仮想的な探索を用いて文脈や時間の経過による番狂わせにも迅速に追従する多腕バンディット手法

三宅 悠介[†,a]　　峯 恒憲[††,b]

† 九州大学 大学院システム情報科学府 情報知能工学専攻/GMO ペパボ株式会社 ペパボ研究所
†† 九州大学 大学院システム情報科学研究院 情報知能工学部門

a) *miyakey@pepabo.com*　b) *mine@ait.kyushu-u.ac.jp*

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 サービスへの応用では，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，文脈や時間の経過を考慮した問題設定への拡張と方策が提案されている．しかし従来の方策は，腕の相対的な有用性が逆転する環境で，不充分な追従性や非効率な探索に起因する機会損失が増加してしまう．本研究では，このような番狂わせを含む環境であっても機会損失を低減可能な方策を提案する．提案手法では，線形カルマンフィルタを用いた継続的な状態推定によって文脈や時間の経過に応じた変化に迅速に追従する．さらに，状態推定の欠損値処理を仮想的な探索に見立て，探索効率を高める．評価では，方策の追従性と探索効率を分析するための新たな指標を導入し，これらが従来の方策と比べて提案手法により改善することを確認した．

キーワード　多腕バンディット問題，線形カルマンフィルタ，非定常，コンテキスト

1　はじめに

適応的なシステムの実現には，利用者と情報システムが互いの状態をよく理解するためのコミュニケーションと，それに応じた振る舞いの変更が必要となる．一方で，そのコミュニケーションから得られる情報や利益の価値を考慮しなければならない環境では，確実な情報に基づく振る舞いを選択しつつ，まだ得られていない情報を引き出すために，価値の低いコミュニケーションを敢行しなければならない．このような，振る舞いの候補に対する利用と評価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．この問題は，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，プレイヤーは試行の結果からこの確率分布を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．この問題に対する方策では，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 Web サービスへの応用では，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，文脈や時間の経過を考慮した問題設定への拡張がなされている．この拡張された問題設定は，文脈付き，かつ，非定常な多腕バンディット問題と呼ばれ，いくつかの方策が提案されている．これらの方策では，文脈や時間の経過に応じて迅速に腕の評価を更新する性能が注目されてきた．一方で，この問題設定では番狂わせが発生する場合がある．ここで番狂わせとは，文脈や時間の経過に応じて腕の相対的な有用性が逆転する状況の中で，最善の腕の有用性は維持されながらも，その他の腕の有用性が向上した結果，逆転する状況を指す．探索を減らして機会損失を抑える方策では，この状況を察知するのは難しく，全ての腕の全ての文脈で一律に継続的に注意を払うような方策では，全体的な機会損失を増加させてしまう．また，これらの不充分な追従性や非効率な探索に加え，試行ごとの評価の更新処理に時間がかかる方策も多い．そのため，この番狂わせな環境に対応するには，評価の更新に加え，変化の検出も迅速かつ効率的に，軽量な機構により行われる必要がある．

本研究では，仮想的な探索を用いて文脈や時間の経過による番狂わせにも効率的かつ迅速に追従する多腕バンディットの方策を提案する．提案手法では，文脈に応じて推定した状態に基づき腕を選定する．状態の推定に，軽量で多変量の特徴量を扱うことができる線形カルマンフィルタを用いることで文脈の考慮と実行時間の短縮を図る．また，線形カルマンフィルタの欠損値処理を仮想的な探索と見立て，選定されなかった腕に対しても評価を更新することで，評価の低い腕に対する長期的な探索を促しながらも，実際の探索を最小限に抑えられる．

評価では，文脈と時間の経過に応じて候補の有用性が変化するシミュレーションを実施し，提案手法と従来の

方策を比較した．比較のため，方策の追従性と探索効率を分析するための新たな指標を導入し，これらが従来の方策と比べて提案手法により改善することを確認した．

本報告の構成を述べる．2 節で多腕バンディット問題における番狂わせの課題について述べる．3 節では，2 節で述べた課題を解決する提案手法について述べる．4 節では提案手法の評価を行い，5 節でまとめる．

2 多腕バンディット問題における番狂わせと，その課題

多腕バンディット問題では，多様かつ継続的に変化する利用者の嗜好傾向のような，報酬分布の変化を想定した問題設定について，文脈と非定常性という 2 種類の観点から拡張が図られている．文脈付き多腕バンディット問題では，複数の要因パラメータからなる文脈に応じて腕から得られる報酬分布が決定される．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．利用者の嗜好傾向が多様かつ継続的に変化する環境において最適な候補を選定するには，文脈付き，かつ，非定常な多腕バンディット問題の方策を用いる必要がある．

これらの方策で非定常性の解決のために採用される様々な方式は，いずれも変化後の報酬分布から得られる報酬のサンプルを一定数必要とする．機会損失の低減を目的とした多腕バンディット問題の設定では，各方策は，ある時点で評価の高い腕を最も多く活用する．そのため，ある時点で最適腕の有用性のみが低くなるような状況では，報酬のサンプルを十分に観測することができ，いずれの方式も有効に働く．反対に，評価の低かった腕が，ある時点で有効性が高くなるような番狂わせの状況では，新たな最適腕に対する報酬のサンプルを得るまでに時間が掛かり，以前の最適腕を選定し続ける機会損失が発生してしまう．そこで，番狂わせの検出のために一定割合の探索機会を設ける方式 [1] も提案されているが，探索を一律に増加させるため，相対的な腕の評価が逆転しない期間での機会損失につながってしまう．そのため，方策を番狂わせに対応させるためには，評価の低い腕の有効性の変化を素早く察知できるよう探索を行いつつ，その探索に伴う機会損失を減らすことが求められる．このような方策は，推薦手法のコールドスタートのように，一定の嗜好情報が蓄積されるまでその有効性が現れない選択肢を比較評価する状況で必要になると考えられる [2]．以降では，従来の方策での，番狂わせへの対応の課題を個別に分析する．

Time varying Thompson sampling (TVTP) [3]，Adaptive Thompson Sampling (AdTS) [4] は評価の更新のみに着目した手法である．これらは，非定常性を前提としたモデルの導入や履歴の削除によって，過去に観測した報酬に捉われずに腕の再評価を迅速に行うが，評価の低い腕に探索を促す仕組みを備えていない．また，迅速な評価の更新のため導入される粒子フィルタやブートストラップといった機構の実行に時間がかかるという課題がある．Decay LinUCB [5] は報酬に対する重み付けにより腕の再評価を迅速に行う．この手法では，腕の相対的な選定回数の差に基づき始めは探索が促されるが，選定回数に対する減衰操作により徐々に探索が減ってしまう．

Dynamic Linear UCB (dLinUCB) [6] と Dynamic Ensemble of Bandit Experts (DenBand) [7] は報酬予測の誤差の変化を検出し，新たな報酬分布に適したバンディットのモデルを追加する．追加されたモデルでは変化後の報酬を利用するため，過去に観測した報酬に捉われないが，同時に，全ての腕に対する評価がやり直しとなる．また，これらの手法は，徐々に変化検出の閾値を下げることから変化の有無によらず定期的にモデル追加と再評価を発生させる．よって間接的に評価の低い腕に探索を促すと見なせるものの，不要な探索が発生する．

腕の報酬が確率的ではなくプレイヤーの方策を知る敵対者によって決定される敵対的バンディットと呼ばれる問題への方策である，ADA-ILTCB+ [8] は，非定常な問題設定にも用いることができる．この方策は，番狂わせの発生も常に想定する必要があることから評価の低い腕に対する探索を積極的に行う．しかしながら，同様の仮定から慢性的に探索が増加する傾向が見られる．また，文脈を扱う場合，全ての状態と腕の組み合わせから最適なペアを推定するため，組み合わせが爆発し実行時間が指数的に増加する課題がある．

KF-MANB [9] は，カルマンフィルタを用いて現在の状態を継続的に推定し，確率一致法と組み合わせることで非定常な問題を扱う手法である．この手法では，カルマンフィルタの欠損値処理の仕組みを導入し，選定されなかった腕に対し長期的に探索が促されるよう評価を更新できる．このことから，変化の検出，評価の更新をそれぞれ迅速かつ効率的に行える．さらに，カルマンフィルタの状態推定は軽量であることから，実行時間の課題も生じにくいという利点がある．しかしながら，この手法では一次元の状態しか扱えないため，文脈によって報酬分布が異なるような文脈付きの問題設定の場合に対応できない．

3 提案手法

3.1 確率一致法による線形カルマンフィルタとの統合

線形カルマンフィルタは，観測時の誤差を含む時系列データに対し，時系列の観測値ならびにその背後にある

Algorithm 1: Linear KalmanFilter Bandits

1 **procedure** `MAIN():` ▷ `main entry`
2 Initialize K arms with $\boldsymbol{\mu}_1, P_1, Z(\boldsymbol{x}_t), H, T, R(\boldsymbol{x}_t), Q$.
3 **for** $t \leftarrow 1, T$ **do**
4 Get $\boldsymbol{x}_t$.
5 $a^{(k)} = \arg\max_{j=1,K} SAMPLE(a^{(j)}, \boldsymbol{x}_t)$
6 Receive r_t by pulling arm $a^{(k)}$.
7 **for** $j \leftarrow 1, K$ **do**
8 $FILTER(\boldsymbol{x}_t, a^{(j)}, r_t, j == k)$.
9 $PREDICT(\boldsymbol{x}_t, a^{(j)})$.
10 **procedure** `SAMPLE(`$a^{(j)}, \boldsymbol{x}_t$`):` ▷ `sample for` $a^{(j)}$`, given` $\boldsymbol{x}_t$`.`
11 Sample $\boldsymbol{s}_t^{(j)} \sim \mathcal{N}_m(\boldsymbol{\mu}_t^{(j)}, \beta P_t^{(j)})$.
12 **return** $\boldsymbol{x}_t^{\mathsf{T}} \boldsymbol{s}_t^{(j)}$.
13 **procedure** `FILTER(`$\boldsymbol{x}_t, a^{(j)}, r_t, observed$`):` ▷ `filter the state, given` $\boldsymbol{x}_t, r_t$`.`
14 **if** *observed* **then**
15 $Z_t = Z(\boldsymbol{x}_t)$
16 Get filtered estimator $\boldsymbol{\mu}_{t|t}$, $P_{t|t}$ for $a^{(j)}$ by Equation (3).
17 **else**
18 $\boldsymbol{\mu}_{t|t}^{(j)} = \boldsymbol{\mu}_t^{(j)}$
19 $P_{t|t}^{(j)} = P_t^{(j)}$
20 **procedure** `PREDICT(`$\boldsymbol{x}_t, a^{(j)}$`):` ▷ `predict the state, given` $\boldsymbol{x}_t$`.`
21 $R_t = R(\boldsymbol{x}_t)$
22 Get predictor $\boldsymbol{\mu}_{t+1}$, P_{t+1} for $a^{(j)}$ by Equation (4).

状態の推定を行う手法の一つである．この手法では，対象の時系列を以下の状態空間モデルで表現する．

$$\boldsymbol{y}_t = Z\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_p(\boldsymbol{0}, H) \tag{1}$$

$$\boldsymbol{\alpha}_{t+1} = T\boldsymbol{\alpha}_t + R\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}_r(\boldsymbol{0}, Q) \tag{2}$$

ここで $\boldsymbol{y}_t \in \mathbb{R}^p$ と $\boldsymbol{\alpha}_t \in \mathbb{R}^m$ は t 時点の観測値と状態を，$\boldsymbol{\epsilon}_t$ は平均 $\boldsymbol{0}$ 分散共分散 $H \in \mathbb{R}^{p\times p}$，$\boldsymbol{\eta}_t$ は平均 $\boldsymbol{0}$ 分散共分散 $Q \in \mathbb{R}^{r\times r}$ の多変量正規分布 $\mathcal{N}$ から得られる誤差を表す．なお，本稿では $\mathcal{N}$ の添字は変量の次元数を示す．また，$Z \in \mathbb{R}^{p\times m}$ は，状態空間から観測値空間への写像，$R \in \mathbb{R}^{m\times r}$ は，誤差空間から状態空間への写像，$T \in \mathbb{R}^{m\times m}$ は，時点の経過に伴う状態の推移を表現している．線形カルマンフィルタは，t 時点までに得られた観測値から，この状態空間モデルで表現された時系列ならびにその背後にある状態を逐次的に推定する．推定は，与えられた $t = 1$ 時点の初期状態 $\boldsymbol{\alpha}_1$ の平均 $\boldsymbol{\mu}_1$ と分散共分散 P_1 を起点とし，フィルタリングと一期先予測と呼ばれる操作を交互に行う．

フィルタリングは，t 時点での観測値 $\boldsymbol{y}_t$ と予測した観測値 $Z\boldsymbol{\mu}_t$ の誤差から，その時点での状態を推定する操作である．この操作によって得られる状態 $\boldsymbol{\alpha}_t$ の平均と分散共分散をフィルタ化推定量と呼び，それぞれ $\boldsymbol{\mu}_{t|t}$，$P_{t|t}$ と表す．フィルタ化推定量の算出は以下の通り．

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{t|t}} &= \boldsymbol{\mu}_t + K\boldsymbol{v}_t \\ P_{t|t} &= P_t - KF_tK^{\mathsf{T}} \end{aligned} \tag{3}$$

ここで，$\boldsymbol{v}_t = \boldsymbol{y}_t - Z\boldsymbol{\mu}_t$，$F_t = ZP_tZ^{\mathsf{T}} + H$，$K = P_tZ^{\mathsf{T}}F_t^{-1}$ とした．

一期先予測は，先に求めたフィルタ化推定量を用いて $t+1$ 時点の状態を推定する操作である．一期先予測における状態 $\boldsymbol{\alpha}_{t+1}$ の平均 $\boldsymbol{\mu}_{t+1}$ と分散共分散 P_{t+1} の算出は以下の通り．

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &= T\boldsymbol{\mu}_{t|t} \\ P_{t+1} &= TP_{t|t}T^{\mathsf{T}} + RQR^{\mathsf{T}} \end{aligned} \tag{4}$$

提案手法では，この推定した状態を多腕バンディット問題における腕の選定に利用する．はじめに，腕ごとに線形カルマンフィルタによって推定された状態の平均と分散共分散を得る．次に，それらを平均と分散共分散とする多変量正規分布からサンプリングを行う．なお，多腕バンディット問題としての探索のバランスを調整できるよう，分散共分散に対してスケール項 β を設けた．最後に，サンプリングされた値とコンテキスト $\boldsymbol{x}_t$ との内積が最も大きかった腕を選定する．これらは，線形カルマンフィルタの状態推定値を用いた確率一致法とみなすことができる．なお，この工程は Algorithm1 の 5 行目ならびに 10 から 12 行目に相当する．

3.2 欠損値処理による仮想的な探索

線形カルマンフィルタは，ある時点で観測値が得られない場合も欠損値として適切に扱うことができる．提案手法では，この欠損値処理を多腕バンディット問題において選定されなかった腕に対する評価の更新操作として取り入れる．この仮想的な探索による評価更新は，各時点で選定した腕の評価のみを更新する従来の多腕バンディットの方策に比べ，以下の 2 つの効果が期待できる．

第一は，番狂わせの早期検出である．線形カルマンフィルタでは，欠損値に対するフィルタ化推定量は，式 (3) の結果ではなく t 時点の状態として推定した平均 $\boldsymbol{\mu}_t$ と分散共分散 P_t となる．一方，一期先予測は観測値を得た場合と同様に行う．状態の分散共分散に着目すると，

フィルタリングは分散共分散を小さく，一期先予測は大きくするよう更新する．そのため，選定されない腕では一期先予測のみ行われることで分散共分散が継続的に増加し，確率一致法の仕組みにより該当する腕の探索が促される．

第二は，評価の更新に伴う機会損失の低減である．線形カルマンフィルタでは，先の欠損値処理により観測値を用いずとも一期先予測による状態の推定が可能である．すなわち，状態のモデルにトレンドや周期的成分が含まれ，これを正しく推定できている場合，探索を経ずに状態の推移傾向を捉えることができる．この探索が不要な予測機能により，腕の有用性の変化を検出して評価が逆転するまでの期間における途中経過の評価を把握するための探索を減らし，探索効率の向上が期待できる．これらは Algorithm1 の 18 から 19 行目に相当する．

3.3 時変行列による文脈に応じた腕の評価更新

多次元の要素からなる状態を扱う線形カルマンフィルタでは，この状態と観測値，誤差の次元数の整合性を取るため，式 (1) における Z や式 (2) における R のような行列を必要とする．これらの行列 Z, R は，式 (3) のフィルタリングや式 (4) の一期先予測において状態の平均や分散共分散のどの要素を更新するかを決定している．多様な文脈を扱う多腕バンディットにおいて，常時同じ行列を用いて，文脈に関連しない状態の要素を更新することは，推定の精度を低下させ非効率な探索が発生してしまう．このような状況として，排他的に発生する文脈において同じ腕であっても有用性が異なる状況が挙げられる．この状況では，ある文脈での試行に対する欠損値処理が，異なる文脈も含んだ分散共分散の継続的な増加に繋がり，本来不要な探索を引き起こす．

提案手法では，文脈に応じて行列 Z, R を切り替えることでこの問題を解消する．これらの時変の行列 Z_t, R_t は，t 時点のコンテキスト情報 $\boldsymbol{x}_t$ のうち値が 0 より大きい要素と対応する成分のみ 1 を設定して得られる．例として，状態のモデルにトレンドや周期的成分を含まず $\boldsymbol{x}_t \in \{0,1\}^m$ の場合，時変の行列 Z_t, R_t はそれぞれ $Z_t = \boldsymbol{x}_t^\mathsf{T}$，$R_t = \boldsymbol{x}_t$ となる．これらは Algorithm1 の 15 行目と 21 行目に相当する．

3.4 方策の特性を分析する評価指標

本研究では，方策の追従性と探索効率を定量的に分析するための新たな指標を導入する．この指標では，各時点での方策の判断結果を，選定された腕が最適腕かどうか，その腕を方策が活用として選定したかどうかの四象限に分類し，全ての試行分を数え上げた，表 1 の混同行列を準備する．この混同行列では②と③の値から，最適な腕が切り替わる状況での誤判定の数を把握できる．番狂わせではない場合，各方策は最適腕の有用性が低下したことを③に該当する試行のみから再評価する．一方，番狂わせの場合，新たな最適腕を再評価するため②に該当する試行が必要となる．そして，どちらの場合でも，最適腕を正しく切り替えられたならば①に該当する試行が増加する．よって，最適な腕が切り替わる状況に迅速かつ効率的に追従可能な方策は①に対し，②と③の値を小さく保つと期待される．

表 1 方策の腕の選定に関する混同行列

	活用	探索
最適腕	①	②
最適腕以外	③	④

この仮定に基づき，変化への追従性と探索効率の指標としての $F_{0.5}$ 値を表 1 の混同行列から算出する．これは分類問題の評価に使われる同指標を先の混同行列に適用したもので，Precision($P = ①/(①+③)$) と Recall($R = ①/(①+②)$) の重み付き調和平均 ($F_{0.5} = 1.25\frac{P \cdot R}{0.25P+R}$) である．Precision を重視する $F_{0.5}$ 値とするのは，②が最適腕への探索であり，誤判定ではあるものの試行の増加に対して短期的な機会損失が生じにくいためである．

従来の評価指標に用いられる累積リグレットは各時点で最善の腕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．この指標では，腕の選定における方策の判断は区別されておらず，値の増加が②③④のどれに起因するものかを把握できない．結果として，方策による追従性の差異の分析が定性的な考察に留まるか，分析の観点が不統一なことが多かった．本指標により，変化に対する各方策の挙動の特性を定量的に示すことが可能になり，本分野での有意義な議論の促進につながると考える．

4 評価

4.1 評価方法

本報告では，文脈と時間の経過に応じて腕の有用性が変化する状況のシミュレーションを通して提案手法の有効性を評価する．評価は 10 本の候補より選定した腕から得られる期間中のクリック数のシミュレーションによって行う．ここで，腕は多腕バンディットの方策によって各時点ごとに選定され，各腕は設定したクリック率のベルヌーイ分布に従いクリックされるものとする．各腕は 4 次元のパラメータ $\boldsymbol{\theta}$ を持ち，選定時点 t のコンテキスト情報 $\boldsymbol{x}_t$ との線形和によってクリック率を算出する．各腕のパラメータを時間の経過に応じて変化させることで，文脈と時間の経過に応じて腕の有用性が変化する状況を再現する．本評価ではコンテキスト情報として

$$\boldsymbol{X} = \{(0,0,0,1)^\mathsf{T}, (0,0,1,0)^\mathsf{T}, (0,0,1,1)^\mathsf{T}, (0,1,0,0)^\mathsf{T}, (1,0,0,0)^\mathsf{T}, (1,1,0,0)^\mathsf{T}\}$$

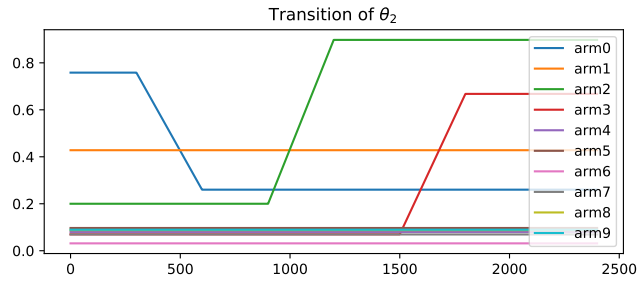

図 1 腕ごとの線形パラメータ $\boldsymbol{\theta}_2$ の推移

から時点 t ごとに無作為に選択された $\boldsymbol{x}_t$ を用いる．各腕の線形パラメータのうち，2 次元目の値 $\boldsymbol{\theta}_2$ の変動を図 1 に示す．約 500 時点で発生する最善な腕の有用性のみが低下する状況と，約 1,000 時点で発生する番狂わせの状況を変動として含めた．なお，パラメータの他の次元は期首に設定した値を維持する．これにより，コンテキスト情報 $\boldsymbol{x}_t \in \{(0,1,0,0)^\mathsf{T}, (1,1,0,0)^\mathsf{T}\}$ の場合にのみ上述のクリック率の変動が発生する．

本評価では，乱数を用いた確率の計算結果を平均化するために，異なる乱数シードを用いてシミュレーションを 50 回行い，この平均を結果として用いた．また，腕の評価が定まった後における変化への対応の性能を調査するため，各方策へ各腕の各コンテキスト $\boldsymbol{x}_t$ ごとに 50 回分の試行結果をシミュレーション実施前に予め与えている．比較する方策には 2 節で紹介した文脈付き，かつ，非定常な方策である Decay LinUCB，AdTS，dLinUCB，DenBand，TVTP，ADA-ILTCB+を用いる．また，提案手法の Linear KalmanFilter Bandits(LKF) では，状態の次元数をコンテキスト $\boldsymbol{x}_t$ と揃えた状態モデルを採用する．なお，評価時とは異なる乱数シードでの予備評価で各方策のパラメータ調整を実施した．パラメータ調整には GP-UCB [10] を用い，方策ごとに設けたパラメータの組み合わせから累積クリック数が最大となるものを探索した．

4.2 機会損失の評価

図 2 に方策ごとの累積リグレットの推移を示す．上段は全ての文脈の合計，下段は図 1 の変化が起きるコンテキスト $\boldsymbol{x}_t = (0,1,0,0)^\mathsf{T}$ での結果である．破線は最適腕の切替が発生する時点を表している．なお，ADA-ILTCB+については本設定における腕と文脈の組み合わせ数の増加により評価作業中に実行が完了しなかったため以降の結果から除いている．シミュレーション全体を通して，累積リグレットが最も少ない方策は DenBand で 94.3，次いで提案手法の LKF がほぼ同程度の 94.8 となった．図より，これらの方策が，番狂わせに追従したことで全体的な累積リグレットの増加を抑えたことがわかる．各方策の切替の様子を確認するため，図 3 に

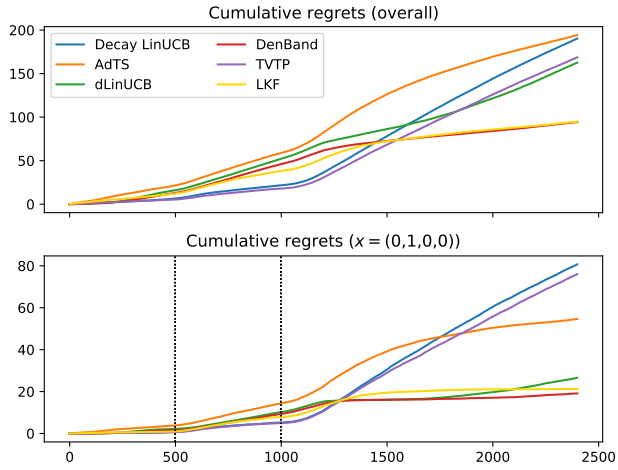

図 2 方策ごとの累積リグレットの推移

$\boldsymbol{x}_t = (0,1,0,0)^\mathsf{T}$ における各方策の腕ごとの累積選定数の推移を示した．ただし，Decay LinUCB，dLinUCB はそれぞれ TVTP と DenBand と同じ推移特性であったことから図から省略した．なお，最上段は理想的な腕の選定がなされた場合の推移である．また，括弧内の数値は各腕の累積選定数を示している．図より Decay LinUCB，AdTS，TVTP は，番狂わせではない最初の変化には迅速に追従できた一方，二度目の番狂わせの検出が遅れ，新たな最適腕への切替が遅れたことがわかる．これは，これらの方策に弱い腕に対する探索の仕組みが備わっていないことに起因する．反対に，dLinUCB，DenBand，LKF では新たな最適腕への切替が見て取れる．

4.3 変化への追従性と探索効率の評価

図 4 に $\boldsymbol{x}_t = (0,1,0,0)^\mathsf{T}$ における方策ごとの $F_{0.5}$ 値と Precision，Recall の値を示す．提案手法の LKF が，高い水準で Precision と Recall を両立したことで $F_{0.5}$ 値が最も大きくなり，誤判定の期間と新たな腕への探索を最小限に抑えたことがわかる．反対に，番狂わせへの追従が遅れた Decay LinUCB と TVTP では，新たな最適腕への探索が極端に少なかったことから Recall は大きいが，以前の最適腕を使い続けたため Precision が小さくなり，結果として $F_{0.5}$ 値も小さくなっている．Decay LinUCB と TVTP より早いものの，やはり番狂わせへの追従が遅れた AdTS では，$F_{0.5}$ 値はこれらと LKF の中間を示した．しかしながら，番狂わせの検出のため混同行列の④の探索が多くなり，変化のない期間でのリグレットの増加が観測されている．本報告時点では $F_{0.5}$ 値は変化時の追従性と探索効率の評価指標であり，変化のない期間の評価は今後の課題である．dLinUCB と DenBand では，図 2 の下段からリグレットを LKF と同程度に抑えているにも関わらず $F_{0.5}$ 値が低く算出された．これは，混同行列の②の極端な増加に起因してい

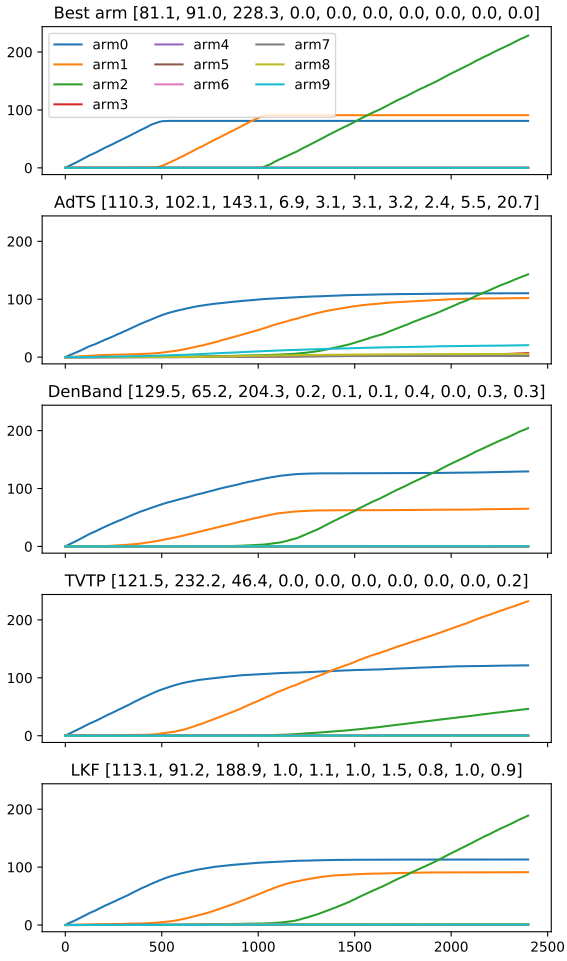

図 3 腕ごとの累積選定数の推移

る．ここから，新たな最適腕への探索を積極的に行いながらも評価の更新が緩やかな特性を持つ方策であることがわかる．すなわち，番狂わせが短期間に高頻度で発生するような環境に課題がある可能性が考慮された値とみなせる．

5 おわりに

本報告では，線形カルマンフィルタを用いた継続的な状態推定と欠損値処理によって，番狂わせを含む環境であっても機会損失を低減可能な多腕バンディット手法を提案した．評価では，機会損失の低減の観点において従来の最先端の方策と遜色のない性能を示した．また，方策の追従性と探索効率の分析のために新たな指標を提案し，この観点でも従来の方策に比べ改善を確認した．今後は，線形カルマンフィルタにおけるトレンドや周期変動モデルの仮想的な探索アプローチに対する有効性の評価や，誤差の分散共分散の逐次的な推定による探索効率の向上，ならびに提案指標の改善を進めたい．

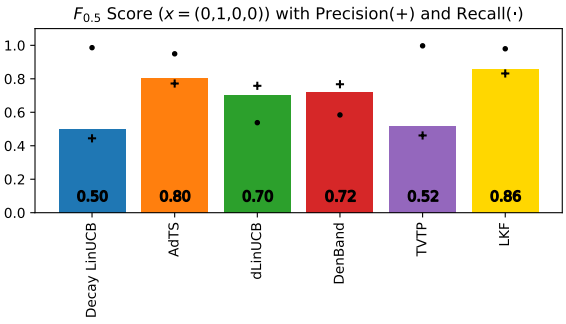

図 4 方策ごとの $F_{0.5}$ 値と Precision，Recall

参考文献

[1] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 418–427. PMLR, 2019.

[2] 三宅悠介, 峯恒憲. Synapse: 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. 電子情報通信学会論文誌 D, Vol. 103, No. 11, pp. 764–775, 2020.

[3] Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 2025–2034, 2016.

[4] Negar Hariri, Bamshad Mobasher, and Robin Burke. Adapting to user preference changes in interactive recommendation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

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

[6] Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 495–504, 2018.

[7] Qingyun Wu, Huazheng Wang, Yanen Li, and Hongning Wang. Dynamic ensemble of contextual bandits to satisfy users' changing interests. In *The World Wide Web Conference*, pp. 2080–2090, 2019.

[8] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726. PMLR, 2019.

[9] Ole-Christoffer Granmo and Stian Berg. Solving non-stationary bandit problems by random sampling from sibling kalman filters. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 199–208. Springer, 2010.

[10] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.