時間変化に関する外部情報を考慮した非定常多腕バンディット問題

Non-stationary Stochastic Multi-armed Bandit Problems with External Information on Stationarity

難波 博之
Hiroyuki Namba

株式会社日立製作所
Hitachi, Ltd.
hiroyuki.namba.dx@hitachi.com

keywords: multi-armed bandit problem, sequential optimization

Summary

Multi-armed bandit problem is a fundamental mathematical problem in sequential optimization and reinforcement learning that has a variety of application such as online recommendation system and clinical trial design. Multi-armed bandit problem can describe a situation in which a player tries to select a good choice sequentially from given candidate choices to maximize the cumulative reward. In this paper, we consider the non-stationary multi-armed bandit problems. Non-stationary means the reward distribution of each arm varies with time. We point out that in some real application, we can utilize information on the change of reward distribution. Especially we consider the type of information that may restrict the rounds at which the reward distribution changes. Against such scenario, we propose a novel strategy called PM policy. The proposed policy is based on existing CUSUM-UCB policy and M-UCB policy that do not consider external information. Though such existing policies monitor all arms to detect the change of reward distribution, our policy monitors only important arms and rounds. As a result, the ratio of unnecessary monitoring is reduced, and an efficient search can be performed. The regret bound of the proposed policy is described. We also show the effectiveness of the proposed method by numerical experiments.

1. はじめに

多腕バンディット問題 (Multi-armed Bandit Problem) とは、与えられた複数の選択肢の中から、良いものを逐次的に探索する戦略について考える問題である。多腕バンディット問題は、オンライン広告配信の最適化や推薦システム [Li 10], モンテカルロ木探索 [Kocsis 06], 診療 [Gittins 89, Villar 15] など幅広い応用があり、古くから現在に至るまで広く研究されている。一般的な確率的多腕バンディット問題の構成要素は、1 人のプレイヤーおよび複数のアームと呼ばれる選択肢である。プレイヤーは、各ラウンドで、過去の経験に基づいてアームの中から 1 つを選ぶ。各アームには報酬分布と呼ばれる確率分布が対応しており、プレイヤーは選んだアームに対応する報酬分布に従って報酬を得る。プレイヤーは、各アームの報酬分布を知らず、各ラウンドで得られる報酬という限られた情報に基づいて動的に良いアームを探す。例えば、広告配信の最適化における定式化の一例では、配信する広告の候補がアームに対応し、選択された広告を 1 人のユーザへ配信することが 1 ラウンドに対応し、広告がクリックされたかどうかを表す 0/1 の値が報酬に対応し、報酬分布はベルヌーイ分布である。

プレイヤーの目標は、継続的に高い報酬を獲得し続けられるような選択をすることである。継続的に高い報酬を得るためには、過去の経験に基づいて報酬が高そうな選択をする活用 (Exploit) と良い選択肢を探すための探索 (Explore) のバランスを取る必要がある。活用と探索のバランスを取る戦略として、UCB (Upper Confidence Bound) 戦略 [Auer 02a, Lai 85] や、トンプソン抽出 [Agrawal 12, Thompson 33] などが知られている。これらの戦略は、リグレットと呼ばれる評価指標の上界が $O(\log T)$ であるという意味で、継続的に高い報酬を獲得し続けられる良い戦略といえる。

上記の良い戦略は、各アームについての報酬分布が時間に依存せず一定であることを前提としている。ところが、実問題では、報酬分布が時間によって変化する場合も多い。例えば、オンライン広告配信の最適化では、ユーザの興味がトレンドによって変化することで、同じ広告であってもクリックされる確率が時間の経過とともに変化することが考えられる。他にも、広告に掲載されている商品の人気の変化や、広告のサイズやデザインの修正など、報酬分布を変化させる要因は多数考えられる。このような問題意識のもと、近年、報酬分布の時間変化を考慮した非定常多腕バンディット問題の研究が盛んに行

われている [Besbes 14, Lattimore 18]. そして、非定常多腕バンディット問題に対する戦略の中でも、変化点検知アルゴリズムを用いて、報酬分布が変化したアームを能動的に探しに行く CUSUM-UCB 戦略 [Liu 18] や M-UCB 戦略 [Cao 19] は、実験的に良い性能であることが知られている。ところが、CUSUM-UCB 戦略や M-UCB 戦略では、報酬分布が変化したかどうかを調べるために定期的にすべてのアームを選択するため、実問題に適用しようとするとは効率的でない場合がある。例えば、実問題では、以下の例に示すように、報酬分布が変化するアームやタイミングについての外部情報が使える場合があり、すべてのアームを定期的に同じ割合で調べるのは無駄である場合が多い。

例 1. (トレンドの影響を受けない広告と受ける広告の混在) オンライン広告配信の最適化において、配信する広告の候補のうち、トレンドの影響を受けず定常とみなせるものと、トレンドの影響を受けるものが混在した状況が考えられる。例えば、あるアームは、長期的に人気のある特定商品を推薦する広告に対応し、別のアームは、その時点で人気の商品を動的に推薦する広告に対応するという状況が考えられる。このような状況を多腕バンディット問題として定式化する場合、報酬分布が変化するアームがトレンドの影響を受ける広告に制限された、非定常多腕バンディット問題となる。

例 2. (一部の広告の修正) オンライン広告配信の最適化において、配信する各広告がトレンドの影響を受けず定常とみなせる場合であっても、いくつかのデザインやサイズなどを途中で修正する状況が考えられる。このような状況を多腕バンディット問題として定式化する場合、修正前の広告と修正後の広告を別のアームとみなす方法と、同じアームとみなす方法が考えられる。別のアームとみなす方法では、修正のタイミングで、修正前のアームについて得た報酬の情報をすべて捨ててしまうために、修正がわずかである場合は、非効率的である。一方、同じアームとみなす方法では、報酬分布が変化するタイミングが広告を修正するタイミングに制限された、非定常多腕バンディット問題となる。

そこで、本論文では、報酬分布が変化するアームやタイミングが制限されるという外部情報を考慮した多腕バンディットの定式化について述べる。そして、そのような外部情報が使える場合において、外部情報を考慮しない既存手法よりも効率が良い戦略を提案する。具体的には、CUSUM-UCB 戦略や M-UCB 戦略ではすべてのアームについて報酬分布の変化を定期的に調べていたのに対し、外部情報を用いて報酬分布の変化が起きる可能性が高いアームのみを調べる PM (Partially Monitored) 戦略を提案する。

本稿の構成は以下の通りである。まず、2 章で一般的な非定常多腕バンディット問題について述べ、3 章で先行研究について述べる。4 章で、報酬分布が変化するア

ームやタイミングが制限できるという外部情報を考慮した定式化をする。5 章で、外部情報を考慮した PM 戦略を提案する。6 章で、提案手法のリグレット上界について述べる。7 章で、提案手法の有効性を検証するための数値実験について述べる。最後に、8 章で本稿の内容をまとめる。

2. 非定常多腕バンディット問題

まず、一般的な確率的多腕バンディットの問題設定を述べる。アームの集合を $\mathcal{K} = \{1, \dots, K\}$ 、ラウンドの集合を $\mathcal{T} = \{1, \dots, T\}$ とする。ラウンド t においてアーム k を選んだ場合に得られる報酬が従う確率分布を $P_{t,k}$ とする。報酬は 0 以上 1 以下の値を取るものとする。プレイヤーは各ラウンド $t \in \mathcal{T}$ において、戦略に基づいてアームの中から $a_t \in \mathcal{K}$ を選ぶ。そして、選んだアームに対応する報酬分布 P_{t,a_t} の実現値である報酬 r_{t,a_t} を得る。プレイヤーは、真の報酬分布 $P_{t,k}$ は知らず、得られた報酬 r_{t,a_t} という情報をもとに次回以降の選択を行う。真の報酬分布 $P_{t,k}$ が t に依存せず k のみに依存する場合は、定常的な多腕バンディット問題となる。本稿では $P_{t,k}$ が t にも依存する非定常多腕バンディット問題を考える。戦略とは、各ラウンド t において、過去の経験に基づいて、 a_t を決定する操作のことである。ただし、過去の経験とは、ラウンド $t-1$ までに選択したアームと得られた報酬の情報 $\{(a_i, r_{i,a_i}) \mid i = 1, \dots, t-1\}$ のことである。多腕バンディット問題における目標は、継続的に高い報酬を獲得し続けられる戦略の構成である。すなわち、戦略の評価指標として、ラウンド T までの累積報酬の期待値 $E[\sum_{t=1}^T r_{t,a_t}]$ が考えられる。実際には、累積報酬の期待値が、理想的な状況と比べてどれだけ低いかを表すリグレットという評価指標を用いることが多い。つまり、戦略 S についてのラウンド T までのリグレット $\mathcal{R}(T)$ を以下で定義する：

$$\mathcal{R}(T) := \sum_{t=1}^T \max_{k \in \mathcal{K}} E[r_{t,k}] - E \left[\sum_{t=1}^T r_{t,a_t} \right]$$

第一項は、最適なアームを選択し続けた場合の報酬の期待値を表し、第二項は、戦略 S を用いた場合の報酬の期待値を表す。すなわち、リグレットが小さいほど、戦略 S を用いた場合が最適なアームを選択し続けた理想的な状況と近く、継続的に高い報酬を獲得し続けることができる良い戦略といえる。なお、リグレットの定義として、

$$\tilde{\mathcal{R}}(T) := \max_{k \in \mathcal{K}} \sum_{t=1}^T E[r_{t,k}] - E \left[\sum_{t=1}^T r_{t,a_t} \right]$$

が用いられることもある。定常な場合は $\mathcal{R}(T) = \tilde{\mathcal{R}}(T)$ である。また、 $\mathcal{R}(T)$ では第一項において最適なアーム $k \in \mathcal{K}$ をラウンドごとに選べるため、 $\mathcal{R}(T) \geq \tilde{\mathcal{R}}(T)$ が成立する。そのため、 $\mathcal{R}(T)$ の方がより厳しい評価指標といえる。

3. 先行研究

定常的な多腕バンディット問題に対する代表的な戦略の1つとして、UCB 戦略が挙げられる。UCB 戦略では、各アームを1回ずつプレイしたあと、各ラウンド t において、 $\hat{\mu}_i + a\sqrt{\log t/n_i}$ を最大にするようなアーム i を選択する。ここで、 $\hat{\mu}_i$ はアーム i の報酬の経験期待値であり、 $\hat{\mu}_i$ が高いアームを選ぶ戦略は、過去の経験の活用を重視した戦略といえる。一方、 n_i はラウンド t の開始以前において、アーム i を選択した回数である。第二項の $\sqrt{\log t/n_i}$ は経験による推定の自信の無さを表し、この値が大きいアームを選ぶ戦略は、探索を重視した戦略といえる。そして、 a は活用と探索のトレードオフを制御するパラメータであり、例えば $a = \sqrt{2}$ とする [Auer 02a]。UCB 戦略は、リグレット上界が $O(K \log T)$ であるという意味で良い戦略である。

一方、非定常多腕バンディット問題に対する既存手法は、受動的な手法と能動的な手法の2種類に分類できる。受動的な手法とは、報酬分布が変化したアームやタイミングを探さず、単に過去の結果よりも直近の結果を重視することで報酬分布の時間変化に追従する手法である。受動的な手法の代表例としては、D-UCB 戦略、SW-UCB 戦略 [Garivier 11]、Exp3.S 戦略 [Auer 02b] が挙げられる。D-UCB 戦略と SW-UCB 戦略は、いずれも UCB 戦略を非定常な場合に適用できるように改良したものである。具体的には、 $\hat{\mu}_i$ や n_i の代わりに、D-UCB 戦略では直近のラウンドを重視した重みつき平均を用い、SW-UCB 戦略では直近の一定ラウンドのみを考慮した値を用いる。一方、能動的な手法とは、変化点検知アルゴリズムを用いて、報酬分布が変化したアームやタイミングを把握することで、変化に追従する手法である。能動的な手法の代表例としては、CUSUM-UCB 戦略、M-UCB 戦略、Exp3.R 戦略 [Allesiardo 15]、Adapt-EvE 戦略 [Hartland 07] が挙げられる。CUSUM-UCB 戦略や M-UCB 戦略では、定期的にすべてのアームを選び、報酬分布が過去と変化していないかを、変化点検知アルゴリズムを用いて確認する。変化点が検知された場合には、変化前の古い結果を忘却することで時間変化に追従する。Exp3.R 戦略と Adapt-EvE 戦略も類似の手法であるが、Exp3.R 戦略は敵対的多腕バンディット問題に対する Exp3 戦略をもとにしている。また、Adapt-EvE 戦略では最適なアームの報酬分布の悪化のみを考慮する戦略である。一方、[Yu 09] では、報酬をもらうアーム以外のアームについても、コストを払うことで報酬の値を観測できるという問題設定と戦略が提案されている。[Yu 09] の問題設定は、報酬をもらうアームの報酬値以外の情報も利用できるという意味で本論文の問題設定と似ているが、利用できる追加情報が外生的かどうか異なる。さらに、[Fouché 19] では、非定常な問題への UCB 戦略などの拡張が提案されており、アーム選択の計算に用いる直近のラウンド数を

動的に決定するアルゴリズムが紹介されている。

受動的な手法と能動的な手法を比較すると、変化点が検出できる範囲では、能動的な手法の方が優れている場合が多い。そこで、本稿でも、外部情報を考慮した問題設定に対して能動的な手法を提案する。

4. 外部情報を考慮した非定常多腕バンディット問題

本章では、報酬分布の時間変化に関する外部情報を考慮した定式化について述べる。本稿で考える外部情報は、報酬分布が変化するアームやタイミングに関する制限である。そこで、 $t = 1, 2, \dots, T-1$ に対して、ラウンド t から $t+1$ の間での報酬分布の変化の制限を表すベクトル $c_t \in \{0, 1\}^K$ を導入する。すなわち、 $c_{t,k} = 0$ である場合は、定常、つまり $P_{t,k} = P_{t+1,k}$ と予想されていることを表し、 $c_{t,k} = 1$ である場合は、 $P_{t,k} \neq P_{t+1,k}$ となる可能性があることを表す。なお、外部情報 c_t はあくまで外部の状況から推定された情報であり、正しくない場合もある。すなわち、実際の問題では定常だと予想していた部分で報酬分布が変化してしまうことも考えられる。プレイヤーは、ラウンド t において、過去の経験 $\{(a_i, r_{i,a_i}) \mid i = 1, \dots, t-1\}$ に加えて、 t_d 期先までの外部情報 $\{c_i \mid i = 1, \dots, \min(t-1+t_d, T-1)\}$ に基づいてアームを選択する。 t_d は、外部情報がどの程度事前に得られるかを表す値である。例えば、 $t_d = 0$ というのは、ラウンド t においては現時点までの外部情報 $\{c_i \mid i = 1, \dots, t-1\}$ しか使えないことを表し、 $t_d = T$ というのは、ラウンド 1 の時点からすべての外部情報 $\{c_i \mid i = 1, \dots, T-1\}$ を使える状況を表す。外部情報 c_t および t_d を用いることで、1章で述べた例を含む様々な非定常多腕バンディット問題を記述できる。

例 1. (トレンドの影響を受けない広告と受ける広告の混在) 各アームを、トレンドの影響を受けにくいアームの集合 K_S と、受けやすいアームの集合 $K_N = K \setminus K_S$ に分類できる場合を考える。例えば、長期的に人気のある特定の商品を推薦する広告を K_S に分類し、その時点での人気商品を推薦する動的な広告を K_N に分類する。各アーム $k \in K$ に対して、 $k \in K_N$ ならばアーム k の報酬分布が変化しやすいと予想されるため $c_{t,k} = 1$ ($t = 1, \dots, T-1$) とし、 $k \in K_S$ ならば $c_{t,k} = 0$ ($t = 1, \dots, T-1$) とする。また、 K_N と K_S の分類が事前にできる状況では、外部情報はすべてラウンド 1 から使えるため、 $t_d = T$ とする。このように、非定常なアームが制限された状況を記述できる。

例 2. (一部の広告の修正) 広告 k のデザインをラウンド t から $t+1$ の間で修正した場合、 $P_{t,k} \neq P_{t+1,k}$ である確率が高いと見込まれるため、 $c_{t,k} = 1$ とする。また、基本的にはトレンドの影響を受けず、修正のタイミング以外では定常とみなせる場合、修正のタイミング以外では

$c_{t,k} = 0$ とする. また, 広告を修正するタイミングが事前に分かっている状況では外部情報はすべてラウンド 1 から使えるため, $t_d = T$ とする. 一方, 修正のタイミングが直前まで分からない状況では $t_d = 0$ とする. このように, 非定常なラウンドが制限された状況を記述できる.

例 3. (一般的な非定常多腕バンディット問題) 報酬分布が変化するアームおよびラウンドが制限されない一般的な非定常多腕バンディット問題は, すべての要素について, $c_{t,k} = 1$ とすることで表現できる.

なお, 最初は外部情報を持っていなかったが, 途中で報酬分布が変化するアームやタイミングの制限を発見した状況も表すことができる. 例えば, ラウンド t_0 までは外部情報を持っていなかったが, t_0 までの結果を見て特定のアーム集合 K_S が定常であると予想される状況は, $t \geq t_0$ かつ $k \in K_S$ の場合に $c_{t,k} = 0$ とし, そうでない場合に $c_{t,k} = 1$ とすることで表せる.

5. 提案手法

前章で述べた, 外部情報を考慮した非定常多腕バンディット問題に対する戦略を提案する. 提案する PM 戦略は, 外部情報を考慮していない非定常多腕バンディット問題に対する既存手法である CUSUM-UCB 戦略および M-UCB 戦略に基づいた戦略である. まず, CUSUM-UCB 戦略および M-UCB 戦略の特徴を以下に示す:

- (1) 基本的には, 通常の UCB 戦略でアームを選択する.
- (2) ただし, 報酬分布の変化を調べるために, 定期的にすべてのアームを選択する.
- (3) 報酬を得た後, 選択したアームについて報酬分布が過去と変化していないかを変化点検知アルゴリズムを用いて調べる. 変化点が検知されたらそのラウンド以前の経験を捨てる.

CUSUM-UCB 戦略と M-UCB 戦略の違いは, (3) で用いる変化点検知アルゴリズムの違いである. CUSUM-UCB 戦略および M-UCB 戦略は, 一般的な区分定常多腕バンディット問題において, 理論的にも実験的にも良い戦略であることが知られている. ところが, 報酬分布が変化するアームやタイミングが制限された状況では, (2) においてすべてのアームを調べるのは効率的ではない. 具体的には, 定常である状態が続いていると見込まれるアームについては, (2) をスキップすることで, 無駄な調査を減らせると期待できる. 以上の観察をもとに, PM 戦略を提案する. PM 戦略の概要を Algorithm 1 に示す. Algorithm 1 における 2 行目が CUSUM-UCB 戦略および M-UCB 戦略の (1) に対応し, 3 行目から 7 行目が (2) に対応し, 8 行目から 11 行目が (3) に対応する.

ここで, q は通常の戦略 (1) と定期的な調査 (2) の比率を制御するパラメータである. q は K より大きい整数であり, 3 行目において, 各アーム k_0 は q ラウンドに 1 回の割合で定期的な調査 (2) の対象となる. 定常戦略 S_0 と

Algorithm 1 PM 戦略

Require: パラメータ $q(> K)$, 定常戦略 S_0 , 変化点検知関数 D

Initialization : $\tau = 0$, $C = \{c_1, \dots, c_{t_d-1}\}$

```

1: for  $t = 1, 2, \dots, T$  do
2:    $\tau$  より後の経験に基づいて, 定常戦略  $S_0$  を用いて
     アーム  $a_t$  を仮選択
3:   if  $t + k_0$  が  $q$  の倍数となる  $k_0 \in K$  が存在 then
4:     if  $\text{IsCheckNecessary}(C, k_0, t)$  then
5:       選択するアーム  $a_t$  を  $k_0$  に更新する
6:     end if
7:   end if
8:   アーム  $a_t$  をプレイし, 報酬  $r_t$  を得る
9:   if  $D$  を  $a_t$  に適用した結果, 変化点が検知された
     then
10:     $\tau = t$ 
11:   end if
12:    $C$  に新たに使える外部情報  $c_{t+t_d}$  を追加
13: end for

```

は, 定常的な多腕バンディット問題に対する戦略であり, 例えば 2 章で述べた UCB 戦略を用いることができる. 変化点検知関数 D とは, 任意の長さのスカラ値の系列 $r = (r_{t_0}, \dots, r_{t_{n-1}})$ を入力とし, その系列に変化点が存在するか否かを予測して出力する 2 値関数である. また, τ は直近で変化点が検知されたラウンドを表す. 9 行目において D を a_t に適用するとは, ラウンド τ 以降においてアーム a_t を選択して得た報酬の系列 r を入力として $D(r)$ を計算することを表す. すなわち, アーム a_t の報酬分布が直近変化したかどうかを判定する. また, C は現時点で参照できる外部情報を表す. PM 戦略と M-UCB 戦略の違いは, 外部情報 C を用いる点および S_0 と D を一般化した点である. すなわち, PM 戦略において, 外部情報を利用する 4, 6, 12 行目を除いて, S_0 として UCB 戦略を用い, D として, [Cao 19] の Algorithm 1 によって定義される $\tilde{D}_{b,w}((r_{t_0}, \dots, r_{t_{n-1}}))$ を用いると, M-UCB 戦略と一致する. ここで, $\tilde{D}_{b,w}((r_{t_0}, \dots, r_{t_{n-1}}))$ とは, $n \geq w$ かつ $\left| \sum_{i=1}^{w/2} r_{t_{n-i}} - \sum_{i=w/2+1}^w r_{t_{n-i}} \right| > b$ のときに True を返し, そうでないなら False を返す関数である. b と w は変化点検知関数のパラメータであり, b は正の実数, w は正の偶数である. また, 2 行目における, τ より後の経験とは, $\{(a_i, r_{i,a_i}) \mid i = \tau + 1, \dots, t - 1\}$ を表す. 4 行目では, 定期的な調査の対象の候補となったアーム k_0 について, 調査が必要かどうかを外部情報 C を用いて判定する. すなわち, アーム k_0 がラウンド t の周辺で報酬分布が変化するかもしれないと予想される場合には True を返す関数 IsCheckNecessary を用いる. 例えば, IsCheckNecessary としては, 以下で定義される $\text{ICN}_l(C, k_0, t)$ を用いる. すなわち, $c_{t',k_0} = 1$ となる t' で, $c_{t'} \in C$ かつ $t - l \leq t' <$

$t+l$ を満たすものが存在する場合は、ラウンド t の周辺で変化が見込まれるため、 $\text{ICN}_l(C, k_0, t) = \text{True}$ とする。そうでない場合はラウンド t の周辺で定常であると見込まれるため、 $\text{ICN}_l(C, k_0, t) = \text{False}$ とする。なお、4章の例3で述べたような、報酬分布が変化するアームおよびラウンドが制限されない一般的な非定常多腕バンディット問題においては、 IsCheckNecessary は常に True になるため、すべてのアームが常に定期的な調査の対象となる。

6. リグレット上界

本章では、PM 戦略のリグレット上界について述べる。

[定理 1] $q = q_0$ とし、 S_0 として $a = \sqrt{2}$ である UCB 戦略を用い、 $D = \tilde{D}_{b_0, w_0}$ とし、 IsCheckNecessary として $\text{ICN}_{\frac{w_0 q_0}{2}}(c, k_0, t)$ を用いた PM 戦略のリグレットの期待値を $\mathcal{R}(T)$ とする。後述する仮定 1 および仮定 2 が成立するとき、

$$\mathcal{R}(T) \leq O(MK \log T + \sqrt{(1-p)KMT \log KT})$$

が成立する。

ここで、 q_0, w_0, b_0 は、[Cao 19] の 5 章で提案されているパラメータである。詳細は付録 A で述べる。また、 p は、Algorithm 1 の 4 行目において、外部情報のおかげで定期的な調査をスキップできる割合を表す。つまり、 IsCheckNecessary が False であった回数を IsCheckNecessary が呼ばれる回数で割った値を表す。例えば、4 章で述べた例 1 のように、報酬分布が変化するアームの集合が、 K_N に制限される場合、 $p = 1 - \frac{|K_N|}{|K|}$ となる。報酬分布が変化するアーム数 $|K_N|$ が小さいほど、定期的な調査をスキップできる割合 p が大きくなる。また、4 章の例 3 で述べたような、報酬分布が変化するアームやタイミングが制限されない一般的な非定常多腕バンディット問題においては、 $p = 0$ となる。そして、 M は報酬分布が変化せず定常である区間の個数を表す。つまり、ある $k \in K$ が存在して $P_{t,k} \neq P_{t+1,k}$ となる t の個数に 1 を加えた値を表す。 M が大きいほど、実際に報酬分布が変化する回数が多く、より難しい問題といえる。定理 1 のリグレット上界において $p = 0$ とすると、M-UCB 戦略のリグレット上界である $O(MK \log T + \sqrt{KMT \log KT})$ と一致する。また、 p が大きいほど PM 戦略が有効であるといえる。

次に、定理 1 の仮定 1 および仮定 2 について述べる。

仮定 1. $c_{t,k} = 0$ ならば、 $P_{t,k} = P_{t+1,k}$ 。

仮定 1 は、外部情報によって報酬分布が定常であると予想されたアームおよびタイミングでは、実際に報酬分布が変化しないことを表している。なお、PM 戦略は、外部情報 C を利用した戦略であり、PM 戦略の性能は C の精度に依存する。外部情報 C が正しくない場合、仮定 1 が崩れたり p が小さくなったりする。すなわち、報酬分布が実際に変化するタイミングで定常であると予測して

しまった場合、仮定 1 が崩れる。逆に、報酬分布が実際には変化しないタイミングで変化するかもしれないと予測してしまった場合、定期的な調査をスキップできる割合 p が小さくなるため、リグレット上界が大きくなる。

仮定 2. パラメータ w_0 と q_0 について、以下の条件を満たす。(a) 各定常区間の幅が、 $w_0 q_0$ より大きい。(b) $\frac{q_0 w_0}{2} \leq t_d$ 。

仮定 2 の (a) は [Cao 19] の Assumption 1 と全く同じである。(a) は、報酬分布の変化の頻度がさほど高くないことを表す。また、(b) は外部情報がある程度早くから分かっていることを表す。例えば、4 章で述べた例 1 では、すべての外部情報がラウンド 1 から使えるため仮定 2(b) は満たされる。

仮定 1 や仮定 2 は、あくまで定理 1 が成立するために必要な仮定に過ぎず、これらが成立していなくても PM 戦略は実行可能である。実際、これらの仮定が成立していない状況でも、外部情報が実態から大きく外れていなければ、PM 戦略が他の戦略よりも効率が良いことを 7 章の数値実験で確認する。定理 1 の証明は、M-UCB 戦略のリグレット上界の導出 [Cao 19] と同様である。証明の概要は付録 B で述べる。

7. 数 値 実 験

本章では、提案手法の有効性を検証するために行った数値実験について述べる。具体的には、複数の非定常多腕バンディット問題に対して提案手法と既存手法を適用し、リグレットを比較する。提案手法としては、2 種類の PM 戦略を用いる。1 種類目は、変化点検知関数 D として、CUSUM-UCB 戦略と同じものを用いた P-CUSUM-UCB 戦略である。2 種類目は、 D として、M-UCB 戦略と同じものを用いた P-M-UCB 戦略である。定常戦略 S_0 は、いずれも UCB 戦略を用いた。また、既存手法としては、D-UCB 戦略、SW-UCB 戦略、Exp3.S 戦略、Exp3.R 戦略、CUSUM-UCB 戦略、M-UCB 戦略の 6 種類を用いた。そして、以下の 2 種類の状況について、実験を行った。

- 実験 1: 4 章の例 1 のように、報酬分布が変化するアームを限定できる状況。定常なアームと非定常なアームが混在する状況。詳細は 7.1 節で述べる。

- 実験 2: 4 章の例 2 のように、報酬分布が変化するタイミングを限定できる状況。詳細は 7.2 節で述べる。

そして、2 種類の実験それぞれについて、正しい外部情報を用いた場合（実験 1.A, 実験 2.A）と一部が誤った外部情報を用いた場合（実験 1.B, 実験 2.B）を検証した。いずれの場合も報酬分布はベルヌーイ分布とし、ラウンド数は $T = 5000$ とした。既存手法の実装およびパラメータ設定は、SMPyBandits[Besson 18] を用いた。実験を公平にするため、M-UCB 戦略と P-M-UCB 戦略、CUSUM-UCB 戦略と P-CUSUM-UCB 戦略で共通のパラメータは同じ値とした。具体的なパラメータ値を表 1 に示す。表

表 1 各手法のパラメータ値. ただし, γ_D は D-UCB 戦略における割引率 γ , τ_{SW} は SW-UCB 戦略における窓幅 τ , γ_{3S} は Exp3.S 戦略における探索割合 γ , γ_{3R} は Exp3.R 戦略における探索割合 γ , ε_{CU} は CUSUM-UCB 戦略および P-CUSUM-UCB 戦略において変化点検出で用いるパラメータ ε , そして w_M, b_M, γ_M は M-UCB 戦略および P-M-UCB 戦略において変化点検出で用いるパラメータ w, b と探索割合 γ を表す.

	実験 1.A	実験 1.B	実験 2.A	実験 2.B
γ_D	0.95			
τ_{SW}	206		58~184	75
γ_{3S}	0.14~0.20	0.18	0.18~0.54	0.42
γ_{3R}	0.10~0.17	0.14	0.12	
ε_{CU}	0.1			
w_M	200~400	300	250	
b_M	43~63	54	49	
γ_M	0.23~0.39	0.31	0.27	

1 において値に幅があるマスが存在するのは, アーム数 K や報酬分布が変化する回数 $M-1$ に応じてパラメータを決定しているためである. 特に, SW-UCB 戦略の窓幅 τ_{SW} と Exp3.S 戦略の探索割合 γ_{3S} は, 実問題では事前に分からない M を用いて決定しており, 他の手法よりも若干有利である. そして, 各設定および各手法についてそれぞれ 50 回実験を行い, リグレットの平均の推移を比較した. 各実験の設定と結果の詳細を以下に示す.

7.1 定常なアームと非定常なアームが混在する状況

実験 1.A (外部情報が正しい場合) 問題設定は, SMPy-Bandits[Besson18] の PROBLEM2 を参考に作成した. すなわち, 報酬分布が変化するアームの数は 3 とし, 各アームの報酬分布は途中で 4 回変化するものとした. また, 非定常な 3 つのアームに加え, 定常なアームを 1, 3, 5 個追加して 3 通り実験を行った. 定常なアームの割合は, それぞれ $p = 1/4, 1/2, 5/8$ となる. 定常なアームを 3 個追加した場合の真の報酬分布の期待値の推移を図 1 の左上図に示す.

実験の結果, すなわちリグレットの平均の推移を図 2 に示す. 外部情報を考慮した提案手法である P-CUSUM-UCB 戦略および P-M-UCB 戦略は, 外部情報を考慮していない既存手法である CUSUM-UCB 戦略および M-UCB 戦略よりもそれぞれリグレットが小さく, 良い戦略といえる. 特に, P-M-UCB 戦略は M-UCB 戦略よりも大幅に改善している. また, P-CUSUM-UCB 戦略は 6 種類のどの既存手法よりもリグレットが小さい. そして, 定常なアームの割合 p が大きいほど提案手法の効果が高い. これは, p が大きいほど, 余分な探索をスキップできる割合が大きくなるためと考えられる.

実験 1.B (外部情報の一部が正しくない場合) 本実験は, 難易度の異なる 2 種類の設定で行った. まず, 難易度の低い 1 種類目について述べる. 1 種類目の設定は, 最良のアームが切り替わる際に, 切り替わり前に最良であった

アームの報酬分布が低下するため, 切り替わりを検知しやすいという意味で難易度が低い. 具体的には, 各アームの報酬分布として, 実験 1.A における $p = 1/2$ の場合と同じもの, すなわち図 1 の左上図に示すものを用いた. すなわち, 定常なアーム $K_S = \{0, 1, 2\}$ と非定常なアーム $K_N = \{3, 4, 5\}$ が 3 本ずつの状況である. また, 一部が正しくない外部情報の与え方として, 以下のような計 6 通りの条件で実験を行った.

- 実際に定常であるアーム 3 本のうち, 定常であるという外部情報を一部のみ与えた場合 3 通り. つまり, $c_{t,k} = 0 \iff k \in K_E$ (ただし, K_E として $\emptyset, \{0\}, \{0, 1\}$) の 3 通り.
- 実際に非定常であるアーム 3 本のうち, 誤って定常であるという情報を追加で与えた場合 3 通り. つまり, $c_{t,k} = 0 \iff k \in K_E$ (ただし, K_E として $\{0, 1, 2, 3\}, \{0, 1, 2, 3, 4\}, \{0, 1, 2, 3, 4, 5\}$) の 3 通り.

リグレットの平均の推移を図 3 に示す. 図 3 に示すように, 外部情報の一部が正しくない場合においても, 外部情報を考慮した提案手法はリグレットが小さく効果的な場合がある. $|K_E|$ が小さく定常であるアームが把握できていない状況では, 提案手法の効果は小さい. 一方, $|K_E|$ が大きく, 定常でないアームを定常であると予想してしまった状況でも, 提案手法の効果は大きい. すなわち, 少なくとも本実験の設定では, 調査を大幅に減らすことによるリグレットの改善効果が, 調査を減らしたことによる変化検出の遅れによる悪影響よりも強いと考えられる.

次に, 難易度の高い 2 種類目について述べる. 2 種類目の設定は, 最良のアームが切り替わる際に, 切り替わり前に最良であったアームの報酬分布は変化しないため, 切り替わりを検知しにくいという意味で難易度が高い. 具体的には, 各アームの報酬分布として, 図 1 の右上図に示すものを用いた. また, 一部が正しくない外部情報の与え方として, 以下のような 2 通りの条件で実験を行った.

- $c_{t,k} = 0 \iff k \in \{0, 1, 2, 3, 4, 5\}$ という, すべてのアームが定常であるという外部情報を与えた場合.
- $c_{t,k} = 0 \iff k \in \{3, 4, 5\}$ という, 完全に間違えた外部情報を与えた場合.

いずれの場合も, 途中で最良に変化するアーム $\{3, 4, 5\}$ の調査はスキップされるため, 提案手法にとって難易度の高い設定である. リグレットの平均の推移を図 4 に示す.

図 4 において, いずれの場合も提案手法のリグレットは大きい. 特に, 定常なアームと非定常なアームを完全に間違えた外部情報を与えた場合, 提案手法は, M-UCB 戦略や CUSUM-UCB 戦略よりもリグレットが大きい. これは, 提案手法において, 最良アームに変化したアームの調査をスキップしたため, 変化の検出が遅れたためと考えられる. しかし, このように難易度が高い問題で外部情報が大きく誤っている場合でも, 提案手法と, M-UCB 戦略や CUSUM-UCB 戦略の差はさほど大きくはない. また, 2 種類目の設定においては, SW-UCB 戦略および

図1 問題設定. (左上) 実験 1.A における $p=1/2$ の場合の報酬分布の期待値, (右上) 実験 1.B の難易度の高い設定における報酬分布の期待値, (下) 実験 2.A における $n=5$ の場合の報酬分布の期待値.

図2 実験 1.A におけるリグレットの推移. (左上) 定常なアーム数が 1 の場合, (右上) 定常なアーム数が 3 の場合, (下) 定常なアーム数が 5 の場合.

図3 実験 1.B の難易度が低い設定におけるリグレットの推移. (左上) $\mathcal{K}_E = \{0, 1\}$ の場合, (左中央) $\mathcal{K}_E = \{0\}$ の場合, (左下) $\mathcal{K}_E = \emptyset$ の場合, (右上) $\mathcal{K}_E = \{0, 1, 2, 3\}$ の場合, (右中央) $\mathcal{K}_E = \{0, 1, 2, 3, 4\}$ の場合, (右下) $\mathcal{K}_E = \{0, 1, 2, 3, 4, 5\}$ の場合.

図4 実験 1.B の難易度が高い設定におけるリグレットの推移. (左) すべてのアームが定常であるという外部情報を与えた場合, (右) 定常なアームと非定常なアームを完全に間違えた外部情報を与えた場合.

Exp3.S 戦略という受動的な戦略の性能が良かった。

7.2 報酬分布が変化するタイミングを限定できる状況

実験 2.A (外部情報が正しい場合) 本実験において、アームの数は $K = 5$ とした。また、報酬分布が変化する合計数 n を固定し、ランダムに n 箇所の変化点 $(k, t) \in K \times T$ を発生させ、アーム k においてラウンド t から $t+1$ の間で報酬分布の期待値を -0.2 以上 0.2 以下の範囲で変化させた。そして、 n を 5 から 50 の間で変化させて実験した。 $n = 5$ の場合の各アームの報酬分布の期待値の推移を図 1 の下図に示す。

リグレットの平均の推移を図 5 に示す。 $n = 5, 20$ のように、 n が小さい場合には、実験 1 と同様に、提案手法の方が既存手法よりもリグレットが小さく良い手法であるといえる。一方、 $n = 50$ のように n が大きい場合には、提案手法が外部情報を考慮しない既存手法よりもわずかに悪い場合があった。すなわち、外部情報が正しい場合であっても、報酬分布が変化する回数が多いほど提案手法の効果は小さいと考えられる。これは、変化点が多い場合には調査をあまり減らすことができず、結果的に外部情報が無い場合とほとんど同じ戦略になってしまうためと考えられる。

実験 2.B (外部情報の一部が正しくない場合) 本実験において、各アームの報酬分布は実験 2.A と同様のものを用いた。ただし、報酬分布が変化する合計数は $n = 30$ とした。また、一部が正しくない外部情報の与え方として、以下のような 2 通りの実験を行った。

- 実際は定常であるにもかかわらず、定常でないかもしれないという外部情報を与えた場合。具体的には、定常である部分で正しくない余分な外部情報を 30 箇所与えた場合。
- 実際に報酬分布が変化する 30 箇所のうち、定常でないかもしれないという正しい外部情報を 15 箇所のみ与えた場合。

リグレットの平均の推移を図 6 に示す。いずれの場合も、外部情報を考慮した提案手法はリグレットが小さく良い手法であるといえる。特に、外部情報について、余分に調査をさせる誤りよりも、必要な調査をさせない誤りがある場合の方が提案手法の効果が大きい。すなわち、実験 1.B と同様に、調査を大幅に減らすことによるリグレットの改善効果が、調査を減らしたことによる変化検出の遅れによる悪影響よりも強いと考えられる。

以上すべての実験をもとに、提案手法が有効な条件について考察する。まず、提案手法は、外部情報の一部が誤っている場合も含め、多くの場合に既存手法よりも良い結果であった。特に、外部情報が正しく、かつ調査をスキップできる割合が大きい場合、提案手法は有効である。一方、外部情報が誤っている場合、提案手法の効果は小さい。特に、誤りにより余分に調査が必要になる場合や、最良アームに変化したアームの調査をスキップし

てしまう場合は、提案手法の効果が小さい。また、外部情報が正しい場合でも、報酬分布が変化する回数が多いために調査をスキップできる割合が小さい場合、提案手法の効果は小さい。実際、外部情報の誤りや、報酬分布が変化する回数の多さにより、調査をスキップできる割合が小さい場合、提案手法は既存手法とほぼ同じ戦略になってしまうため効果が小さいと考えられる。

8. おわりに

本論文では、各アームの報酬分布が時間変化する非定常多腕バンディット問題を扱った。実問題においては、報酬分布の時間変化に関する外部情報が使える場合があることを指摘した。特に、報酬分布が変化するアームやタイミングが制限されるという種類の外部情報を考慮した定式化を行った。そして、定式化した問題に対して、外部情報を考慮していない既存手法である CUSUM-UCB 戦略および M-UCB 戦略に基づいた PM 戦略を提案した。PM 戦略は、定常である状態が続いていると見込まれるアームについては定期的な調査をスキップするため、効率的に探索をすることができる。また、提案手法のリグレット上界について述べ、数値実験で提案手法の有効性を検証した。今後の課題として、他の種類の外部情報の活用が挙げられる。例えば、各アームやタイミングにおける報酬分布が変化する確率や変化量の大小関係といった外部情報を活用することで、より効率的に探索できる問題もあると考えられる。

◇ 参考文献 ◇

- [Agrawal 12] Agrawal, S. and Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem, in *Conference on Learning Theory*, pp. 39.1–39.26 (2012)
- [Allesiardo 15] Allesiardo, R. and Féraud, R.: Exp3 with drift detection for the switching bandit problem, in *IEEE International Conference on Data Science and Advanced Analytics*, pp. 1–7 (2015)
- [Auer 02a] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time analysis of the multiarmed bandit problem, *Machine Learning*, Vol. 47, No. 2, pp. 235–256 (2002)
- [Auer 02b] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E.: The nonstochastic multiarmed bandit problem, *SIAM Journal on Computing*, Vol. 32, No. 1, pp. 48–77 (2002)
- [Besbes 14] Besbes, O., Gur, Y., and Zeevi, A.: Stochastic multi-armed-bandit problem with non-stationary rewards, in *Advances in Neural Information Processing Systems*, pp. 199–207 (2014)
- [Besson 18] Besson, L.: SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python, Code at <https://GitHub.com/SMPyBandits/SMPyBandits/>, documentation at <https://SMPyBandits.GitHub.io/> (2018)
- [Cao 19] Cao, Y., Wen, Z., Kveton, B., and Xie, Y.: Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit, in *Proceedings of Machine Learning Research*, pp. 418–427 (2019)
- [Fouché 19] Fouché, E., Komiyama, J., and Böhm, K.: Scaling multi-armed bandit algorithms, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1449–1459 (2019)
- [Garivier 11] Garivier, A. and Moulines, E.: On upper-confidence

図5 実験 2.A におけるリグレットの推移. (左上) $n = 5$ の場合, (右上) $n = 20$ の場合, (下) $n = 50$ の場合.

図6 実験 2.B におけるリグレットの推移. (左) 実際は定常であるにもかかわらず定常でないかもしれないという外部情報を与えた場合, (右) 実際に報酬分布が変化する 30 箇所のうち定常でないかもしれないという外部情報を一部のみ与えた場合.

- bound policies for switching bandit problems, in *International Conference on Algorithmic Learning Theory*, pp. 174–188, Springer (2011)
- [Gittins 89] Gittins, J.: *Multi-armed Bandit Allocation Indices*, John Wiley and Sons (1989)
- [Hartland 07] Hartland, C., Baskiotis, N., Gelly, S., Sebag, M., and Teytaud, O.: Change point detection and meta-bandits for online learning in dynamic environments, in *CAp 2007: 9^e Conférence francophone sur l'apprentissage automatique*, pp. 237–250 (2007)
- [Kocsis 06] Kocsis, L. and Szepesvári, C.: Bandit based monte-carlo planning, in *European Conference on Machine Learning*, pp. 282–293, Springer (2006)
- [Lai 85] Lai, T. L. and Robbins, H.: Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, Vol. 6, No. 2, pp. 4–22 (1985)
- [Lattimore 18] Lattimore, T. and Szepesvári, C.: *Bandit Algorithms*, Cambridge University Press (2018)
- [Li 10] Li, L., Chu, W., Langford, J., and Schapire, R. E.: A contextual-bandit approach to personalized news article recommendation, in *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670 (2010)
- [Liu 18] Liu, F., Lee, J., and Shroff, N.: A change-detection based framework for piecewise-stationary multi-armed bandit problem, in *32nd AAAI Conference on Artificial Intelligence* (2018)
- [Thompson 33] Thompson, W. R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, Vol. 25, pp. 285–294 (1933)
- [Villar 15] Villar, S. S., Bowden, J., and Wason, J.: Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges, *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, Vol. 30, No. 2, pp. 199–215 (2015)
- [Yu 09] Yu, J. Y. and Mannor, S.: Piecewise-stationary bandit problems with side observations, in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1177–1184 (2009)

〔担当委員：榊 剛史〕

2020 年 8 月 26 日 受理

◇ 付 録 ◇

A. PM 戦略における最適パラメータ

6 章で述べた q_0, w_0, b_0 は, [Cao 19] の 5 章で提案されているパラメータである。具体的には, w_0 は

$$w_0 \geq \left(\frac{4}{\delta^2} \right) \{ (\log(2KT^2))^{\frac{1}{2}} + (\log(2T))^{\frac{1}{2}} \}^2,$$

を満たす最小の偶数であり,

$$b_0 = \left\{ \frac{1}{2} w_0 \log(2KT^2) \right\}^{\frac{1}{2}},$$

$$q_0 = \left\lceil \frac{K}{\gamma_0} \right\rceil,$$

とする。ただし,

$$\delta = \min_i \max_{k \in K} |\mu_k^{i+1} - \mu_k^i|,$$

$$\gamma_0 = \sqrt{\frac{1}{2T} (M-1) K \cdot \min \left(\frac{w_0}{2}, \left\lceil \frac{b_0}{\delta} \right\rceil + 3\sqrt{w_0} \right)}$$

である。

B. 定理 1 の証明の概要

定理 1 の証明は, [Cao 19] の Theorem 1 の証明と同様である。ただし, Lemma 1 の証明でスキップの割合 p を考慮した式変形が必要になる。具体的には, 以下の手順で証明できる。

- (1) 実際は定常であるような状況に P-M-UCB 戦略を適用したときのリグレット上界について述べた補題 1 を証明する。ここで, 補題 1 とは [Cao 19] の Lemma 1 において γT の項を $\gamma T(1-p) = (K/q)T(1-p)$ に変更したものである。補題 1 の証明は [Cao 19] の Lemma 1 の証明と同様である。ただし, 本論文の設定では, アーム k において定期的な調査がスキップされる場合があるため, 最適なアームが k^* であるのに誤ってアーム k を選んでしまう回数についての不等式が異なる。すなわち, [Cao 19] の (11) 式の右辺においてアーム k が実際に調査された回数として, $\lceil T\gamma/K \rceil$ の代わりに, 外部情報を考慮した上で本当に調査された回数 n_k を用いる必要がある。より正確には, n_k とは, アーム k が定期的な調査の対象となり, しかも IsCheckNecessary=True となる回数である。この不等式を k について足し上げることで, 補題 1 として得られる不等式の第二項は $\sum_k \Delta_k n_k \leq \sum_k n_k \leq (K/q)T(1-p)$ となる。
- (2) 実際は定常であるような状況に P-M-UCB 戦略を適用したときに, 変化点検知アルゴリズムが誤って発報する確率について述べた補題 2 を証明する。ここで, 補題 2 とは [Cao 19] の Lemma 2 と同じものであり, 証明も同様である。
- (3) 実際に報酬分布が 1 回のみ変化する状況に P-M-UCB 戦略を適用したときに, 変化点検知アルゴリズムが一定時間内に正しく発報できる確率について述べた補題 3, および発報するまでにかかる時間の期待値について述べた補題 4 を証明する。ここで, 補題 3, 4 とは [Cao 19] の Lemma 3, 4 と同じものであり, 証明もほぼ同様である。ただし, [Cao 19] における Lemma 3, 4 の証明では, 実際の変化点の前で定期的な調査が $w_0/2$ 回ずつ行われていることを使っているため, 本論文の設定においても変化点の前で $w_0/2$ 回以上ずつサンプルが得られていることを保証する必要がある。実際, アーム k がラウンド t^* から t^*+1 の間で報酬分布が変化すると仮定すると, 仮定 1 より $c_{t^*,k} = 1$ である。さらに, 仮定 2(b) よりこの外部情報は $t^* - q_0(w_0/2)$ 以前から知っている。よって, ICN の定め方により, t^* の前後 ($w_0/2$) 回ずつは ICN = True となり, 定期的な調査はスキップされない。
- (4) 補題 1 から補題 4 をもとに, P-M-UCB 戦略のリグレット上界について述べた定理 1 を証明する。ただし, (1) で述べたように補題 1 の不等式が一部異なるため, ここで得られる不等式は, [Cao 19] の Theorem 1 において γT の項を $\gamma T(1-p) = (K/q)T(1-p)$ に変更した以下の式になる:

$$\begin{aligned} \mathcal{R}(T) \leq & \sum_{i=1}^M \tilde{C}_i + \frac{K}{q} T(1-p) \\ & + \sum_{i=1}^{M-1} 2q \min \left(\frac{w_0}{2}, \left\lceil \frac{b_0}{\delta} \right\rceil + 3\sqrt{w_0} \right) + 3M \end{aligned}$$

ただし, $\tilde{C}_i$ は定常な状況における UCB 戦略のリグレット上界である。すなわち, $\tilde{C}_i \leq O(K \log T)$ を満たす。よって, 付録 A で述べたパラメータ w_0, b_0, q_0 を用いることで,

$$\mathcal{R}(T) \leq O(MK \log T + \sqrt{(1-p)KMT \log KT})$$

となり, 定理 1 が証明される。

—— 著 者 紹 介 ——

難波 博之(正会員)

2016 年東京大学大学院情報理工学系研究科博士前期課程修了。同年, 株式会社日立製作所に入社。日本オペレーションズ・リサーチ学会会員。