# 変化検出と要約データ構造を用いた 利用者の嗜好の変化に迅速に追従する多腕バンディット手法

三宅 悠介[1,2,a)]　栗林 健太郎[1,b)]

**概要**：多腕バンディット問題は，腕と呼ばれる複数の候補から得られる報酬を最大化する問題である．同問題の Web サービスにおける広告配信や推薦システムへの応用では，腕となる利用者の嗜好傾向が多様である課題に対処するため，利用者の文脈を考慮した線形な問題設定への拡張と解法が提案されている．一方で，時間の経過に従い報酬分布が変化する非定常な課題に対処するため，変化検出の手法を組み合わせ報酬の変化を観察することで変化に追従する解法が提案されている．しかしながら，線形な問題設定にこの解法を適用する場合，文脈の増加に伴い各文脈での報酬の観測回数が低下するため，文脈ごとに報酬の変化を観測する方式では，変化の検出と追従が遅れてしまう．加えて，変化の検出と追従に必要な文脈ごとの報酬の履歴データのサイズも文脈の増加に伴い肥大化する．本研究では，多様な文脈であっても腕の報酬分布の変化に迅速に追従可能な，線形かつ非定常な多腕バンディット問題の解法を提案する．提案手法では，文脈ごとの報酬からではなく，文脈の数によらない固定数の値の推移のみから報酬分布の変化を検出することで，腕の報酬分布の変化に迅速に追従する．さらに，過去期間の値を要約するデータ構造を導入することで，報酬分布の変化検出と追従に必要な履歴データのサイズの肥大化を抑制する．評価では，線形かつ非定常な多腕バンディット問題を設定し，提案手法を用いない場合と比較して変化への追従性が高いこと，履歴データのサイズの肥大化が抑えられることを確認した．

# Multi Armed Bandit Method for Following Preference Changes using Change Detection and Summary Data Structures

***Abstract:*** A multi-armed bandit problem is a problem that maximizes reward from making choices between candidates called arms. In the application of advertisement or recommendation system of this problem, the linear extension of problem setting and policy is proposed to deal with users' various preferences. On the other hand, the policy that follows changes by observing reward change using change detection methods is proposed to deal with non-stationary problems that change reward distribution according to time. However, this policy delays the following changes because the number of reward observations for each context decreases if the policy applies to a linear problem setting. In addition, the data size of reward history required for change detection and follow changes increases as the number of contexts increases. In this paper, we propose a non-stationary linear multi-armed bandit policy that can quickly follow changes in the reward distribution in various contexts. The proposed policy follows changes by detecting a change of reward distribution from two value transition that is independent of the number of contexts. Also, the policy suppresses the size of the historical data by introducing a data structure that summarizes the value of the past period. We set up a non-stationary and linear multi-armed bandit problem for the evaluation and confirmed our policy makes the performance increase and suppress the size of historical data.

## 1. はじめに

消費者向け電子商取引（以下 EC）の市場規模拡大 [1] に伴い，取り扱う商品種類は増大している．EC サイト利用者の通常の行動では全ての商品を見て回ることは困難であるため，多くの EC サイトには推薦システムが導入されて

[1] GMO ペパボ株式会社 ペパボ研究所
Pepabo R&D Institute, GMO Pepabo, Inc., Tenjin, Chuo ku, Fukuoka 810-0001 Japan
[2] 九州大学 大学院システム情報科学府 情報知能工学専攻
Department of Advanced Information Technology, Graduate School of ISEE, Kyushu University
[a)] miyakey@pepabo.com
[b)] antipop@pepabo.com

いる．多様な利用者が訪問する EC サイトでは，全ての利用者に対する画一的な提案は必ずしも有用ではないことから，多くの推薦システムでは提案の個別化を図っている．

このような個別化した提案では，利用者ごとに有用な提案を予め知ることは難しい．そこで，推薦システムは利用者の嗜好を蓄積し，その時点で有用と考えられる情報を活用した提案を行う一方で，より有用な提案につながる情報の探索が求められる．この活用と探索のトレードオフの最適な解を求める問題は，多腕バンディット問題として知られている [2]．この問題は，ある確率分布に従い報酬を生成する腕と呼ばれる複数の候補から得られる報酬を最大化する問題であり，同問題に対する解法が推薦システムやインターネット広告の分野で利用されている [3][4]．一方で，基本的な多腕バンディット問題では，報酬の確率分布が常に同じであるという仮定が置かれている．推薦システムにおいて腕となる，利用者の商品に対する嗜好傾向は様々な要因によって変化することから，問題設定について 2 種類の拡張が図られている．線形な多腕バンディット問題では，文脈によって腕から得られる報酬の確率分布が決定される [5]．ここで文脈とは，例えば利用者の性別，年代，行動カテゴリといった複数の要因のパラメータの組み合わせによって表現される状態のことを指す．非定常な多腕バンディット問題では，例えば記事の目新しさと利用者の反応の大きさの関係のように，腕から得られる報酬の確率分布が時間経過によって変化する [6]．

利用者の嗜好が多様かつ継続的に変化する環境において，推薦システムが利用者の要求に応えるためには，できるだけ多くの文脈と報酬分布の変化を考慮できることが望ましい．そのため，線形かつ非定常な問題設定とその解法が求められる．非定常な解法では，変化検出の手法を組み合わせ，報酬の変化を観察することで報酬分布の変化に追従する．一方で，線形な問題設定では，要因の数に対する文脈の指数的な増加に伴い試行回数が文脈ごとに分散することから，各文脈で報酬を観測できる回数が急激に低下する．そのため，従来の方式では，報酬分布の変化を検出するために一定期間の報酬の履歴の蓄積が求められることから，変化の検出と追従が遅れてしまう．よって，線形かつ非定常な解法には，文脈や要因の数によらずに報酬分布の変化を検出可能な指標が必要となる．加えて，変化の検出と追従のために文脈ごとの報酬に関する履歴を利用することから，履歴データのサイズの肥大化も課題となる．

本研究では，変化検出と要約データ構造を用いた，多くの文脈と報酬分布の変化を考慮可能な，線形かつ非定常な多腕バンディット問題の解法を提案する．提案手法では，多様な文脈における報酬分布の変化を迅速にとらえるため，文脈ごとの報酬からではなく，要因の組み合わせ数によらない固定数の値の推移から報酬分布の変化を検出する．このために，まず各腕に対する試行回数と報酬から要因のパラメータに対する係数を推定する．次に，この求めた値をベクトルとみなしその方向データの異常度と大きさの推移のみから報酬分布の変化検出を行う．最後に，変化以前の試行結果を取り除くことで，新しい報酬分布に追従させる．また，これらの変化検出と追従に必要な複数の履歴データの保存に，過去期間の値を要約するデータ構造を導入し，データサイズの肥大化を抑制する．なお，提案手法は既存の解法に対し，変化検出と要約データ構造による拡張を施すため，理論保証があり実績のある既存の線形な解法を利用できる．評価では，線形かつ非定常な多腕バンディット問題を設定し，既存の線形な解法に対し提案手法を適用することで変化への追従性が向上すること，履歴データのサイズの肥大化が抑えられることを確認した．

本論文の構成を述べる．2 章で多腕バンディット問題の関連研究を紹介し，推薦システムでの応用における課題について述べる．3 章では，線形かつ非定常な多腕バンディット問題を解決する提案手法について述べる．4 章では提案手法の評価を行い，5 章でまとめる．

## 2. 関連研究

### 2.1 多腕バンディット問題

多腕バンディット問題は，腕と呼ばれる複数の候補から得られる報酬を最大化する問題である．プレイヤーは各試行で 1 つの腕を選択し，その腕から報酬を得る．各腕はある確率分布に従い報酬を生成するが，プレイヤーはこの確率分布を試行の結果から推測しなければならない．そのため，プレイヤーはある時点の腕ごとの評価に基づき，最も評価の高い腕を用いながらも，真に評価の高い腕の探索を並行して行う．この問題に対する解法では，ある時点で最も評価の高い腕を用いることを活用，各腕の評価を行うことを探索と呼び，これらの活用と探索，報酬による評価の見直しを繰り返し行うことで，短期的には探索による機会損失を，長期的には腕の固定化による機会損失を低減する．

同問題の最も単純な解法に $\epsilon$-Greedy アルゴリズム [7] がある．この解法では，腕の評価に報酬の標本平均を用いる．探索の割合を $\epsilon \in [0, 1]$ で指定し，活用時はその時点で最も評価の高い腕を，探索時にはその他の評価の低い腕を均等に選択する．UCB1 アルゴリズム [8] は，腕の選択に報酬の標本平均の値に，選択回数が少ないほど値が大きくなる項を加えたスコアを用いる．Thompson Sampling[9] は，各腕の期待値が最大である確率に従い腕を選定する．この解法では，各腕の期待値をベイズ推定によって求め，この分布からの乱数の値が最も大きかった腕を選定する．

### 2.2 線形な多腕バンディット問題

上述の問題設定では，文脈は常に一つであり，報酬の確率分布が変わらないという仮定を置いていた．線形な多腕バンディット問題は，複数の要因パラメータからなる文脈

に応じて報酬分布が決定される問題である．この問題では，各腕は線形パラメータと呼ばれる要因のパラメータに対する係数をベクトルとして持つ．文脈に応じた腕の報酬は，この線形パラメータと要因パラメータとの内積の結果に誤差を加えた値として求める．

文脈の種類が少ない場合には，単一の文脈を仮定した解法を文脈ごとに適用することでも対応できるが，要因のパラメータが増えるに従い，その組み合わせ結果である文脈の種類が指数的に増えてしまう．この場合，各文脈での試行回数は急激に低下し，腕の評価が充分に行えない．同問題の解法では，腕ごとの線形パラメータを推定しながら，活用と探索のトレードオフを解決する必要がある．

同問題の解法には，UCB1 アルゴリズムを拡張した LinUCB[5] が提案されている．この解法では，$t$ 回目の試行における各腕 $a$ のスコア，

$$LinUCB_a(t) = \boldsymbol{b}(t)^\mathsf{T}\hat{\boldsymbol{\theta}}_a(t) + \alpha\sqrt{\boldsymbol{b}(t)^\mathsf{T}\boldsymbol{B}_a^{-1}\boldsymbol{b}(t)} \quad (1)$$

が最も大きくなる腕を選択する．ここで $\boldsymbol{b}$ は $d$ 次元の要因パラメータ，$\hat{\boldsymbol{\theta}}_a$ は腕 $a$ に対して推定した線形パラメータである．$\hat{\boldsymbol{\theta}}_a = \boldsymbol{B}_a^{-1}\boldsymbol{f}_a$ であり，腕 $a$ における各要因のパラメータの累積の試行回数を記録する $\boldsymbol{B}_a \in \mathbb{R}^{d\times d}$ と各要因のパラメータの累積の報酬 $\boldsymbol{f}_a \in \mathbb{R}^d$ から推定する．なお，ハイパーパラメータ $\alpha(\alpha \geq 0)$ が小さいほど腕の探索よりも活用が重視される．また，Thompson Sampling を同様に拡張した Linear Thompson Sampling[10] も提案されている．腕の報酬分布が正規分布に従う場合，この解法では $\hat{\boldsymbol{\mu}}$ を平均，$v^2\boldsymbol{B}_a^{-1}$ を分散とする多変量正規分布から $\tilde{\boldsymbol{\mu}}$ を求め，要因パラメータである $\boldsymbol{b}(t)$ との内積が最も大きくなる腕を選定する．なお，$\hat{\boldsymbol{\mu}} = \boldsymbol{B}_a^{-1}\boldsymbol{f}_a$ であり LinUCB と同様に推定される．ハイパーパラメータも同様に $v^2(v^2 \geq 0)$ が小さいほど腕の探索よりも活用が重視される．

### 2.3 非定常な多腕バンディット問題

ここまでの問題設定は，腕ごとの報酬分布が文脈によって定まり，同じ文脈であれば変わらないという仮定を置いていた．非定常な多腕バンディット問題は，同じ文脈においても報酬分布が時間経過によって変化する問題である．報酬分布の変化が周期的な場合，この周期を線形パラメータに含め要因パラメータで指定することで適切に扱える．一方で，変化が不規則である場合にはこの限りではない．同問題の解法では，腕の報酬分布が変化した際に，不利な腕を使い続ける機会損失を抑えるため，過去に観測した報酬に捉われずに腕の評価を迅速に更新する必要がある．

同問題の解法には，大きく二つのアプローチが見られる．一つ目は，腕の報酬分布の変化を前提に，継続的に腕の評価を更新するものである．UCB1 アルゴリズムや Thompson Sampling をこの問題に適用した提案 [11][12][13] がなされている．これらの手法では，各腕の評価に用いる試行回数と報酬に割引の概念が導入され，過去の観測された値は継続的に更新され，新しく観測された報酬が重視される．

二つ目は，腕の報酬分布の変化を契機に，腕を再評価するものである．このアプローチでは，多腕バンディット問題の解法とは別に，変化検出の手法を利用するため，既存の解法を非定常な環境に適用できる．このアプローチでは，UCB1 アルゴリズムの一種である UCB1-Tuned[8] に変化検出の手法である Page-Hinkley test 法を組み合わせた手法が提案されている [14]．S-TS-ADWIN[15] は，Thompson Sampling に変化検出の手法である ADWIN[16] を組み合わせた手法である．ADWIN は直近からウィンドウと呼ばれる期間の値を記録する．記録時にウィンドウを過去のサブウィンドウ $W_0$ と現在のサブウィンドウ $W_1$ に分割し，この平均の差が $\epsilon_{cut}$ 以上であれば，その時点で変化があったとみなし，サブウィンドウ $W_0$ を取り除く．ここで，

$$\epsilon_{cut} = \sqrt{\frac{2}{m}\cdot\sigma_W^2\cdot\ln\frac{2}{\delta'}} + \frac{2}{3m}\ln\frac{2}{\delta'}, \quad (2)$$

と定義される．また，$\delta' = \frac{\delta}{|W|}$，$m$ は $|\,W_0\,|$ と $|\,W_1\,|$ に対する調和平均である．なお，$\delta \in (0,1)$ はこの統計的仮説検定の信頼度である．S-TS-ADWIN では，各腕の報酬の推移を ADWIN を用いて記録する．いずれかの腕の報酬に対する変化を検出した場合，全ての腕を通して最も近い時点以降に観測された試行回数と報酬を ADWIN より求め，これを用いて各腕の評価を更新する．

### 2.4 線形かつ非定常な多腕バンディット問題

上述の基本的または非定常な問題に対する解法は，単一の文脈を前提としている．これらの解法を複数の文脈を持つ環境に適用するには，2.2 節で述べたように指数的に増加する文脈への対策を講じなければならない．

そのため，複数の文脈を前提とする線形な解法を非定常な問題に対応させる方式が提案されている．Decay LinUCB[17] は線形かつ非定常な問題に対する解法である．LinUCB に割引の概念を導入し，式 1 における $\boldsymbol{B}_a$ と $\boldsymbol{f}_a$ に減衰パラメータ $\gamma \in (0,1)$ を乗じた結果を腕の評価に利用する．このアプローチでは，試行結果を常に定数倍する操作のみで線形な解法に追従性を付与できるため，導入が容易である．しかしながら，この減衰操作により試行回数は $\gamma^i$ の数列に対する無限級数として $1/(1-\gamma)$ に収束する．線形な解法ではこれを総数として，要因パラメータの次元ごとに試行回数を分配する．結果として，要因パラメータごとの試行回数は小さく，かつ一定数から増加せず，探索が継続してしまう．そのため，安定した運用のためには $\gamma$ を大きくする必要があり，追従性を高めることができない．

我々は，非定常な問題設定へのもう一つのアプローチである変化検出と線形な解法との組み合わせを検討した．このアプローチでは，問題の解法とは別に，変化検出の手法を利用するため，既存の線形な解法を非定常な環境に適用

できる．しかしながら，線形な問題設定では，要因の数に対する文脈の指数的な増加に伴い，各文脈で報酬を観測できる回数が急激に低下する．そのため，文脈ごとに報酬の変化を観測する従来の方式では，報酬の履歴の蓄積が遅れ，変化への追従性に課題が残る．また，変化検出のためには，観測対象となる値を一定期間保持する必要がある．加えて，変化検出後に腕の評価を更新するため，試行回数と報酬についても同様に保持しなければならない．特に，線形な解法では，腕の評価に各要因パラメータの試行回数と報酬が必要となる．すなわち，複数の要因パラメータからなる線形な問題では，要因のパラメータ数が増えるに従い，変化検出と追従に必要な履歴データの次元数も増え，データサイズが肥大化してしまう課題がある．

## 3. 提案手法

本研究では，推薦システムの利用者の嗜好傾向のような，多様な文脈があり，文脈に対応する報酬分布が時間経過によって変化する環境に対して有効な，線形かつ非定常な多腕バンディット問題の解法を提案する．そのために，非定常な問題設定において，腕の評価を継続的に更新するアプローチに対しても最高水準の評価を得た [15]，変化検出アプローチである S-TS-ADWIN の線形な解法への適用を図る．提案手法では，文脈ごとの報酬からではなく，文脈を横断して推定する，腕の線形パラメータの値から腕の報酬分布の変化を検出する．これにより，文脈の増加に伴う文脈ごとの報酬の観測回数の低下の影響を避け，多様な文脈における報酬分布の変化を迅速にとらえることができる．また，変化検出に，線形パラメータの値をベクトルとみなしその方向データの異常度と大きさのみを用いることで，線形パラメータごとに変化を検出する場合と比べて変化検出に必要な履歴データの数を減らすことができる．加えて，過去期間の値を要約するデータ構造を導入することで，報酬分布の変化検出と追従に必要な複数の履歴データのサイズの肥大化を抑制する．なお，提案手法は既存の解法に対し，変化検出と要約データ構造による拡張を施すため，既存の線形な解法を利用できる．

### 3.1 文脈数によらない指標を用いた変化検出

提案手法では，文脈ごとの報酬ではなく，その報酬から推定される $d$ 次元の線形パラメータの値を用いて腕の報酬分布の変化を検出する．検出の指標には，方向データの異常検知に用いられる異常度 [18] を採用する．ここで方向データとは，L2 正規化されたベクトルデータを指す．この方向データの分布にフォンミーゼス・フィッシャー分布を仮定したとき，任意の方向ベクトル $\boldsymbol{x}'$ の平均ベクトル $\hat{\boldsymbol{\mu}}$ に対する異常度 $a(\boldsymbol{x}')$ は $1-\hat{\boldsymbol{\mu}}^{\mathsf{T}}\boldsymbol{x}'$ で表される．ここで，$\hat{\boldsymbol{\mu}}$ はフォンミーゼス・フィッシャー分布の平均方向の最尤推定値であり，方向データの標本平均を L2 正規化したものとなる．よって，異常度 $a(\boldsymbol{x}')$ は L2 正規化前の任意のベクトル $\boldsymbol{x}'$ の平均ベクトル $\hat{\boldsymbol{\mu}}$ とのコサイン類似度を 1 から除したものに等しい．ただし，方向データの異常度のみでは，方向が同じで大きさのみが変化する場合に変化を検出できない．多腕バンディット問題の解法では，ある腕の線形パラメータの値が増減し，他の腕に対する有効性が変化した際に，腕の選定基準を改める必要がある．そこで，提案手法では，線形パラメータをベクトルと見なしたときの大きさである L2 ノルムの値の変化も考慮することで該当する状況での検出漏れを防ぐ．

提案手法では，この方向データの異常度と大きさの推移に対して，ADWIN の統計的仮説検定による変化検出を行う．ADWIN では，変化検出の閾値の根拠として，確率の分布によらない確率の近似的な評価である確率不等式を用いており，提案手法の入力値である方向データの異常度と大きさの変化を同一の方式で検出が可能かつ，異なる値での評価も容易である．また，サブウィンドウ間での比較を行うため，学習の工程が不要である．これらの実用性の観点から多腕バンディットのシステム実装を想定した提案手法との親和性が高いと考え採用した．

提案手法における，推定した線形パラメータに対する変化検出のアルゴリズムを Algorithm1 に示す．我々は本方式を ADWIN-V と名付けた．また，実装は OSS として公開済みである[*1]．ADWIN-V では，入力となるベクトルデータから，方向データの異常度と大きさを求め，これらにスケーリング定数 $s$ を乗じた値を個々の ADWIN へ記録する．なお，スケーリング定数 $s$ は ADWIN が入力値の大きさに応じて変化検出の精度が異なる挙動に対する調整パラメータである．個々の ADWIN で変化が検出された場合，ADWIN は変化前の系列データを削除する．これらは Algorithm1 の 4 行目から 10 行目に相当する．なお，方向データの異常度の算出に必要な，平均ベクトルはこれまでの入力ベクトルの平均から求める．ただし，変化前の系列データの削除に伴う平均ベクトルの再計算は行わず，直近の入力値を平均ベクトルとして利用する．これは，変化検出に必要な複数の履歴データの数を減らすよう，平均ベクトルは履歴を持たず逐次計算で求めるためである．これらは Algorithm1 の 11，13 行目に相当する．

### 3.2 データ構造による履歴データサイズの低減

提案手法では，3.1 節の方式によって腕の報酬分布の変化を検出した後，該当する腕における変化前の試行の結果を取り除くことで環境の変化に追従させる．ここで試行の結果とは，任意の時点での要因パラメータと報酬額を指す．2.2 節で述べたように線形な解法では，$t$ 時点までの試行の結果の累積値から線形パラメータを推定できる．しか

[*1] https://github.com/monochromegane/adwin-v

**Algorithm 1** ADWIN-V

**Require:** ADWIN confidence value $\delta_m$, $\delta_a$. And scaling value parameter $s_m$,$s_a$.
1: $A_m \leftarrow$ instance of ADWIN for vector magnitude with $\delta_m$
2: $A_a \leftarrow$ instance of ADWIN for vector angle with with $\delta_a$
3: **for all** $t = 1, 2, ...,$ **do**
4: Add $s_m \cdot ||\boldsymbol{x}_t||_2$ to $A_m$
5: **if** $A_m$ detects change **then**
6: $A_m$ shrinks window.
7: **end if**
8: Add $s_a \cdot 1 - (\boldsymbol{m}_{t-1}^{\mathsf{T}} \boldsymbol{x}_t / (||\boldsymbol{m}_{t-1}||_2 \cdot ||\boldsymbol{x}_t||_2))$ to $A_a$
9: **if** $A_a$ detects change **then**
10: $A_a$ shrinks window.
11: $\boldsymbol{m}_t = \boldsymbol{x}_t$
12: **else**
13: $\boldsymbol{m}_t = (t \cdot \boldsymbol{m}_{t-1} + \boldsymbol{x}_t)/(t+1)$
14: **end if**
15: **end for**

し，この累積値から変化前の試行の結果を取り除くためには，各時点での試行結果の値を履歴として保持する必要がある．また，変化検出のために，各時点における方向データの異常度と大きさも同様に保持する必要がある．

これらの変化検出と追従に必要な複数の履歴データのサイズを小さく抑えるためには，履歴の期間の長さに対して小さいデータサイズで保存できることが望ましい．そこで提案手法では，過去期間の値を要約するデータ構造である指数ヒストグラム [19] を利用する．このデータ構造は，時点ごとの入力値を保存するが，古い値についてはある期間で合算した値で保持する．この合算する期間を指数的に大きくすることで，データ構造のサイズを小さく保つ．このデータ構造では，この合算された期間のデータをバケットと呼び，各バケットは期間の長さと期間中の合算値との 2 つの値で表現される．

提案手法で利用する ADWIN では変化検出の履歴データに，この指数ヒストグラムを用いる ADWIN2[16] も提案されており，S-TS-ADWIN でも利用されている [15]．そこで，3.1 節で提案した ADWIN-V における，方向データの異常度と大きさを記録する ADWIN インスタンスを ADWIN2 に差し替えることで変化検出に必要な履歴データのサイズを抑えることができる．以降，ADWIN2 を用いる報酬分布の変化検出を ADWIN2-V と呼ぶ．

提案手法では，追従に必要な試行結果の値に関する複数の履歴データも指数ヒストグラムを用いて記録する．線形な解法では，任意の時点での $d$ 次元の要因パラメータ $\boldsymbol{b}$ の直積 $B = \boldsymbol{b}\boldsymbol{b}^{\mathsf{T}}$，ならびに要因パラメータに報酬 $r$ を乗じた $\boldsymbol{f} = r\boldsymbol{b}$ が試行結果にあたる．ここで，$B$ は $d$ 個の元を持つ集合から 2 個の重複あり組み合わせとして $(d+1)!/2!/(d-1)!$ 次元，$\boldsymbol{f}$ は $d$ 次元のベクトルデータとみなせる．ただし，従来の指数ヒストグラムは単一の値を記録することを想定しており，ベクトルデータの記録のためには各次元に対してこのデータ構造を用意しなければならない．この場合，各バケットで期間の長さを扱う値が重複し，結果としてデータサイズが増加してしまう．そこで，多次元の値を保持するための指数ヒストグラムとして，各バケットで単一の値ではなくベクトルを用いるデータ構造 ExpHist-V を考案，OSS として公開した[*2]．提案するデータ構造では，各バケットは期間の長さと，ベクトルデータに対する期間中の合算値からなる．なお，期間中の値の合算はバケットごとのベクトル同士の和によって行われる．このデータ構造により，$d'$ 次元のデータ保持に必要なバケットあたりのデータサイズが $2d'$ から $d'+1$ へと縮小する．

### 3.3 既存の線形な多腕バンディット問題の解法の拡張

上述した ADWIN2-V と ExpHist-V を用いて，従来の線形な解法の環境変化への追従性を高める提案手法を説明する．提案手法である Adaptive-Linear-MAB のアルゴリズムを Algorithm2 に示す．はじめに，累積試行回数，累積報酬を記録する $\boldsymbol{B}$，$\boldsymbol{f}$，腕の報酬分布の変化を検出するため推定した線形パラメータを記録する ADWIN2-V である $\boldsymbol{A}$，変化前の試行の結果を記録する ExpHist-V である $\boldsymbol{E}^{(B)}$，$\boldsymbol{E}^{(\boldsymbol{f})}$ を腕ごとに用意する．次に $t$ 時点の試行において，任意の従来の線形な解法 $P$ を用いて，腕 $a(t)$ を選定し，報酬 $r_t$ を得る．提案手法では，この解法 $P$ として，従来の線形な解法に加え，これらに減衰操作を加える Decay LinUCB も利用できる．これは Algorithm2 の 7 行目に当たる．次にこの試行の結果を受け，選定した腕に関する累積試行回数 $\boldsymbol{B}_i$，累積報酬 $\boldsymbol{f}_i$ を更新する．合わせて，推定した線形パラメータ $\boldsymbol{B}_i^{-1}\boldsymbol{f}_i$ を ADWIN2-V の $\boldsymbol{A}_i$ へ，試行の結果を ExpHist-V である $\boldsymbol{E}^{(B)}$，$\boldsymbol{E}^{(\boldsymbol{f})}$ へ記録する．これは Algorithm2 の 11 から 14 行目にあたる．ADWIN2-V によって腕の報酬分布の変化が検出された場合，変化前の期間データとして指数ヒストグラムの最も古いバケットが取り除かれる．このとき $\boldsymbol{E}^{(B)}$，$\boldsymbol{E}^{(\boldsymbol{f})}$ の最も古いバケットから，この削除された期間と同期間の試行の結果を求め，選定した腕に関する累積試行回数 $\boldsymbol{B}_i$，累積報酬 $\boldsymbol{f}_i$ から除することで変化前の試行の結果を取り除く．これは Algorithm2 の 15 から 19 行目にあたる．なお，ADWIN2-V と同期させるため $\boldsymbol{E}^{(B)}$，$\boldsymbol{E}^{(\boldsymbol{f})}$ の最も古いバケットも取り除く．

Decay LinUCB のような試行の結果の減衰を伴う解法の場合，累積試行回数，累積報酬を減衰する．減衰操作は Algorithm2 の 9 行目にあるように，累積試行回数 $\boldsymbol{B}$，累積報酬 $\boldsymbol{f}$ に $\gamma$ を乗じる．変化前の試行の結果である $\boldsymbol{E}^{(B)}$，$\boldsymbol{E}^{(\boldsymbol{f})}$ では各バケットの合算値に対して $\gamma$ を乗じる．なお，減衰を伴わない解法では $\gamma = 1$ を用いる．

## 4. 評価

本章では，提案する，推定した線形パラメータに対する

[*2] https://github.com/monochromegane/exponential-histograms

**Algorithm 2** Adaptive-Linear-MAB

**Require:** Set of arms [K], Linear MAB Policy $P$.
Set $\boldsymbol{B}_i = I_d, \forall_i \in [K]$
Set $\boldsymbol{f}_i = 0_d, \forall_i \in [K]$
$\boldsymbol{A}_i \leftarrow$ instance of ADWIN2-V with $\delta_m, \delta_a, s_m, s_a, \forall_i \in [K]$
$\boldsymbol{E}_i^{(B)} \leftarrow$ instance of ExpHist-V for $B_i, \forall_i \in [K]$
$\boldsymbol{E}_i^{(\boldsymbol{f})} \leftarrow$ instance of ExpHist-V for $\boldsymbol{f}_i, \forall_i \in [K]$
**for all** $t = 1, 2, ...,$ **do**
  Play arm $a(t) = P(\boldsymbol{B}, \boldsymbol{f}, \boldsymbol{b}(t))$, and observe reward $r_t$.
  **for all** $i = 1, 2, ..., K$ **do**
    Decay $\boldsymbol{B}_i, \boldsymbol{f}_i, \boldsymbol{E}_i^{(B)}, \boldsymbol{E}_i^{(\boldsymbol{f})} = \gamma \boldsymbol{B}_i, \gamma \boldsymbol{f}_i, \gamma \boldsymbol{E}_i^{(B)}, \gamma \boldsymbol{E}_i^{(\boldsymbol{f})}$
    **if** $i = a(t)$ **then**
      Update $\boldsymbol{B}_i = \boldsymbol{B}_i + \boldsymbol{b}(t)\boldsymbol{b}(t)^\mathsf{T}$
      Update $\boldsymbol{f}_i = \boldsymbol{f}_i + r_t \boldsymbol{b}$
      Add $\boldsymbol{B}_i^{-1} \boldsymbol{f}_i$ to $\boldsymbol{A}_i$
      Add $\boldsymbol{b}(t)\boldsymbol{b}(t)^\mathsf{T}, r\boldsymbol{b}(t)$ to $\boldsymbol{E}_i^{(B)}, \boldsymbol{E}_i^{(\boldsymbol{f})}$
      **if** $\boldsymbol{A}_i$ detects change **then**
        $\boldsymbol{A}_i$ drops oldest bucket.
        Get $\boldsymbol{B}_i^{(old)}, \boldsymbol{f}_i^{(old)}$ from $\boldsymbol{E}_i^{(B)}, \boldsymbol{E}_i^{(\boldsymbol{f})}$.
        Update $\boldsymbol{B}_i = \boldsymbol{B}_i - \boldsymbol{B}_i^{(old)}, \boldsymbol{f}_i - \boldsymbol{f}_i^{(old)}$
      **end if**
    **end if**
  **end for**
**end for**

方向データの異常度と大きさを用いた変化検出によって，線形かつ非定常な問題設定での追従性が高まることを確認する．確認は，従来の線形な解法と，これに提案手法を適用し追従性を高めた解法を比較して行う．また，提案手法で採用したデータ構造により，これを採用しない場合と比べ，報酬分布の変化検出・追従に必要な履歴データのサイズの肥大化が抑えられることを確認する．

### 4.1 評価環境

提案手法による，線形かつ非定常な問題に対する性能を評価するため，シミュレーションを行った．シミュレーションでは時間あたりの購入数が多い商品を選定する推薦システムを想定した．ここで，文脈によって購入数が異なる状況を表現するため，購入数を性別，年代，行動カテゴリのような要因に対する線形パラメータから算出する．また，線形パラメータをある時点で変化させることで非定常な環境を再現する．このような状況には，例えば，特定の年代に向けたメディアへの露出による突発的な購買傾向の変化などが考えられる．シミュレーションでは，数百程度の文脈を想定し，次元数 $d = 8$ の線形パラメータを持つ腕 $a_i \in \{a_0, a_1\}$ に対し提案手法を含む複数の解法を用いて 2,000 時点までの累積報酬と累積リグレットを計測した．ここで累積リグレットは試行時の文脈において候補の腕のうち最大の期待値と選択した腕の期待値の差を期間までに合計したものである．腕の線形パラメータはそれぞれ，$\boldsymbol{\theta}_0 = [14, 15, 16, 17, 18, 19, 20, 4]$，$\boldsymbol{\theta}_1 = [12, 13, 14, 15, 16, 17, 18, 20]$ とするが，非定常な環境とするため，腕同士の要因ごとの係数の多寡を入れ替えるよう，500 時点目に $\boldsymbol{\theta}_1 = [20, 24, 28, 32, 2, 4, 6, 8]$ に変更する．また，これらの値は解法によって腕の選定回数が偏ることで，腕ごとの線形パラメータを正しく推定するまでの期間に差が出ることを防ぐため，線形パラメータの変更前後で腕 $a_0$ と $a_1$ の選定割合がほぼ同等となるよう設定した．要因パラメータ $\boldsymbol{b}$ は，各次元が 0 と 1 の離散値から成り，各次元の値は 1 となる確率 $p = 0.5$ のベルヌーイ分布に従い得られることとする．$t$ 時点の試行で選択した腕 $i$ から得られる報酬は $\boldsymbol{\theta}_i^\mathsf{T} \boldsymbol{b}_t + \epsilon_t$ となる．ここで誤差項 $\epsilon_t$ は平均 0，分散 $\sigma^2 = 2$ の正規分布に従う乱数を用いた．なお，乱数を用いた確率の計算結果を平均化するために上述のシミュレーションを 500 回行い，この平均を結果として用いた．

本評価では，この線形かつ非定常な問題に対し 3.3 節で提案した提案手法を従来の線形な解法に適用したものを利用する．また，比較のため，提案手法を適用しない従来の解法も評価する．従来の解法には，線形な解法である LinUCB ならびに，Linear Thompson Sampling(以下，LTS) を用いる．また，線形かつ非定常な解法である Decay LinUCB，ならびに同様の減衰操作を LTS に施した Decay LTS も評価する．以降，これらの 4 つの解法に対して提案手法を適用した解法には Adaptive の接頭辞を付けて区別する．

各解法のハイパーパラメータには，探索の重視，減衰の度合い，変化検出の感度に関するものがある．これらの効果が大きいとき，線形パラメータに変化がなくとも累積リグレットが継続的に増加し，評価期間の長さで結果が異なる場合が見られた．これを加味し，各解法のハイパーパラメータは線形パラメータの変化を伴わない予備実験によって求めた．予備実験では，線形パラメータの推定に対して，不要な探索を避けつつ一定の追従性を持つものを選定した．この基準として 500 時点と 2000 時点の累積リグレットの増加率を採用し，複数の候補値のうち，この値が最も小さいものを選定する．なお，試行履歴の減衰や変化検出の効果を明確にするため，基礎となる解法と共通するハイパーパラメータの値には同じものを設定した．探索に関するハイパーパラメータの予備実験では $5, 10, 20, 30, 40, 50, 75, 100, 150, 200, 250$ のうち，LTS で $v^2 = 150$，LinUCB で $\alpha = 20$ を選定した．減衰に関するハイパーパラメータは，上記の探索に関する値のもとで $0.95, 0.99, 0.999$ から選定した．Decay LTS と Decay LinUCB でともに $\gamma = 0.999$ であった．変化検出に関するハイパーパラメータは，上記の探索に関する値のもとで $0.1, 0.01, 0.001, 0.0001$ から選定した．Adaptive LTS，Adaptive LinUCB でともに $\delta_m = 0.0001$，$\delta_a = 0.0001$ であった．スケーリング定数 $s$ は異常度 $a$ と大きさ $m$ が同程度の範囲となるよう $s_m = 0.1$，$s_a = 1.0$ を用いた．減衰と変化検出をともに行う Adaptive Decay LTS，Adaptive Decay LinUCB では，これまでに選定した減衰と変化検出の値を利用する．なお，予備実験のシミュレーション回数は 100 回である．

表 1　累積報酬と累積リグレット
**Table 1**　Cumulative reward and regret.

| 解法 | 累積報酬 | 累積リグレット |
|---|---|---|
| LTS | 133624.88 | 2420.73 |
| Adaptive LTS | 134014.42 | 2030.73 |
| Decay LTS | 134001.87 | 2043.92 |
| Adaptive Decay LTS | 134253.28 | 1792.76 |
| LinUCB | 134256.95 | 1788.08 |
| Adaptive LinUCB | 134644.75 | 1401.28 |
| Decay LinUCB | 134639.68 | 1406.06 |
| Adaptive Decay LinUCB | 134847.37 | 1198.89 |

### 4.2　評価結果

シミュレーション結果を表 1 に示す．提案手法の適用によって，全ての解法において，適用前と比較して累積報酬が増加，累積リグレットが低下していることが見て取れる．ここで，非定常な変化への追従性を比較するため**図 1** に示す累積リグレットの推移を確認する．点線が従来の解法，同色の実線が提案手法を適用したものである．減衰や変化検出を用いない LTS や LinUCB の場合には，500 時点での線形パラメータの変化に対して 2000 時点においても累積リグレットが増加していることから，非定常な環境においては過去の観測結果が悪影響を及ぼすことがわかる．Decay LTS と Decay LinUCB では，過去の観測結果の減衰により累積リグレットの増加は抑えられたが，十分に減衰されるまでは累積リグレットの増加が見られた．これらの 4 つの解法と比較して，提案手法では，推定した線形パラメータに対する変化検出が有効に働き，累積リグレットの収束までの期間が短縮している．なお，LinUCB と比較して LTS の累積リグレットが全体として高いのは，LinUCB では推定した線形パラメータと試行回数から決定的に腕が選定されるが，LTS では推定した線形パラメータから確率的に腕が選定され，常に探索される可能性があるためである．

### 4.3　考察

非定常な変化があった腕 $a_1$ における各解法の振る舞いを**図 2** に示す．なお，図 2 では各解法の挙動の特性を明確にするため，500 回のシミュレーションの平均ではなく，基準となる LTS と LinUCB で最も累積リグレットの低かったシミュレーションの回の結果を示している．図の 1 段目と 2 段目はそれぞれ推定した線形パラメータの大きさと異常度の推移である．黒色の点線が真の線形パラメータから算出した値の推移を示している．500 時点の線形パラメータの変化後に，従来の解法では推定値が緩やかに真の値に近づき，提案手法では急激に収束した．これは，提案手法が，ADWIN-V によるウィンドウサイズの縮小によって直近の観測結果のみを利用できた効果である．図の 3 段目は提案手法における履歴の期間数の推移であり，線形パラメータの推定にどの程度直近までの試行結果を利用したかを表す．図の 1 段目と 2 段目で確認できた値の収束後に再度ウィンドウサイズが増加し，非定常な変化後の値が線形パラメータの推定に利用されていることがわかる．図の 4 段目は提案手法における報酬分布の変化検出と追従に必要な履歴データの要素数の推移を示している．点線が指数ヒストグラムを用いない場合の要素数，同色の実線は，用いた場合の要素数である．指数ヒストグラムを用いることで，履歴の期間数に対して必要なデータ数を大幅に削減できるため，履歴データの全体の要素数を抑え，データサイズの肥大化を抑制できた．一方で，変化検出時に履歴の期間が短くなった時に要素数の差が見られなくなった．全腕を通して変化検出の頻度が多いような環境においては平均ウィンドウ数を事前に計測した上で ExpHist-V の導入を検討する必要があると考えられる．

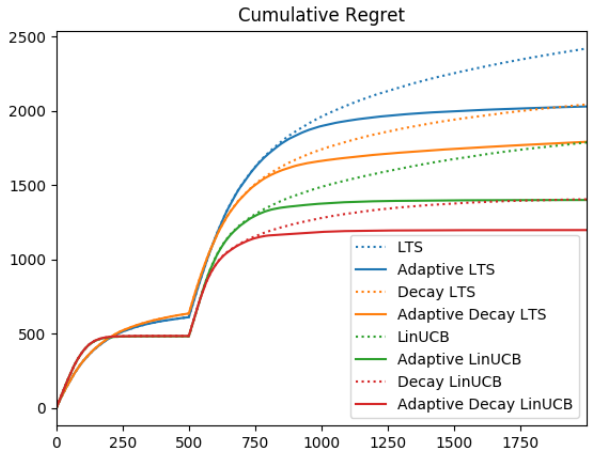


図 1　試行回数と累積リグレットのシミュレーション間比較
**Fig. 1**　Comparation of cumulative regrets.

## 5.　まとめ

本研究では，利用者の嗜好が多様かつ継続的に変化する環境において，推薦システムが利用者の要求に応えるため，多腕バンディット問題を線形かつ非定常な問題設定に拡張し，その解法を提案した．提案手法では，従来の線形な解法に変化検出の手法を組み合わせ変化への追従性を高めた．また，変化を観測する報酬の系列数が指数的に増加する課題を解決するため，推定した線形パラメータの値から方向データの異常度と大きさを求めることで，要因の組み合わせ数によらない変化検出を行った．加えて，過去期間の値を要約するデータ構造を導入し，報酬分布の変化検出と追従に必要な履歴データのサイズを低減した．

評価では，提案手法の適用により従来の解法と比較して，線形パラメータの変化に素早く追従し，累積報酬の増加と累積リグレットの減少が確認された．一方で，履歴の期間が短い場合，提案するデータ構造がデータサイズの低減に

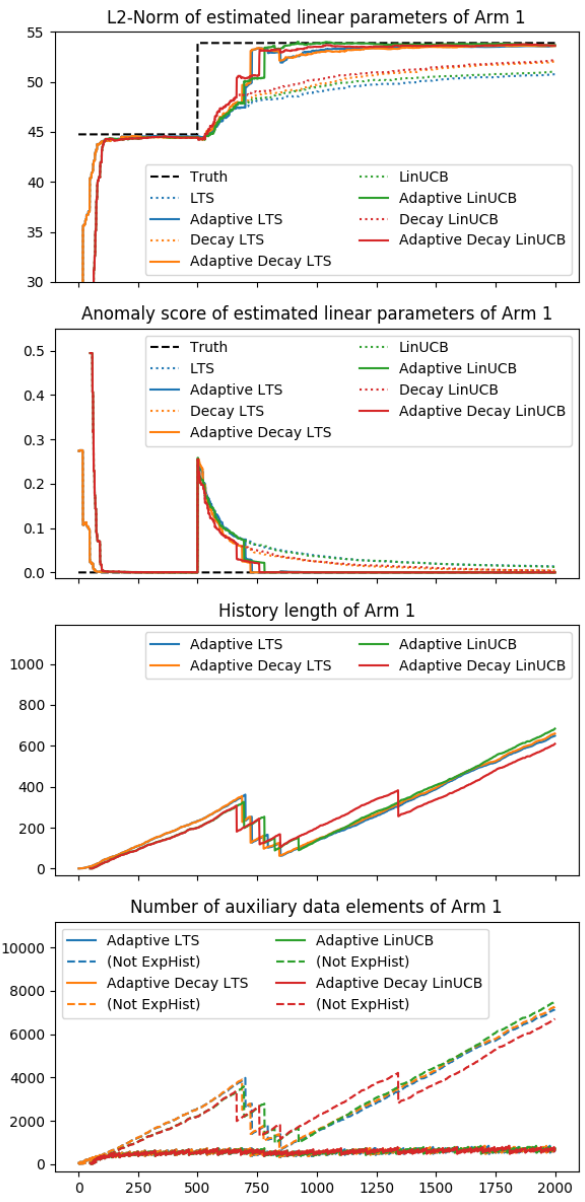


**図 2** 腕 $a_1$ における推定した線形パラメータとデータサイズのシミュレーション間比較

**Fig. 2** Comparation of estimated linear parameters, data size of arm $a_1$.

寄与できない課題も示された．今後は，更なるデータ表現の改善の検討と実システムでの有効性の評価を進めていく．

## 参考文献

[1] 経済産業省 商務情報政策局情報経済課. 平成 30 年度我が国におけるデータ駆動型社会に係る基盤整備（電子商取引に関する市場調査）, 2019.

[2] Michael N Katehakis and Arthur F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, Vol. 12, No. 2, pp. 262–268, 1987.

[3] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th international conference on World wide web*, pp. 21–30, 2009.

[4] 三宅悠介, 松本亮介. Synapse: 利用者の文脈に応じて継続的に推薦手法の選択を最適化する推薦システム. 研究報告インターネットと運用技術 (IOT), Vol. 2019-IOT-45, pp. 1–7, 2019.

[5] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

[6] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Explore/exploit schemes for web content optimization. In *2009 Ninth IEEE International Conference on Data Mining*, pp. 1–10. IEEE, 2009.

[7] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

[9] 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.

[10] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013.

[11] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pp. 174–188. Springer, 2011.

[12] Neha Gupta, Ole-Christoffer Granmo, and Ashok Agrawala. Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, Vol. 1, pp. 484–489. IEEE, 2011.

[13] Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

[14] Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michele Sebag. Multi-armed bandit, dynamic environments and meta-bandits, 2006. In *NIPS-2006 workshop, Online trading between exploration and exploitation, Whistler, Canada*, 2006.

[15] Edouard Fouché, Junpei Komiyama, and Klemens Böhm. Scaling multi-armed bandit algorithms. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1449–1459, 2019.

[16] Albert Bifet and Ricard Gavalda. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining*, pp. 443–448. SIAM, 2007.

[17] Muhammad Ammar Hassan. *Non-Stationary Contextual Multi-Armed Bandit with Application in Online Recommendations*. PhD thesis, University of Virginia, 2015.

[18] Tsuyoshi Idé and Hisashi Kashima. Eigenspace-based anomaly detection in computer systems. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 440–449, 2004.

[19] Mayur Datar, Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Maintaining stream statistics over sliding windows. *SIAM journal on computing*, Vol. 31, No. 6, pp. 1794–1813, 2002.