this happens $\lesssim K\log(1/\delta)/\Delta^2$.

Phase 1: find the best M arms

The multiplayer case

First problem: can't do round-robin.

Phase 1: find the best M arms

The multiplayer case

First problem: can't do round-robin.
Solution: do random exploration

Phase 1: find the best M arms

The multiplayer case

First problem: can't do round-robin. Do random exploration.

Second problem: can't get unbiased estimator for means, because of collisions.

Phase 1: find the best M arms

The multiplayer case

First problem: can't do round-robin. Do random exploration.

Second problem: can't get unbiased estimator for means, because of collisions.

expected reward from arm i = mean of arm $i \times (1 - 1/K)^{M-1}$, so

$\frac{\text{average reward from arm } i}{(1-1/K)^{M-1}}$ is unbiased estimator for mean of arm i

Phase 1: find the best M arms

The multiplayer case

First problem: can't do round-robin.

Second problem: can't get unbiased estimator for means, because of collisions. Divide by $(1 - 1/K)^{M-1}$.

Third problem: if some arm switches to Phase 2 earlier, the no-collision probability is wrong!

Phase 1: find the best M arms

The multiplayer case

First problem: can't do round-robin.

Second problem: can't get unbiased estimator for means, because of collisions. Divide by $(1-1/K)^{M-1}$.

Third problem: if some arm switches to Phase 2 earlier, the no-collision probability is wrong!

$\tau :=$ time a player discovers the M best arms. Then, $\tau \in [K\log(1/\delta)/\Delta^2, 25K\log(1/\delta)/\Delta^2]$.

Multiplayer multi-armed bandits

Our algorithm for invisible collisions

M players, K arms, $\Delta =$ gap between arm M and $M+1$

The Algorithm

1. Pull arms randomly and keep confidence intervals, until the gap is discovered at time τ.
2. Pull arms randomly for 24τ rounds.
3. Run musical chairs.

Analysis. $\delta = 1/MT$

Rounds to stabilize $\lesssim K\log(1/\delta)/\Delta^2 + M\log(M/\delta)/\Delta$

Regret $\lesssim MK\log(MT)/\Delta^2 + M^2\log(M^2T)/\Delta + 1$

Invisible collisions: known results

M players, K arms, Δ = gap between arm M and $M+1$,
μ = known lower bound for all means

Instance-dependent upper bounds for regret

1. $(KM/\Delta^2)\log T$ [Lugosi, M'18]
2. $(KM/\Delta + K^2 M/\mu)\log T$ [Boursier, Perchet'18]

Best known lower bound: $(K/\Delta)\log T$
[Anantharam, Varaiya, Walrand'87]

Invisible collisions: known results

M players, K arms, Δ = gap between arm M and $M+1$,
μ = known lower bound for all means

Instance-dependent upper bounds for regret

1. $(KM/\Delta^2)\log T$ [Lugosi, M'18]
2. $(KM/\Delta + K^2M/\mu)\log T$ [Boursier, Perchet'18]

Best known lower bound: $(K/\Delta)\log T$
[Anantharam, Varaiya, Walrand'87]

General upper bounds for regret

3. $K^2M\log^2(T)/\mu + KM\sqrt{T\log T}$ [Lugosi, M'18]
4. $K^2M\log T/\mu + K\sqrt{MT\log T}$ [Boursier, Perchet'18]

Best known lower bound (for $M=1$): $\sqrt{KT}$
[Auer, Cesa-Bianchi, Freund, Schapire'95]

II: Visible Collisions

Visible collisions: known results

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Instance-dependent upper bounds for regret

1. $\zeta(M, K, \Delta) \log T$ [Liu and Zhao'10]
2. $(KM/\Delta^2) \log T$ [Rosenski, Shamir, Szlak'16]
3. $(KM/\Delta) \log T$ [Lugosi, M'18]

Best known lower bound: $(K/\Delta) \log T$

[Anantharam, Varaiya, Walrand'87]

Visible collisions: known results

M players, K arms, $\Delta =$ gap between arm M and $M + 1$

Instance-dependent upper bounds for regret

1. $\zeta(M, K, \Delta) \log T$ [Liu and Zhao'10]
2. $(KM/\Delta^2) \log T$ [Rosenski, Shamir, Szlak'16]
3. $(KM/\Delta) \log T$ [Lugosi, M'18]

Best known lower bound: $(K/\Delta) \log T$
[Anantharam, Varaiya, Walrand'87]

General upper bounds for regret

4. $KM\sqrt{T \log T}$ [Lugosi, M'18]

Best known lower bound (for $M = 1$): $\sqrt{KT}$
[Auer, Cesa-Bianchi, Freund, Schapire'95]

Our algorithm for visible collisions

The single-player case

Epoch-based arm-elimination algorithm

1. All arms are alive initially
2. In epoch i:
   2.1 pull each alive arm 2^i times.
   2.2 update confidence intervals.
   2.3 if interval of some arm lies below another active arm, kill it.

Our algorithm for visible collisions

The single-player case

Epoch-based arm-elimination algorithm

1. All arms are alive initially
2. In epoch i:
   2.1 pull each alive arm 2^i times.
   2.2 update confidence intervals.
   2.3 if interval of some arm lies below another active arm, kill it.

Analysis. An arm with gap Δ will be pulled $\lesssim 4\log(T)/\Delta^2$ times, hence its contribution to regret $\lesssim \min\{4\log(T)/\Delta, \Delta T\} \leq 2\sqrt{T\log T}$,

Regret $\lesssim 2K\sqrt{T\log T}$

Our algorithm for visible collisions

The single-player case

Epoch-based arm-elimination algorithm

1. All arms are alive initially
2. In epoch i:
   2.1 pull each alive arm 2^i times.
   2.2 update confidence intervals.
   2.3 if interval of some arm lies below another active arm, kill it.

Difficulties for multiplayer case:

1. not enough to kill bad arms; must also pull the discovered good arms
2. coordinate the explorations

Our algorithm for visible collisions

The multiplayer case

Epoch-based arm-elimination algorithm

Each arm is either golden, silver, or dead.

1. All arms are silver initially
2. In epoch i:
   - 2.1 pull each silver arm 2^i times.
     (distribute silver arms between players via MC).
   - 2.2 update confidence intervals.
   - 2.3 mark arms as golden or dead as necessary.
   - 2.4 try to occupy new golden arms (using MC).

Our algorithm for visible collisions

The multiplayer case

Epoch-based arm-elimination algorithm

Each arm is either golden, silver, or dead.

1. All arms are silver initially
2. In epoch i:
   2.1 pull each silver arm 2^i times.
   (distribute silver arms between players via MC).
   2.2 update confidence intervals.
   2.3 mark arms as golden or dead as necessary.
   2.4 try to occupy new golden arms (using MC).

Regret $\lesssim M \min\{K \log(T)/\Delta, K\sqrt{T \log T}\}$

Visible collisions: known results

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Instance-dependent upper bounds for regret

1. $(KM/\Delta^2)\log T$ [Rosenski, Shamir, Szlak'16]
2. $(KM/\Delta)\log T$ [Lugosi, M'18]

Best known lower bound: $(K/\Delta)\log T$
[Anantharam, Varaiya, Walrand'87]

General upper bounds for regret

4. $KM\sqrt{T\log T}$ [Lugosi, M'18]

Best known lower bound (for $M=1$): $\sqrt{KT}$
[Auer, Cesa-Bianchi, Freund, Schapire'95]

Visible collisions: known results

M players, K arms, $\Delta =$ gap between arm M and $M+1$

Instance-dependent upper bounds for regret

1. $(KM/\Delta^2)\log T$ [Rosenski, Shamir, Szlak'16]
2. $(KM/\Delta)\log T$ [Lugosi, M'18]
3. $(KM + K/\Delta)\log T$ [Boursier, Perchet'18]

Best known lower bound: $(K/\Delta)\log T$
[Anantharam, Varaiya, Walrand'87]

General upper bounds for regret

4. $KM\sqrt{T\log T}$ [Lugosi, M'18]
5. $K\sqrt{T\log T}$ [Boursier, Perchet'18]

Best known lower bound (for $M=1$): $\sqrt{KT}$
[Auer, Cesa-Bianchi, Freund, Schapire'95]

Adversarial case

Known results

Upper bounds for the regret (visible collisions):

1. $K^2 T^{2/3}$ [Alatur, Levy, Krause'19]
2. $K^2 T^{1/2}$ for $M = 2$ [Bubeck, Li, Peres, Selke'19]

Upper bounds for the regret (invisible collisions):

1. $KT^{3/4}$ for $M = 2$ [Bubeck, Li, Peres, Selke'19]

Open questions

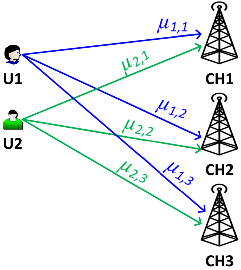
Simpler algorithms? Such as UCB, EXP3?

Better lower bounds?

III: Visible Collisions, Heterogeneous Setting

Multiplayer multi-armed bandits

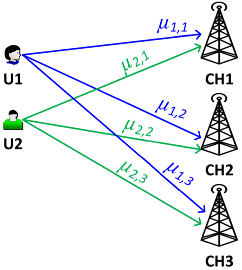
Heterogeneous setting



Distributed online stochastic maximum-weight matching

Multiplayer multi-armed bandits

Heterogeneous setting



Distributed online stochastic maximum-weight matching

Cooperative game-theoretic situation:

	channel 1	channel 2	channel 3
Player 1	1	0.9	0.2
Player 2	1	0.1	0.3

Heterogeneous setting

Known results

M players, K arms, $\Delta =$ gap between value of best matching and second best value, $\varepsilon > 0$ arbitrary

Instance-dependent upper bounds

1. $\zeta(M, K, \Delta, \varepsilon)(\log T)^{1+\varepsilon}$ [Bistritz and Leshem'19]
2. $\zeta(\varepsilon) M^3 K (\log T/\Delta)^{1+\varepsilon}$ [Boursier, Perchet, Kaufmann, M'19]
3. $M^3 K \log(T)/\Delta$ if the maximum matching is unique.

General upper bounds

4. $KM^2\sqrt{T \log T}$ [Boursier, Perchet, Kaufmann, M'19]

Heterogeneous setting

Known results

M players, K arms, $\Delta =$ gap between value of best matching and second best value, $\varepsilon > 0$ arbitrary

Instance-dependent upper bounds

1. $\zeta(M, K, \Delta, \varepsilon)(\log T)^{1+\varepsilon}$ [Bistritz and Leshem'19]
2. $\zeta(\varepsilon) M^3 K (\log T/\Delta)^{1+\varepsilon}$ [Boursier, Perchet, Kaufmann, M'19]
3. $M^3 K \log(T)/\Delta$ if the maximum matching is unique.

General upper bounds

4. $KM^2 \sqrt{T \log T}$ [Boursier, Perchet, Kaufmann, M'19]

Conjecture: if collisions are invisible, regret is linear.

Algorithm description

Leader election and implicit communication

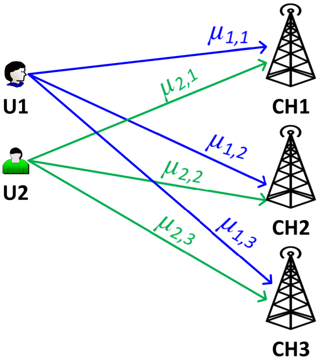
Leader election

1. Players start by running musical chairs.
2. Player occupying smallest chair becomes the leader.
3. Players will use their arms to communicate with the leader via collisions.

Each communicated bit adds M to regret.

Algorithm description

eliminating edges



1. Players explore the edges, get better estimates for the means, communicate to leader.
2. Leader eliminates useless edges gradually.

Algorithm outline

1. Leader is elected.
2. $E \leftarrow$ all edges
3. For epoch $i = 1, 2, \ldots,$
   3.1 Leader: for each $e \in E$, find max matching containing e, send these matchings to players.
   3.2 Players: pull each received matching 2^i times, send updated mean estimates to leader.
   3.3 Leader: for each $e \in E$, find max matching containing e, using updated estimates. Eliminate e if its gap is large.

Algorithm outline

1. Leader is elected.
2. $E \leftarrow$ all edges
3. For epoch $i = 1, 2, \ldots,$
   3.1 Leader: for each $e \in E$, find max matching containing e, send these matchings to players.
   3.2 Players: pull each received matching 2^i times, send updated mean estimates to leader.
   3.3 Leader: for each $e \in E$, find max matching containing e, using updated estimates. Eliminate e if its gap is large.

Analysis (unique maximum matching). A matching with gap Δ is detected to be non-optimal as soon as edge mean accuracy $\leq \Delta/M$, i.e., epoch $\log_2\left(\frac{\log(T)}{(\Delta/M)^2}\right)$.
The matching is pulled $\lesssim \frac{M^2}{\Delta^2}\log(T) \times KM$ times.
Regret $\lesssim \min\{KM^3\log(T)/\Delta, KM\Delta T\} \leq KM^2\sqrt{T\log T}$.

Analysis

Multiple optimal matchings

Number of bits to send in epoch $i = \Theta(i)$, so
total communication bits $= \sum_{i=1}^{\log_2(T)} \Theta(i) = \Theta(\log^2 T)$.

Can make this $(\log T)^{1+1/c}$ by epoch sizes 2^{i^c}
Final regret bound $\leq 2^{2^{c^c}} MK(M^2 \log(T)/\Delta)^{1+1/c}$

Heterogeneous setting

Known results

M players, K arms, $\Delta =$ gap between value of best matching and second best value,

T rounds, $\varepsilon > 0$ arbitrary

Instance-dependent upper bounds

1. $\zeta(M, K, \Delta, \varepsilon)(\log T)^{1+\varepsilon}$ [Bistritz and Leshem'19]
2. $2^{2^{2^{1/\varepsilon}}} M^3 K (\log T/\Delta)^{1+\varepsilon}$ [Boursier, Perchet, Kaufmann, M'19]
3. $M^3 K \log(T)/\Delta$ if the maximum matching is unique.

General upper bounds

4. $KM^2\sqrt{T \log T}$ [Boursier, Perchet, Kaufmann, M'19]

Heterogeneous setting

Known results

M players, K arms, $\Delta =$ gap between value of best matching and second best value,

T rounds, $\varepsilon > 0$ arbitrary

Instance-dependent upper bounds

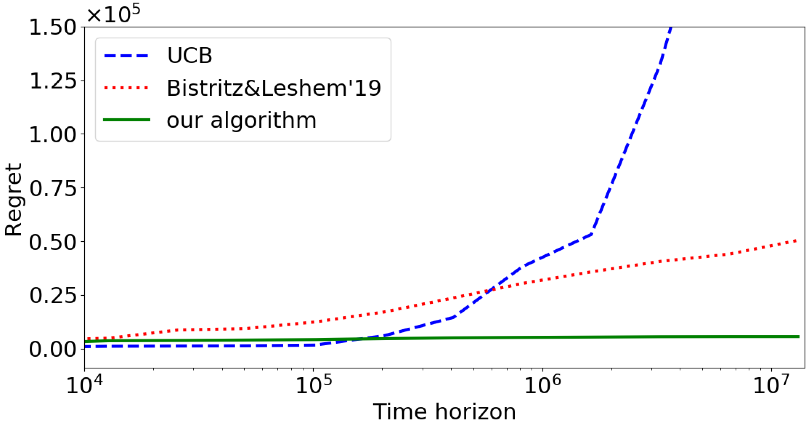
1. $\zeta(M, K, \Delta, \varepsilon)(\log T)^{1+\varepsilon}$ [Bistritz and Leshem'19]
2. $2^{2^{2^{1/\varepsilon}}} M^3 K (\log T/\Delta)^{1+\varepsilon}$ [Boursier, Perchet, Kaufmann, M'19]
3. $M^3 K \log(T)/\Delta$ if the maximum matching is unique.

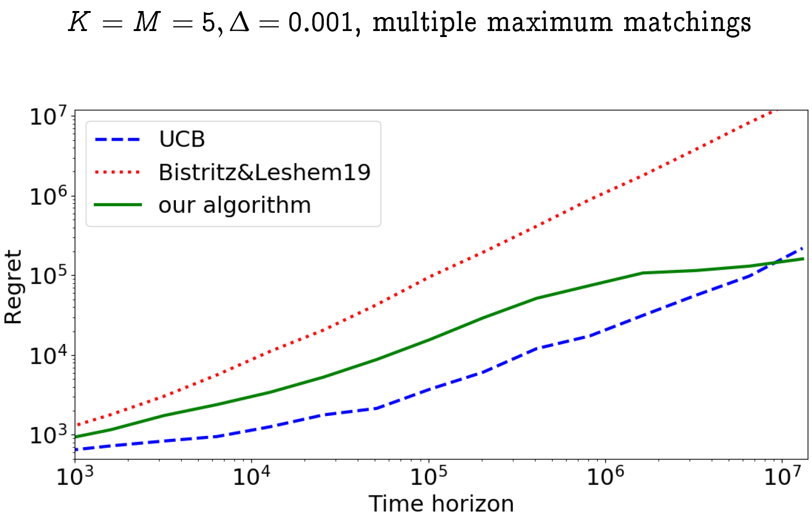
General upper bounds

4. $KM^2\sqrt{T \log T}$ [Boursier, Perchet, Kaufmann, M'19]

Question: Regret $O(\log T)$ while multiple optimal matchings?

$K = M = 3, \Delta = 0.35$, unique maximum matching





$K = M = 5, \Delta = 0.001$, multiple maximum matchings
UCB
Bistritz&Leshem19
our algorithm
Regret
10^7
10^6
10^5
10^4
10^3
Time horizon


WOO HOO! PRESENTATION OVER
ANY QUESTIONS?
memegenerator.net