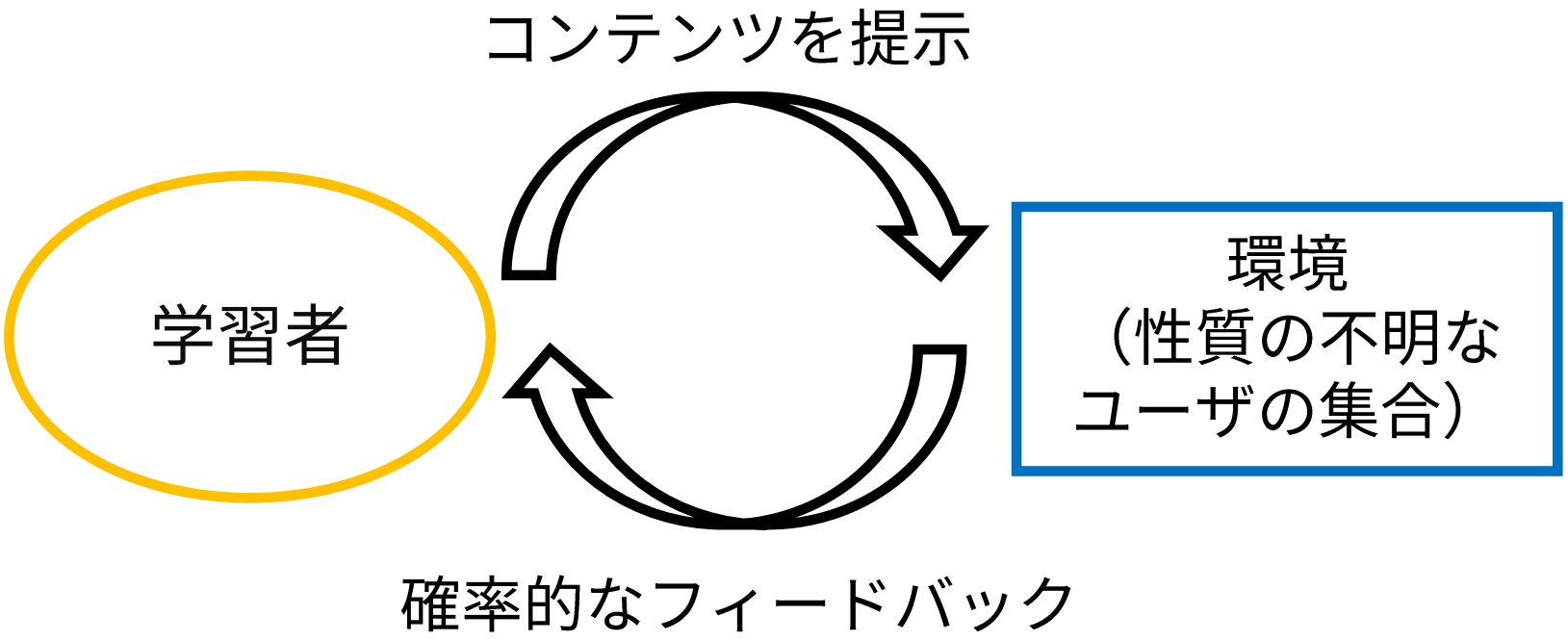
Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem (COLT15')

Junpei Komiyama [1], Junya Honda [1],

Hisashi Kashima [2], Hiroshi Nakagawa [1]

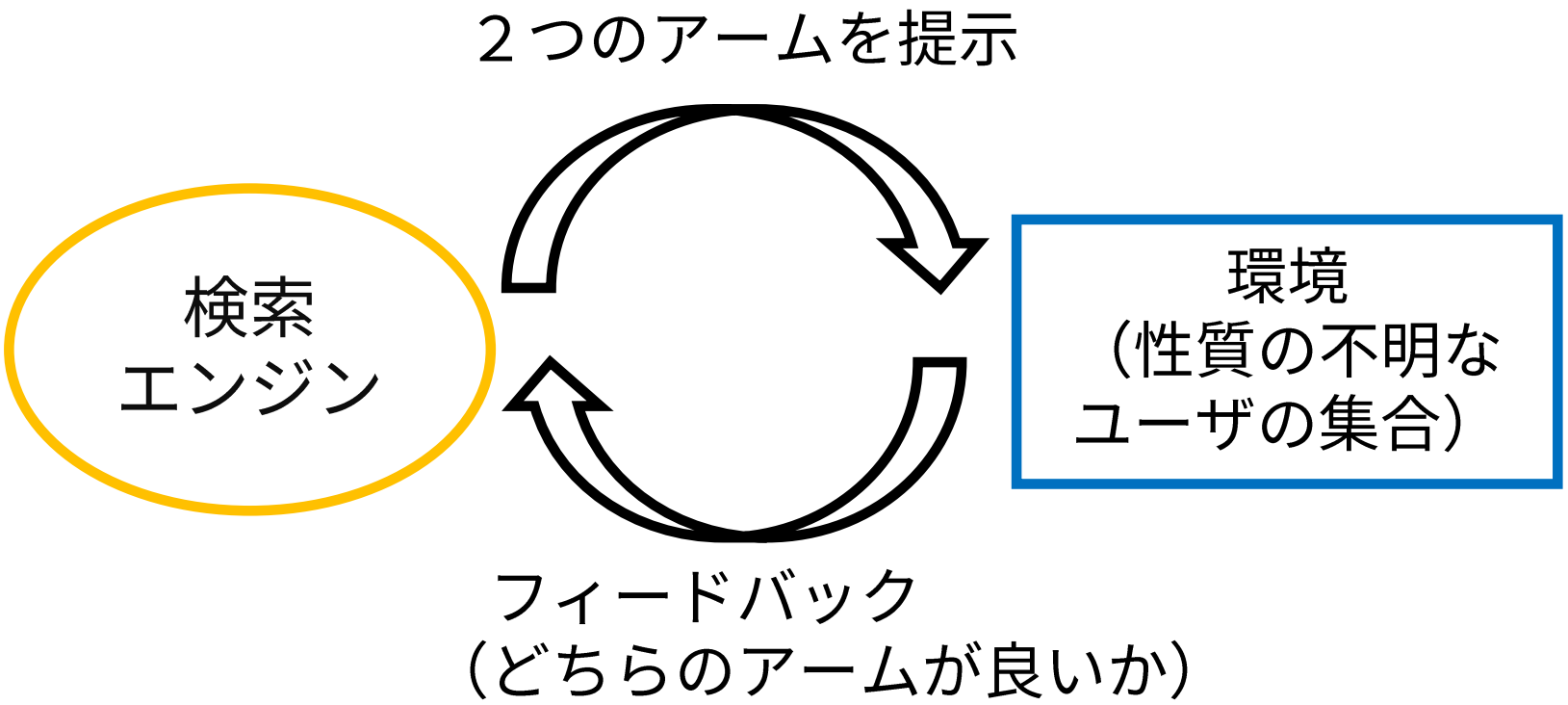
1. Univ. of Tokyo 2. Kyoto Univ.

最適コンテンツ提示問題



- 未知の環境、確率的なフィードバックの元で、最適なコンテンツが何なのかを学習
- 統計的な機械学習

比較バンディット問題 (dueling bandit prob): 比較フィードバックによる最適コンテンツ提示



- 本研究の内容：提示するアームをどのように選ぶのが最適か？→アーム選択アルゴリズム（比較バンディット問題）

例１：寿司の嗜好抽出

- 寿司ネタのうち、最も好まれるものはどれか？

Image from http://www.sharetoyou168.com/

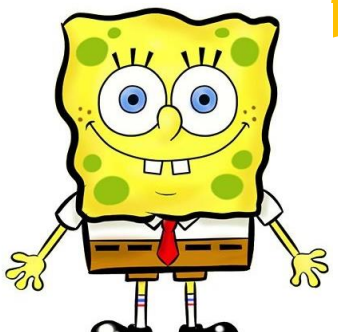
寿司の嗜好抽出

I prefer left (fatty tuna) to right (eel).

Alice

寿司の嗜好抽出

I prefer left (flatfish) to right (shrimp).



Bob

寿司の嗜好抽出

I prefer right (shrimp) to left (fatty tuna).

Charlie

寿司の嗜好抽出

I prefer right (shrimp) to left (fatty tuna).

何人のユーザに
もっとも好まれる寿司を
提供できるか？

例２：検索エンジンのランキング

- 複数のランキング手法のどれが最も良いかを比較したい
- 従来法：専門のチームによる評価
  - 高コスト、ユーザ評価との差異
- 代案：インターリービング[Joachims+02]：
  - ユーザに２つのランキングを混ぜて提示
  - どちらのランキング由来の文書をクリックしたかどうかで、ユーザが好むランキング手法がどちらであるかを推定

Team Draft Interleaving
(Comparison Oracle for Search)

Slide from
[Radlinski et al. 2008,
Yue et al. 2011]

Ranking A

1.Napa Valley – The authority for lodging...
www.napavalley.com
2.Napa Valley Wineries - Plan your wine...
www.napavalley.com/wineries
3.Napa Valley College
www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley
www.ivebeenther
81
5. Napa Valley Wine
www.napavintne
6. Napa Country, Ca
Wikipedia
en.wikipedia.org/

Ranking B

1. Napa Country, California – Wikipedia
en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
www.napavalley.com
he Story of an American
oogl uk/books?isbn=...
ls – Bed and
palinks.com
lley.org
apavalley.org
Marathon
marathon.org

Presented Ranking

1.Napa Valley – The authority for lodging...
www.napavalley.com
2. Napa Country, California – Wikipedia
en.wikipedia.org/wiki/Napa_Valley
3. Napa: The Story of an American Eden...
books.google.co.uk/books?isbn=...
4.Napa Valley Wineries – Plan your wine...
www.napavalley.com/wineries
5. Napa Valley Hotels – Bed and Breakfast...
www.napalinks.com
6.Napa Balley College

Click

Click

B wins!

概要

- 問題設定：比較バンディット問題
  - 性能指標：Regret
- アルゴリズムの性能限界←主結果１
- アルゴリズム：RMED ←主結果２
  - 性能限界を達成する現在唯一のアルゴリズム
- 数値実験による性能比較
  - Sushiデータセット
  - MSLRデータセット

比較バンディット問題
Dueling bandit problem [Yue+ COLT2009]

- K個のアーム $[K] = \{1, \ldots, K\}$
- 各ラウンドt=1,⋯,Tに、２つのアーム$l(t), m(t)$を提示
  - ２値フィードバック：$l(t)$ or $m(t)$ のほうが良い
  - 無情報だが、$l(t) = m(t)$の比較も可能
- 確率的仮定：選好行列$M = \{\mu_{i,j}\} \in (0,1)^{K \times K}$が存在し、ペア$(i, j)$を比較したときに$i$が好まれる確率が$\mu_{i,j}$.
- Mは歪対称：$1 - \mu_{j,i} = \mu_{i,j}$.

比較バンディット問題：選好行列の例

- arXivにおける６つの検索エンジンのランキングアルゴリズムの間での選好行列 [Yue+ ICML2011]

	1	2	3	4	5	6
1	0.50	0.55	0.55	0.54	0.61	0.61
2	0.45	0.50	0.55	0.55	0.58	0.60
3	0.45	0.45	0.50	0.54	0.51	0.56
4	0.46	0.45	0.46	0.50	0.54	0.50
5	0.39	0.42	0.49	0.46	0.50	0.51
6	0.39	0.40	0.44	0.50	0.49	0.50

比較バンディット問題における最も良いアームの定義

- 先ほどの場合は、アームに順序が存在
  - 順序: 1>2>3>4>5>6であり、
  - $i > j$なら$\mu_{i,j} > 0.5$（iがjより好まれる）
- 順序の仮定は実データでは成立しないことが多い
  - MS検索エンジンのデータセット[MSR 2010, Zoghi+WSDM2014]では、ランキング手法128個の間に完全な順序が存在しない

比較バンディット問題：選好行列の例

下の例（MS検索エンジンのランキング手法間の比較、部分行列）は、順序が存在しない

	1	2	3	4	5	6
1	0.50	0.36	0.36	0.34	0.36	0.36
2	0.64	0.50	0.49	0.46	0.48	0.47
3	0.64	0.51	0.50	0.48	0.49	0.51
4	0.66	0.54	0.52	0.50	0.53	0.52
5	0.64	0.52	0.51	0.47	0.50	0.49
6	0.64	0.53	0.49	0.48	0.51	0.50

コンドルセ勝者 (Condorcet winner)

- 最低限、”一番良い”アームが定義できるのはどのような仮定を置けばいい？
- コンドルセ勝者 [Urvoy+ ICML2013]：あるアーム（アーム1とおく）が存在し、$\mu_{1,j} > 1/2$
  - 実データ：MS検索エンジンのランキングアルゴリズム間ではコンドルセ勝者は存在
- 本研究では、コンドルセ勝者を仮定し、それを探すためのアルゴリズムを提案

比較バンディット問題：選好行列の例

- 順序のある例では、アーム１（順序１位）がコンドルセ勝者

	1	2	3	4	5	6
1	0.50	0.55	0.55	0.54	0.61	0.61
2	0.45	0.50	0.55	0.55	0.58	0.60
3	0.45	0.45	0.50	0.54	0.51	0.56
4	0.46	0.45	0.46	0.50	0.54	0.50
5	0.39	0.42	0.49	0.46	0.50	0.51
6	0.39	0.40	0.44	0.50	0.49	0.50

比較バンディット問題：選好行列の例

- この例では、アーム4がコンドルセ勝者
- 以降のページではアーム1をコンドルセ勝者に

	1	2	3	4	5	6
1	0.50	0.36	0.36	0.34	0.36	0.36
2	0.64	0.50	0.49	0.46	0.48	0.47
3	0.64	0.51	0.50	0.48	0.49	0.51
4	0.66	0.54	0.52	0.50	0.53	0.52
5	0.64	0.52	0.51	0.47	0.50	0.49
6	0.64	0.53	0.49	0.48	0.51	0.50

評価手法：Regret

- 可能な限り多くのユーザに最も良いアームを提示したい
- 最も良いアームを提示できない＝後悔（Regret）の増加→これを最小化したい
- $\mathrm{Regret(T)} = \sum_{t=1}^{T} \frac{\mu_{1,l(t)} + \mu_{1,m(t)} - 1}{2}$
  - $\mu_{1,l(t)} > 1/2$: アーム1と$l(t)$の間の良さの差分
  - $(l(t), m(t)) = (1,1)$でない限りregretが増加
  - アーム1をasapで発見し、$(1,1)$を残りのラウンドで提示したい

強一致性のあるアルゴリズム

- 「良い」アルゴリズムはどういう性質を持つべきか
- 「良くない」：ある選好行列には上手く学習できるが、別の選好行列に対しては学習ができない
- 強一致性：任意の選好行列$M, a > 0$に対し、Regretの期待値が$o(T^a)$.
  - 通常のバンディット問題の理論はLai&Robbins [1985]が示した→比較バンディット問題では？

主結果１：Regret下限（アルゴリズムの性能限界）

- K-1個のアームがコンドルセ勝者でないことを確認
- 最小比較回数 $\frac{\log T}{d\left(\mu_{i,j},\frac{1}{2}\right)}$：アーム$i$が$j$より好まれない（$\mu_{i,j} < 0.5$）ことを確認
- $r_{i,j} := \frac{\mu_{1,i}+\mu_{1,j}-1}{2} > 0$ ：iとjの一比較あたりのregret増分
- 強一致なアルゴリズムのRegretについて、以下が成立：

$$\mathrm{E}[\mathrm{Regret}(T)] \geq \sum_{i\neq 1} \min_{\mathrm{j}:\mu_{i,j}<1/2} \frac{r_{i,j}}{d\left(\mu_{i,j},\frac{1}{2}\right)} \log T - o(\log T)$$

アーム$i \neq 1$に関わるRegret

Regret下限の難しさ

- それぞれのアーム$i \neq 1$に対し、$b^*(i) =$

  $\underset{\mathrm{j}:\mu_{i,j}<\frac{1}{2}}{\operatorname{argmin}} \frac{r_{i,j} \log T}{d\left(\mu_{i,j}, \frac{1}{2}\right)}$を探してそれと比較する＝最小Regret
  を達成

  - $b^*(i) = 1$であることが大半だが、そうでないケースも

- そもそもどれと比較すればいいか分かっていればiがコンドルセ勝者であることが分かるので、これはぱっと見難しいのでは？

  - RMEDはこの問題を上手く解決

主結果２：Relative Minimum Empirical Divergence (RMED) アルゴリズム

- 通常のバンディット問題向けのアルゴリズム DMED [Honda&Takemura 2010]を比較バンディット問題に拡張
- Let $I_i(t) = \sum_{j \in [K]: \hat{\mu}_{i,j} \leq \frac{1}{2}} N_{i,j}(t) d\left(\hat{\mu}_{i,j}, \frac{1}{2}\right)$.

iが負けている相手j

iとjの比較回数

Bernoulli(0.5)からのKL距離（どれだけ負けているか）

- $\exp(-I_i(t))$はアームiがコンドルセ勝者である”尤度”

アルゴリズムRMEDの擬似コード：

while $t < T$ do

・候補集合の生成：すべての$\exp(-I_i(t)) > 1/t$を満たすアームを候補集合に入れる

・候補集合の中のそれぞれのアームに関して、

(a) 今のところ最もコンドルセ勝者に見えるアームと比較する（RMED ver.1）、もしくは、

(b) そのアームが最も負けているアームと比較する（RMED ver.2）

- 理論性能（Regret上限）：RMED1は$O(K \log T)$、RMED2は$O(K \log T)$で定数係数も最適

RMEDの理論的枠組

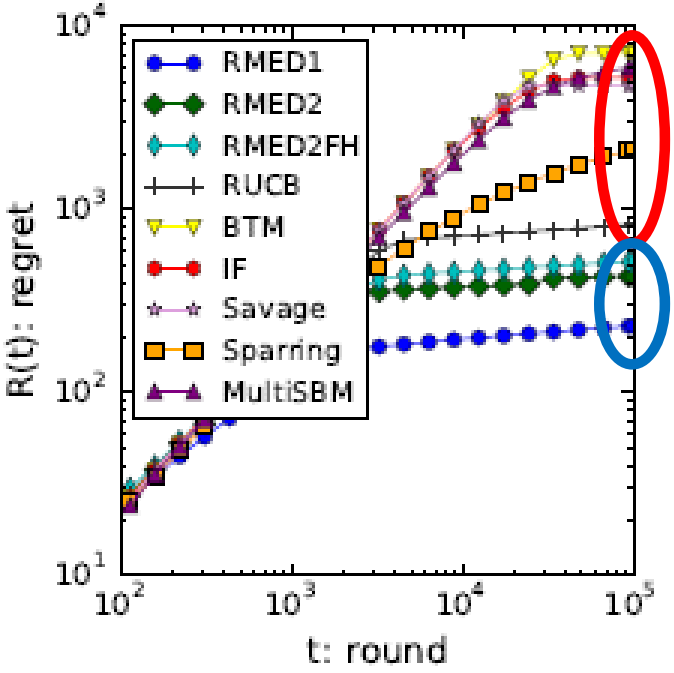
- RMED2: $O(K \log T)$かつ最適なアルゴリズム
  - 下準備：各ペアを最低$O(\log \log T)$回程度比較
  - 最適な比較相手$b^*(i)$を$1 - O(\mathrm{poly}(1/\log T))$の確率で正しく推定→最適regret
  - 推定に失敗したときも$O(\log T \log \log T)$程度の

    Regret→$O\left(\log T \log \log T \times \mathrm{poly}\left(\frac{1}{\log T}\right)\right) = o(\log T)$.

数値実験

- 提案手法および既存の比較バンディット問題の手法を５つのデータセットで比較
  - Sushi：Kamishima [KDD2003]による寿司の好みのデータセット（アーム＝寿司）
    - 寿司16種に絞る、コンドルセ勝者は中トロ
  - Microsoft Learning to Rank (MSLR)データセット [Microsoft 2010]：検索エンジンのランキング手法同士のインターリービング比較（アーム＝ランキング手法）

実験での性能比較：Sushi



(*d*) Sushi

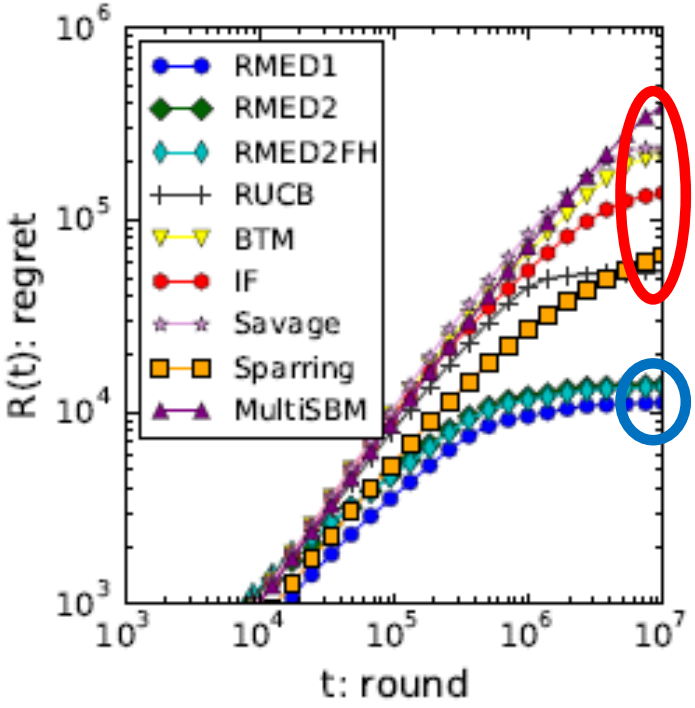
既存手法

提案手法（RMED1/2）

既存手法と比べ
regretが**およそ1/4**

＝ベストでない選択肢を
提示した回数がおよそ1/4

実験での性能比較：MSLR



(f) MSLR $K = 64$

既存手法

提案手法
（RMED1/2）

既存手法と比べ
regretが**およそ1/5**

まとめ

- 一対比較のフィードバックを通じて最も良い選択肢（アーム）を探す問題である比較バンディット問題を扱った
- 主結果１：理論性能限界（Regret下限）の導出を行った
- 主結果２：通常のバンディット問題のアルゴリズムDMEDを拡張し、最適な理論性能を達成するアルゴリズムRMED1/2を提案した
  - 比較相手をどうやってうまく見つけてくるかという、比較バンディット問題特有の難しさを解決した