

複数チャネル通信を用いた衝突検知情報なしマルチプレイヤー多腕バンディットアルゴリズム

Multi-Channel Communication Algorithm for Multi-Player Multi-Armed Bandits without Collision Sensing

泉 知成, 中村篤祥 / 北海道大学大学院 情報科学院

研究の背景・動機

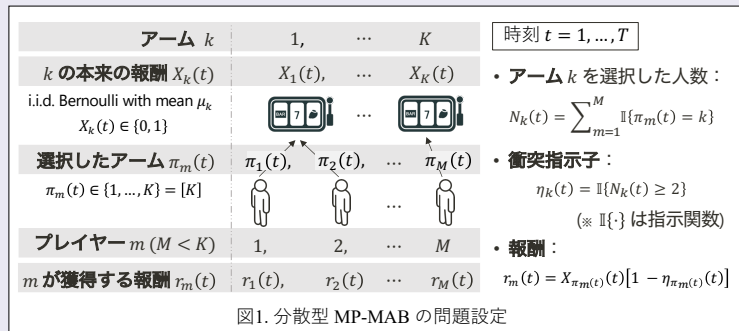
- 近年, IoT や **認知無線** の発展に伴いリソース割当問題の重要性が増す中, 中央制御を持たない 分散型 **マルチプレイヤー多腕バンディット (MP-MAB)** が注目されている [Boursier 24].
- 「**衝突検知情報なし**」という他プレイヤーとの衝突を直接観測できない最も困難な設定に対し, 期待値の高い1つの有望なアーム (Good Arm) を通信チャネルとする手法が提案されている [Huang 22].
- しかし, 単一チャネルでの逐次通信は他プレイヤーの待機時間がボトルネックとなり, プレイヤー数の増加に伴い通信リグレットが悪化する問題があった.

研究の目的

- 既存手法を拡張し, **n 本の Good Arm を通信チャネル**として活用する並列分散アルゴリズムを提案する.
- 同期・通信プロセスの並列化により, 待機時間のボトルネックを解消し, 特定の環境下において提案手法が優れたリグレット上界を達成することを示す.

問題設定

- 本研究では, アームが K 本, プレイヤーが M 人 ($M < K$) からなる 分散型 MP-MAB を扱う (図1).

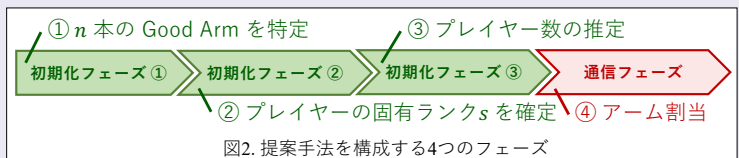


- 衝突が発生していない時の各アームの期待報酬を降順に並べたものを $\mu_{(1)}, \dots, \mu_{(K)}$ と定義し, $\mu_{(1)} \geq \mu_{(2)} \geq \dots \geq \mu_{(K)}$ が成り立つとする. (ただし, プレイヤーはこの順序や具体的な値を知らない.)
- プレイヤーは, 各時刻で必ず **1つだけ** アームを引く.
- 衝突情報検知なし**: 各プレイヤーは, 自身の報酬 $r_m(t)$ のみを観測し, それ以外の情報 (他者の行動, **衝突指示子 $\eta_k(t)$** , 通信など) を一切受け取らない. (報酬が0だったとき, その原因が**衝突によるものか**, 単にアームの報酬が外れただけなのかをプレイヤーは区別できない.)
- 目的**: 以下の累積リグレット $R(T, \mu)$ を最小化する:

$$R(T, \mu) = \sum_{t=1}^T \sum_{m=1}^M \mu_{(m)} - \mathbb{E}[r_m(t)]$$

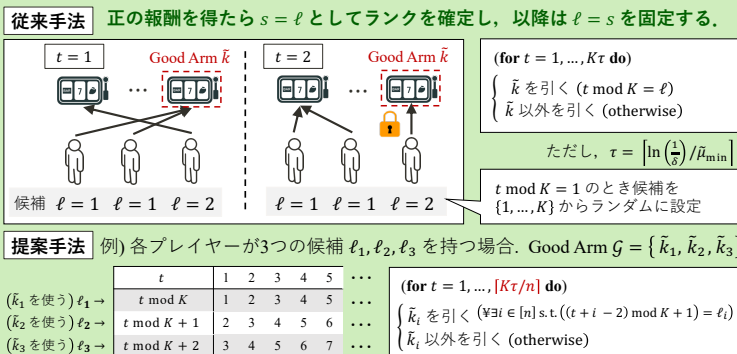
提案するアルゴリズム

- 提案手法は, 4つのフェーズによって構成される (図2).



- 初期化フェーズ ①** では, 通信路として使用可能な n 本の Good Arm を特定する. (特定した Good Arm $\tilde{k}_i$ は番号の小さい順から i 番目.)

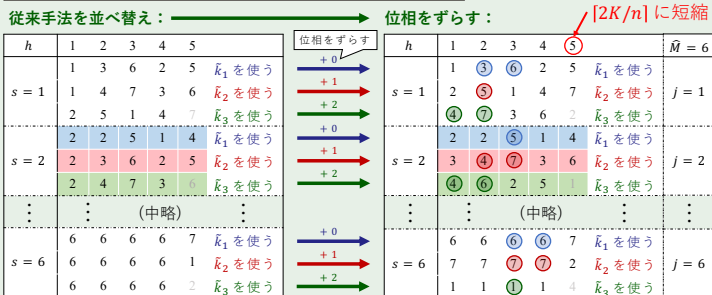
■ 初期化フェーズ ②: 椅子取りゲーム (以降, 従来手法は [Huang 22])



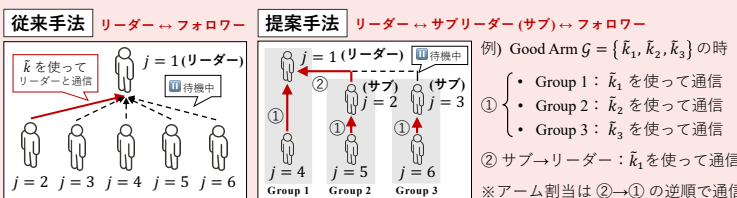
■ 初期化フェーズ ③: 他者と必ず1度だけ衝突し, プレイヤー数を推定



提案手法: $K = 7, M = 6$, Good Arms: $\mathcal{G} = \{\tilde{k}_1, \tilde{k}_2, \tilde{k}_3\}$



■ 通信フェーズ: フォロワーの観測結果をリーダーに集約・アーム割当



結論 (リグレット上界)

- 第2項 (初期化リグレット) が支配的な場合, 提案手法は従来手法に劣る.
- 一方で, 第3項 (通信リグレット) がシステム全体のボトルネックになる場合は優位性をもつ. 具体的には, アームの期待値の減衰が緩やかで $n\mu_{(n)} > \mu_{(1)}$ を満たす環境下では, 学習効率を向上させる.

従来手法: $R_{\text{prev}}(T, \mu) \leq O\left(\sum_{k>M} \frac{\ln T}{\Delta_k}\right) + O(K^2 M \ln T) + O\left(KM^2 \ln\left(\frac{1}{\Delta_M}\right)^2 \ln T\right)$

理論限界 (ただし, $\Delta_k = \mu_{(M)} - \mu_{(k)}$, $\Delta_M = \mu_{(M)} - \mu_{(M+1)}$)

提案手法: $R_{\text{prop}}(T, \mu) \leq O\left(\sum_{k>M} \frac{\ln T}{\Delta_k}\right) + O\left(K^2 M \frac{\mu_{(1)}}{\mu_{(n)}} \ln(nT)\right) + O\left(\frac{KM^2}{n} \frac{\mu_{(1)}}{\mu_{(n)}} \ln\left(\frac{1}{\Delta_M}\right)^2 \ln(nT)\right)$

参考文献

- [Boursir 24] Boursir, E. and Perchet, V.: A Survey on Multi-player Bandits, *JMLR*, Vol. 25, No.137, pp. 1–45, 2024.
- [Huang 22] Huang et al.: Towards Optimal Algorithms for Multi-player Multi-armed Bandits without Collision Sensing Information, *COLT*, Vol. 178, pp. 1990–1012, 2022.

Appendix: 数値実験

■ 提案アルゴリズムのパラメータ n (グループサイズ) と **累積リグレット** のトレードオフを定量評価.

■ ベースライン [Huang 22] との比較により改善効果を検証.

表1. 数値実験時の各パラメータ

パラメータ	値
アーム数 K , プレイヤー数 M	$K = 20, M = 8$
ホライズン T	$T = 10^6$
失敗確率 δ	$\delta = 7.24 \times 10^{-8}$

■ **アーム期待報酬** : 0.95, 0.90, 0.86, 0.82, 0.78, 0.70, 0.64, 0.58, ..., 0.03

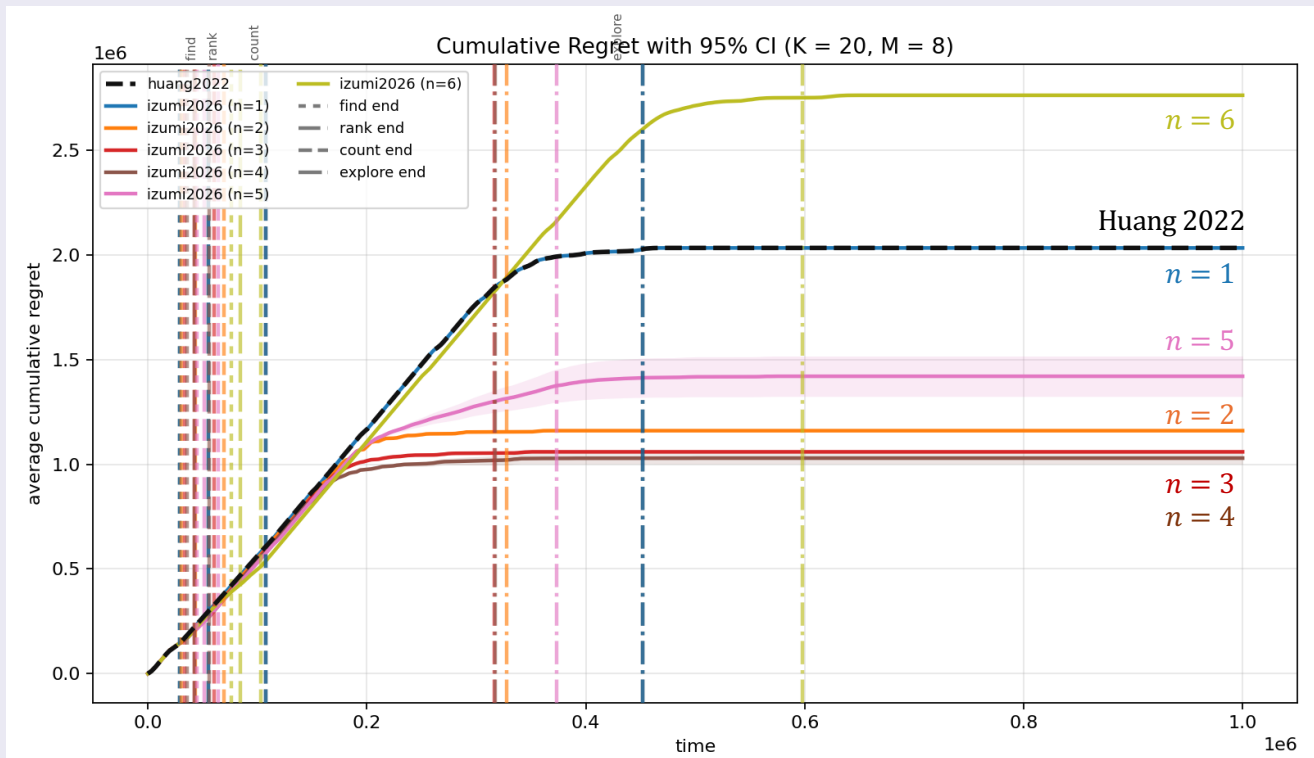


図1. 数値実験の結果 (100回の独立試行平均)

■ $n = 1$: ベースライン [Huang 22] と一致.

■ $n = 4$ が最良 : ベースライン [Huang 22] 比でリグレットを約 **49%** 削減.

■ $n \leq 4$: n の増加とともにリグレット減少, n 間での改善幅は逓減.

■ $n \geq 5$ で劣化 : **Find god arms** フェーズのコストが増大.

($n = 1$: 28,800 [steps] $\rightarrow$ $n = 6$: 76,050 [steps])