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 *1　　中村 篤祥 *1
Tomonari Izumi　　Atsuyoshi Nakamura

*1北海道大学大学院 情報科学院
Graduate School of Information Science and Technology, Hokkaido University

衝突検知情報なしマルチプレイヤー多腕バンディットは，プレイヤーが衝突発生を検知できない最も困難な分散学習設定である．Huang ら (2022) は同問題に対し，単一アームを介した通信手法を提案した．この手法は事前情報を必要としない利点を持つが，単一通信路に起因する大規模環境での非効率性が課題であった．本研究では，通信用アームを 2 本以上に拡張し，通信の非効率性を解消することを目的とする．我々の提案手法は，初期探索のステップ数が増加するものの，その後の通信フェーズでの並列実行による高速化を可能にする．本稿では，先行研究と比較した場合，大規模システムにおける有効性を示すことを目指す．

1. はじめに

近年，IoT (Internet of Things) デバイスの爆発的な普及やコグニティブ無線ネットワークの発展に伴い，中央集権的な制御を持たない分散システムにおけるリソース割り当て問題が注目を集めている．このような問題は，**マルチプレイヤー多腕バンディット** (Multi-Player Multi-Armed Bandits: MP-MAB) として定式化され，数多くのアルゴリズムが提案されてきた [Boursier 24]. MP-MAB は，複数のプレイヤーが未知の確率分布に従うアーム (リソース) を繰り返し選択し，プレイヤー間の衝突を回避しながらシステム全体の累積報酬最大化を目指す問題である．本研究では，中央サーバによる指示が存在せず，各プレイヤーが自律的に意思決定を行う**分散型 (Decentralized)** の設定を対象とする．

分散型 MP-MAB の中でも，協調学習を行う上で最も困難な設定として，プレイヤーが「他者と衝突した」という事実を直接観測できない **衝突検知情報なし** (No Collision Sensing) という制約がある．同問題に対し Huang ら [Huang 22] は，期待値の高いアーム (Good Arm) を 1 つ特定し，そのアームを通信チャネルとして利用する手法を提案した．彼らの手法は，最悪ケースのアーム期待値に依存しないという優位性をもつ一方で，全ての同期と通信を単一の Good Arm を介して行うため，処理が逐次的に行われる．そのため，アーム数 K やプレイヤー数 M が増大すると，通信の待ち時間や同期にかかるステップ数が $O(K^2)$ のオーダーで増大し，学習の収束が遅れるという構造的なボトルネックを抱えていた．

そこで本研究では，Huang らの手法を拡張し，**複数の通信チャネルを並列に利用する分散アルゴリズム**を提案する．本手法の新規性は，初期の探索フェーズにおいて単一ではなく n 本の Good Arm を特定し，それらを並列な通信路として活用する点にある．これにより，従来は逐次的に行われていた同期プロセス (各プレイヤーのランク付けや総人数の推定) を n 並列で実行することが可能となる．本稿では，提案アルゴリズムの詳細な設計を示し，理論解析を通じて，特定の環境下において提案手法が従来手法よりも優れたリグレット上界を達成することを示す．

2. 問題設定

本研究では，K 本のアーム (選択肢) が存在し，プレイヤーが M 人 ($M < K$) の分散型 MP-MAB 問題を扱う．各時刻 $t = 1, \dots, T$ において，プレイヤー $m \in \{1, \dots, M\}$ は $\pi_m(t) \in \{1, \dots, K\}$ で表されるアームを必ず選択する．あるアーム k を選択したプレイヤーの総数を $N_k(t) = \sum_{m=1}^{M} \mathbb{1}\{\pi_m(t) = k\}$ と定義する（ここで $\mathbb{1}\{\cdot\}$ は指示関数である）．同一のアーム k が 2 人以上のプレイヤーによって選択された場合（すなわち $N_k(t) \geq 2$ の場合），そのアームにおいて衝突が発生したという．アーム k における**衝突指示子** (collision indicator) を $\eta_k(t) = \mathbb{1}\{N_k(t) \geq 2\}$ と定義すると，プレイヤー m が受け取る**報酬** (reward) $r_m(t)$ は次式で表される：

$$r_m(t) = X_{\pi_m(t)}(t) \left[1 - \eta_{\pi_m(t)}(t)\right]$$

ここで，$X_k(t) \in \{0, 1\}$ はアーム k 本来の報酬を表すベルヌーイ確率変数であり，期待値 μ_k を持つ独立同分布 (i.i.d.) 過程に従う．つまり，プレイヤー m は衝突に巻き込まれれば報酬は 0 で，そうでなければ $\mu_{\pi_m(t)}$ を平均とする二値報酬を受け取る．衝突が発生していないときの各アームの平均報酬を降順に並べたものを $\mu_{(1)}, \dots, \mu_{(K)}$ と定義し，$\mu_{(1)} \geq \mu_{(2)} \geq \cdots \geq \mu_{(K)}$ が成り立つものとする．ただし，プレイヤーはこの順序や具体的な値を知らない．

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 $\eta_k(t)$ や**通信** (communication) など）を一切受け取らない．ここで重要な点は，プレイヤー m が $r_m(t) = 0$ を観測した際，その原因が衝突によるものなのか（$\eta_{\pi_m(t)}(t) = 1$），単にアームの報酬が外れただけなのか（$X_{\pi_m(t)}(t) = 0$）を区別できないことである．一方，$r_m(t) = 1$ を観測した場合は，衝突が発生していないことが確定する．これが「衝突検知情報なし」という制約の本質的な困難さである．

システム全体の目標は，全プレイヤーの累積報酬を最大化すること，言い換えれば，以下の累積リグレット $R(T, \mu)$ を最小化することである．

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

理想的な状態は，期待値の高い上位 M 本のアームを，M 人のプレイヤーが重複なく分担して選択し続ける状態である．

3. 先行研究とその課題

本研究の基礎となる Huang ら [Huang 22] の手法は，大きく 4 つのフェーズからなる．第一に，期待値が $\mu_{\tilde{k}} \geq \frac{1}{8}\mu_{(1)}$ を満たすアーム (Good Arm) $\tilde{k}$ を 1 本特定し，以降の唯一の通信路として利用する (**FindGoodArm**)．ここで，全プレイヤーは同一のアーム $\tilde{k}$ を Good Arm として認識・共有する．第二に，プ

連絡先: 泉 知成，北海道大学大学院 情報科学院，〒 060-0814 北海道札幌市北区北 14 条西 9 丁目，+81-80-3455-9812，izumi.tomonari.n6@elms.hokudai.ac.jp

レイヤーを一意に定めるための疎な外部ランク $s \in \{1, \ldots, K\}$ を割り当てる (**VirtualMusicalChairs**)．第三に，外部ランク s を用いて，総プレイヤー数 $\hat{M}$ の推定と，より密な内部ランク $j \in \{1, \ldots, M\}$ の導出を行う (**VirtualNumberPlayers**)．最後に，プレイヤーごとに計算した平均報酬の全情報をリーダー 1 名に集約し，全プレイヤーに対して最適なアームを割り当てる (**DistributedExploration**)．

この一連の手法における最大の課題は，同期および情報伝達の全てを「たった 1 本の Good Arm」に依存している点である．ランク割り当てやプレイヤー数推定において，全プレイヤーが単一の通信路を順番に利用して状態を確認するため，$O(K)$ の処理を繰り返す必要があり，全体として $O(K^2)$ の初期化コストが発生する．このコストは大規模システムにおいて致命的であり，リグレットが対数オーダーで収束し始めるまでの時間を遅らせる原因となる．

4. 提案手法

提案手法では，並列数 n をパラメータとして導入し，以下の 4 つの拡張フェーズとそれをまとめる 1 つのアルゴリズムにより構成される．なお，以降で提案される全てのフェーズ・アルゴリズムは，プレイヤー $m = 1, \ldots, M$ が独立して実行するものとする．

4.1 FindMultipleGoodArms

通信路として使用可能な n 本の Good Arm を特定する (Algorithm 1)．まず全プレイヤーがランダム探索を行い，期待報酬が閾値 2^{1-p} 以上のアームを特定することを，$p = 1$ から開始して p の値をインクリメントしながら n 本見つかるまで繰り返す．特定されたアームは集合 $\mathcal{G}$ に追加され，探索候補 $\mathcal{K}$ から除外される．このフェーズの計算量は増大するが，後続フェーズの効率化によって全体の性能向上を図る．

なお，本アルゴリズムが終了するための前提条件として、環境内に $\mu_{(n)} > 0$ を満たすアームが存在することを仮定している．また，以降のアルゴリズムにおいて，特定した Good Arm $\tilde{k}_i \in \mathcal{G}$ $(1 \le i \le n)$ はアームのインデックスが小さい順から i 番目の要素とし，全プレイヤー間で共通の認識であると仮定する．

4.2 ParallelVirtualMusicalChairs

各プレイヤーに一意な外部ランク $s \in \{1, \ldots, K\}$ を割り当てる (Algorithm 2)．先行研究では，時刻 $t \bmod K$ が 1 から K まで反復する長さ K のサイクルにおいて，各プレイヤーは自身が選んだ 1 つのタイミングでしか空席確認 (報酬による観測) を行えなかった．これに対し提案手法では，特定した n 本の Good Arm $\tilde{k}_i$ $(i = 1, \ldots, n)$ それぞれに対して，個別の空席を確認するタイミング ℓ_i を割り当てることで，1 サイクル中に最大 n 回の空席確認を行う．これにより，本アルゴリズムでは，先行研究と比較して探索機会は n 倍に増加する．

ここで重要な点は，同一時刻 t に複数のアームを引く物理的制約違反を防ぐメカニズムである．各プレイヤーは，新たなタイミング ℓ_i をランダムに選択する際，それ以前に決定したタイミング ℓ_j $(j < i)$ と観測時刻が重ならないように位相をずらす必要がある．アルゴリズム内では，この制約を禁止集合 $\mathcal{F}_i$ として動的に計算し，サンプリング候補から除外することで，同一時刻における複数アームの同時選択を回避している．

結果として制約を満たしつつ安全に n 回の探索を実行でき，ランク s 確定に必要な総反復回数は，従来の $K\tau$ から $\lceil K\tau/n \rceil$ へと短縮される．ただし，Algorithm 2 および後続のアルゴリズムにおいて，$\tau = \lceil \ln(1/\delta)/\tilde{\mu}_{\min} \rceil$ と定義される．ここで，$\tilde{\mu}_{\min} = \min_{k \in \mathcal{G}} \tilde{\mu}[k]$ である．

4.3 ParallelVirtualNumberPlayers

各プレイヤーが総プレイヤー数 $\hat{M}$ と，密な順位付けである内部ランク $j \in \{1, \ldots, \hat{M}\}$ を推定する (Algorithm 3)．各プレイヤーは，第二フェーズで獲得した疎な外部ランク $s \in \{1, \ldots, K\}$ に基づき，意図的に他プレイヤーと衝突させることで互いの存在を確認する．

この存在確認は，仮想的な時刻 v に基づく「待機」と「巡回」によって行われる．各プレイヤーは，自身の仮想時間 v が $2s$ に達するまでは，外部ランク s に対応する位置に留まり (待機)，$2s$ 以降は他ランクを順番に訪問する (巡回)．この設計により，各プレイヤーは自分以外の全プレイヤーと正確に 1 度ずつ衝突を起こすため，衝突回数を計測すれば 総プレイヤー数 $\hat{M}$ を推定できる．さらに，外部ランク s が小さい (上位の) プレイヤーほど先に巡回を開始する性質を利用し，自身が待機している期間 $(v \le 2s)$ に発生した衝突回数を計測することで，自身より上位のプレイヤー数を高い確率で把握し，内部ランク j を決定することができる．

先行研究では，仮想時間 v に対して 1 ステップずつ逐次検証していたため $2K^2\tau$ 回のサンプリングを要した．これに対し提案手法では，v を n 個のブロックに分割し，n 本の Good Arm を用いて並列に検証を行う．各反復ステップ h において，プレイヤーは n 個の仮想時間に対する目標位置 l_i を同時に計算し，Algorithm 2 と同様の位相ずらしを適用することで，同時刻に複数本のアームを同時に引けないという制約を満たしつつ状態を観測できる．結果として，全プレイヤーの存在確認とランク圧縮に必要な総ステップ数は，従来の $2K^2\tau$ から $\lceil 2K/n \rceil K\tau$ へと短縮される．

4.4 HierarchicalDistributedExploration

これまでのフェーズで推定した情報を用いて分散探索を行い，最終的なアーム割り当てを決定する (Algorithm 4)．先行研究では，単一のリーダーが全プレイヤーから情報を順次収集する

Algorithm 1 FindMultipleGoodArms

Input: K: total number of arms, δ: confidence parameter, n: number of good arms to find ($n < K - M$).
Output: $\mathcal{G}$: set of find good arms, $\tilde{\mu}$: lower bounds map.

```
G ← ∅, K ← {1,...,K} # Initialization
p ← 0 # Initialize phase counter
while |G| < n and |K| > 1 do
  p ← p + 1, R[k], N[k] ← 0 for k ∈ K
  # Sub-phase 1: Explore active arms uniformly
  for t = 1,...,6|K|2^p ⌈ln 2n/δ⌉ do
    Select arm k ∈ K uniformly at random, observe reward r, R[k] ← R[k] + r, N[k] ← N[k] + 1
  end for
  # Sub-phase 2: Confirm accepted arms
  for each ℓ ∈ K in ascending order do
    R'[k] ← 0 for k ∈ K # rewards of samples
    # if arm ℓ was accepted sample arms uniformly
    if N[ℓ] > 0 and R[ℓ]/N[ℓ] ≥ 2^{1−p} then
      for t = 1,...,|K|2^p ⌈ln 2n/δ⌉ do
        Select arm k ∈ K uniformly at random, observe reward r, R'[k] ← R'[k] + r
      end for
      if R'[ℓ] ≥ 1 then
        Add ℓ to G, μ̃[ℓ] ← 2^{−p}
        if |G| ≥ n then break end if # Exit FOR loop
      end if
    else
      for t = 1,...,|K|2^p ⌈ln 2n/δ⌉ do Select arm ℓ observe reward r, R'[ℓ] ← R'[ℓ] + r end for
    end if
  end for
  K ← K\G # Update the active set at the end of the phase
end while
```

Algorithm 2 ParallelVirtualMusicalChairs

Input: K: total number of arms, $\mathcal{G}$: set of good arms, τ: sampling times.
Output: s: external rank of the player.

$s \leftarrow -1$ # *Rank of the player is initially unset*
$n = |\mathcal{G}|$ # *Number of the good arms*
# *Musical chairs on the good arms* $\mathcal{G} = \{\tilde{k}_1, \ldots, \tilde{k}_n\}$
for $t = 1, \ldots, \lceil K\tau/n \rceil$ **do**
  # *Select sampling slots at each block start*
  if $t \bmod K = 1$ **then**
    if $s = -1$ **then**
      $\mathcal{F}_1 = \emptyset$ # *Initialization of the sets*
      for $i = 1, \ldots, n$ **do**
        # *Choose the i-th candidate arm*
        Draw $\ell_i \in \{1, \ldots, K\} \setminus \mathcal{F}_i$ uniformly at random
        $\mathcal{F}_{i+1} \leftarrow \{\big((\ell_j + (i-j)) \bmod K\big) + 1 \mid 1 \le j \le i\}$
      end for
    else
      $\ell_j \leftarrow s$ for all $j \in \{1, \ldots, n\}$ # *Wait at rank s*
    end if
  end if
  # *sample the corresponding time slot*
  if $\exists i \in \{1, \ldots, n\}$ s.t. $((t+i-2) \bmod K)+1 = \ell_i$ **then**
    Select arm $\tilde{k}_i$, and observe reward r_i
    # *set the rank if it was not set yet and a non zero reward was obtained*
    if $r_i > 0$ and $s = -1$ **then**
      $s \leftarrow \ell_i$
      $\ell_j \leftarrow s$ for all $j \in \{1, \ldots, n\}$
    end if
  else
    Select an arbitrary arm in $\{1, \ldots, K\} \setminus \mathcal{G}$
  end if
end for

1 対多の通信を採用しているため，通信中以外のフォロワーは待機状態となり，通信の非効率性が生じていた．

この非効率性を削減するため，提案手法では n 本の Good Arm を用いた**階層的リーダー制**を導入する．具体的には，各プレイヤーが自身の内部ランク j に基づき，第 $\big((j-1) \bmod n\big)+1$ グループに所属する．ここで，内部ランクが $j \le n$ のプレイヤーは第 j グループの**サブリーダー**を務め，残りのプレイヤー $(j > n)$ は**フォロワー**として各グループに配置される．n 人のサブリーダーは同時に，自身に割り当てられたグループのフォロワーから探索結果を集約する．

ここで，内部ランク $j = 1$ のプレイヤーは**グランドリーダー**と呼ばれる特別な役割を持つ．グランドリーダーは，まず自身も第 1 グループのサブリーダーとして自グループの情報を収集した後に，他の全サブリーダー $(j = 2, \ldots, n)$ から各グループで集約された情報を収集する．最後に，グランドリーダーが全情報を統合して各プレイヤーへの最適なアーム割り当てを決定し，情報の収集時とは逆の経路で指令を伝達する．

なお，本論文では各階層の具体的な通信プロトコル関数 (`ComGrandLeader`, `ComSubLeader`, `ComFollower`) の擬似コードによる定義を割愛する．

4.5 提案するアルゴリズム

以上の 4 つの拡張フェーズを組み合わせることで，提案手法は Algorithm 5 のようになる．

5. 理論的解析

本節では，提案手法の計算量およびリグレット上界を理論的に評価し，先行研究 [Huang 22] との関係を明らかにする．

Algorithm 3 ParallelVirtualNumberPlayers

Input: K: total number of arms, $\mathcal{G} = \{\tilde{k}_1, \ldots, \tilde{k}_n\}$: set of good arms, s: external rank, τ: sampling times.
Output: $\hat{M}$: estimated number of players, j: internal rank.

$\hat{M} \leftarrow 1$, $j \leftarrow 1$, $n \leftarrow |\mathcal{G}|$ # *Initialization*
for $h = 1, \ldots, \lceil 2K/n \rceil$ **do**
  for $i = 1, \ldots, n$ **do**
    $v \leftarrow (h-1)n + i$ # *Calculate virtual time*
    if $v > 2K$ **then** $\ell_i \leftarrow -1$; **continue end if**
    if $v > 2s$ **then**
      $\ell_i \leftarrow \big((s + (v-2s) - 1) \bmod K\big) + 1$
    else
      $\ell_i \leftarrow s$ # *Wait at rank s*
    end if
  end for
  for $k = 1, \ldots, K$ **do**
    $R \leftarrow 0$
    if $\exists i \in \{1, \ldots, n\}$ s.t. $\ell_i \neq -1$ **and** $\big((\ell_i - 1 + (i-1)) \bmod K\big) + 1 = k$ **then**
      $\tilde{i} \leftarrow$ the matching index
      for $t = 1, \ldots, \tau$ **do** Select arm $\tilde{k}_{\tilde{i}}$, observe reward r, $R \leftarrow R + r$ **end for**
      if $R = 0$ **then**
        $\hat{M} \leftarrow \hat{M} + 1$; $v \leftarrow (h-1)n + \tilde{i}$
        if $v \le 2s$ **then** $j \leftarrow j + 1$ **end if**
      end if
    else
      for $t = 1, \ldots, \tau$ **do** Select an arbitrary arm in $\{1, \ldots, K\} \setminus \mathcal{G}$ **end for**
    end if
  end for
end for

先行研究における期待リグレット上界 $R_{\mathrm{prev}}(T, \mu)$ は，以下の 3 つの主要項で構成されている．

$$R_{\mathrm{prev}}(T, \mu) \le O\left(\sum_{k>M} \frac{\ln T}{\Delta_k}\right) + O\left(K^2 M \ln T\right) + O\left(KM^2 \ln\left(\frac{1}{\Delta_M}\right)^2 \ln T\right)$$

ここで，$\Delta_k = \mu_{(M)} - \mu_{(k)}$，$\Delta_M = \mu_{(M)} - \mu_{(M+1)}$ である．また，第 1 項は最適アームの探索に伴う純粋な活用リグレット，第 2 項は単一の Good Arm を特定・共有して全プレイヤーのランクと人数を確定させるまでの初期化リグレット，第 3 項は 1 対多の逐次通信に伴う通信リグレットである．ここで，初期化にかかる所要ステップ数自体は $O(K^2/\mu_{(1)} \ln(1/\delta))$ であり分母に最良アームの期待値 $\mu_{(1)}$ を含む．しかし，1 ステップあたりにシステム全体で失われる最大リグレットは最良アーム基準の $M\mu_{(1)}$ であり，所要時間と掛け合わせることで $\mu_{(1)}$ が相殺される．さらに，期待リグレットを定数オーダーに抑えるためのパラメータ設定として $\delta = 1/T$ を代入することで $\ln(1/\delta)$ が $\ln T$ へと変換され，第 2 項や第 3 項のリグレット上界にはアームの絶対的な期待値が現れない構造となっている．

これに対し，提案手法の初期化フェーズでは n 本の Good Arm を並列に探索・確立する．全プレイヤーが一様ランダムにアームを選択した際，他者との衝突を免れて報酬を得られる期待値は $\rho_k = (1 - 1/K)^{M-1} \mu_k$ となる．提案手法のアルゴリズムが終了するためのステップ数は，期待値が n 番目に大きいアーム $\mu_{(n)}$ に対応する $\rho_{(n)}$ の逆数に依存して増大する．さらに，n 本全てのアームを確率 $1 - \delta$ 以上で正しく特定するためには，各アーム単体の誤判定確率を δ/n 以下に抑える必要があり，サ

Algorithm 4 HierarchicalDistributedExploration

Input: K: number of arms, j: internal rank of a player, $\hat{M}$: estimated number of players, $\mathcal{G} = \{\tilde{k}_1, \ldots, \tilde{k}_n\}$: set of good arms, τ: sampling times

Output: f: an arm amongst the M best arms assigned to the player

$p \leftarrow 0$, $f \leftarrow -1$, $n \leftarrow |\mathcal{G}|$ # *Initialization*
$R[k], v[k] \leftarrow 0$ for $k = 1, \ldots, K$ # *Rewards and number of samples for each arm*
# *Initialize memory for Grand Leader and Sub-Leaders*
if $j \leq n$ **then**
  for $m = 1, \ldots, \hat{M}$ and $k = 1, \ldots, K$ **do**
    $\hat{\mu}[k, m], N[k, m] \leftarrow 0$ **end for**
end if
$M' \leftarrow \hat{M}$, $\mathcal{K} \leftarrow \{1, \ldots, K\}$ # *Number of active players and set of active arms*
while $f = -1$ **do**
  $p \leftarrow p + 1$
  # *Sub-phase 1: Explore arms by sequential hopping*
  $k \leftarrow j$
  for $t = 1, \ldots, |\mathcal{K}| 2^p \lceil \ln \frac{1}{\delta} \rceil$ **do**
    $k \leftarrow (k + 1) \bmod |\mathcal{K}|$
    Select arm k, observe reward r, $R[k] \leftarrow R[k] + r$, $v[k] \leftarrow v[k] + 1$, $E[k] \leftarrow R[k]/v[k]$
  end for
  # *Sub-phase 2: Hierarchical Information Exchange*
  $Q \leftarrow \lceil \frac{p}{2} + 3 \rceil$ # *Message length*
  if $j = 1$ **then**
    # *Grand-Leader*
    $(f, \mathcal{K}, M', \hat{\mu}, N)$
      $\leftarrow$ `ComGrandLeader`$(\hat{\mu}, N, E, \mathcal{G}, M', Q, \tau, p)$
  else if $j \leq n$ **then**
    # *Sub-Leader: Send to Grand-Leader*
    $(f, \mathcal{K}, M', \hat{\mu}, N)$
      $\leftarrow$ `ComSubLeader`$(\hat{\mu}, N, E, j, \mathcal{G}, M', Q, \tau, p)$
  else
    # *Follower: Send to Sub-Leader*
    $(f, \mathcal{K}, M') \leftarrow$ `ComFollower`$(E, j, \mathcal{G}, M', Q, \tau)$
  end if
end while

ンプリング回数が増加する．したがって，提案手法の初期化ステップ数は $O\left(\frac{K^2}{\mu_{(n)}} \ln \frac{n}{\delta}\right)$ となり，先行研究よりも時間を要する．この期間中に生じる 1 ステップあたりの最大リグレットは，先行研究と変わらず $M\mu_{(1)}$ である．両者を掛け合わせると，先行研究のように分母と分子が完全には相殺されず，リグレットの上界には $\mu_{(1)}/\mu_{(n)}$ という期待値の比率が残る．

一方で，こうして特定された n 本の通信チャネルは，通信フェーズ（Algorithm 4）における並列情報集約に用いられる．サブリーダーを介して n 並列の通信を行うことで，1 フェーズあたりの通信ステップ数は $O(M)$ から $O(M/n)$ へと削減される．

これらの所要時間と最大リグレットの関係性を統合し，$\delta = 1/T$ を代入すると，提案手法の期待リグレット上界 $R_{\mathrm{prop}}(T, \mu)$ は以下のように導出される．

定理 1. 提案手法を用いた際の T ステップ後の期待リグレットは以下の上界を持つ．

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

Algorithm 5 ProposedParallelAlgorithm

Input: K: number of arms, δ: confidence level, n: number of good arms to find

Output: $\bar{k}$: an arm amongst the M best arms assigned to the player

$(\mathcal{G}, \tilde{\mu}) \leftarrow$ `FindMultipleGoodArms`(K, δ, n)
$\tilde{\mu}_{\min} \leftarrow \min_{k \in \mathcal{G}} \tilde{\mu}[k]$
$\tau \leftarrow \lceil \ln(1/\delta)/\tilde{\mu}_{\min} \rceil$ # *Calculate sampling times*
$s \leftarrow$ `ParallelVirtualMusicalChairs`$(K, \mathcal{G}, \tau)$
$(\hat{M}, j) \leftarrow$ `ParallelVirtualNumberPlayers`$(K, \mathcal{G}, s, \tau)$
$\bar{k} \leftarrow$ `HierarchicalDistributedExploration`$(K, j, \hat{M}, \mathcal{G}, \tau)$

ここで，第 2 項が初期化リグレット，第 3 項が並列化された通信リグレットを表す．先行研究の上界と比較すると，提案手法の第 2 項および第 3 項には，アームの期待値の比 $\gamma = \mu_{(1)}/\mu_{(n)}$（$\gamma \geq 1$）というペナルティ係数が乗じられている．これは，$n$ 本の並列通信チャネルを確立するために，最も期待値の高いアームだけでなく，より期待値の低い n 番目のアーム $\mu_{(n)}$ までを探索・利用しなければならないことに起因する．

システム全体の支配項が第 2 項（初期化リグレット）となる環境においては，本手法は先行研究に劣る．この場合，アルゴリズム内部の確定的な同期待ち時間を並列化によって $1/n$ に削減できたとしても，チャネル探索ステップ数の増大（ペナルティ γ）に吸収されるため，第 2 項全体としては先行研究の $O(K^2 M \ln T)$ よりも確実に大きくなる．

一方で，先行研究に対する優位性を持つのは，第 3 項（通信リグレット）がシステム全体のボトルネックとなる環境である．上位アーム間のギャップ Δ_M が適度に小さく，第 3 項がシステム全体の支配項へと逆転する環境においては，提案手法は支配項である通信リグレット自体を並列化により $1/n$ に圧縮する．したがって，アームの期待値の減衰が緩やかであり，初期投資のペナルティ増大が通信の短縮効果を打ち消さない（すなわち $n\mu_{(n)} > \mu_{(1)}$）という条件を満たせば，提案手法は大規模システムにおける全体の学習効率を理論的に改善する．

6. 結論

本研究では，衝突検知情報なしの分散型マルチプレイヤー多腕バンディット問題に対し，複数の通信チャネルを動的に確立して並列処理を行うアルゴリズムを提案した．提案手法は，n 本の Good Arm を用いることで，従来手法の課題であった初期化フェーズにおける同期待ち時間を $1/n$ に圧縮し，さらに探索情報の通信にかかるステップ数を $1/n$ に並列化する．理論解析の結果，初期化リグレット全体は従来手法と比較して増大するものの，通信コストがシステム全体のボトルネックとなる環境においては，本手法がトータルのリグレットを改善できることを明らかにした．具体的には，アームの期待値の減衰が緩やかで $n\mu_{(n)} > \mu_{(1)}$ を満たす条件下において，提案手法は大規模システムにおける学習効率の向上に寄与する．今後の展望として，シミュレーションによる実証と，より厳密な論理的解析を行いたい．

参考文献

[Boursier 24] Boursier, E. and Perchet, V.: A Survey on Multi-player Bandits, *Journal of Machine Learning Research*, Vol. 25, No. 137, pp. 1–45 (2024)

[Huang 22] Huang, W., Combes, R., and Trinh, C.: Towards Optimal Algorithms for Multi-Player Bandits without Collision Sensing Information, in *Proceedings of the 35th Conference on Learning Theory (COLT)*, Vol. 178 of *Proceedings of Machine Learning Research*, pp. 1990–2012, PMLR (2022)