Vol.2020-IOT-50 No.2
2020/7/10

非定常な多腕バンディット問題における変化検出アプローチの線形モデルへの拡張

三宅 悠介[1,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かつ継続的に変化する課題に対処するため，利用者の文脈を考慮した問題設定への拡張と解法が提案されている．時間の経過に従い報酬分布が変化する非定常な問題設定の解法では，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変化に追従する．しかしながら，腕の文脈を複数の要因のパラメータの組み合わせで表現し，文脈に応じて報酬分布が決定する線形な問題設定にこの解法を適用する場合，要因の数に対して指数的に増加する全ての報酬の変化を観察しなければならない．本報告では，要因の組み合わせ数によらない単一の値の推移のみから報酬分布の変化を検出・追従することで，従来の線形な解法を利用可能でありながら，汎用的でメモリ効率に優れた非定常かつ線形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，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試行時点での係数を推定し，この値の和の推移から報酬分布の変化を検出する．また，報酬分布の変化に合わせた動的なハイパーパラメータ調整により迅速に変化に追従する．評価では，非定常かつ線形な多腕バンディット問題を設定し，変化検出を行わない場合と比較して性能が上回ることを確認した．

Extension of Change Detection Method for Non-Stationary Linear Multi Armed Bandit

Abstract: A multi-armed bandit problem is a problem that maximizes reward from making choices between candidates called arms. For the application of advertisement or recommendation system of this problem, the contextual extension of problem setting and policy is proposed to deal with variously and continuously changing user preferences. For non-stationary problems, policies of making choices follow changes by observing reward change using change detection methods. On the other hand, for a linear problem that changes reward distribution according to factors, the decision-maker of the policies must observe all reward changes that increase exponentially. In this report, we propose a non-stationary linear multi-armed bandit policy to detect a change of reward distribution from a single value transition that is independent of the number of reward changes. The proposed policy estimates coefficients of the factors from the number of trials and rewards and detects a change of reward distribution from the sum of the estimated values. Also, the policy can follow change rapidly using dynamic hyper-parameter turning. We set up a non-stationary and linear multi-armed bandit problem for the evaluation and confirmed our policy makes the performance increase.

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いる．推薦システムは，EC サイトが取り扱う多くの商品の中から，何らかの方策に基づき利用者が興味を持つ商品を提案することで，閲覧や購入行為を支援するシステムである．多様な利用者が訪問する EC サイトにおいては，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

[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
[a)] miyakey@pepabo.com
[b)] antipop@pepabo.com

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，事前に定めた要因の組み合わせ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]．この問題の解法では，要因ごとの報酬に対する重みを活用と探索によって求める．非定常な多腕バンディット問題では，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]．この問題の解法では，確率分布の変化を踏まえ，変化後の報酬を重視して変化に追従する．

この拡張された 2 つの問題設定は，多腕バンディット問題における文脈の考慮と捉えることができる．ここで，文脈とは，複数の要因のパラメータによって定まる状態のことを指す．基本的な多腕バンディット問題では，要因のパラメータを用いず，結果として常に一つの文脈に決定される．すなわち腕の報酬分布は一定である．線形な多腕バンディット問題では，いくつかの要因のパラメータから文脈が定まる．そして，文脈に対応して腕の報酬分布が決定する．要因のパラメータの組み合わせにより，表現できる文脈は指数的に増加する．非定常な多腕バンディット問題では，なんらかの理由により，時間の経過によって文脈に対応する報酬分布が変化していく．この意味で，線形な多腕バンディット問題は，文脈に対応する腕の報酬分布が時間の経過により変化しない問題設定と言い換えることができる．従来の解法ではいずれかの問題設定を扱うが，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．

文脈が複数の要因のパラメータによって決定され，時間経過によっても報酬分布が変化する環境に対して，多腕バンディット問題を適用するためには，従来の線形な解法と非定常な解法を組み合わせる必要がある．しかしながら，線形な多腕バンディット問題においては，要因のパラメータの組み合わせごとに文脈が異なるため，要因の数に対して報酬分布のパターンが指数的に増加する．非定常な多腕バンディット問題では，報酬分布の変化を観測するため，線形な多腕バンディット問題を扱う場合に，観測する報酬分布の系列数に比例して推薦システムのメモリ使用量が増加してしまう．また，利用者の文脈に応じて推薦システムが適応的に振る舞うためには，報酬分布の変化を迅速に検出し，文脈と報酬分布の新しい関係性を速やかに学習する必要がある．加えて，多腕バンディット問題には理論保証があり実績のある既存の解法が多数存在するため，これらの解法を非定常かつ線形な多腕バンディット問題に適用できることが望ましい．

本報告では，要因の組み合わせ数によらない単一の値の推移のみから報酬分布の変化を検出・追従することで，従来の線形な解法を利用可能でありながら，汎用的でメモリ効率に優れた非定常かつ線形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，各腕に対する試行回数と報酬から，それぞれの要因のパラメータに対する試行時点での係数を推定する．次に，この求めた値の和の推移のみから報酬分布の変化検出を行う．最後に，線形な解法に対して変化以前の観測値を取り除くことで，新しい報酬分布に追従させる．また，記録された観測値の数の減少に合わせて探索が重視されるよう動的にハイパーパラメータを調整する．評価では，非定常かつ線形な多腕バンディット問題を設定し，変化検出を行わない場合と比較して性能が上回ることを確認した．

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，非定常かつ線形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 1 回の試行で 1 つの腕を選択し，選択した腕から報酬を得る．それぞれの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として ϵ-Greedy アルゴリズム [7] が挙げられる．この解法では，腕の評価に報酬の標本平均を用いる．探索を行う割合は $0 \leq \epsilon \leq 1$ で指定され，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．この解法では，候補

の評価の差が明らかな場合に探索による機会損失が発生する．また，腕の評価において探索された回数を考慮しないことから特に探索初期において誤った腕の活用が起こる可能性がある．UCB1 アルゴリズム [8] は，選択回数の少ない腕を積極的に選択することでこの問題に対応する．この解法では腕の選択に，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．これにより，初期は積極的に探索し，充分な試行を経た後は，評価が高い腕が活用され，誤った判断や探索による機会損失を抑えることができる．UCB1 アルゴリズムでは，ある時点までの報酬から選択する腕が一意に定まる．しかしながら，報酬が遅れて反映される環境においては選定する腕が固定され，報酬がそれに対して反映されない期間に不利な腕を使い続ける可能性がある．Thompson Sampling[9] は，各腕が期待値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が最も大きかった腕を選定する．腕の選定が確率的に行われることから，報酬が遅れて反映される環境に起因する機会損失を低減することができる．

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，複数の文脈があり，文脈に応じて報酬分布が決定される多腕バンディット問題である．この問題では，各腕はそれぞれの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，要因のパラメータの値ベクトルとの内積の結果に誤差を加えた値として求める．本稿では，腕が持つ各要因に対する係数ベクトルを線形パラメータ，腕の選択時の要因のパラメータの値ベクトルをコンテキスト情報と呼ぶ．

文脈の種類が少ない場合には，基本的な多腕バンディット問題の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，

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

が最も大きくなる腕を選択する．ここで $\boldsymbol{b}$ はコンテキスト情報，$\hat{\boldsymbol{\theta}}_a$ は腕 a に対して推定した線形パラメータである．$\hat{\boldsymbol{\theta}} = \boldsymbol{B}_a^{-1}\boldsymbol{f}_a$ 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 $\boldsymbol{B}_a$ と各要因のパラメータの累積の報酬 $\boldsymbol{f}_a$ 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，$v^2\boldsymbol{B}^{-1}$ を分散とする多変量正規分布から $\tilde{\boldsymbol{\mu}}$ を求め，コンテキスト情報である $\boldsymbol{b}_i(t)$ との内積が最も大きくなる腕を選定する．$\hat{\boldsymbol{\mu}} = \boldsymbol{B}^{-1}\boldsymbol{f}$ であり，各要因のパラメータの累積の試行回数を記録する $\boldsymbol{B}$ と各要因のパラメータの累積の報酬 $\boldsymbol{f}$ から推定される．なお，LinUCB と同様にハイパーパラメータ $v^2(v^2 \geq 0)$ 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多腕バンディット問題である．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コンテキスト情報で指定することで適切に扱うことができる．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．Discounted UCB と Sliding Window UCB は，UCB1 アルゴリズムをこの問題に適用した提案である [11]．Discounted UCB 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．Sliding Window UCB では，これに加え，評価に利用する報酬系列に対してウィンドウを設け，ある時点以降の報酬のみを評価に利用する．Dynamic Thompson Sampling は，Thompson Sampling をこの問題に適用した提案である [12]．指数平滑法を用いて，直近 C 回までの結果を指数的に減衰しながら腕の評価に利用する．Discounted TS[13] は，Discounted UCB と同様に試行回数と報酬に割引の概念を導入した．

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ことため，既存の解法を非定常な環境に適用することができる．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]．この手法では，変化検出後に，腕の評価を初期状態に戻したものと戻さないものを一定期間比較し，変化検出の誤りに備え

るメタ・バンディットを採用した．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 はウィンドウ W に t 時点の値 x_t を記録する．記録時には，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 ϵ_{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，

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

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．

ただし，これらの非定常な多腕バンディット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，基本的な多腕バンディット問題を線形に拡張するのと同様に，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．また，これらの解法は活用と探索のトレードオフの解消を従来の解法の枠組みに従っている．すなわち，報酬分布の変化に伴い腕の試行回数が変更され，Thompson Sampling では腕の選定時の分散が，UCB1 では補正項が増加することで探索が重視される．しかしながら，腕の報酬分布の変化が頻繁に発生する環境では，変化時の積極的な探索と，変化がない時の試行結果の蓄積が重要であり，変化への迅速な追従に改善の余地が残ると考えられる．

3. 提案手法

本報告では，推薦システムの利用者の嗜好傾向のような，複数の文脈があり，文脈に対応する報酬分布が時間経過によっても変化する環境に対して適用可能な，非定常かつ線形な多腕バンディット問題の解法を提案する．そのために，非定常な多腕バンディット問題において，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 の線形な問題設定への拡張を図る．提案手法では，腕の報酬分布の変化を文脈ごとの報酬からではなく，推定した線形パラメータの値の和から検出する．これにより，要因のパラメータの数やその組み合わせである文脈の数によらず単一の値の推移から変化の検出が可能となる．また，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ことで，既存の線形な多腕バンディット問題の解法を非定常な環境へ適用することができる．加えて，腕の報酬分布の変化に従い，既存の線形な多腕バンディット問題の解法における探索と活用のバランスを調整するハイパーパラメータを動的に調整するアニーリング手法を導入する．これによって，変化時の探索と，変化がない時の活用を積極的に行い変化への迅速な追従を実現する．

ここで，提案手法の位置付けを整理するため，多腕バンディット問題設定と解法の適用領域の関係を表 1 に示す．1 列目は多腕バンディット問題設定が仮定する報酬の確率分布である．2 列目の解法では基本的な多腕バンディット問題である報酬分布が変化しない環境を扱う．3 列目の線形な解法ではこれに加え，要因のパラメータによって文脈が定まり，文脈に対応する報酬分布が決定する環境を扱うが時間の経過による報酬分布の変化は扱わない．これと対照的に，4 列目の非定常な解法では時間の経過による報酬分布の変化のみを扱っている．提案手法となる 5 列目の解法では，複数の要因のパラメータによる文脈の決定と文脈に対応する報酬分布の変化が同時に起こる，非定常かつ線形な多腕バンディット問題への適用を実現する．

3.1 推定した線形パラメータによる変化検出と探索ハイパーパラメータの動的な調整

提案手法のアルゴリズムを Algorithm1 に示す．d 次元のコンテキスト情報を持つ腕 i それぞれに，$\boldsymbol{B}_i$，$\boldsymbol{f}_i$，$\boldsymbol{A}_i$ が設定される．ここで $\boldsymbol{B}_i$ は次元数をコンテキスト情報の次元数 d と同じにする単位行列，$\boldsymbol{f}_i$ は d 次元のベクトルである．$\boldsymbol{B}_i$ はその腕 i におけるコンテキスト情報の各次元が試行された累積回数を，$\boldsymbol{f}_i$ はコンテキスト情報の各次元で得た累積報酬が記録される．また，$\boldsymbol{A}_i$ は腕 i に対して推定した線形パラメータの和の推移を ADWIN をデータ構造に用いて記録する．なお，$\boldsymbol{B}$，$\boldsymbol{f}$，$\boldsymbol{A}$ は，全ての腕の対応する行列，ベクトルを保持する配列である．

提案手法は，既存の線形な多腕バンディット問題の解法によって選定された腕 i とその報酬 r を得た後に呼ばれる一連の処理として定義される．はじめに，腕の報酬分布の変化検出のため，線形パラメータの推定を行う．このために，選定した腕についてコンテキスト情報 $\boldsymbol{b}$ と報酬 r から，試行回数 $\boldsymbol{B}_i$ と累積報酬 $\boldsymbol{f}_i$ を更新する．そして，$\boldsymbol{B}_i^{-1}\boldsymbol{f}_i$ によって得られた d 次元のベクトルを線形パラメータの推定値として扱う [5][12]．提案手法では，この推定値の各次元の和 M の推移を用いて腕の報酬分布の変化を検出する．これは Algorithm1 の 4 行目に該当する．推定した線形パラメータの和を用いることで，文脈の種類によらない単一の時系列による省メモリな変化検出を実現する．次に，この値の推移に対して ADWIN の統計的仮説検定による変化検出を行う．変化を検出した場合，ADWIN は変化前の系列データを削除する．提案手法では，変化後の系列データ長 $|\, \boldsymbol{A}_i \,|$ を腕 i の新しい試行回数とみなし，その期間に記録されたコンテキスト情報と報酬から試行回数 $\boldsymbol{B}_i$ と累積報酬 $\boldsymbol{f}_i$ を更新する．

提案手法では，腕の報酬分布の変化を検出した際，腕の

表 1　多腕バンディット問題の問題設定と解法
Table 1　Problem settings and policies of multi-armed bandit problem.

報酬の確率分布	ϵ-Greedy[7] UCB1[8] Thompson Sampling[9]	LinUCB[5] Linear- Thompson Sampling[10]	Discounted UCB[11] Sliding Window UCB[11] Dynamic- Thompson Sampling[12] Discounted TS[13] Adapt-EvE[14] S-TS-ADWIN[15]	LTS- ADWIN-Anneal LinUCB- ADWIN-Anneal
一定	✓	✓	✓	✓
文脈により決定		✓		✓
時間経過で変化			✓	✓

試行回数が一旦減少する点に着目し，腕の試行回数に従った，探索と活用のバランスを調整するハイパーパラメータの動的な調整を行う．これは Algorithm1 の 10 行目に該当する．この式では，腕ごとの試行回数と反する形で値 τ が増減する．既存の多腕バンディット問題の解法における，探索と活用のバランスを調整するハイパーパラメータにこの値を用いることで，腕の試行回数の増加に伴い活用を，試行回数の減少が起きた場合に，探索が重視することが可能となる．なお，本手法におけるハイパーパラメータである $\beta(>0)$ により最終的に求まる値の大きさを調整する．

Algorithm 1 Linear ADWIN Anneal
function LINEAR-ADWIN-ANNEAL($i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{f}, \boldsymbol{b}, r, \beta$)
  Update $\boldsymbol{B}_i = \boldsymbol{B}_i + \boldsymbol{b}\boldsymbol{b}^\mathsf{T}$
  Update $\boldsymbol{f}_i = \boldsymbol{f}_i + \boldsymbol{b}r$
  $M = \sum^d (\boldsymbol{B}_i^{-1}\boldsymbol{f}_i)_d$
  Add M into $\boldsymbol{A}_i$
  if $\boldsymbol{A}_i$ detects change **then**
    $\boldsymbol{A}_i$ shrinks window
    Update $\boldsymbol{B}_i = \sum_{t-|\boldsymbol{A}_i|}^{t} \boldsymbol{b}\boldsymbol{b}^\mathsf{T}$
    Update $\boldsymbol{f}_i = \sum_{t-|\boldsymbol{A}_i|}^{t} \boldsymbol{b}r$
  end if
  $\tau = \sum_{i=0}^{|\boldsymbol{A}|} \frac{\beta}{\ln|\boldsymbol{A}_i|+1}$
  return $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{f}, \tau$
end function

3.2　既存の線形な多腕バンディット問題の解法の拡張

3.1 節で提案した手法は，変化検出と動的なハイパーパラメータの算出を，利用する多腕バンディット問題の解法に依存しない．そのため，既存の線形な多腕バンディット問題の解法を拡張することができる．本節では，線形な多腕バンディット問題の解法である，Linear Thompson Sampling，LinUCB に対して，提案手法を適用し，非定常かつ線形な多腕バンディット問題の解法へと拡張する．

拡張した解法のアルゴリズムを Algorithm2 と Algorithm3 に示す．d 次元のコンテキスト情報を持つ K 本の腕それぞれに，$\boldsymbol{B}_i$，$\boldsymbol{f}_i$，$\boldsymbol{A}_i$ が設定される．Linear Thompson Sampling では $\boldsymbol{B}^{-1}\boldsymbol{f}$ を平均，$\boldsymbol{B}^{-1}$ を分散とする多変量正規分布から $\tilde{\boldsymbol{\mu}}$ を求め，t 回目の試行における腕 i への d 次元のコンテキスト情報である $\boldsymbol{b}_i(t)$ との線形和が最も大きくなる腕を選定する．LiUCB では腕ごとに式 1 から求めた値が最も大きくなる腕を選定する．報酬 r を得た後に，提案手法を用いて腕の線形なパラメータの推定と報酬分布の変化検出を行う．提案手法では，これらの解法で用いられていた線形なパラメータの推定方法を用いているため，推定に用いた $\boldsymbol{B}$，$\boldsymbol{f}$ の結果をそのまま用いることができ，提案手法との親和性が高い．また，腕の選定には提案手法により得られた動的なハイパーパラメータ値 τ を用いることで，試行回数の増加に伴い活用，試行回数の減少が起きた場合に探索を重視することが可能となる．

Algorithm 2 LTS-ADWIN-Anneal
Require: Set of arms [K], d-dimentional context
ADWIN confidence value δ
Anneal parameter β
Set $\boldsymbol{B}_i = I_d \quad \forall_i \in [K]$
Set $\boldsymbol{f}_i = 0_d \quad \forall_i \in [K]$
$\boldsymbol{A}_i \leftarrow$ instance of ADWIN with δ, $\forall_i \in [K]$
for all $t = 1, 2, ...,$ **do**
  for all $i = 1, 2, ..., K$ **do**
    Sample $\tilde{\boldsymbol{\mu}}_i(t)$ from distribution $\mathcal{N}(\boldsymbol{B}_i^{-1}\boldsymbol{f}_i, \tau\boldsymbol{B}_i^{-1})$.
  end for
  Play arm $a(t) := argmax_i \ \boldsymbol{b}(t)^\mathsf{T}\tilde{\boldsymbol{\mu}}_i(t)$, and observe reward r_t.
  Update $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{f}, \tau$ = Linear-ADWIN-Anneal$(a(t), \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{f}, \boldsymbol{b}(t), r_t, \beta)$
end for

4.　評価

4.1　評価環境

提案手法による，非定常で線形な多腕バンディット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，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

Algorithm 3 LinUCB ADWIN Anneal

Require: Set of arms [K], d-dimentional context
ADWIN confidence value δ
Anneal parameter β
Set $\boldsymbol{B}_i = I_d \quad \forall_i \in [K]$
Set $\boldsymbol{f}_i = 0_d \quad \forall_i \in [K]$
$\boldsymbol{A}_i \leftarrow$ instance of ADWIN with δ, $\forall_i \in [K]$
for all $t = 1, 2, ...,$ **do**
  for all $i = 1, 2, ..., K$ **do**
    $\hat{\boldsymbol{\theta}}_i(t) = \boldsymbol{B}_i^{-1}\boldsymbol{f}_i$
    $LinUCB_i(t) = \boldsymbol{b}(t)^\top \hat{\boldsymbol{\theta}}_i(t) + \tau\sqrt{\boldsymbol{b}(t)^\top \boldsymbol{B}_i^{-1}\boldsymbol{b}(t)}$
  end for
  Play arm $a(t) := argmax_i\ LinUCB_i(t)$, and observe reward r_t.
  Update $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{f}, \tau$ = Linear-ADWIN-Anneal$(a(t), \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{f}, \boldsymbol{b}(t), r_t, \beta)$
end for

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，$\boldsymbol{\theta}_0 = [14, 15, 16, 17, 18, 19, 20, -10]$，$\boldsymbol{\theta}_1 = [12, 13, 14, 15, 16, 17, 18, 10]$ とするが，非定常な環境とするため，500 時点目に $\boldsymbol{\theta}_1$ の 8 次元目のみ-11 に変更する．すなわち $\boldsymbol{\theta}_1 = [12, 13, 14, 15, 16, 17, 18, -11]$ となる．コンテキスト情報 $\boldsymbol{b}$ 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 $\boldsymbol{\theta}_i^\top \boldsymbol{b}_t + \epsilon_t$ となる．ここで誤差項 ϵ_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，この非定常で線形な多腕バンディット問題に対し，3.2 節で提案した提案手法 (以下，LinUCB ADWIN Anneal，LTS ADWIN Anneal) を解法として利用する．また，比較のため，線形な多腕バンディット問題の解法である LinUCB ならびに，Linear Thompson Sampling(以下，LTS) を評価する．合わせて，非定常な多腕バンディット問題の解法である S-TS-ADWIN を評価するが，本評価の設定に合わせ，報酬分布に正規分布を仮定し，試行時の腕の選択本数の判定処理を除いた解法（以下，TS ADWIN) を用いる．また，同様の拡張を UCB1 に施した解法（以下，UCB1 ADWIN）も評価する．なお，文脈を扱えない TS ADWIN と UCB1 ADWIN については，文脈ごとに別の腕として計測することで擬似的に文脈を考慮した．

各解法のハイパーパラメータは予備実験によって求めた．予備実験では，5，10，20，30，40，50，75，100，150，200 のうち，累積報酬が最も高く，累積リグレットが最も低くなった値を利用した．各解法におけるハイパーパラメータは LinUCB で $\alpha = 20$，LTS で $v^2 = 50$，LinUCB ADWIN Anneal では $\beta = 30$，LTS ADWIN Anneal では $\beta = 100$ となった．UCB1 ADWIN と TS ADWIN ではハイパーパラメータが存在しない．また，ADWIN のハイパーパラメータ δ には誤検知が少なくなるような値として $\delta = 0.001$ を用いた．ただし，UCB1 ADWIN と TS ADWIN について，文脈ごとに試行が分散することから同じ設定で変化検出が行われなかったため，$\delta = 0.5$ とした．なお，予備実験のシミュレーション回数は 100 回である．

表 2　累積報酬と累積リグレット
Table 2　Cumulative reward and regret.

解法	累積報酬	累積リグレット
LTS	108248.94	4003.44
TS ADWIN	107526.03	4725.26
LTS ADWIN Anneal	111431.18	821.68
LinUCB	108731.40	3522.99
UCB1 ADWIN	108092.71	4157.86
LinUCB ADWIN Anneal	111699.39	552.76

4.2　評価結果

シミュレーション結果を表 2 に示す．提案手法が共に，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．シミュレーションの累積リグレットの推移を**図 1** に示す．また，ハイパーパラメータの推移を**図 2** に示す．図 1 より，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 2,000 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．一方で，変化検出を用いたとしても文脈ごとに異なる腕を用いた TS ADWIN と UCB1 ADWIN では，文脈ごとに試行が分散することから評価の蓄積が遅れ，リグレットの収束に時間がかかっている．これらと比較して，提案手法では，推定した線形パラメータの値の和での変化検出が有効に働き，おおよそ 575 時点において累積リグレットの増加が収束傾向になっている．また，図 2 より，提案手法によるハイパーパラメータの動的な調整によって線形パラメータに変化がない期間中は探索を抑え，変化時に探索を重視することで不要な探索を省き，期間全体で累積リグレットを低く抑えたことが見て取れる．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ることで常に探索が行われる可能性があるためである．

4.3　考察

それぞれの腕における各解法の振る舞いを**図 3** と**図 4** に示す．図の上段は推定した線形パラメータの和の推移である．黒色の点線が真の線形パラメータの和を示している．図の下段はウィンドウサイズの推移であり，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

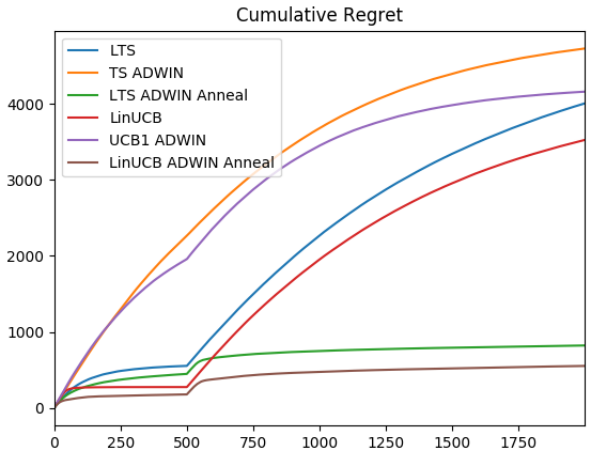

図 1 試行回数と累積リグレットのシミュレーション間比較
Fig. 1 Comparation of cumulative regrets.

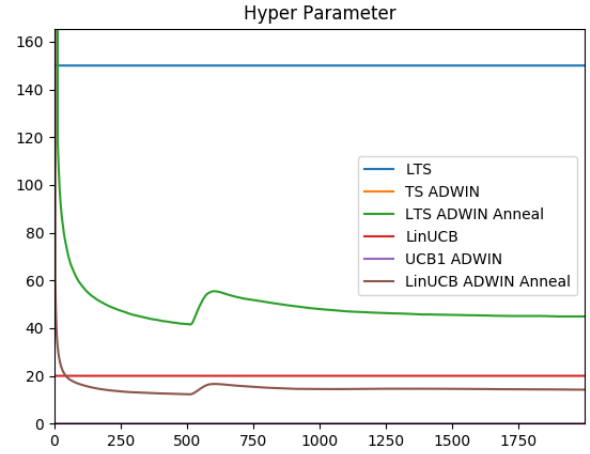

図 2 ハイパーパラメータの推移のシミュレーション間比較
Fig. 2 Comparation of hyper parameter.

す．なお，ADWIN を用いない LTS と LinUCB では過去の試行が全て反映されるためウィンドウサイズは腕の試行回数に等しい．

図 3 の腕 a_0 の線形パラメータの推定について，期間中に線形パラメータに変化がなかった腕 a_0 では各解法で真の値に収束した．また，提案手法でのウィンドウサイズについて，腕の試行回数に対して少ないことから継続的に小幅な縮小が行われていることがわかる．報酬誤差を変化として検出していると考えられるため，報酬誤差がより大きくなる環境では，より小さい δ の値を検討する必要がある．図 4 の腕 a_1 では，500 時点の線形パラメータの変化後に，LTS と LinUCB は緩やかな低下を示し，提案手法では急激な低下を示した．これは，提案手法が，ADWIN 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を利用できた効果である．しかしながら，今回のシミュレーションでは提案手法は線形パラメータの真の値に対して低い値を

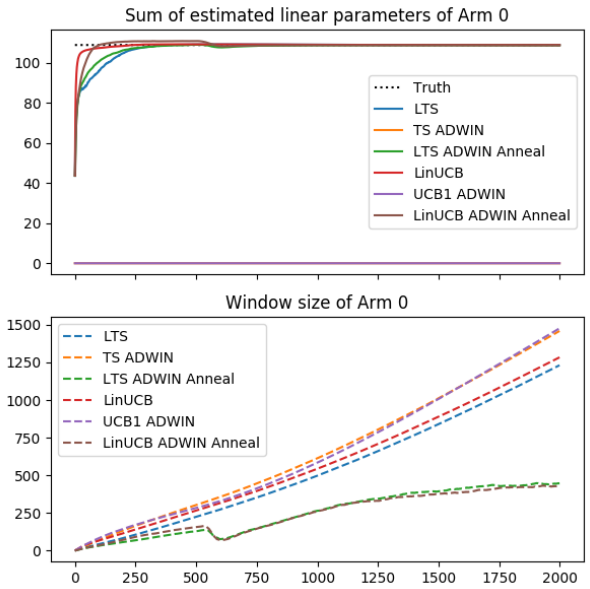

図 3 腕 a_0 における推定した線形パラメータとウィンドウサイズのシミュレーション間比較
Fig. 3 Comparation of estimated linear parameters, window size of arm a_0.

推定した．これは，腕 a_1 において線形パラメータの推定に必要な十分な試行が行われなかったことに起因している．今回のシミュレーション設定では，線形パラメータの変化後に腕 a_1 はどのようなコンテキスト情報であっても腕 a_0 より少ない報酬となるため，各解法において選択される回数は少なくなる．実際に，500 時点以降の試行期間を通した腕 a_1 の累積選択数は LTS ADWIN Anneal で 70 回，LinUCB ADWIN Anneal で 80 回程度であった．また，十分なウィンドウサイズを確保できなかったことも影響したと考えられる．図 4 の下段より，500 時点以降のウィンドウサイズが LTS ADWIN Anneal で 40，LinUCB ADWIN Anneal で 30 程度を推移しており，腕の選択回数の約半分程度の試行結果から推定を行っていることがわかる．

5. まとめ

本報告では，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非定常かつ線形な問題設定に拡張し，その解法を提案した．提案手法では，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和を用い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．また，変化時の積極的な探索と変化がない時の活用を効果的に切り替えるため，探索に関するハイパーパラメータの動的な調整手法を提案した．

非定常かつ線形な多腕バンディット問題の評価において，この環境を想定しない従来手法と比較して，提案手法が累積報酬を増やし，累積リグレットを減少させることが確認

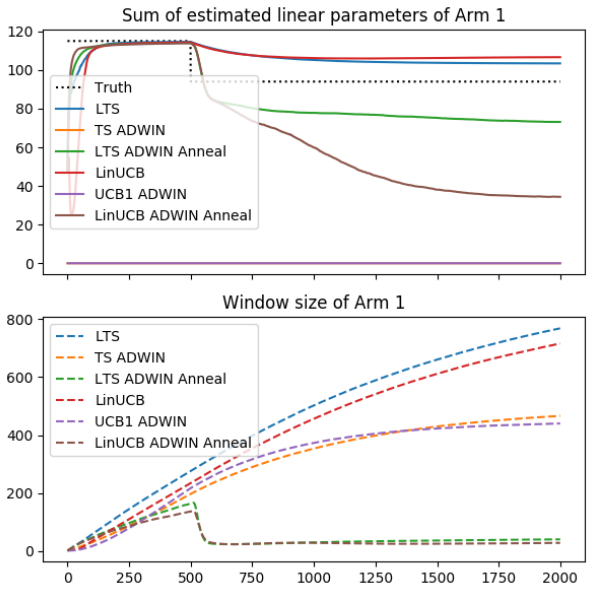

図 4　腕 a_1 における推定した線形パラメータとウィンドウサイズのシミュレーション間比較

Fig. 4　Comparation of estimated linear parameters, window size of arm a_1.

された．考察の結果，精度の高い線形パラメータの推定にはウィンドウサイズの調整に課題が残ることが示された．

研究報告時点では，線形パラメータの値の和が多次元の時系列データの変化検出に有効である理論的な裏付けは行えていない．多変量時系列に対する変化検出の更なる調査が必要である．また，ADWIN に記録する値を ADWIN のウィンドウサイズに依存した値から求めていることで，変化検出に関する安定性の低下を招いている可能性もあることから，この点についても方式の改善を図っていきたい．今後は，上述の課題の解決に加え，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.