観測可能な閲覧深度を用いたカルーセルUIのためのランキングバンディット

安田　琢真†　　中村　篤祥††

† 北海道大学工学部情報エレクトロニクス学科
†† 北海道大学大学院情報科学研究院
E-mail: †yastar.tkm83@gmail.com, ††atsu@ist.hokudai.ac.jp

あらまし　Web 推薦システムにおいて，カルーセル UI を用いれば，ユーザの閲覧範囲（閲覧深度）はシステムにより直接観測可能である．よって，閲覧深度を直接観測できない設定である従来のランキングバンディット手法（カスケードモデル，位置ベースモデル等）より効率的な学習が可能と考えられる．本研究では，閲覧深度分布の下で，アイテムの期待クリック数を最大化することを目標とする，閲覧深度観測可能な設定のランキングバンディット問題を定式化し，その問題に対する解法アルゴリズムを提案する．また，評価実験を通して，提案手法の有効性を示す．

キーワード　バンディットアルゴリズム, ランキング学習, 推薦システム, 学習理論

1　はじめに

Web 推薦システムでは，ユーザのフィードバック（クリックや滞在など）を用いて提示順序を逐次最適化するオンライン学習が重要である．検索結果，ニュース，EC サイトの推薦など，ユーザに複数アイテムを順位付きで提示する場面は多く，このような設定はランキングバンディットとして定式化されてきた．ランキングバンディットは，探索（推定評価値の信頼度が低いアイテムの提示）と活用（過去の評価値から高い評価が期待できるアイテムの提示）のバランスをとりながら，限られた観測から提示順序を改善する枠組みとして有用であり，推薦・検索の中核要素の一つである．

従来のランキングバンディット研究では，ユーザが上位から順に閲覧する行動をモデル化したカスケードモデルや，位置による閲覧確率を考慮した位置ベースモデル（Position-Based Model; PBM）など，クリックモデルに基づく定式化が広く用いられてきた [9,10]．これらのモデルでは，一般に位置依存の閲覧確率（examination）が潜在変数として扱われ，観測されるのはクリック（あるいは最初のクリック位置）に限られる．その結果，クリックが発生しなかった位置について，(i) ユーザは閲覧したがクリックしなかった負例なのか，(ii) そもそも閲覧されていない未観測なのかが区別しにくい．さらに，PBM のように位置バイアスとアイテム魅力度を分離して推定する必要がある設定では，推定対象が増える分だけ学習効率が低下し得る．このように，従来の枠組みは実用上の重要な状況をカバーする一方で，閲覧状況が追加で観測できる実システムに対して必ずしも最も情報効率のよい学習を与えるとは限らない．

一方，近年の Web サービスではカルーセル UI（横スワイプ型のリスト）やスクロール可能なリスト UI が広く利用されている．図 1 にカルーセル UI の概念図を示す．これらの UI では，ユーザが実際にどこまで表示領域を閲覧したか（最大表示位置，閲覧深度）が，クライアント側のイベントログ等から直接観測可能な場合がある．閲覧深度が観測できれば，学習に利用すべきサンプルを実際に閲覧された位置に限定できる．すなわち，閲覧深度以下の位置ではクリックの有無が観測されるため，クリックしなかったという負例を確定でき，それより深い位置は未閲覧として学習に混ぜない，という分離が可能になる．このとき，潜在閲覧確率（examination）による追加の不確実性を避けられるため，同じ試行回数でもより効率的にランキングを学習できることが期待される．しかし，閲覧深度が観測可能であることを前提としたランキングバンディットの定式化と理論的整理は十分に進んでいない．

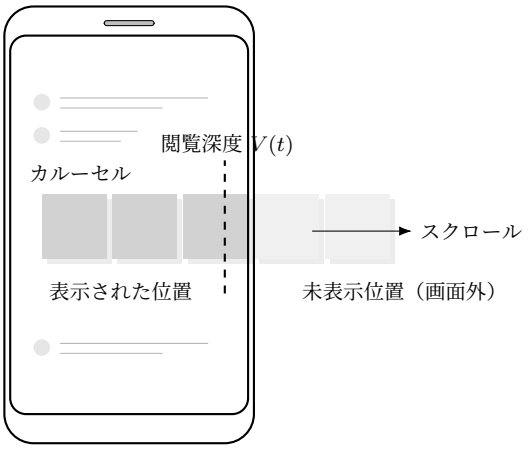

図 1　カルーセル UI の概念図（閲覧深度 $V(t)$ までが観測される）

本研究では，カルーセル UI を想定し，閲覧深度が観測可能な条件下で総クリック数の期待値を最大化するランキングバンディット問題を定式化する．各ラウンドで長さ L のランキングを提示し，ユーザが閲覧した最大位置（閲覧深度）までのクリックのみが観測されるとする．このとき期待報酬は，閲覧深度分布に由来する位置ごとの閲覧確率と，アイテム固有の魅力度（閲覧された場合のクリック率）の積の和として表される．本稿ではまず，閲覧深度観測の効果を明確化するため，クリッ

ク生成をシンプルな形（観測された閲覧深度により閲覧範囲が決まり，閲覧された位置ではアイテムの魅力度に従ってクリックが生起する）で扱う．より複雑な位置効果を同時に含むモデル化は今後の拡張として議論する．

上記の問題に対し，本研究は閲覧深度の観測を直接用いるランキング学習アルゴリズムを提案する．提案法は，そのラウンドで実際に閲覧された位置に置かれたアイテムのみを，露出（観測）されたとみなし，その露出回数に基づいて推定量を更新する．具体的には，露出に基づく信頼区間を用いて探索を行う Observable-Depth UCB と，露出に基づく事後分布からサンプリングする Observable-Depth TS を提示する．さらに，提案 UCB 法についてギャップ依存の期待リグレット上界を導出し，閲覧深度が観測可能であることが学習効率に与える影響を理論的に示す．加えて，シミュレーション実験により，PBM/Cascade 系の既存ベースライン（UCB/TS）と比較して提案法が累積リグレットを改善することを確認する．

本研究の主な貢献は以下の通りである．

- カルーセル UI を想定し，閲覧深度が観測可能なランキングバンディット問題を定式化した．
- 閲覧深度観測に基づき，未観測位置を学習に混ぜないオンライン学習アルゴリズム（Observable-Depth UCB/TS）を提案した．
- 提案 UCB 法についてギャップ依存の期待リグレット上界を与え，深度観測の効果を理論的に整理した．
- シミュレーション実験により，PBM/Cascade 系ベースラインに対する学習効率の改善を示した．

2 関 連 研 究

本節では, 本研究と関係の深い研究領域として, (i) バンディットアルゴリズムの基礎，(ii) クリックモデルに基づくランキングバンディット（カスケードモデル，PBM）と理論保証，(iii) カルーセル等の複数リスト UI（スクロール・閲覧深度）に関する行動モデル化，を概観し，本研究の位置づけを明確化する．

2.1 バンディットアルゴリズムの基礎

バンディットアルゴリズムは，不確実な報酬を持つ選択肢（腕）から逐次的に選択し，累積報酬を最大化するオンライン学習の枠組みである [3,11]．代表的手法として，上側信頼区間（Upper Confidence Bound: UCB）に基づく手法 [1] や事後分布からサンプルして探索する Thompson Sampling [14] などがある．本研究はこれらの枠組みをランキング提示（複数アイテム同時提示）に拡張した設定を扱う．

2.2 クリックモデルとランキング学習

検索・推薦におけるクリックログは，アイテムの真の魅力度だけでなく表示位置や UI の影響を強く受けるため，クリック確率を生成モデルとして記述するクリックモデルが研究されてきた．位置バイアス（position bias）の存在とそのモデル化は古くから議論されており [6]，クリックモデルの体系的な整理として [4] がある．これらのモデルに基づきオンラインでランキングを学習する研究として，以下の 2.3-2.5 節で説明するようなランキングバンディットが発展している．

2.3 カスケードモデルに基づくランキングバンディット

カスケードモデルは，ユーザが上位から順に閲覧し，最初に魅力的なアイテムをクリックするとそこで閲覧を停止する，という逐次閲覧を仮定するクリックモデルである．Kveton らはカスケードモデル下でのランキングバンディットを定式化し，部分観測（最初のクリックまで）に基づく学習アルゴリズムとリグレット解析を与えた [9]．本研究の設定はランキング提示という点では同様であるが，カルーセル UI 等ではどこまで閲覧されたか（閲覧深度）がログから観測可能である点が異なる．この追加情報により，未閲覧位置を負例として誤って扱うことを避けつつ，実際に露出した位置に限定した統計量で推定を進められることが期待される．

2.4 位置ベースモデル（PBM）に基づく複数選択バンディット

位置ベースモデル（Position-Based Model: PBM）は，位置 j の閲覧確率（examination）とアイテムの魅力度の積でクリック確率を表す代表的モデルである．Lagrée らは PBM 下での複数スロット提示（multiple-play）を扱うランキングバンディットを研究し [10]，Komiyama らは更に位置バイアスが未知である設定での理論解析とアルゴリズムを与えた [8]．これらの研究では一般に examination は観測できず，クリックのみから位置バイアスと魅力度を分離推定する必要があるため，推定対象が増える分だけ学習が難しくなり得る．一方，本研究はカルーセル UI 等において閲覧深度が観測可能である状況に着目し，examination を潜在として推定するのではなく，露出（観測）された位置を閲覧深度から直接判定して学習に反映する点で立場が異なる．

従来モデルと本研究の違いは表 1 に整理する．

2.5 カルーセル UI と複数リスト提示

近年の推薦システムでは，単一のランキングリストだけでなく，複数のリスト（カルーセル）を並べる UI が用いられる．このような UI では，ユーザ行動が 2 次元的になり，従来の単一リスト向けクリックモデルでは捉えにくい側面が生じる．Rahdari らはランキングリストからカルーセルへの拡張としてカルーセル向けのクリックモデルを提案し [13]，さらにカルーセル UI を想定したシミュレーション評価の枠組みを検討している [12]．カルーセルとバンディットを組み合わせた応用研究としては，Bendada らが音楽ストリーミングのカルーセル最適化を文脈バンディットとして扱い，実データで有効性を示した [2]．これらは主に行動モデル化・評価手法の側面を中心に扱うのに対し，本研究は閲覧深度が観測可能という実システムで成り立ちやすい情報を前提として，オンラインランキング学習（バンディット）を定式化し，アルゴリズムおよび理論保証（リグレット解析）を与える点に特徴がある．

2.6 本研究の位置づけ

以上を踏まえると，本研究はランキングバンディット（カス

観点	カスケードモデル	位置ベースモデル（PBM）	提案モデル（深度観測）
閲覧深度の観測	観測不可（潜在）	観測不可（潜在）	観測可能（$V(t)$）
クリック観測	最初のクリックまで	位置ごとのクリック	表示位置のクリック
未閲覧位置の扱い	未観測と負例が混在	未観測と負例が混在	未表示として分離
推定対象	魅力度のみ	魅力度と位置バイアス	魅力度のみ（深度は観測）
学習の性質	部分観測で学習	位置バイアスによる補正が必要	露出に基づく更新
代表文献	[9]	[10]	本研究

表 1　既存モデルと提案モデルの比較

ケード／ PBM）という確立した枠組みを踏まえつつ [8–10]，カルーセル UI における閲覧深度の観測可能性に着目して，(1) 露出（観測）に基づく更新則を持つ学習アルゴリズムを設計し，(2) 深度観測を組み込んだ理論解析を与えることで，クリックのみを観測とする従来設定と比べて何が効率化されるかを明確化することを目的とする．従来モデルと本研究の違いは表 1 に整理する．

3　問 題 設 定

3.1　ランキング提示

アイテム集合を $[K] := \{1, 2, \ldots, K\}$，提示枠数（カルーセル内のスロット数）を L とする。各ラウンド $t = 1, 2, \ldots, T$ において，学習者は重複のない長さ L のランキング（順序付きリスト）

$$\boldsymbol{a}(t) = (a_1(t), a_2(t), \ldots, a_L(t)) \in \mathcal{A}$$

を提示する。ここで選択可能なランキングの集合は

$$\mathcal{A} = \left\{(a_1, \ldots, a_L) \in [K]^L \mid a_{j_1} \neq a_{j_2}\ (j_1 \neq j_2)\right\}$$

である。

3.2　閲 覧 深 度

ユーザはラウンド t においてカルーセルをスクロールし，最大で位置 $V(t) \in \{1, \ldots, L\}$ まで閲覧する。本研究では，この閲覧深度 $V(t)$ がクライアントログ等により直接観測可能である設定を扱う。

閲覧深度列 $V(1), V(2), \cdots$ は独立同分布に従うものとし，その分布を

$$p_v := \Pr(V(t) = v), \qquad \sum_{v=1}^{L} p_v = 1$$

とおく。また，位置 j が表示（閲覧範囲に含まれる）される累積確率を

$$P_j := \Pr(V(t) \geqq j) = \sum_{v=j}^{L} p_v$$

と定義すると，以下が成立する。

$$1 = P_1 \geqq P_2 \geqq \cdots \geqq P_L > 0$$

3.3　クリック生成と観測フィードバック

ラウンド t でアイテム i は閲覧されたとき，クリックされたら 1，されなかったら 0 の値をとる確率変数を $C_i(t)$ とする．本稿では確率変数 $C_i(t)$ はベルヌーイ分布に従う，つまり

$$C_i(t) \sim \mathrm{Bernoulli}(\theta_i)$$

とする．ここでアイテム i のクリック率 (本質的魅力度)$\theta_i \in (0, 1)$ は未知パラメータとする．ラウンド t においては，閲覧されたアイテム $a_j(t)$ $(j \leqq V(t))$ に対してのみ，$C_{a_j(t)}(t)$ のフィードバックが観測される．

3.4　報酬と目標

ランキング $\boldsymbol{a}$ を提示したときの総クリック数 $r(\boldsymbol{a})$ をそのラウンドの報酬とする。$r(\boldsymbol{a})$ は

$$r(\boldsymbol{a}) = \sum_{j=1}^{L} \mathbb{1}[j \leqq V(t)] C_{a_j}(t)$$

と書ける。このときの期待報酬は

$$\mathbb{E}[r(\boldsymbol{a})] = \sum_{j=1}^{L} \mathbb{E}\left[\mathbb{1}[j \leqq V(t)] C_{a_j}(t)\right] = \sum_{j=1}^{L} P_j\, \theta_{a_j}$$

となる。

目標は，未知の $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ および与えられた閲覧深度分布の下で，累積期待報酬を最大化するランキング方策を設計することである。

3.5　最適ランキングとリグレット

魅力度 θ_i を降順に並べたときの i 番目の値を $\theta_{(i)}$ とする．$\boldsymbol{a}^\star$ を

$$\boldsymbol{a}^\star = ((1), (2), \ldots, (L))$$

とすると

$$\mathbb{E}[r(\boldsymbol{a}^\star)] = \sum_{j=1}^{L} P_j\, \theta_{(j)} = \max_{\boldsymbol{a} \in \mathcal{A}} \mathbb{E}[r(\boldsymbol{a})]$$

を満たす．つまり $\boldsymbol{a}^\star$ は最適ランキングである．2 番目に最適なランキングを $\boldsymbol{a}^{\star\star}$ で表す．つまり

$$\boldsymbol{a}^{\star\star} = \max_{\boldsymbol{a} \in \mathcal{A}, \mathbb{E}[r(\boldsymbol{a})] < \mathbb{E}[r(\boldsymbol{a}^\star)]} \mathbb{E}[r(\boldsymbol{a})]$$

とする．学習者がラウンド t で選択したランキングを $\boldsymbol{a}(t)$ とすると，期待リグレットは

$$\mathrm{Reg}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \Big(r(\boldsymbol{a}^\star) - r(\boldsymbol{a}(t)) \Big)\right]$$

で定義される。

最適ランキング $\boldsymbol{a}^\star$ に対するランキング $\boldsymbol{a}$ のギャップ $\Delta_{\boldsymbol{a}}$ を以下の通り定義する．

$$\Delta_{\boldsymbol{a}} = \mathbb{E}[r(\boldsymbol{a}^\star)] - \mathbb{E}[r(\boldsymbol{a})] = \sum_{j=1}^{L} P_j(\theta_{(j)} - \theta_{a_j})$$

任意のランキング $\boldsymbol{a}$ に対して $0 \leqq r(\boldsymbol{a}) \leqq L$ であるから，$0 \leqq \Delta_{\boldsymbol{a}} \leqq L$ が成り立つ．このギャップを用いて $\mathrm{Reg}(T)$ は以下のように表現できる．

$$\mathrm{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)}\right]$$

3.6 既存モデルとの関係

本設定は PBM における位置依存の閲覧確率（examination）を，潜在変数としてではなく観測変数 $V(t)$ として扱う点が特徴である。その結果，表示された位置に置かれた各アイテムについては Bernoulli(θ_i) の独立サンプルが直接得られ，推定と解析が単純化される。

4 提 案 手 法

本節では，第 3 節で定式化した閲覧深度 $V(t)$ が観測可能なランキングバンディットに対し，閲覧深度の観測を直接利用する学習アルゴリズムを提案する．提案法の基本方針は，そのラウンドで実際に表示された（$j \leqq V(t)$）位置に置かれたアイテムのみを露出（観測）として数え，その露出回数に基づき推定を更新することである．これにより，未表示位置（$j > V(t)$）を負例として誤って扱うことを避けつつ，表示された位置に関しては Bernoulli(θ_i) の独立サンプルとして扱って推定することができる．

4.1 統計量（露出回数とクリック回数）

ラウンド t までに，アイテム i が表示（露出）された回数とクリックされた回数を

$$n_i(t) := \sum_{s=1}^{t}\sum_{j=1}^{L} \mathbb{1}\big[a_j(s) = i,\ j \leqq V(s)\big], \tag{1}$$

$$s_i(t) := \sum_{s=1}^{t}\sum_{j=1}^{L} \mathbb{1}\big[a_j(s) = i,\ j \leqq V(s)\big] C_i(s) \tag{2}$$

と定義する．このとき，表示（露出）された位置に限ればクリックは Bernoulli(θ_i) に従うため，$n_i(t) \geqq 1$ のとき，

$$\hat{\theta}_i(t) := \frac{s_i(t)}{n_i(t)}$$

をアイテム魅力度 θ_i の推定量として用いる．

4.2 Observable-Depth UCB（OD-UCB）

UCB 型の探索を行うため，各ラウンド $t \geqq 1$ において

$$\mathrm{UCB}_i(t) := \begin{cases} \hat{\theta}_i(t-1) + \sqrt{\dfrac{\alpha \log t}{n_i(t-1)}} & (n_i(t-1) \geqq 1) \\ +\infty & (n_i(t-1) = 0) \end{cases}$$

Algorithm 1 Observable-Depth UCB

Require: アイテム数 K，スロット数 L，パラメータ $\alpha > 0$

$n_i \leftarrow 0,\ s_i \leftarrow 0 \quad (i = 1, \ldots, K)$
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, \ldots, K$ **do**
    $\mathrm{UCB}_i \leftarrow \begin{cases} s_i/n_i + \sqrt{\alpha \log t / n_i} & (n_i \geqq 1) \\ +\infty & (n_i = 0) \end{cases}$
  end for
  $\boldsymbol{a}(t) \leftarrow$ UCB スコア上位 L 個を降順に並べたランキング
  ランキング $\boldsymbol{a}(t)$ を提示し，閲覧深度 $V(t)$ とクリック $(C_{a_1(t)}(t), \ldots, C_{a_{V(t)}(t)}(t))$ を観測
  for $j = 1, \ldots, V(t)$ **do**
    $i \leftarrow a_j(t)$
    $n_i \leftarrow n_i + 1$
    $s_i \leftarrow s_i + C_i(t)$
  end for
end for

Algorithm 2 Observable-Depth TS

Require: アイテム数 K，スロット数 L，事前パラメータ $a_0, b_0 > 0$

$n_i \leftarrow 0,\ s_i \leftarrow 0 \quad (i = 1, \ldots, K)$
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, \ldots, K$ **do**
    $\tilde{\theta}_i \sim \mathrm{Beta}(a_0 + s_i,\ b_0 + n_i - s_i)$
  end for
  $\boldsymbol{a}(t) \leftarrow \tilde{\theta}$ の上位 L 個を降順に並べたランキング
  ランキング $\boldsymbol{a}(t)$ を提示し，閲覧深度 $V(t)$ とクリック $(C_{a_1(t)}(t), \ldots, C_{a_{V(t)}(t)})$ を観測
  for $j = 1, \ldots, V(t)$ **do**
    $i \leftarrow a_j(t)$
    $n_i \leftarrow n_i + 1$
    $s_i \leftarrow s_i + C_i(t)$
  end for
end for

をアイテム $i \in [K]$ のスコアとして用いる．ただし，$\alpha > 0$ は探索と活用のバランスをとるパラメータとする．

そして $\mathrm{UCB}_i(t)$ の大きい順に上位 L 個のアイテムを選び，それらをスコア降順に並べたランキング $\boldsymbol{a}(t)$ を提示する．観測後，深度 $V(t)$ までの位置に置かれたアイテムのみ統計量 (n_i, s_i) を更新する．疑似コードを Algorithm 1 に示す．

4.3 Observable-Depth TS（OD-TS）

次に，Thompson Sampling（TS）に基づくアルゴリズムを与える．Algorithm 2 に疑似コードを示す．

各アイテム i に対し，事前分布として $\theta_i \sim \mathrm{Beta}(a_0, b_0)$ を仮定し，露出回数 $n_i(t)$ とクリック回数 $s_i(t)$ に基づいて事後分布

$$\theta_i \mid \mathcal{H}_t \sim \mathrm{Beta}\big(a_0 + s_i(t),\ b_0 + n_i(t) - s_i(t)\big)$$

を得る．各ラウンド t で各アイテムから独立にサンプル $\tilde{\theta}_i(t) \sim \mathrm{Beta}(a_0 + s_i(t-1),\ b_0 + n_i(t-1) - s_i(t-1))$ を生成し，$\tilde{\theta}_i(t)$ の大きい順に上位 L 個を選んで提示する．更新

は UCB と同様に $j \leqq V(t)$ の位置に限って行う．

4.4 計 算 量

各ラウンドで全アイテムのスコアを計算し，上位 L 個を選ぶため，ヒープを用い入れば計算量は $O(K \log L)$ である．また，統計量の更新の計算量は $O(V(t))$ であり，$V(t) \leqq K$ であるから，OD-UCB，OD-TS 共に [1]，1 ラウンドの計算量は $O(K \log L)$ である．

5 リグレット解析

本節では，提案手法 **Observable-Depth UCB**（Algorithm 1）の期待リグレット上界を与える．解析の骨格は古典的な UCB 解析（集中不等式＋劣腕が選ばれる回数の上界）に従う [1, 11]．一方で本設定では，閲覧深度 $V(t)$ が観測されるため，(i) $j \leqq V(t)$ の位置に置かれたアイテムのみが露出（観測）される，(ii) 表示回数（ランキングに含めた回数）と露出回数が一致しないという点が通常の PBM/Cascade の解析と異なる [8–10]．

最適集合（魅力度上位 L 個）を

$$\mathrm{Top}_L := \{(i) \in [K] \mid i = 1, 2, \ldots, L\}$$

と定義する．

推定値 $\hat{\theta}_i(t) := s_i(t)/n_i(t)$ に対し，良い事象

$$\mathcal{E}_t := \bigcap_{i=1}^{K} \left\{ \left|\hat{\theta}_i(t) - \theta_i\right| \leqq \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\}$$

を導入する．事象 $\mathcal{E}_t$ の余事象を $\mathcal{E}_t^c$ で表す．

定理 1. $\alpha > 1/2$ とする．*OD-UCB* に関し，任意の自然数 $1 \leqq d \leqq L$ に対して

$$\begin{aligned}\mathrm{Reg}(T) \leqq\ & 32\alpha \log T \left(\frac{\left(\sum_{k=1}^{L} P_k\right)^2}{d} + \left(\sum_{k=1}^{d} P_k\right)^2 \right) \\ & \times \left(\sum_{i=L+1}^{K} \frac{1}{P_L(\theta_{(L)} - \theta_{(i)})} + \frac{L}{\Delta_{\boldsymbol{a}^{\star\star}}} \right) \\ & + 2LK\,\zeta(2\alpha).\end{aligned}$$

ここで $\zeta(s) := \sum_{t=1}^{\infty} t^{-s}$ はリーマンゼータ関数であり，$s > 1$ で収束する．特に $\alpha > 1/2$ のとき $2\alpha > 1$ より $\sum_{t=1}^{\infty} t^{-2\alpha} = \zeta(2\alpha)$ が有限である．

Proof. 事象 $\mathcal{E}_t$，$\mathcal{E}_t^c$ を用いて $\mathrm{Reg}(T)$ は以下のように表現できる．

$$\mathrm{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t^c]\right]$$

第 1 項を補題 5 で，第 2 項を補題 1 で上から抑えることにより，不等式は導かれる． □

1：OD-TS に関しては，Beta 分布からサンプリングするのにかかる計算量は無視するものとする．

補題 1. $\alpha > 1/2$ のとき，以下の不等式が成り立つ．

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t^c]\right] \leqq 2LK \sum_{t=1}^{\infty} t^{-2\alpha} = 2LK\,\zeta(2\alpha).$$

Proof. $\Delta_{\boldsymbol{a}} \leqq L$ が任意のランキング $\boldsymbol{a}$ に対して成り立つので

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t^c]\right] \leqq L \sum_{t=1}^{T} \Pr\left(\mathcal{E}_t^c\right)$$

が成り立つ．$\Pr\{\mathcal{E}_t^c\}$ は，以下の計算より $2Kt^{-2\alpha}$ で上から抑えられることがわかる．

$$\begin{aligned}\Pr(\mathcal{E}_t^c) =& \Pr\left\{ \bigcup_{i=1}^{K} \left\{ \left|\hat{\theta}_i(t) - \theta_i\right| > \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\} \right\} \\ \leqq& \sum_{i=1}^{K} \Pr\left\{ \left|\hat{\theta}_i(t) - \theta_i\right| > \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\} \\ =& \sum_{i=1}^{K} \Pr\left\{ \left\{ \hat{\theta}_i(t) > \theta_i + \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\} \right. \\ & \qquad \left. \cup \left\{ \hat{\theta}_i(t) < \theta_i - \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\} \right\} \\ \leqq& \sum_{i=1}^{K} \Pr\left\{ \hat{\theta}_i(t) > \theta_i + \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\} \\ & + \sum_{i=1}^{K} \Pr\left\{ \hat{\theta}_i(t) < \theta_i - \sqrt{\frac{\alpha \log t}{n_i(t)}} \right\} \\ \leqq& \sum_{i=1}^{K} e^{-2n_i(t) \cdot \frac{\alpha \log t}{n_i(t)}} + \sum_{i=1}^{K} e^{-2n_i(t) \cdot \frac{\alpha \log t}{n_i(t)}} \\ =& 2Ke^{-2\alpha \log t} = 2Kt^{-2\alpha}\end{aligned}$$

ただし，最後の不等号は Hoeffding 不等式を用いた．よって

$$\sum_{t=1}^{T} \Pr(\mathcal{E}_t^c) \leqq 2K \sum_{t=1}^{\infty} t^{-2\alpha}$$

であり，$\alpha > 1/2$ のとき右辺は収束するので補題が成り立つ． □

事象 $\mathcal{F}_t$ を

$$\mathcal{F}_t = \left\{ 0 < \Delta_{\boldsymbol{a}(t)} \leqq 2 \sum_{j=1}^{L} P_j \sqrt{\frac{\alpha \log t}{n_{a_j(t)}(t-1)}} \right\}$$

と定義すると以下の補題が成り立つ．

補題 2. 以下の包含関係が成り立つ．

$$\mathcal{E}_t \cap \{\Delta_{\boldsymbol{a}(t)} > 0\} \subseteq \mathcal{F}_t$$

Proof. ラウンド t においてランキング $\boldsymbol{a}(t)$ が選ばれているので，

$$\sum_{j=1}^{L} P_j \mathrm{UCB}_{a_j(t)}(t) \geqq \sum_{j=1}^{L} P_j \mathrm{UCB}_{(j)}(t)$$

が成り立っている．このとき事象 $\mathcal{E}_t$ が起こっているとする

と，すべての $j \in [K]$ に対して $\theta_j - \sqrt{\frac{\alpha \log t}{n_j(t-1)}} \leqq \hat{\theta}_j(t-1) \leqq \theta_j + \sqrt{\frac{\alpha \log t}{n_j(t-1)}}$ が成り立っているので，

$$\sum_{j=1}^{L} P_j \left(\theta_{a_j(t)} + 2\sqrt{\frac{\alpha \log t}{n_{a_j(t)}(t-1)}} \right) \geqq \sum_{j=1}^{L} P_j \theta_{(j)}$$

が成り立つ．したがって

$$\Delta_{\boldsymbol{a}(t)} = \sum_{j=1}^{L} P_j \theta_{(j)} - \sum_{j=1}^{L} P_j \theta_{a_j(t)} \leqq 2 \sum_{j=1}^{L} P_j \sqrt{\frac{\alpha \log t}{n_{a_j(t)}(t-1)}}$$

が成り立つ．よって事象 $\mathcal{E}_t$ と事象 $\Delta_{\boldsymbol{a}(t)} > 0$ が同時に起こっていれば事象 $\mathcal{F}_t$ が起こっている． □

ラウンド t の位置集合 S_t を以下のように定義する．

$$S_t = \left\{ j \in [L] \mid n_{a_j(t)}(t) \leqq \frac{16\alpha(\sum_{k=1}^{L} P_k)^2 \log t}{\Delta^2_{\boldsymbol{a}(t)}} \right\}$$

このとき，自然数 $0 < d < L$ に対し，以下の 2 つの事象を考える．

$$\mathcal{G}_t(d) = \{|S_t| \geqq d\}$$

$$\mathcal{H}_t(d) = \left\{ \begin{array}{l} |S_t| < d, \exists j \in [L], \\ \\ n_{a_j(t)}(t) \leqq \dfrac{16\alpha(\sum_{k=1}^{d} P_k)^2 \log t}{\Delta^2_{\boldsymbol{a}(t)}} \end{array} \right\}$$

このとき，以下の補題が成立する．

補題 3. 任意の自然数 $0 < d < L$ に対し，以下の包含関係が成り立つ．

$$\mathcal{F}_t \subseteq \mathcal{G}_t(d) \cup \mathcal{H}_t(d)$$

Proof. 背理法で証明する．$\mathcal{F}_t$ が成り立っているのに $\mathcal{G}_t(d) \cup \mathcal{H}_t(d)$ が成り立っていないとする．$\mathcal{G}_t(d) \cup \mathcal{H}_t(d)$ が成り立っていないので

$$|S_t| < d, \forall j \in [L], n_{a_j(t)}(t) > \frac{16\alpha(\sum_{k=1}^{d} P_k)^2 \log t}{\Delta^2_{\boldsymbol{a}(t)}}$$

が成り立つ．よって

$$\begin{aligned}
\Delta_{\boldsymbol{a}(t)} \leqq & 2 \sum_{j=1}^{L} P_j \sqrt{\frac{\alpha \log t}{n_{a_j(t)}(t-1)}} && (\mathcal{F}_t \text{ 成立より}) \\
= & 2 \sum_{j \in S_t} P_j \sqrt{\frac{\alpha \log t}{n_{a_j(t)}(t-1)}} + 2 \sum_{j \in [L] \setminus S_t} P_j \sqrt{\frac{\alpha \log t}{n_{a_j(t)}(t-1)}} \\
< & 2 \sum_{j \in S_t} P_j \frac{\Delta_{\boldsymbol{a}(t)}}{4 \sum_{k=1}^{d} P_k} + 2 \sum_{j \in [L] \setminus S_t} P_j \frac{\Delta_{\boldsymbol{a}(t)}}{4 \sum_{k=1}^{L} P_k} \\
= & \frac{\Delta_{\boldsymbol{a}(t)}}{2} \left(\frac{\sum_{j \in S_t} P_j}{\sum_{k=1}^{d} P_k} + \frac{\sum_{j \in [L] \setminus S_t} P_j}{\sum_{k=1}^{L} P_k} \right) \\
\leqq & \Delta_{\boldsymbol{a}(t)}
\end{aligned}$$

という矛盾が導かれる．したがって $\mathcal{F}_t \subseteq \mathcal{G}_t(d) \cup \mathcal{H}_t(d)$ が成り立つ． □

アイテム $i \in [K]$ を含むランキングの集合を $\mathcal{A}_i$ で表す．以下の既存の補題を使って，補題 5 を証明する．

補題 4 (Lemma 2 in Appedix B.4 [5])**.** 任意のアイテム $i \in [K]$，正数 $C > 0$ に対し，以下の不等式が成り立つ．

$$\sum_{t=1}^{T} \mathbb{1} \left[i \in \boldsymbol{a}(t), n_i(t) \leqq \frac{C}{\Delta^2_{\boldsymbol{a}(t)}} \right] \Delta_{\boldsymbol{a}(t)} \leqq \frac{2C}{\Delta_{\min,i}}$$

ただし，$\Delta_{\min,i} = \min_{\boldsymbol{a} \in \mathcal{A}_i \setminus \{\boldsymbol{a}^\star\}} \Delta_{\boldsymbol{a}}$ とする．

事象 $\mathcal{G}_t(d)$ と $\mathcal{H}_t(d)$ を以下のようにさらに細かく分ける．

$$\mathcal{G}_{i,t}(d) = \mathcal{G}_t(d) \cap \{i \in S_t\}$$

$$\mathcal{H}_{i,t}(d) = \mathcal{H}_t(d) \cap \left\{ i \in \boldsymbol{a}(t), n_i(t) \leqq \frac{16\alpha(\sum_{k=1}^{d} P_k)^2 \log t}{\Delta^2_{\boldsymbol{a}(t)}} \right\}$$

すると，

$$\sum_{i=1}^{K} \mathbb{1}[\mathcal{G}_{i,t}(d)] = \mathbb{1}[\mathcal{G}_t(d)] \sum_{i=1}^{K} \mathbb{1}[i \in S_t] \geqq d \mathbb{1}[\mathcal{G}_t(d)] \tag{3}$$

$$\mathbb{1}[\mathcal{H}_t(d)] \leqq \sum_{i=1}^{K} \mathbb{1}[\mathcal{H}_{i,t}(d)] \tag{4}$$

が成り立つ．不等式 (3)， (4) と補題 4 を用いて以下の補題 5 が証明できる．

補題 5. 以下の不等式が成り立つ．

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t] \right] \leqq & 32\alpha \log T \left(\frac{(\sum_{k=1}^{L} P_k)^2}{d} + \left(\sum_{k=1}^{d} P_k \right)^2 \right) \\
& \times \left(\sum_{i=L+1}^{K} \frac{1}{P_L(\theta_{(L)} - \theta_{(i)})} + \frac{L}{\Delta_{\boldsymbol{a}^{\star\star}}} \right)
\end{aligned}$$

Proof. 以下の式変形により補題が証明できる．

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t] \right] \\
= & \sum_{t=1}^{T} \mathbb{E}\left[\Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{E}_t, \Delta_{\boldsymbol{a}(t)} > 0] \right] \\
\leqq & \sum_{t=1}^{T} \mathbb{E}\left[\Delta_{\boldsymbol{a}(t)} \mathbb{1}[\mathcal{F}_t] \right] && (\text{補題 2 より}) \\
\leqq & \sum_{t=1}^{T} \mathbb{E}\left[\Delta_{\boldsymbol{a}(t)} \left(\mathbb{1}[\mathcal{G}_t(d)] + \mathbb{1}[\mathcal{H}_t(d)] \right) \right] && (\text{補題 3 より}) \\
\leqq & \sum_{t=1}^{T} \mathbb{E}\left[\Delta_{\boldsymbol{a}(t)} \left(\frac{1}{d} \sum_{i=1}^{K} \mathbb{1}[\mathcal{G}_{i,t}(d)] + \sum_{i=1}^{K} \mathbb{1}[\mathcal{H}_{i,t}(d)] \right) \right] \\
& && ((3), (4) \text{ より}) \\
= & \mathbb{E}\left[\frac{1}{d} \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{1}[\mathcal{G}_{i,t}(d)] \Delta_{\boldsymbol{a}(t)} + \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{1}[\mathcal{H}_{i,t}(d)] \Delta_{\boldsymbol{a}(t)} \right] \\
\leqq & \mathbb{E}\left[\frac{1}{d} \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left[i \in \boldsymbol{a}(t), n_i(t) \leqq \frac{16\alpha(\sum_{k=1}^{L} P_k)^2 \log T}{\Delta^2_{\boldsymbol{a}(t)}} \right] \Delta_{\boldsymbol{a}(t)} \right]
\end{aligned}$$

$$+\sum_{i=1}^{K}\sum_{t=1}^{T}\mathbb{1}\left[i\in\boldsymbol{a}(t), n_i(t)\leqq\frac{16\alpha(\sum_{k=1}^{d}P_k)^2\log T}{\Delta_{\boldsymbol{a}(t)}^2}\right]\Delta_{\boldsymbol{a}(t)}\Bigg]$$
$$\leqq 32\alpha\log T\left(\frac{(\sum_{k=1}^{L}P_k)^2}{d}+\left(\sum_{k=1}^{d}P_k\right)^2\right)$$
$$\times\left(\sum_{i=L+1}^{K}\frac{1}{P_L(\theta_{(L)}-\theta_{(i)})}+\frac{L}{\Delta_{\boldsymbol{a}^{\star\star}}}\right)\qquad\text{(補題 4 より)}$$

□

注意 1. *Lagrée* ら *[10]* は，位置ベースモデル *(PBM)* における複数スロット提示（*multiple-play*）を扱うランキングバンディットを研究し, 位置 j の閲覧確率（*examination*） P_j が既知の下で *PBM-UCB* の対数リグレット上界を与えた．同論文の *Theorem 9* によれば，任意の $\epsilon>0$ に対し，ある定数 $C_0(\varepsilon)$ が存在し，任意の自然数 $1\leqq d\leqq L$ に対して

$$\mathrm{Reg}(T)\leqq 16(1+\varepsilon)\log T\left(\frac{\left(\sum_{k=1}^{L}P_k\right)^2}{d}+\left(\sum_{k=1}^{d}P_k\right)^2\right)/P_L^2$$
$$\times\left(\sum_{i=L+1}^{K}\frac{1}{P_L(\theta_{(L)}-\theta_{(i)})}+\frac{L}{\Delta_{\boldsymbol{a}^{\star\star}}}\right)$$
$$+C_0(\varepsilon)\qquad(5)$$

が成り立つ．この式において $\epsilon\to 0$ とした式と定理 *1* で $\alpha\to 1/2$ とした式において, 主要項である第 1 項は P_L^{-2} 倍だけ *PBM-UCB* の方が大きく，表示リスト末尾の閲覧確率 P_L が小さい状況では上界が大きく悪化し得る．

一方，本研究の設定では，各ラウンドでどこまで表示されたかを表す閲覧深度 $V(t)$ が観測できるため，位置 j が露出したかどうか（$V(t)\geqq j$）が直接わかる．これは *PBM* における *examination* が 潜在（*censored*） ではなく 観測（*uncensored*）である状況に対応し，*Lagrée* らも *uncensored PBM* では情報量が単純化することを指摘している *[10]*.

6 実　　験

本節では，提案手法（Observable-Depth UCB/TS；OD-UCB/OD-TS）の有効性を (i) 合成シミュレーション，(ii) 実ログ（RecGaze）から推定したパラメータに基づく実データ駆動実験の 2 つで検証する．いずれも真の環境は第 3 節の閲覧深度つき cutoff 型モデルとし，提案法は各ラウンドで閲覧深度 $V(t)$ を観測できる一方，既存ベースラインは $V(t)$ を観測できない（クリックのみ）という条件で比較する．

6.1 評価指標（累積擬似リグレット）

最適ランキング $\boldsymbol{a}^\star$ に対するアルゴリズムのランキング列 $\boldsymbol{a}(1),\dots,\boldsymbol{a}(T)$ の累積期待リグレット $\sum_{t=1}^{T}\Delta_{\boldsymbol{a}(t)}$ を用いる．合成シミュレーションでは真のパラメータ (θ,P) が既知であるため，各ラウンドの期待報酬差 $\Delta_{\boldsymbol{a}(t)}$ を直接計算して累積する．また，実データ駆動実験ではログから計算した (クリック割合, 閲覧割合) を (θ,P) として用い，合成シミュレーションと同様に各ラウンドの期待報酬差を計算して累積する．

6.2 合成シミュレーション環境

各ラウンド t で閲覧深度 $V(t)\in\{1,\dots,L\}$ を, 以下で述べる分布に従ってサンプルし，$j\leqq V(t)$ の位置のみが露出（表示）される．露出された位置 j では，選ばれたアイテム $i=a_j(t)$ を, 以下の方法で生成したパラメータ θ_i のベルヌーイ分布でクリック $C_i(t)$ を生成する．

a） 閲覧深度分布

露出確率 $P_j=\Pr(V(t)\geqq j)$ を与え，$\Pr(V(t)=j)=P_j-P_{j+1}$（$P_{L+1}=0$）で閲覧深度分布を定める．本稿では, 深度が浅い／深い状況を模すため，$L=5$ のとき以下の 2 つの設定を用いる：

$$\text{(shallow)}\ (P_1,\dots,P_5)=(1.0,0.55,0.30,0.15,0.08),$$

$$\text{(deep)}\ (P_1,\dots,P_5)=(1.0,0.80,0.70,0.60,0.50).$$

b） 実 験 設 定

主設定として $K=50$， $L=5$， $T=200{,}000$ とした．アイテムの魅力度 θ_i は，上位 5 個を 0.18, 0.16, 0.14, 0.12, 0.10 とし，残りの 45 個を [0.09,0.02] の両端を含めた等間隔で生成し, 乱数 seed を固定して一様ランダムにシャッフルして各アイテムに割り当てた．さらに，seed を変えた 5 回の独立試行を行い，平均と標準誤差で評価する．UCB 系アルゴリズムの信頼半径の係数は $\alpha=0.5$ とした．

6.3 実データ駆動実験

公開データセット RecGaze のログを用い，(i) 露出確率列 P（閲覧深度分布）と (ii) アイテム魅力度 θ をデータから算出される閲覧割合，クリック割合に設定し，そのモデル上で各手法の期待リグレットを比較する．本実験の狙いは，閲覧深度観測により P_L が小さい（末尾が見られにくい）状況で理論的に有利となるという定理 1 の含意が，実ログ由来の分布でも観察されるかを確認する点にある．

RecGaze は，カルーセル型 UI におけるユーザ行動を対象としたデータセットであり，視線（eye tracking），クリック，カーソル移動，および選択理由の説明を含む包括的なフィードバックが含まれる．3 つの映画選択タスクにおいて，各ユーザに対して 40 種類のカルーセル画面を提示し，合計 87 名・3,477 回のインタラクションが記録されている [7].

a） 閲 覧 割 合

視線, カーソル移動, クリック等のログから計算した閲覧割合は以下であった：

$$(P_1,\dots,P_{15})=(1.0,0.9031,0.8529,0.7720,0.6879,$$
$$0.2992,0.2991,0.2988,0.2986,0.2968,$$
$$0.2576,0.2574,0.2563,0.2531,0.2432).$$

特に $j=5$ から $j=6$ で P_j が大きく低下しており，上位数件は比較的見られるが，それ以降は露出が急減する浅い閲覧が混在していることがわかる．このような状況では，深度を観測

できない手法は未露出の負例混入により推定効率が低下しやすく，深度観測の利点が顕在化すると期待される．

b） 実 験 設 定

K = 150， L = 15， T = 200,000 とした．RecGaze ログから計算した (θ, P) を固定し，各手法は同一のモデルの下で評価した．ここでクリック割合 θ は各アイテムのクリック数と露出数から計算した．さらに，乱数 seed を変えた 5 回の独立試行を行い，平均と標準誤差で評価した．UCB 系アルゴリズムの信頼半径の係数は $\alpha = 0.5$ とした．

6.4 比 較 手 法

合成・実データ駆動の両実験で，環境は常に深度つき cutoff モデルで固定し，アルゴリズム側が $V(t)$ を利用できるかどうかのみを変えて比較する．

a） 提 案 法

提案法は各ラウンドで $V(t)$ を観測し，$j \leqq V(t)$ の位置のみを露出サンプルとして更新する：

- **OD-UCB**：露出回数に基づく推定と UCB により上位 L 件を提示する．
- **OD-TS**：露出回数に基づく Beta 事後分布からサンプルし，上位 L 件を提示する．

b） ベースライン

ベースラインは深度を観測できない状況を想定し，クリック系列のみから学習を行う：

- **PBM-UCB（known bias）**：位置バイアス P_j が既知として PBM の更新を行う UCB．
- **PBM-TS（known bias）**：同様に P_j 既知の PBM 系 TS [10]．事後分布が Beta に閉じず，θ のサンプルは棄却サンプリングで生成するため，OD-TS より計算コストが大きくなりうる．
- **Cascade-UCB（last-click heuristic）**：最後のクリック位置までを露出範囲として更新する単純ベースライン．

6.5 結 果

図 2，図 3 に合成シミュレーション環境（shallow/deep）での累積リグレットを示す．また図 4 に RecGaze 実ログから計算した閲覧割合 P に基づく累積期待リグレットを示す．各曲線は複数 seed の平均であり，誤差帯は標準誤差を表す．

shallow 設定（図 2）では，OD-TS が最小の累積リグレットを達成し，PBM-TS がこれに僅差で続いた．一方で UCB 系は全体に大きく，特に PBM-UCB（閲覧確率 P 既知）は大きなリグレットを示した．OD-UCB は PBM-UCB および Cascade-UCB を大きく上回り，深度観測を用いることが学習効率の改善に寄与していることが確認できる．また Cascade-UCB は試行間のばらつきが相対的に大きい傾向が見られた．

deep 設定（図 3）でも，OD-TS が一貫して最良であり，PBM-TS が次点となった．OD-UCB は TS 系に比べると大きいものの，PBM-UCB および Cascade-UCB より小さい累積リグレットを示し，shallow と同様に深度観測の効果が確認された．

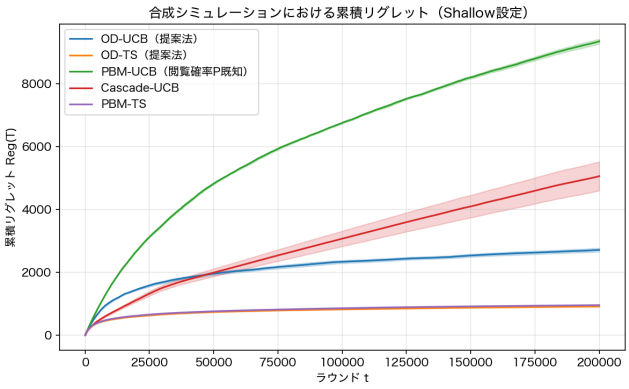

図 2 shallow 設定における累積リグレット．

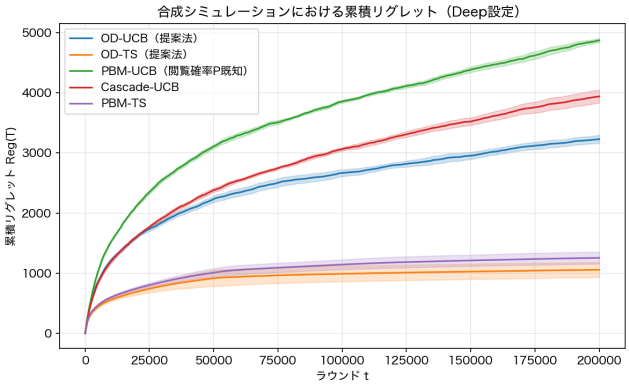

図 3 deep 設定における累積リグレット．

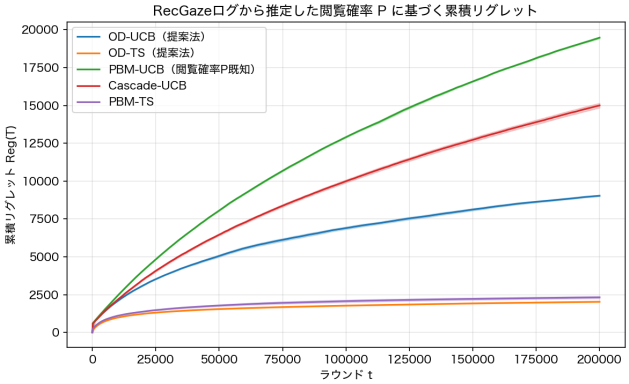

図 4 RecGaze ログから計算した (θ, P) に基づく累積擬似リグレット．

RecGaze を用いた実験（図 4）においても，OD-TS が最小，PBM-TS が次点であり，OD-UCB は PBM-UCB および Cascade-UCB を大きく上回った．特に PBM-UCB は最も大きい累積擬似リグレットを示し，閲覧確率が既知であっても深度を観測しない学習は不利になり得ることが示唆される．

6.6 考 察

提案法（OD-UCB/OD-TS）が改善する主因は，観測された深度 $V(t)$ までの位置のみを学習に用いることで，未露出位置をクリック 0 の負例として混入させない点にある．この効果は shallow 設定および RecGaze を用いた実験で特に顕著であり（図 2, 4），深部位置の露出が希薄な状況では深度を観測できな

い手法ほど負例混入が増え，推定効率が低下しやすい．

一方で deep 設定では多くの位置が露出されるため，深度観測による利得は shallow に比べて相対的に小さくなり得るが，本実験ではそれでも OD-UCB が PBM-UCB や Cascade-UCB を一貫して上回った（図 3）．これは，露出が増えてもどこまで見られたかという情報を明示的に用いることで，学習に用いるサンプルの質（負例混入の抑制）が改善されるためと解釈できる．

また，TS 系（OD-TS / PBM-TS）が UCB 系より大幅に小さい累積リグレットを示した点も重要である．特に OD-TS は shallow/deep/RecGaze の全条件で最良であり，深度観測による情報を活かしつつ事後分布に基づく探索が有効に機能したことを示唆する．一方，OD-UCB は提案枠組みにより PBM-UCB や Cascade-UCB を改善するものの，TS 系ほどの改善幅は得られていないため，有限時間での信頼半径の保守性や探索強度の設計が性能差として現れている可能性がある．

また，第 5 節で述べたように，定理 1 の主要項は露出確率列 P を通じて $\left(\sum_{k=1}^{L} P_k\right)^2$ や $\sum_{k=1}^{d} P_k$，およびギャップ項に現れる $1/P_L$ に依存する．したがって，末尾閲覧確率 P_L が小さいほど（末尾が見られにくいほど），深度観測により未露出位置の負例混入を抑える利得が大きくなり，深度観測なしの既存手法との性能差（特に OD-UCB と PBM-UCB の差）が拡大しやすい．実際，shallow 設定では P_L が相対的に小さいため差が顕著であり（図 2），deep 設定でも shallow ほどではないが末尾確率の有限性に起因する差が残る（図 3）．RecGaze では中位以降で露出確率が大きく低下しており，末尾が見られにくい条件が含まれるため，深度観測の利点が実験でも確認された（図 4）．

RecGaze の結果は推定モデル上のリグレットであり，真のユーザ行動モデルに対する厳密なリグレット保証を与えるものではない．しかし，ログ推定により得られる浅い閲覧を含む条件下で，提案手法（OD-UCB/OD-TS）が一貫して小さい累積期待リグレットを示したことは，深度観測により学習効率が改善するという理論的示唆と整合する．

7 おわりに

本研究では，カルーセル UI を想定し，ユーザの閲覧深度（最大表示位置）が観測可能な条件下でのランキングバンディット問題を定式化した．従来のクリックモデル（カスケードモデル，PBM 等）では閲覧位置（examination）が潜在変数として扱われることが多く，クリックしなかった観測が負例か未観測か判別しにくい．これに対し本研究は，観測された閲覧深度に基づいて実際に露出（観測）された位置のみで学習を更新することで，未観測位置を誤って負例として扱うことを避け，情報効率のよいランキング学習を可能にする枠組みを提示した．

提案手法として，露出回数に基づく信頼区間を用いる Observable-Depth UCB と，同様に露出回数に基づく事後分布からサンプリングする Observable-Depth TS を提案した．さらに Observable-Depth UCB についてギャップ依存の期待リグレット上界を導出し，閲覧深度が観測可能であることが学習効率に寄与することを理論的に整理した．また，シミュレーション実験により，PBM/Cascade 系の既存ベースライン（UCB/TS）と比較して，提案法が累積リグレットを改善することを確認した．

今後の課題としては，まず理論解析の改善が挙げられる．本稿の上界は最悪の閲覧確率として P_L を用いたため保守的であり，位置ごとの閲覧確率 P_j や実際の配置分布を利用したよりタイトな上界への改良が考えられる．また，Observable-Depth TS の理論保証や，下界（深度観測あり設定のミニマックス下界／ギャップ下界）の導出も重要である．さらに，クリック生成に追加の位置効果や文脈情報を含めた拡張，2 次元カルーセル（行×列）や多様性制約を伴う UI 設計に適合したモデル化など，実システムに近い設定への一般化も今後の発展方向である．

謝　辞

研究室内での議論やコメントを通じて貴重な示唆を頂いたアルゴリズム研究室の皆様に感謝いたします．

文　献

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, Vol. 47, No. 2–3, pp. 235–256, 2002.

[2] Walid Bendada, Guillaume Salha, and Théo Bontempelli. Carousel personalization in music streaming apps with contextual bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20)*, pp. 420–425, 2020.

[3] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[4] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. *Click Models for Web Search*. Morgan & Claypool Publishers, 2015.

[5] Richard Combes, Mohammad Sadegh Talebi, Alexandre Proutière, and Marc Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015), Montreal, Quebec, Canada, December 7–12, 2015*, pp. 2116–2124, 2015.

[6] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 87–94, 2008.

[7] Santiago de Leon-Martinez, Jingwei Kang, Robert Moro, Maarten de Rijke, Branislav Kveton, Harrie Oosterhuis, and Maria Bielikova. RecGaze: The first eye tracking and user interaction dataset for carousel interfaces. In *SIGIR 2025: 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3702–3711. ACM, July 2025.

[8] Junpei Komiyama, Junya Honda, and Akiko Takeda. Position-based multiple-play bandit problem with unknown position bias. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[10] Paul Lagrée, Claire Vernade, and Olivier Cappé. Multiple-play bandits in the position-based model. In *Advances in*

Neural Information Processing Systems (NeurIPS), 2016.

[11] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[12] Behnam Rahdari, Peter Brusilovsky, and Branislav Kveton. Towards simulation-based evaluation of recommender systems with carousel interfaces. *ACM Transactions on Recommender Systems*, Vol. 2, No. 1, pp. 9:1–9:25, 2024.

[13] Behnam Rahdari, Branislav Kveton, and Peter Brusilovsky. From ranked lists to carousels: A carousel click model. *CoRR*, Vol. abs/2209.13426, , 2022.

[14] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, Vol. 25, No. 3–4, pp. 285–294, 1933.