A TUTORIAL ON ONLINE PREFERENCE LEARNING AND RANKING

Eyke Hüllermeier
Department of Computer Science
University of Paderborn

ALT-2014, BLED, 9-OCT-2014

OUTLINE

PART 1

Introduction to preference learning

general

PART 2

Ranking problems

specific

PART 3

Preference-based bandit algorithms

more specific

PREFERENCES ARE UBIQUITOUS

Preferences play a key role in many applications of computer science and modern information technology:

- COMPUTATIONAL ADVERTISING
- RECOMMENDER SYSTEMS
- COMPUTER GAMES
- AUTONOMOUS AGENTS
- ELECTRONIC COMMERCE
- ADAPTIVE USER INTERFACES
- PERSONALIZED MEDICINE
- ADAPTIVE RETRIEVAL SYSTEMS
- SERVICE-ORIENTED COMPUTING

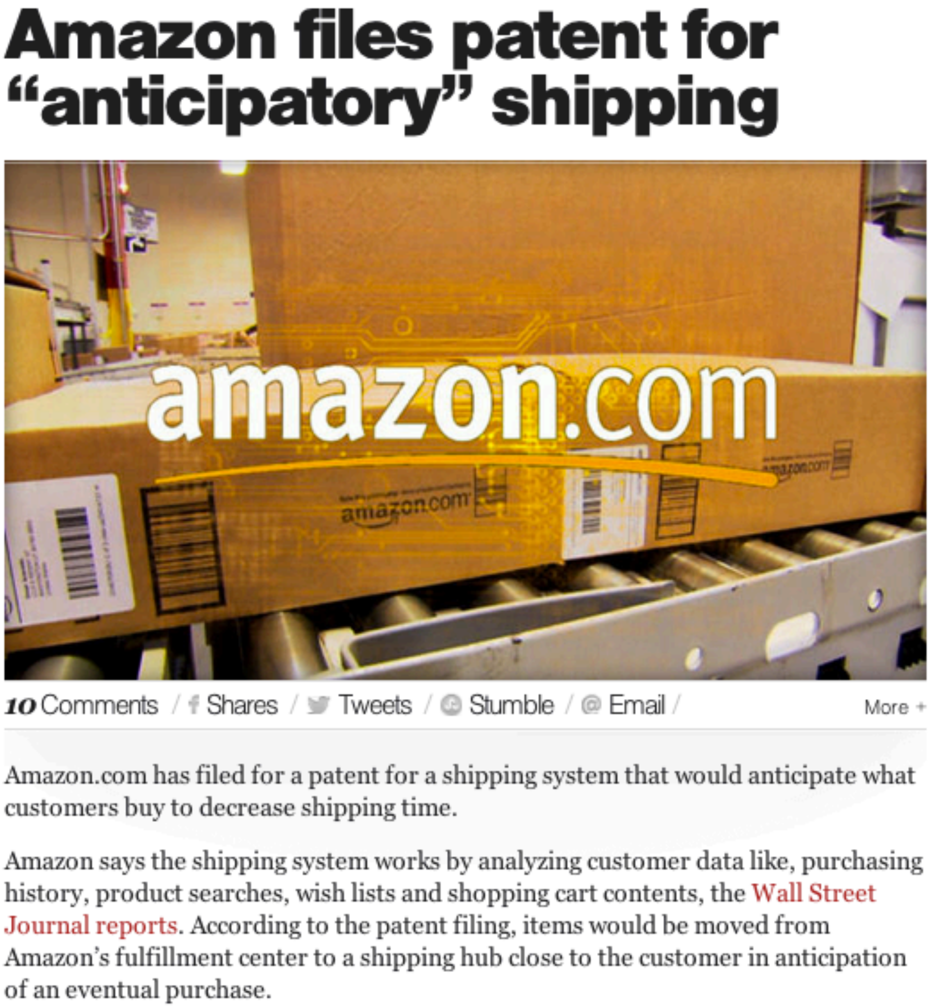
PREFERENCES ARE UBIQUITOUS

Preferences play a key role in many applications of computer science and modern information technology:

- COMPUTATIONAL ADVERTISING
- RECOMMENDER SYSTEMS
- COMPUTER GAMES
- AUTONOMOUS AGENTS
- ELECTRONIC COMMERCE
- ADAPTIVE USER INTERFACES
- PERSONALIZED MEDICINE
- ADAPTIVE RETRIEVAL SYSTEMS
- SERVICE-ORIENTED COMPUTING

medications or therapies specifically tailored for individual patients

COMMERCIAL INTEREST



Amazon files patent for "anticipatory" shipping

10 Comments / Shares / Tweets / Stumble / Email / More +

Amazon.com has filed for a patent for a shipping system that would anticipate what customers buy to decrease shipping time.

Amazon says the shipping system works by analyzing customer data like, purchasing history, product searches, wish lists and shopping cart contents, the Wall Street Journal reports. According to the patent filing, items would be moved from Amazon's fulfillment center to a shipping hub close to the customer in anticipation of an eventual purchase.

PREFERENCES IN AI

> **"Early work in AI focused on the notion of a goal—an explicit target that must be achieved**—and this paradigm is still dominant in AI problem solving. But as application domains become more complex and realistic, it is apparent that **the dichotomic notion of a goal**, while adequate for certain puzzles, **is too crude in general**. The problem is that in many contemporary application domains ... **the user has little knowledge about the set of possible solutions or feasible items,** and what she typically seeks is the best that's out there. But since the user does not know what is the best achievable plan or the best available document or product, she typically cannot characterize it or its properties specifically. **As a result, she will end up either asking for an unachievable goal, getting no solution in response, or asking for too little, obtaining a solution that can be substantially improved."**

[Brafman & Domshlak, 2009]

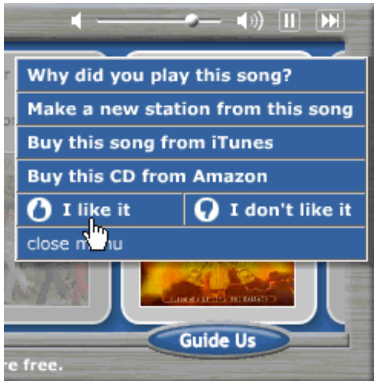
... compared with the dichotomic notion of a ***goal****, preference formalisms significantly increase* ***flexibility*** *in knowledge representation and problem solving!*

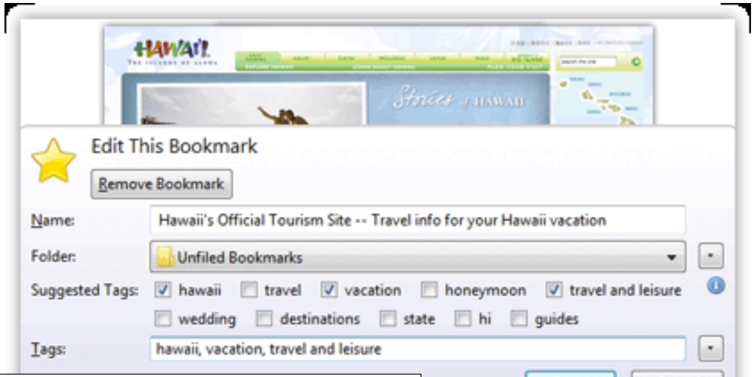
PREFERENCES IN AI

PREFERENCES IN ARTIFICIAL INTELLIGENCE RESEARCH:

- **preference representation** (preference relations, CP nets, GAI networks, logical representations, fuzzy constraints, …)
- **reasoning** with preferences (decision theory, constraint satisfaction, non-monotonic reasoning, …)
- **preference acquisition** (preference elicitation, **preference learning**, ...)

PREFERENCE INFORMATION


Why did you play this song?
Make a new station from this song
Buy this song from iTunes
Buy this CD from Amazon
I like it
I don't like it
close menu
Guide Us
e free.

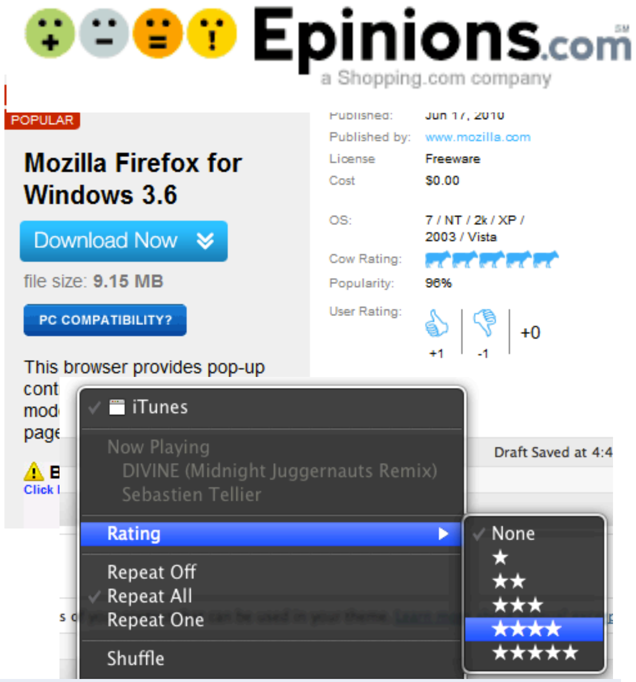


Name: Hawaii's Official Tourism Site -- Travel info for your Hawaii vacation

Folder: Unfiled Bookmarks

Suggested Tags: ☑ hawaii □ travel ☑ vacation □ honeymoon ☑ travel and leisure □ wedding □ destinations □ state □ hi □ guides

Tags: hawaii, vacation, travel and leisure



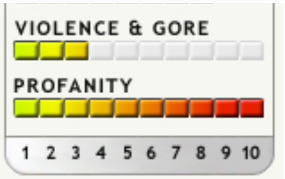
Mozilla Firefox for Windows 3.6

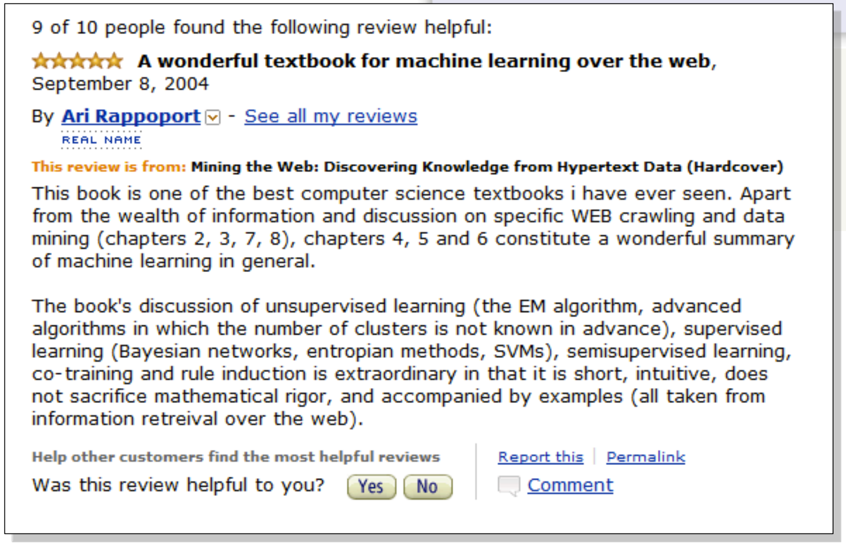
file size: 9.15 MB

This browser provides pop-up cont mod page

Published:	Jun 17, 2010
Published by:	www.mozilla.com
License	Freeware
Cost	$0.00
OS:	7 / NT / 2k / XP / 2003 / Vista
Cow Rating:	
Popularity:	96%
User Rating:	+1 -1 +0

Now Playing DIVINE (Midnight Juggernauts Remix) Sebastien Tellier





9 of 10 people found the following review helpful:

★★★★★ **A wonderful textbook for machine learning over the web**, September 8, 2004

By Ari Rappoport - See all my reviews
REAL NAME

This review is from: **Mining the Web: Discovering Knowledge from Hypertext Data (Hardcover)**

This book is one of the best computer science textbooks i have ever seen. Apart from the wealth of information and discussion on specific WEB crawling and data mining (chapters 2, 3, 7, 8), chapters 4, 5 and 6 constitute a wonderful summary of machine learning in general.

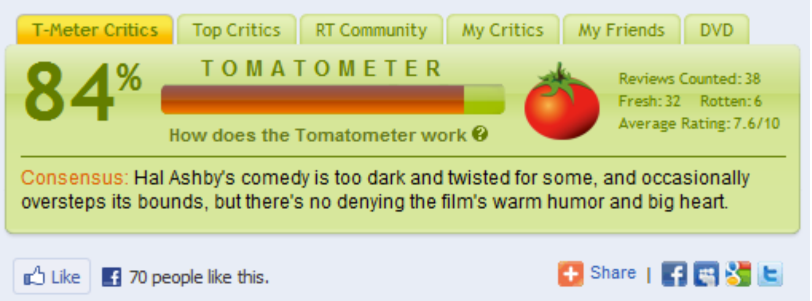
The book's discussion of unsupervised learning (the EM algorithm, advanced algorithms in which the number of clusters is not known in advance), supervised learning (Bayesian networks, entropian methods, SVMs), semisupervised learning, co-training and rule induction is extraordinary in that it is short, intuitive, does not sacrifice mathematical rigor, and accompanied by examples (all taken from information retreival over the web).

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report this | Permalink

Comment

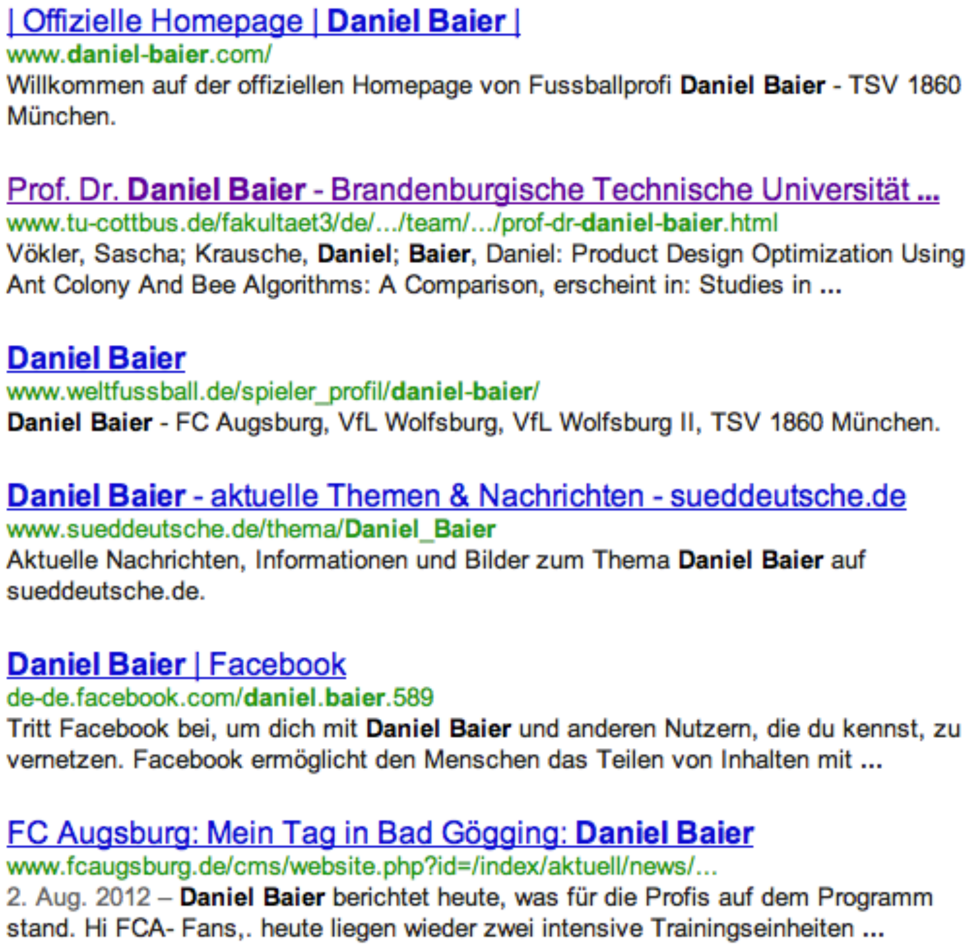


Reviews Counted: 38
Fresh: 32 Rotten: 6
Average Rating: 7.6/10

Consensus: Hal Ashby's comedy is too dark and twisted for some, and occasionally oversteps its bounds, but there's no denying the film's warm humor and big heart.

70 people like this.

PREFERENCE INFORMATION



| Offizielle Homepage | **Daniel Baier** |
www.**daniel-baier**.com/
Willkommen auf der offiziellen Homepage von Fussballprofi **Daniel Baier** - TSV 1860 München.

Prof. Dr. **Daniel Baier** - Brandenburgische Technische Universität **...**
www.tu-cottbus.de/fakultaet3/de/.../team/.../prof-dr-**daniel-baier**.html
Vökler, Sascha; Krausche, **Daniel**; **Baier**, Daniel: Product Design Optimization Using Ant Colony And Bee Algorithms: A Comparison, erscheint in: Studies in **...**

Daniel Baier
www.weltfussball.de/spieler_profil/**daniel-baier**/
Daniel Baier - FC Augsburg, VfL Wolfsburg, VfL Wolfsburg II, TSV 1860 München.

Daniel Baier - aktuelle Themen & Nachrichten - sueddeutsche.de
www.sueddeutsche.de/thema/**Daniel_Baier**
Aktuelle Nachrichten, Informationen und Bilder zum Thema **Daniel Baier** auf sueddeutsche.de.

Daniel Baier | Facebook
de-de.facebook.com/**daniel.baier**.589
Tritt Facebook bei, um dich mit **Daniel Baier** und anderen Nutzern, die du kennst, zu vernetzen. Facebook ermöglicht den Menschen das Teilen von Inhalten mit **...**

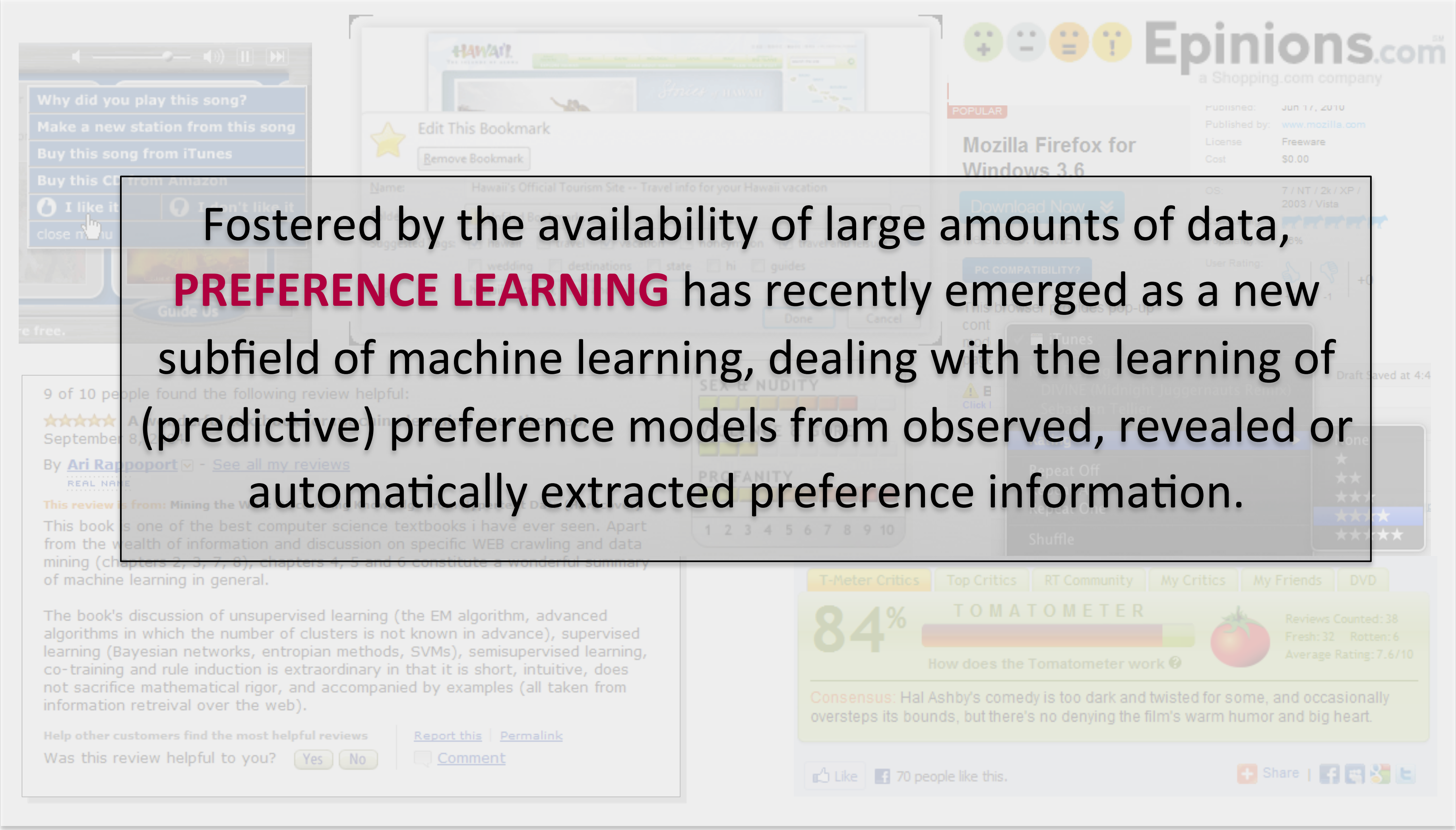
FC Augsburg: Mein Tag in Bad Gögging: **Daniel Baier**
www.fcaugsburg.de/cms/website.php?id=/index/aktuell/news/...
2. Aug. 2012 – **Daniel Baier** berichtet heute, was für die Profis auf dem Programm stand. Hi FCA- Fans,. heute liegen wieder zwei intensive Trainingseinheiten **...**

NOT CLICKED ON

CLICKED ON

Preferences are not necessarily expressed explicitly, but can be extracted implictly from people's behavior!

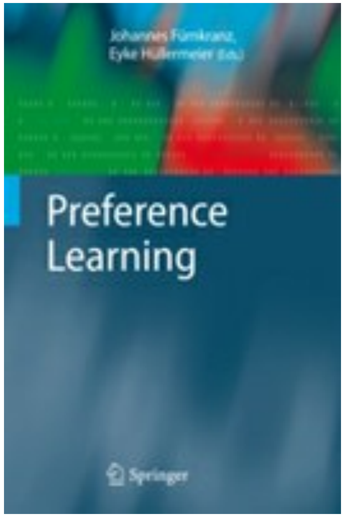
PREFERENCE LEARNING


Why did you play this song?
Make a new station from this song
Buy this song from iTunes
I like it
I don't like it
Edit This Bookmark
Remove Bookmark
Epinions.com
a Shopping.com company
Mozilla Firefox for Windows 3.6
Download Now
9 of 10 people found the following review helpful:
By Ari Rappoport - See all my reviews
Was this review helpful to you? Yes No
T-Meter Critics
Top Critics
RT Community
TOMATOMETER
84%
Repeat Off
Shuffle

Fostered by the availability of large amounts of data, **PREFERENCE LEARNING** has recently emerged as a new subfield of machine learning, dealing with the learning of (predictive) preference models from observed, revealed or automatically extracted preference information.

PL IS AN ACTIVE FIELD

Special Issue on Representing, Processing, and Learning Preferences: Theoretical and Practical Challenges (2011)



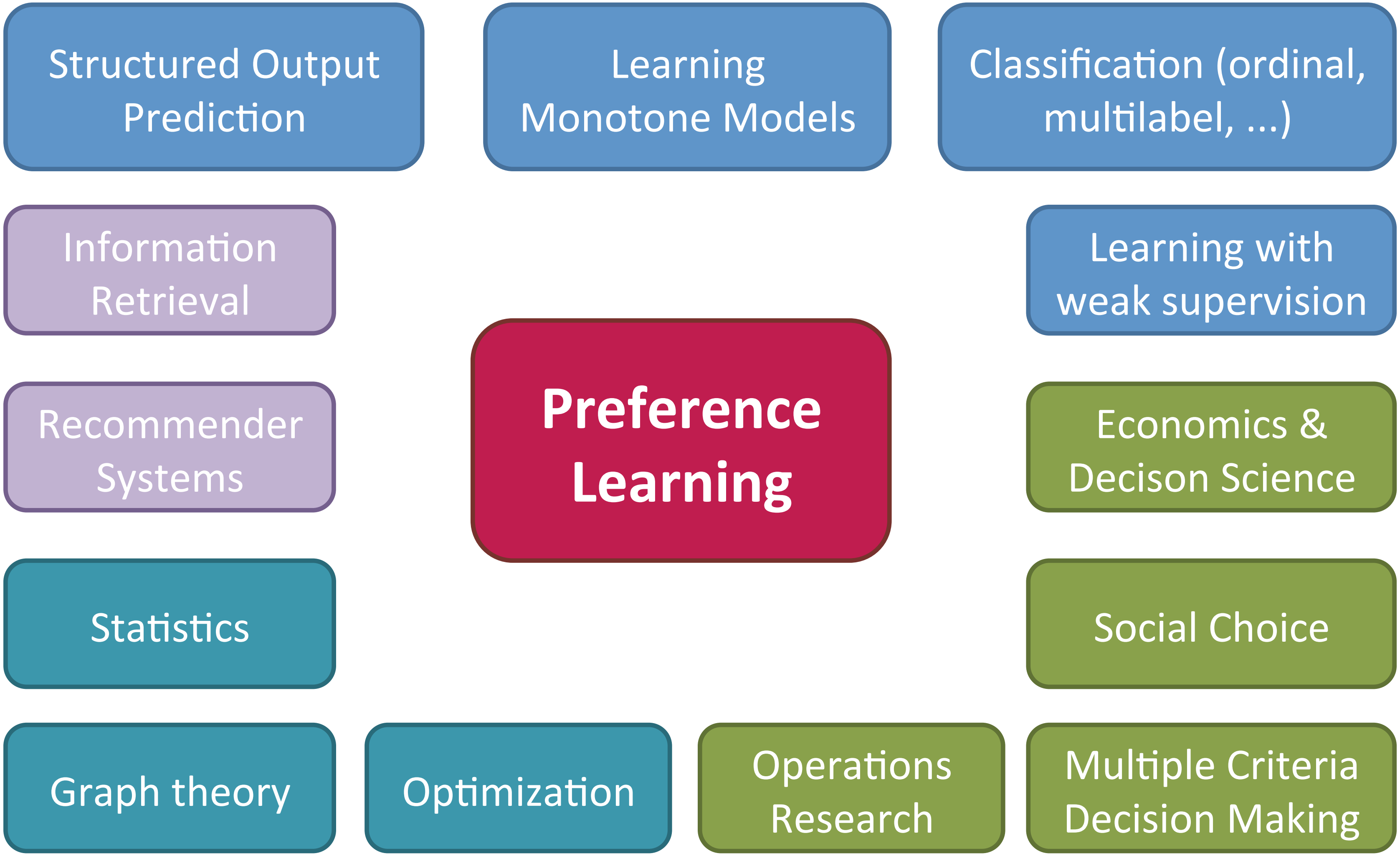
J. Fürnkranz &
E. Hüllermeier (eds.)
Preference Learning
Springer-Verlag 2011

Special Issue on Preference Learning (2013).

PL IS AN ACTIVE FIELD

- NIPS–01: New Methods for Preference Elicitation
- NIPS–02: Beyond Classification and Regression: Learning Rankings, Preferences, Equality Predicates, and Other Structures
- KI–03: Preference Learning: Models, Methods, Applications
- NIPS–04: Learning with Structured Outputs
- NIPS–05: Workshop on Learning to Rank
- IJCAI–05: Advances in Preference Handling
- SIGIR 07–10: Workshop on Learning to Rank for Information Retrieval
- ECML/PDKK 08–10: Workshop on Preference Learning
- NIPS–09: Workshop on Advances in Ranking
- American Institute of Mathematics Workshop in Summer 2010: The Mathematics of Ranking
- NIPS-11: Workshop on Choice Models and Preference Learning
- EURO 2009-12: Special Track on Preference Learning
- ECAI-12: Workshop on Preference Learning: Problems and Applications in AI
- Dagstuhl Seminar on Preference Learning (2014)

CONNECTIONS TO OTHER FIELDS



Structured Output Prediction

Learning Monotone Models

Classification (ordinal, multilabel, ...)

Information Retrieval

Learning with weak supervision

Preference Learning

Recommender Systems

Economics & Decison Science

Statistics

Social Choice

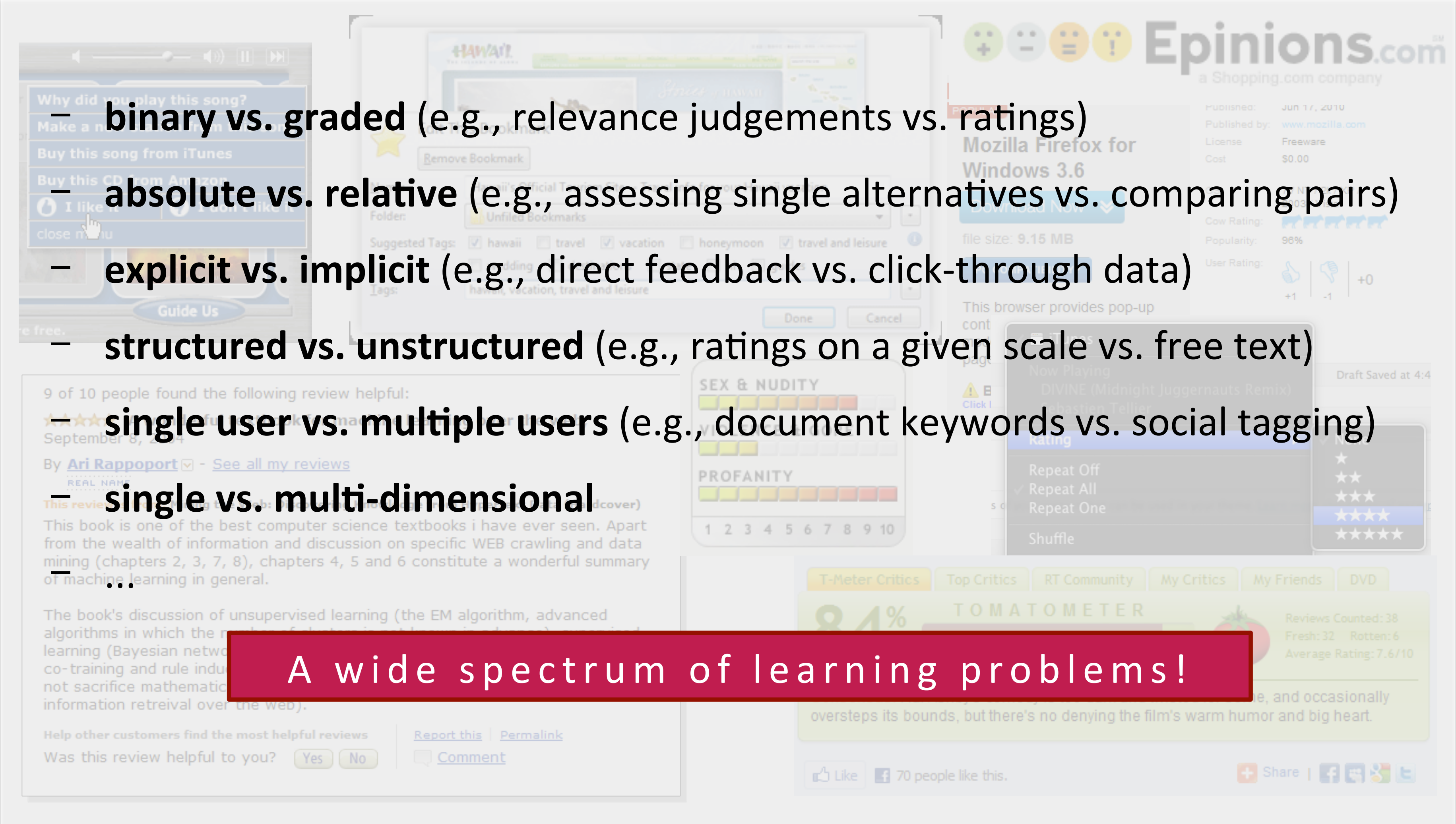
Graph theory

Optimization

Operations Research

Multiple Criteria Decision Making

MANY TYPES OF PREFERENCES


Why did you play this song?
Buy this song from iTunes
I like it
I don't like it
Guide Us
Epinions.com
a Shopping.com company
Mozilla Firefox for Windows 3.6
SEX & NUDITY
PROFANITY
1 2 3 4 5 6 7 8 9 10
Repeat Off
Repeat All
Repeat One
Shuffle
TOMATOMETER
9 of 10 people found the following review helpful:
By Ari Rappoport - See all my reviews
Was this review helpful to you? Yes No

- **binary vs. graded** (e.g., relevance judgements vs. ratings)
- **absolute vs. relative** (e.g., assessing single alternatives vs. comparing pairs)
- **explicit vs. implicit** (e.g., direct feedback vs. click-through data)
- **structured vs. unstructured** (e.g., ratings on a given scale vs. free text)
- **single user vs. multiple users** (e.g., document keywords vs. social tagging)
- **single vs. multi-dimensional**
- ...

A wide spectrum of learning problems!

COLLABORATIVE FILTERING

PRODUCTS

USERS

	P1	P2	P3	…	P38	…	P88	P89	P90
U1	☆		☆☆☆	…		…		☆☆☆	
U2		☆☆	☆	…		…	☆		
…				…		…			
U46	?	☆☆	?	…	?	…	?	?	☆☆☆
…				…		…			
U98	☆☆☆			…		…	☆☆☆		
U99			☆	…		…	☆☆		

- absolute preferences
- graded ratings (on an ordinal scale)
- direct feedback
- multiple users
- no feature description of users or products

OUTLINE

PART 1

Introduction to preference learning

PART 2

Ranking problems

PART 3

Preference-based bandit algorithms

OBJECT RANKING [Cohen et al., 1999]

TRAINING

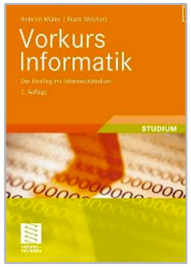
$$\begin{array}{ccc}
(0.74, 1, 25, 165) & \succ & (0.45, 0, 35, 155) \\
(0.47, 1, 46, 183) & \succ & (0.57, 1, 61, 177) \\
(0.25, 0, 26, 199) & \succ & (0.73, 0, 46, 185)
\end{array}$$

Pairwise preferences between objects

$\succ$

Heinz-Peter Gumm | Manfred Sommer
Einführung in die Informatik
9. Auflage
Oldenbourg

$\succ$


Heinrich Müller | Frank Weichert
Vorkurs Informatik
Der Einstieg ins Informatikstudium
2. Auflage
STUDIUM

OBJECT RANKING [Cohen et al., 1999]

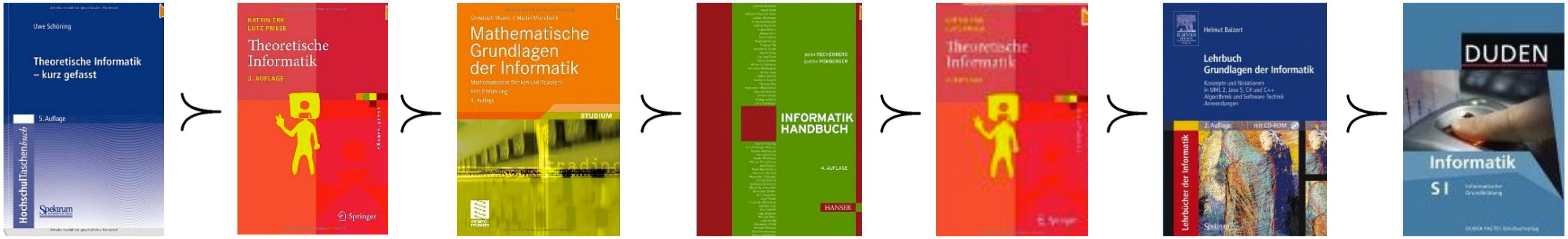
PREDICTION (ranking a new set of objects)

$$\mathcal{Q} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}\}$$

$$\boldsymbol{x}_{10} \succ \boldsymbol{x}_4 \succ \boldsymbol{x}_7 \succ \boldsymbol{x}_1 \succ \boldsymbol{x}_{11} \succ \boldsymbol{x}_2 \succ \boldsymbol{x}_8 \succ \boldsymbol{x}_{13} \succ \boldsymbol{x}_9 \succ \boldsymbol{x}_3 \succ \boldsymbol{x}_{12} \succ \boldsymbol{x}_5 \succ \boldsymbol{x}_6$$

$\mathcal{Q} =$

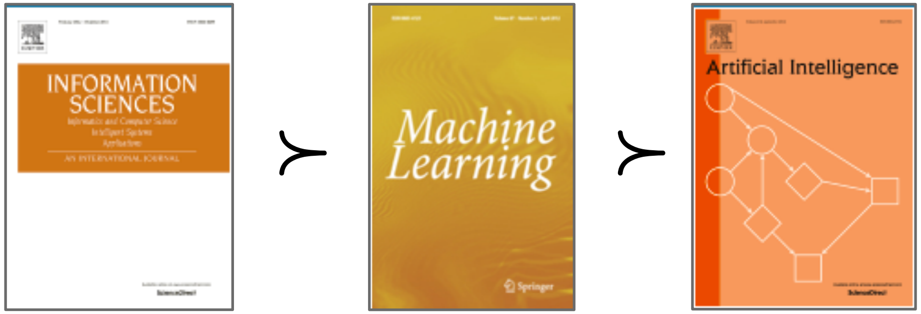
Theoretische Informatik
Mathematische Grundlagen der Informatik
Theoretische Informatik – kurz gefasst
INFORMATIK HANDBUCH
DUDEN Informatik
Lehrbuch Grundlagen der Informatik
Theoretische Informatik


Theoretische Informatik – kurz gefasst
≻
Theoretische Informatik
≻
Mathematische Grundlagen der Informatik
≻
INFORMATIK HANDBUCH
≻
Theoretische Informatik
≻
Lehrbuch Grundlagen der Informatik
≻
DUDEN Informatik

LABEL RANKING [E.H. et al., 2008]

... mapping instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

(35, 1, 187, 325) $\mapsto$



instance $x \in \mathcal{X}$
(e.g., features of a person)

ranking of labels/alternatives
$\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$

LABEL RANKING: TRAINING DATA

TRAINING

X_1	X_2	X_3	X_4	preferences
0.34	0	10	174	$A \succ B$, $C \succ D$
1.45	0	32	277	$B \succ C \succ A$
1.22	1	46	421	$B \succ D$, $A \succ D$, $C \succ D$, $A \succ C$
0.74	1	25	165	$C \succ A \succ D$, $A \succ B$
0.95	1	72	273	$B \succ D$, $A \succ D$
1.04	0	33	158	$D \succ A \succ B$, $C \succ B$, $A \succ C$

Instances are associated with preferences between labels

... no demand for full rankings!

LABEL RANKING: PREDICTION

PREDICTION				A	B	C	D
0.92	1	81	382	?	?	?	?

new instance

ranking ?

LABEL RANKING: PREDICTION

PREDICTION				A	B	C	D
0.92	1	81	382	4	1	3	2

A ranking of all labels

new instance

$\pi(i) =$ position of i-th label

LABEL RANKING: PREDICTION

PREDICTION

0.92	1	81	382	4	1	3	2

A ranking of all labels

LOSS

GROUND TRUTH

0.92	1	81	382	2	1	3	4

LABEL RANKING: PREDICTION

PREDICTION

0.92	1	81	382	4	1	3	2

A ranking of all labels

LOSS

GROUND TRUTH

0.92	1	81	382	2	1	3	4

KENDALL

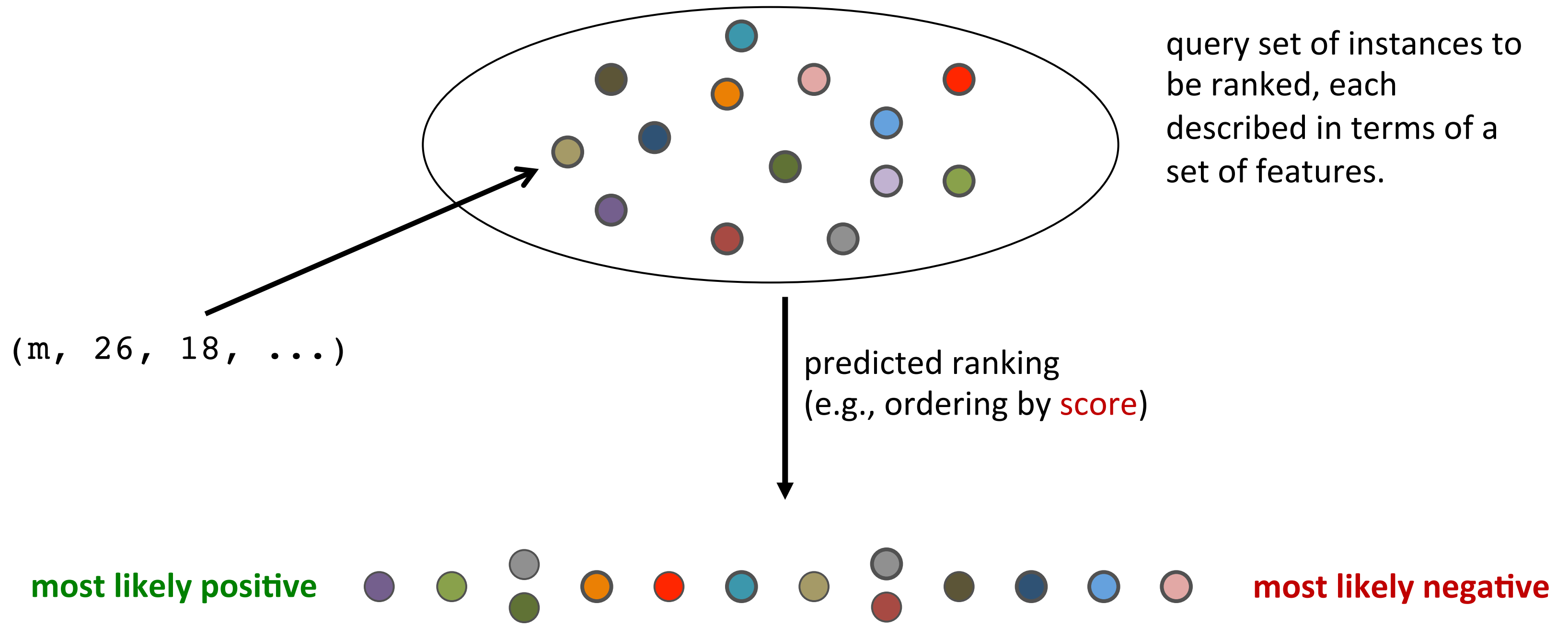
$$\mathcal{L}(\pi,\pi^*) = \sum_{1\leq i<j\leq M} \left[\!\left[(\pi(i)-\pi(j))(\pi^*(i)-\pi^*(j)) < 0 \right]\!\right]$$

LOSS

$$\tau = 1 - \frac{4D(\pi,\pi^*)}{M(M-1)}$$

RANK CORRELATION

BIPARTITE RANKING



query set of instances to be ranked, each described in terms of a set of features.

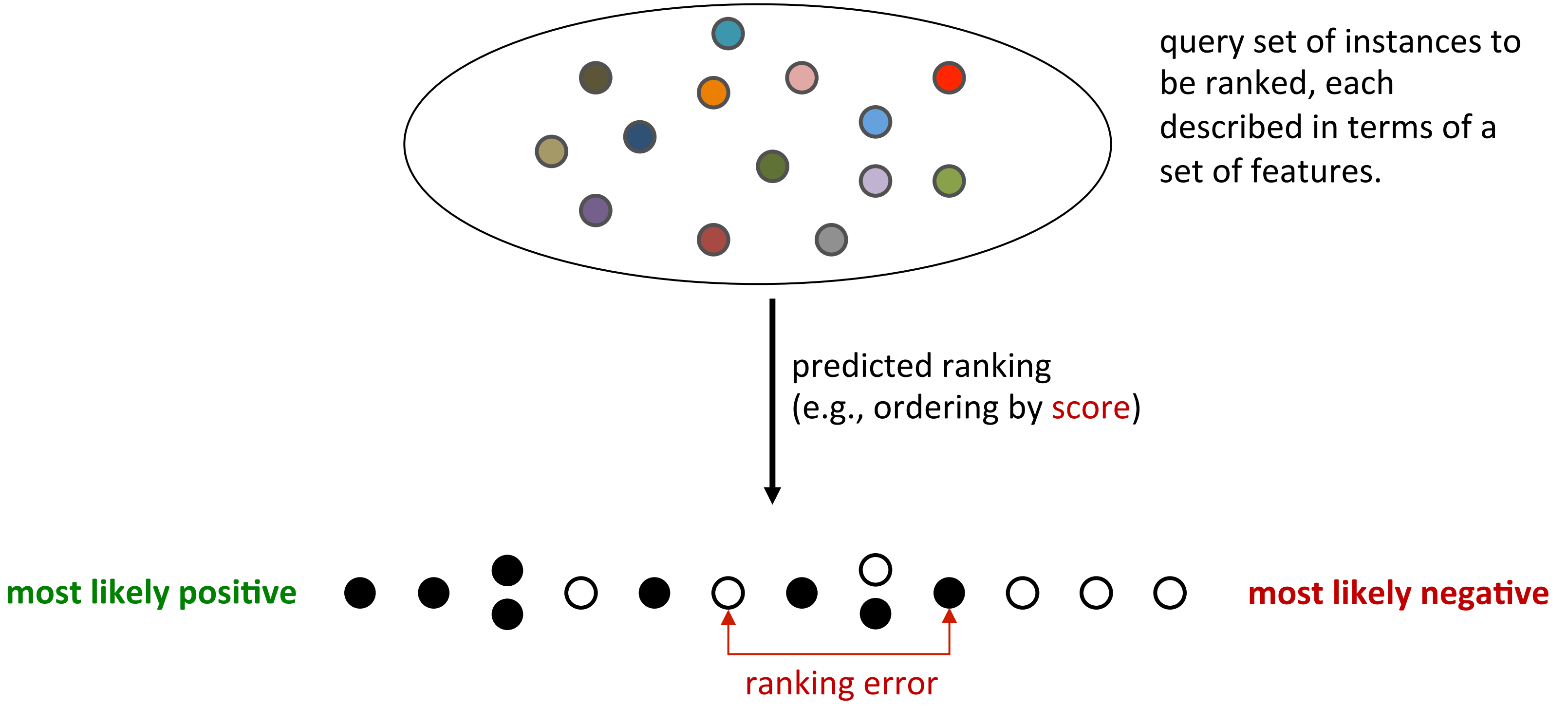
(m, 26, 18, ...)

predicted ranking
(e.g., ordering by score)

most likely positive

most likely negative

BIPARTITE RANKING



query set of instances to be ranked, each described in terms of a set of features.

predicted ranking (e.g., ordering by score)

most likely positive

most likely negative

ranking error

$$\text{rank-loss} = \frac{1}{|P| \cdot |N|} \sum_{(p_i, n_j)} \begin{cases} 1 & f(p_i) < f(n_j) \\ 1/2 & f(p_i) = f(n_j) \\ 0 & f(p_i) > f(n_j) \end{cases}$$

BIPARTITE RANKING

Training

	X1	X2	X3	X4	class
$\boldsymbol{x}_1$	0.34	0	10	174	–1
$\boldsymbol{x}_2$	1.45	0	32	277	+1
$\boldsymbol{x}_3$	0.74	1	25	165	+1
	…	…	…	…	…
$\boldsymbol{x}_n$	0.95	1	72	273	–1

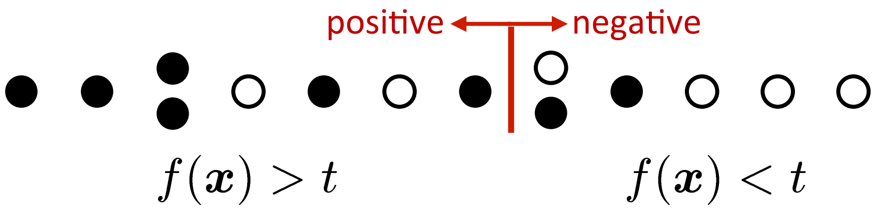
class information
(positive or negative)

$$(\boldsymbol{x}, y) \in X \times \{-1, +1\}$$

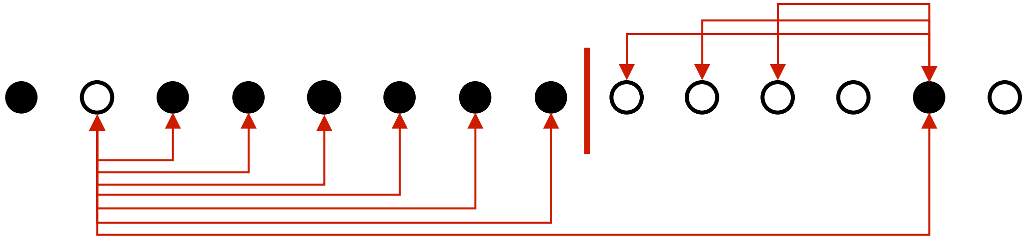
Just the same as classification?

RANKING VERSUS CLASSIFICATION

A ranker (based on scores) can be turned into a classifier via thresholding:



A good classifier is not necessarily a good ranker:



2 classification but
10 ranking errors

→ *learning scoring functions that minimize rank loss* !

RankSVM AND RELATED METHODS

- The idea is to minimize a convex upper bound on the empirical ranking error over a class of (kernelized) ranking functions:

$$f^* \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{|P| \cdot |N|} \underbrace{\sum_{\boldsymbol{x} \in P} \sum_{\boldsymbol{x}' \in N}}_{\substack{\text{check for all positive/} \\ \text{negative pairs}}} L(f, \boldsymbol{x}, \boldsymbol{x}') + \lambda \cdot R(f) \right\}$$

RankSVM AND RELATED METHODS

- The idea is to minimize a convex upper bound on the empirical ranking error over a class of (kernelized) ranking functions:

convex upper bound on
$\mathbb{I}\left(f(\boldsymbol{x}) < f(\boldsymbol{x}')\right)$

$$f^* \in \arg\min_{f\in\mathcal{F}} \left\{ \frac{1}{|P|\cdot|N|} \underbrace{\sum_{\boldsymbol{x}\in P}\sum_{\boldsymbol{x}'\in N}} L(f,\boldsymbol{x},\boldsymbol{x}') + \lambda \cdot R(f) \right\}$$

check for all positive/
negative pairs

regularizer

RankSVM AND RELATED METHODS

- The bipartite RankSVM algorithm [Herbrich et al. 2000, Joachims 2002]:

$$f^* \in \arg \underset{\substack{\uparrow \\ \text{reproducing kernel} \\ \text{Hilbert space (RKHS) with} \\ \text{kernel } \mathbf{K}}}{\min_{f \in \mathcal{F}_K}} \left\{ \frac{1}{|P| \cdot |N|} \sum_{\boldsymbol{x} \in P} \sum_{\boldsymbol{x}' \in N} (1 - (f(\boldsymbol{x}) \underset{\substack{\uparrow \\ \text{hinge loss}}}{-} f(\boldsymbol{x}'))_+ + \frac{\lambda}{2} \cdot \overset{\substack{\text{regularizer} \\ \downarrow}}{\|f\|_K^2} \right\}$$

→ learning comes down to solving a QP problem (expensive)
→ issues with statistical consistency (e.g., Duchie et al. (2010))

OUTLINE

INTELLIGENT
SYSTEMS

PART 1

Introduction to preference learning

PART 2

Ranking problems

PART 3

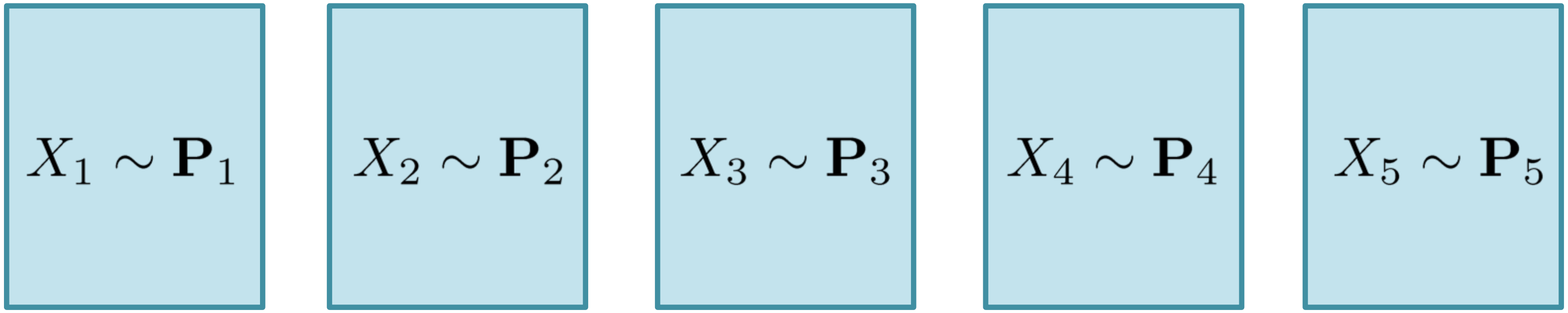
Preference-based bandit algorithms

MULTI-ARMED BANDITS

„pulling an arm“ ↔ choosing an option

partial information online learning
sequential decision process

MULTI-ARMED BANDITS



$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

„pulling an arm“ ↔ choosing an option

choice of an option/strategy (arm) yields a **random reward**

partial information online learning
sequential decision process

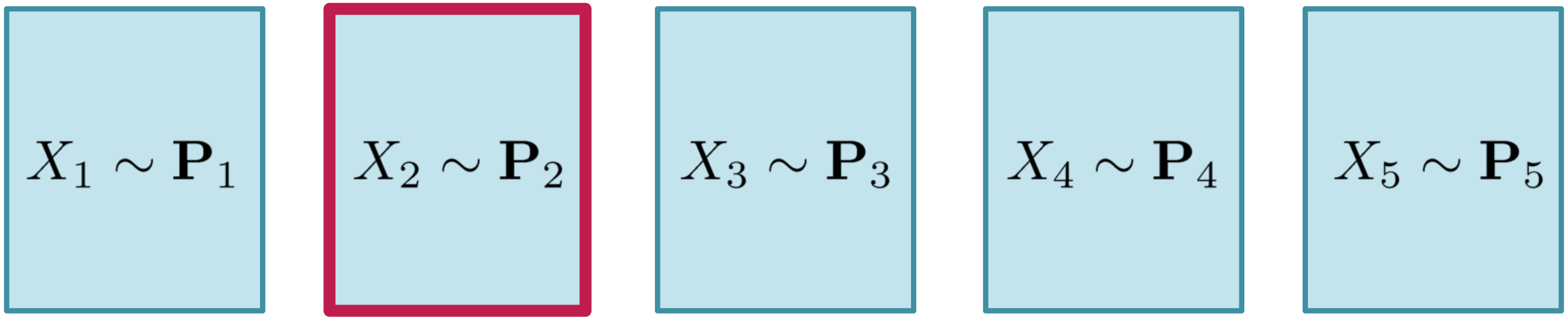
MULTI-ARMED BANDITS

„pulling an arm“ ↔ putting an advertisement on a website

choice of an option/strategy (arm) yields a **random reward**

partial information online learning
sequential decision process

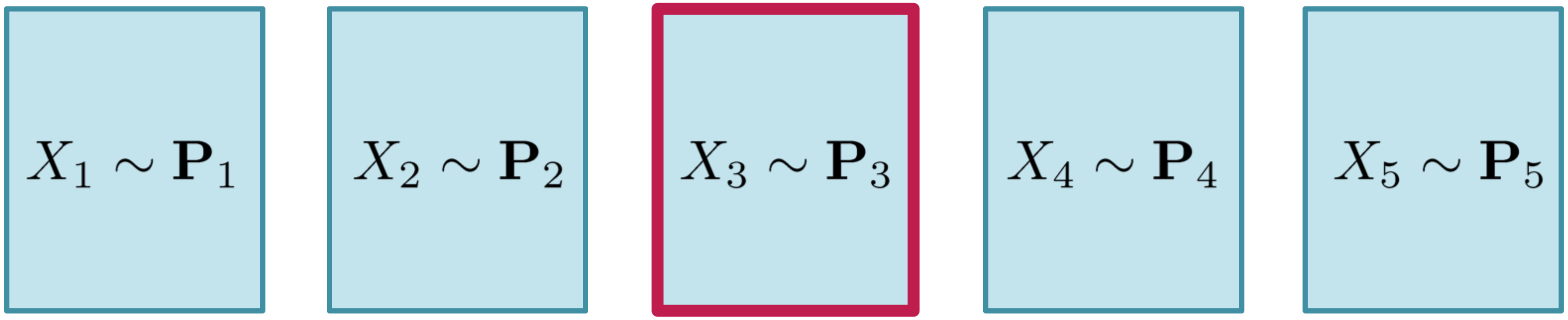
MULTI-ARMED BANDITS



$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

Immediate reward:	2.5
Cumulative reward:	2.5

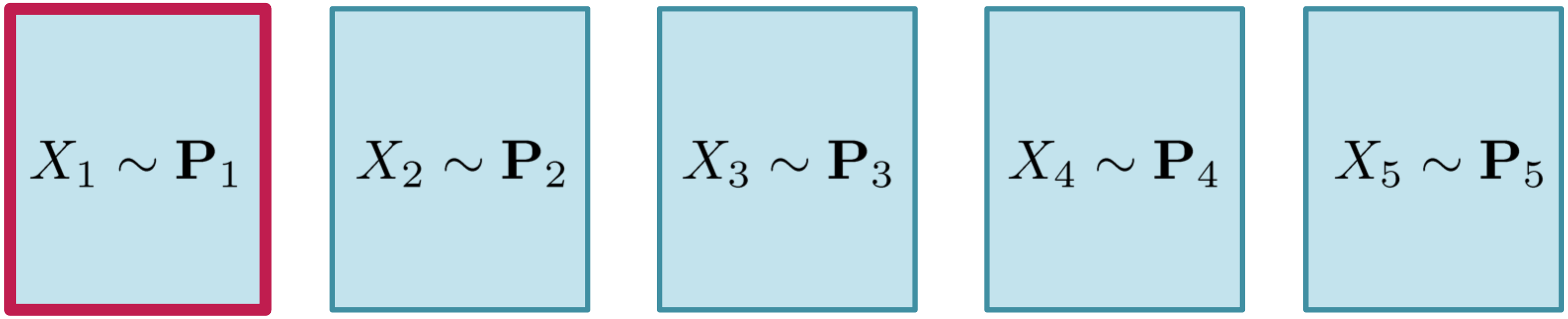
MULTI-ARMED BANDITS



$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

Immediate reward: `2.5 3.1`
Cumulative reward: `2.5 5.6`

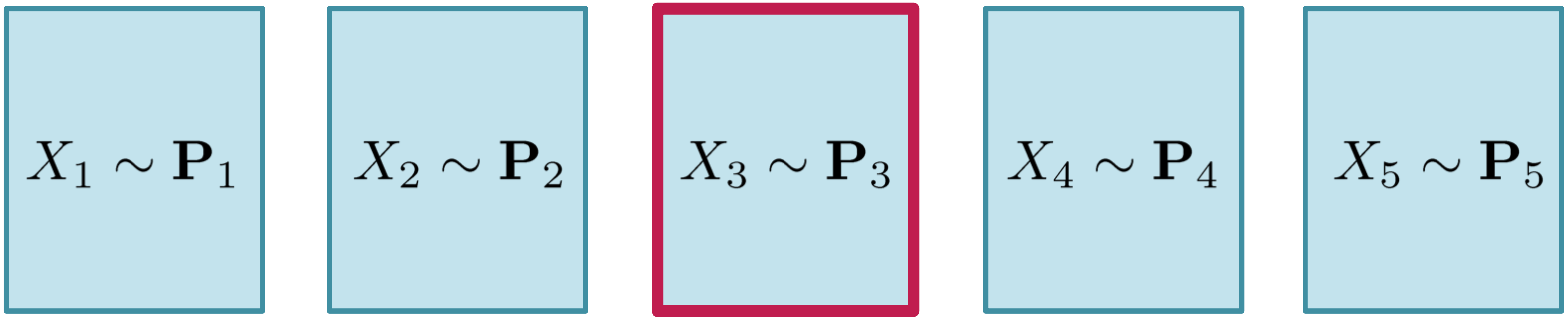
MULTI-ARMED BANDITS



$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

Immediate reward:	2.5	3.1	1.7
Cumulative reward:	2.5	5.6	7.3

MULTI-ARMED BANDITS



Immediate reward:	2.5	3.1	1.7	3.7	...
Cumulative reward:	2.5	5.6	7.3	11.0	...

maximize cumulative reward → *explore and exploit (tradeoff)*

find best option → *pure exploration*

BOUND ON EXPECTED REGRET

- A **policy** is an algorithm that prescribes an arm to be played in each round, based on the outcomes of the previous rounds.

- Denote by $\mu_i = \mathbf{E}(X_i)$ the expected reward of arm a_i and

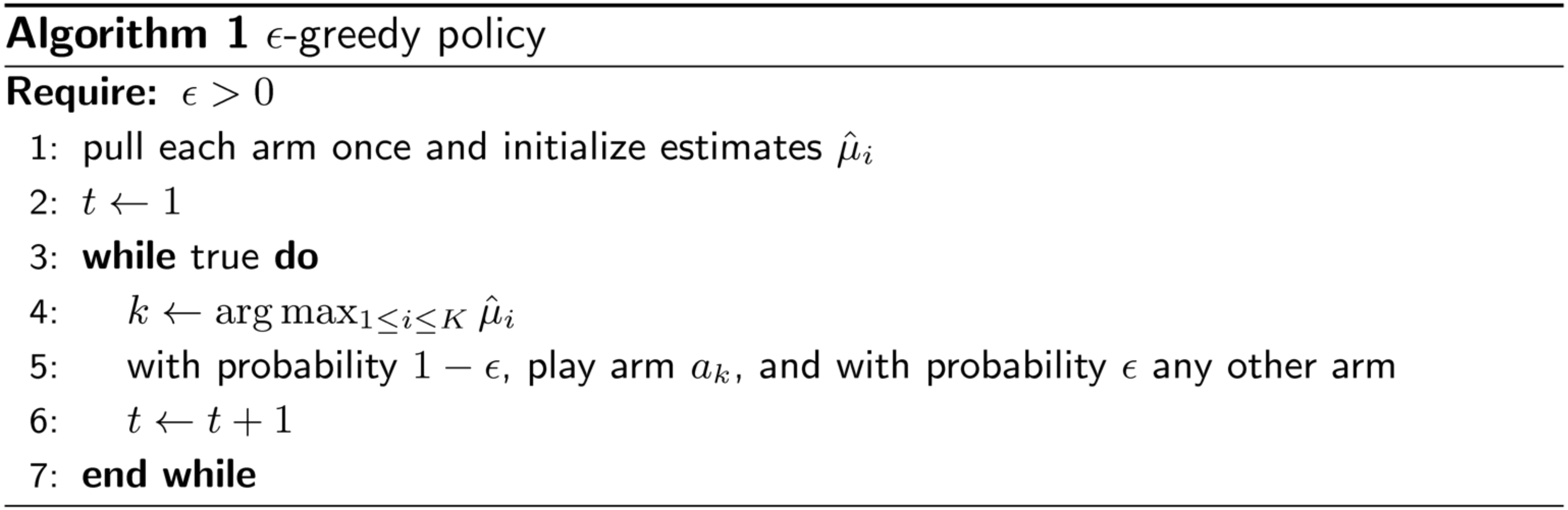
$$\mu^* = \max_{1 \leq j \leq K} \mu_j \quad .$$

- Define the **regret** and **cumulative regret**, respectively, as

$$r_t = \mu^* - x_{i(t)}, \quad R^T = \sum_{t=1}^{T} r_t \quad ,$$

where $i(t)$ is the index of the arm played in round t.

THE EPS-GREEDY ALGORITHM

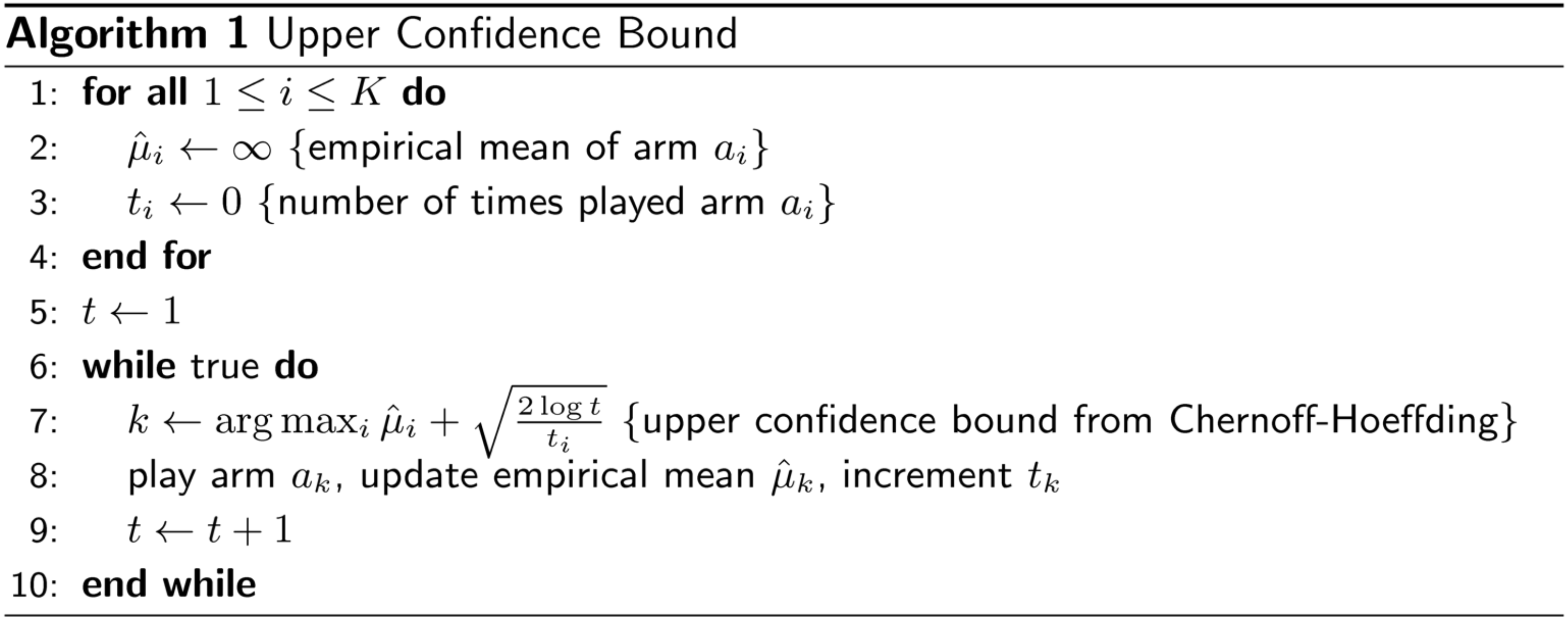


Algorithm 1 ϵ-greedy policy

```
Require: $\epsilon > 0$
pull each arm once and initialize estimates $\hat{\mu}_i$
$t \leftarrow 1$
while true do
    $k \leftarrow \arg\max_{1 \le i \le K} \hat{\mu}_i$
    with probability $1 - \epsilon$, play arm $a_k$, and with probability $\epsilon$ any other arm
    $t \leftarrow t + 1$
end while
```

- For this policy, the cumulative regret is obviously $\mathcal{O}(T)$.
- Presumably optimal arm should be selected with an increasing probability, depending on confidence in the arm, while presumably suboptimal arms should be eschewed (suggesting sequence $\epsilon(t) \searrow 0$)
- Logarithmic regret for $\epsilon(t) = \min\left(1, \frac{6K}{\Delta^2 t}\right)$ with Δ the smallest (positive) suboptimality $\mu^* - \mu_i$.

THE UCB ALGORITHM

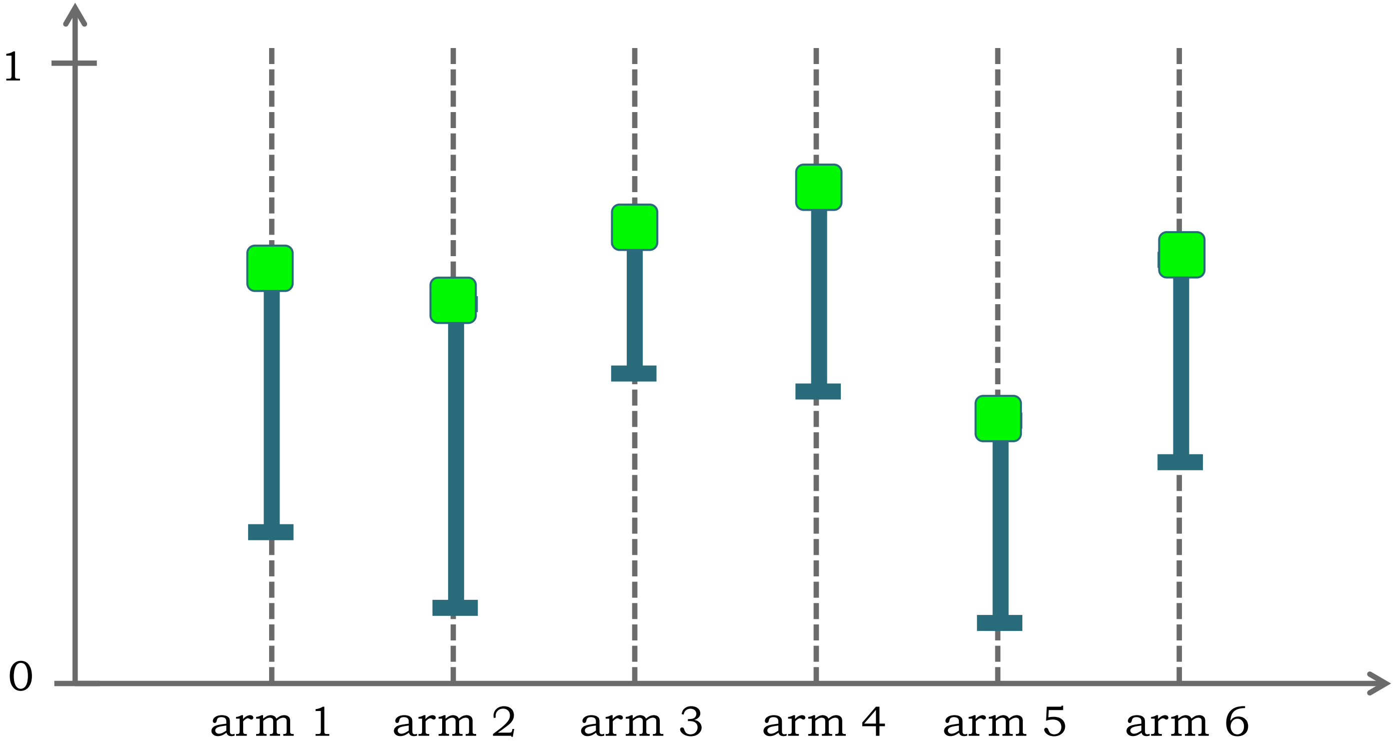


Algorithm 1 Upper Confidence Bound

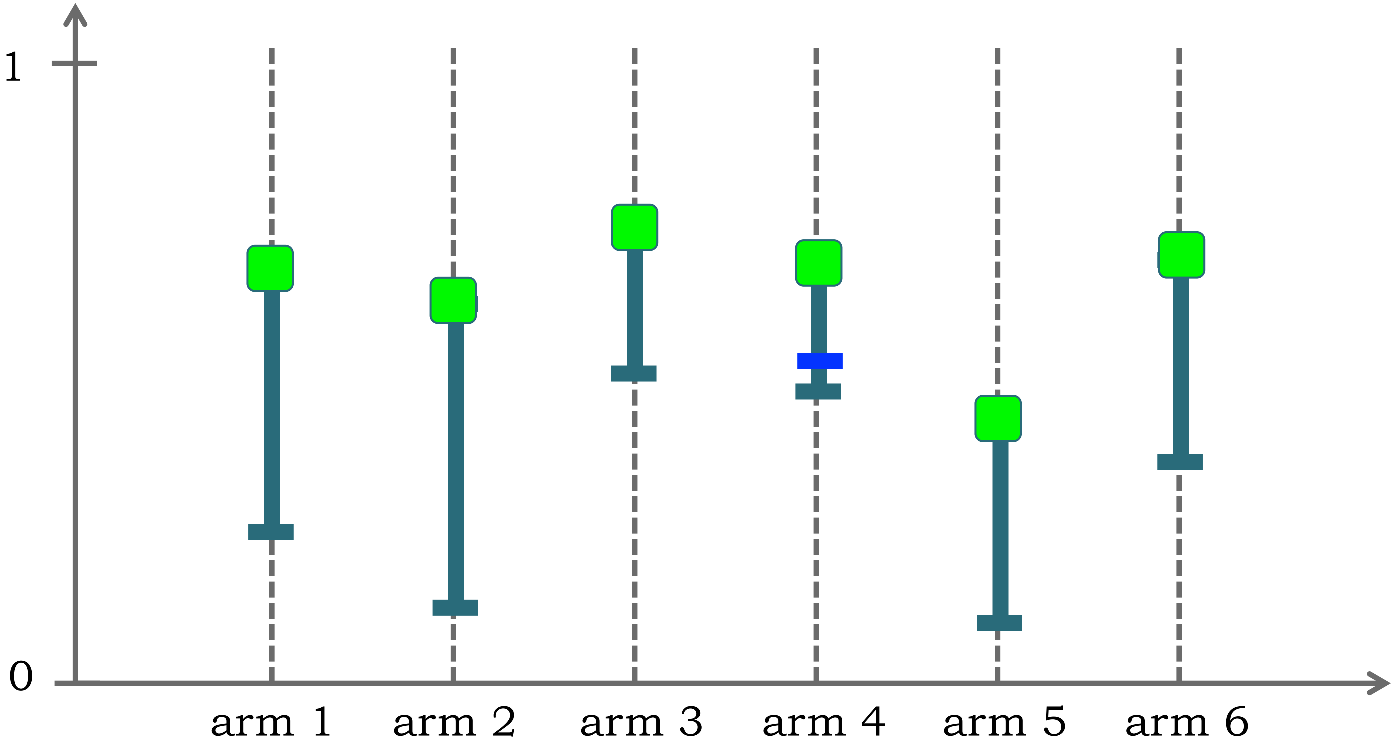
```
for all $1 \leq i \leq K$ do
    $\hat{\mu}_i \leftarrow \infty$ {empirical mean of arm $a_i$}
    $t_i \leftarrow 0$ {number of times played arm $a_i$}
end for
$t \leftarrow 1$
while true do
    $k \leftarrow \arg\max_i \hat{\mu}_i + \sqrt{\frac{2 \log t}{t_i}}$ {upper confidence bound from Chernoff-Hoeffding}
    play arm $a_k$, update empirical mean $\hat{\mu}_k$, increment $t_k$
    $t \leftarrow t + 1$
end while
```

The UCB algorithm, introduced by Auer et al. (2002), implements the **optimism in the face of uncertainty** principle.

THE UCB ALGORITHM



THE UCB ALGORITHM



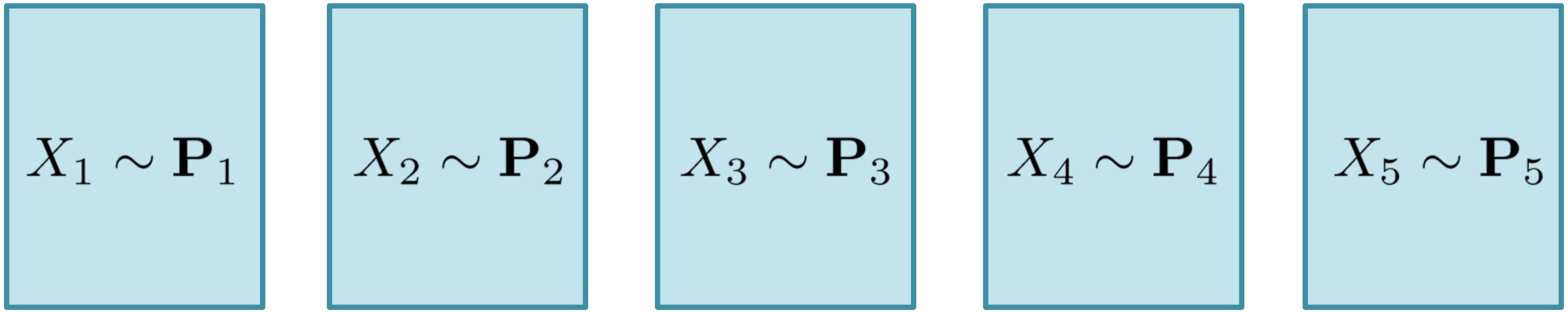
BOUND ON EXPECTED REGRET

Theorem: Assume rewards in $[0,1]$ (i.e., distributions $\mathbf{P}_1, \ldots, \mathbf{P}_K$ with support in $[0,1]$). The expected cumulative regret of UCB after any number of rounds T is upper-bounded by

$$\left[8 \sum_{i\colon \mu_i < \mu^*} \left(\frac{\log T}{\Delta_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{j=1}^{K} \Delta_j \right) \in \mathcal{O}(K \log T) \ ,$$

where $\Delta_i = \mu^* - \mu_i$.

PREFERENCE-BASED BANDITS

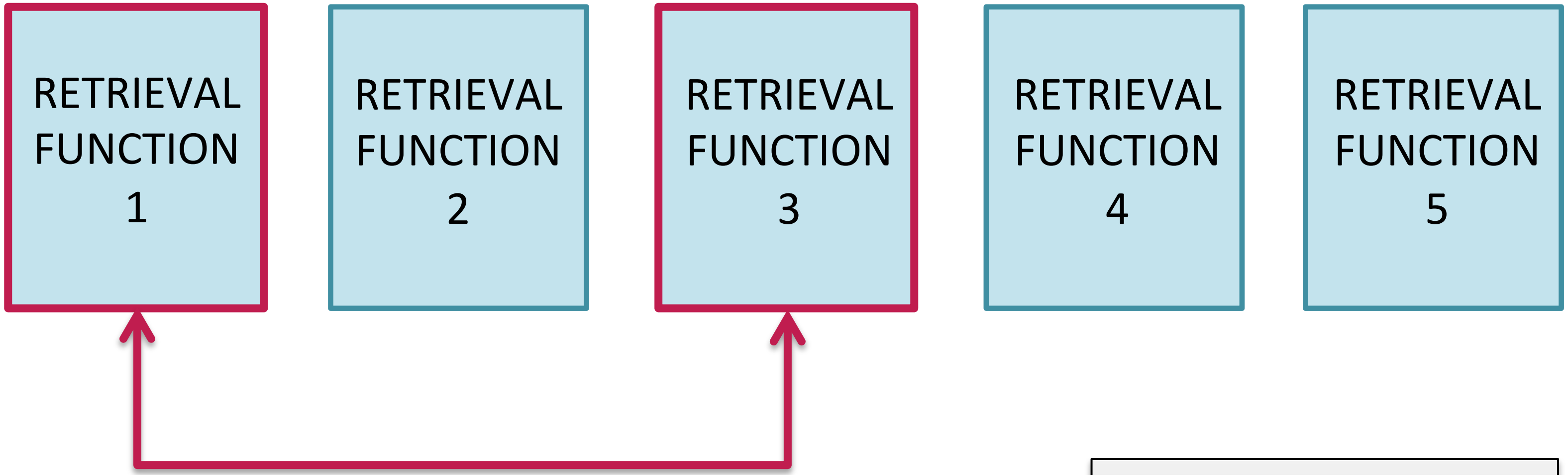


$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

In many applications,

- the assignment of (numeric) **rewards to single outcomes** (and therefore the assessment of individual options on an absolute scale) is difficult,
- while the **qualitative comparison between pairs of outcomes** (arms/ options) is more feasible.

PREFERENCE-BASED BANDITS

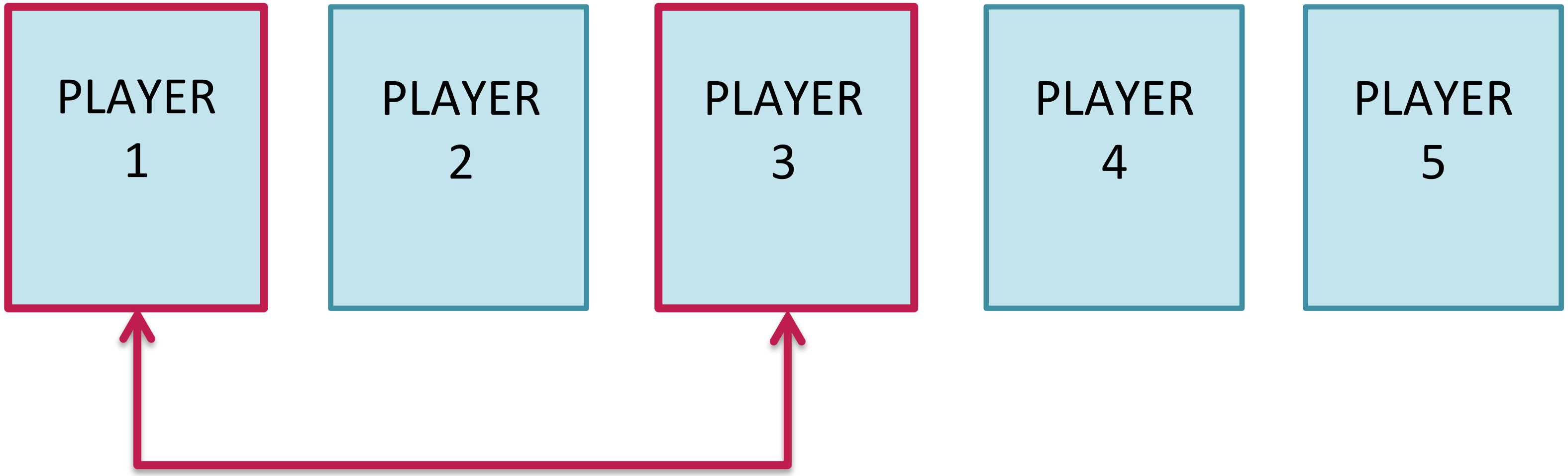


$$X_3 \succ X_1$$

The result returned by the third retrieval function, for a given query, is preferred to the result returned by the first search engine.

Noisy preference can be inferred from how a user clicks through an **interleaved** list of documents [Radlinski et al., 2008].

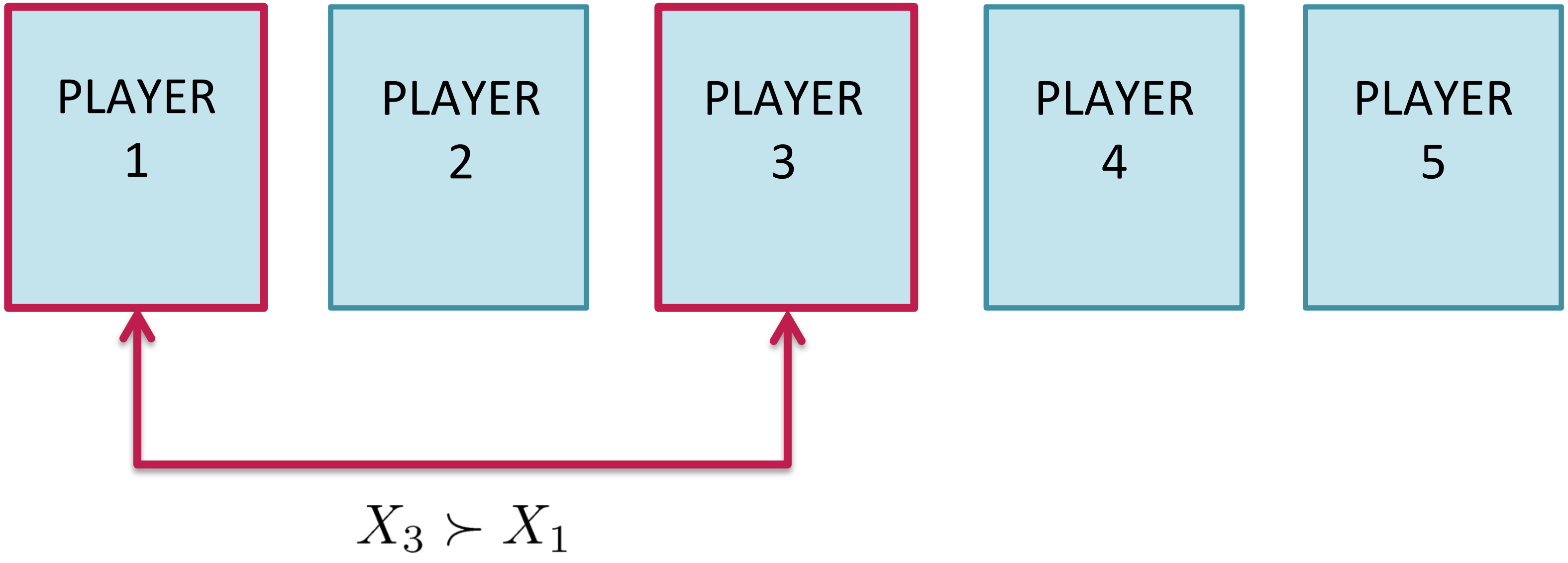
PREFERENCE-BASED BANDITS



$$X_3 \succ X_1$$

Third player has beaten first player in a match.

PREFERENCE-BASED BANDITS



- *This setting has first been introduced as the* ***dueling bandits problem*** *(Yue and Joachims, 2009).*
- *More generally, we shall speak of* ***preference-based multi-armed bandits*** *(PB-MAB).*

FORMAL SETTING

- fixed set of arms (options) $\mathcal{A} = \{a_1, \ldots, a_K\}$
- **action space** of the learner (agent) $= \{\, (i,j) \,|\, 1 \leq i \leq j \leq K \,\}$ (compairing pairs of arms a_i and a_j)
- feedback generated by an (unknown, time-stationary) probabilistic process characterized by a **preference relation**

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K,1} & q_{K,2} & \cdots & q_{K,K} \end{bmatrix} ,$$

where

$$q_{i,j} = \mathbf{P}\left(a_i \succ a_j\right)$$

- typically, $\mathbf{Q}$ is reciprocal $(q_{i,j} = 1 - q_{j,i})$

THE PREFERENCE RELATION

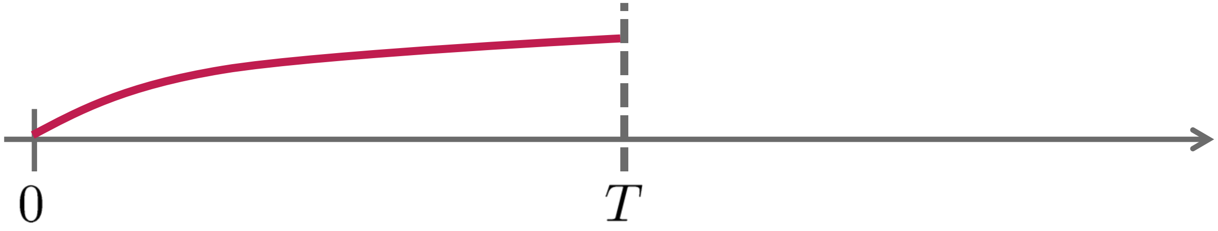
- We say arm a_i beats arm a_j if $q_{i,j} > 1/2$.
- The degrees of **distinguishability**

$$\Delta_{i,j} = q_{i,j} - \frac{1}{2}$$

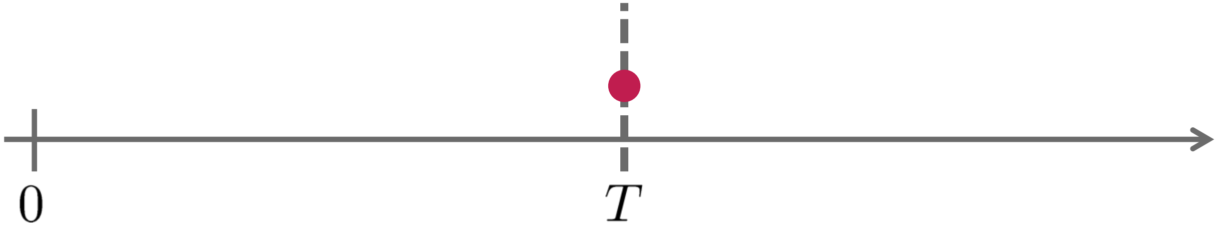
  quantify the hardness of a PB-MAB task.
- Assumptions on properties of $\mathbf{Q}$ are crucial for learning.
- **Coherence:** The pairwise comparisons need to provide hints (even if "noisy" ones) on the target.

TIME HORIZON

– Process iterates in discrete steps, either through a finite ($\mathbb{T} = [T] = \{1, \ldots, T\}$) or infinite ($\mathbb{T} = \mathbb{N}$) time horizon.



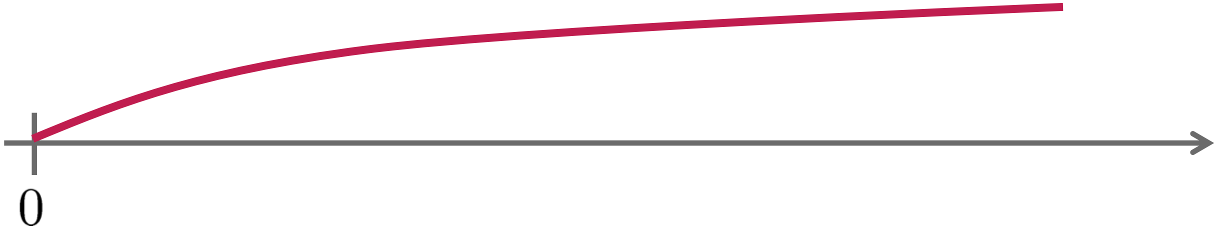
finite time horizon,
cumulative regret



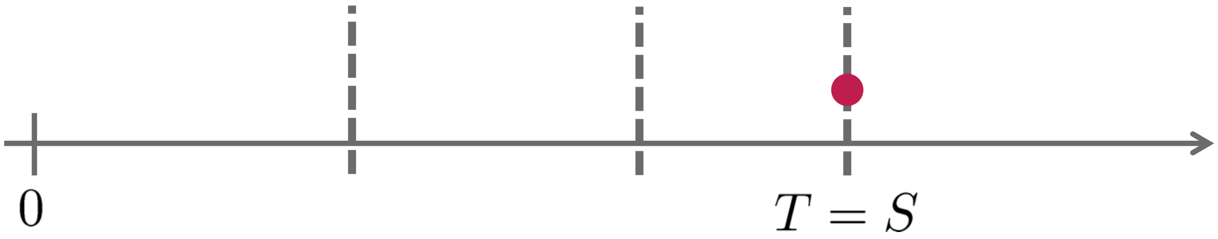
finite time horizon,
simple regret

TIME HORIZON

- Process iterates in discrete steps, either through a finite ($\mathbb{T} = [T] = \{1, \ldots, T\}$) or infinite ($\mathbb{T} = \mathbb{N}$) time horizon.



infinite time horizon, cumulative regret



pure exploration, termination decided by algorithm

PROBABILITY ESTIMATION

- In each iteration $t \in \mathbb{T}$, the learner selects $(i(t), j(t))$ and observes

$$\begin{cases} a_{i(t)} \succ a_{j(t)} & \text{with probability } q_{i(t),j(t)} \\ a_{j(t)} \succ a_{i(t)} & \text{with probability } q_{j(t),i(t)} \end{cases}$$

- Probability $q_{i,j}$ can be estimated by the proportion of wins of a_i against a_j up to iteration t:

$$\widehat{q}_{i,j}^{\,t} = \frac{w_{i,j}^t}{n_{i,j}^t} = \frac{w_{i,j}^t}{w_{i,j}^t + w_{j,i}^t}$$

- As samples are i.i.d., this is a plausible estimate; yet, it might be biased, since $n_{i,j}^t$ depends on the choice of the learner and hence on the data ($n_{i,j}^t$ is a random quantity).

PROBABILITY ESTIMATION

– A high probability confidence interval of the form

$$\left[\,\widehat{q}^{\,t}_{i,j} - c^t_{i,j},\ \widehat{q}^{\,t}_{i,j} + c^t_{i,j}\,\right]$$

can be obtained based on concentration inequalities like Hoeffding:

Let $X_1, \ldots, X_m$ be i.i.d. random variables with values in $[0,1]$, $\mu = \mathbf{E}(X_i)$ and $\bar{X} = (X_1 + \ldots + X_m)/m$. Then, for any $\epsilon > 0$,

$$\mathbf{P}\big(|\bar{X} - \mu| > \epsilon\big) \leq 2\exp\big(-2\epsilon^2 m\big) \ .$$

Thus,

$$\mu \leq \bar{X} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

with probability at least $1 - \delta$.

PROBABILITY ESTIMATION

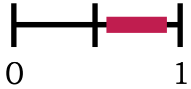
- A high probability confidence interval of the form

$$\left[\widehat{q}_{i,j}^{t} - c_{i,j}^{t} ,\ \widehat{q}_{i,j}^{t} + c_{i,j}^{t} \right]$$

  can be obtained based on concentration inequalities like Hoeffding.

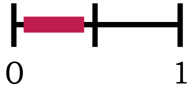
- Option a_i beats a_j with high probability if

$$\widehat{q}_{i,j}^{t} - c_{i,j}^{t} > 1/2 \ .$$

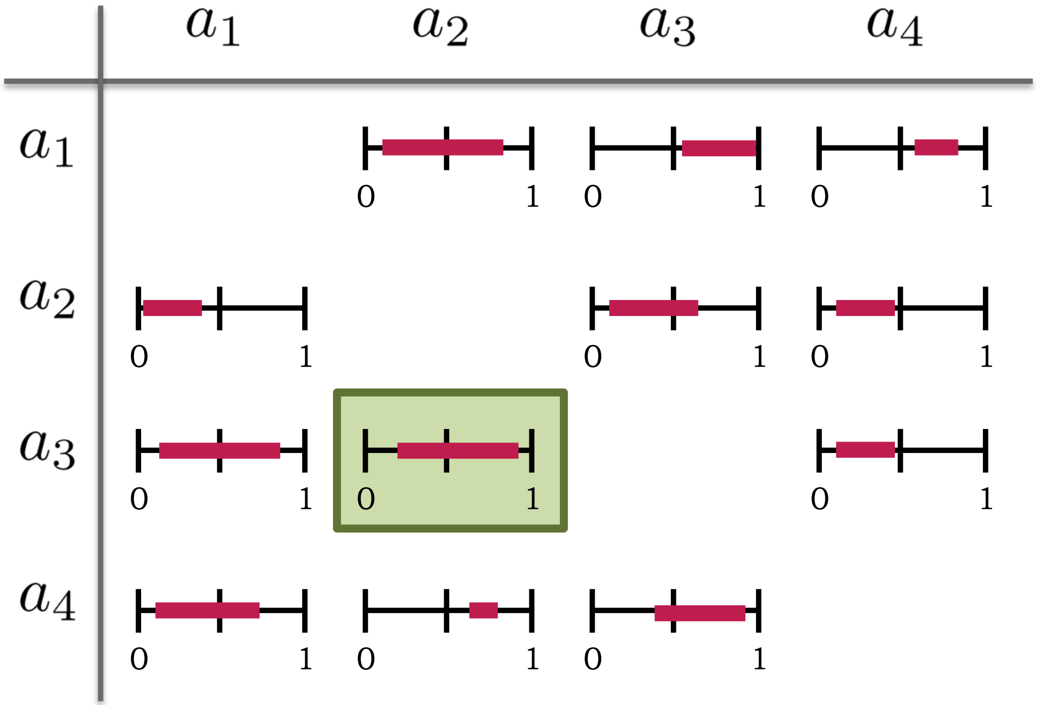


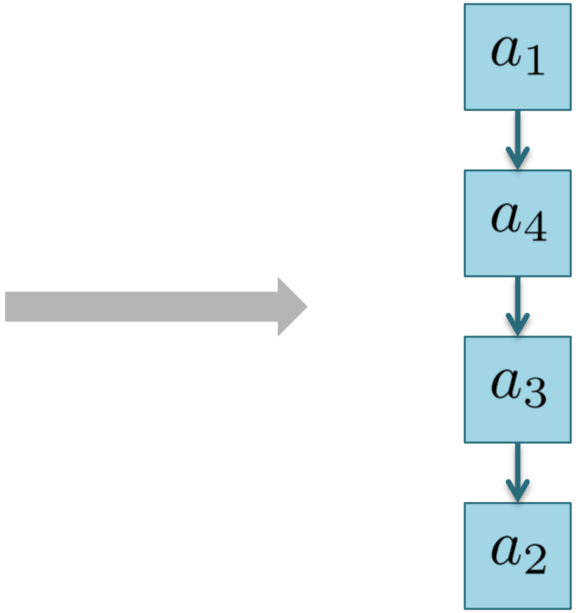
- Option a_j beats a_i with high probability if

$$\widehat{q}_{i,j}^{t} + c_{i,j}^{t} < 1/2 \ .$$



PAIRWISE SAMPLING





uncertainty about pairwise preferences

translates into → uncertainty about ranking

SAMPLE COMPLEXITY

- In each iteration of the **pure exploration** setting, the learner either selects a pair of arms to be compared or terminates and return its recommendation.

- A **recommendation** could be

  - a single best arm,
  - a complete ranking of all arms,
  - a probability distribution over all rankings,
  - the subset of top-k arms,
  - ...

SAMPLE COMPLEXITY

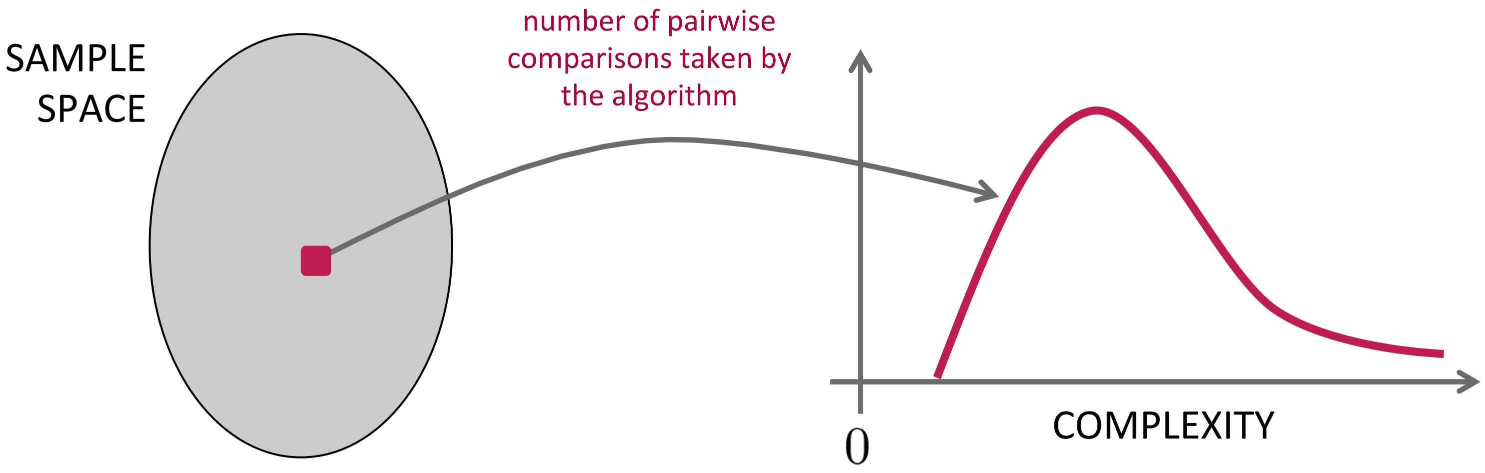
- The **sample complexity** S of the learner is the number of comparisons prior to termination.
- A bound on this complexity is of the form

$$S \leq B(\mathbf{Q}, K, \delta) \ ,$$

with $1 - \delta$ a lower bound on the probability that the learner terminates and returns the correct solution.

- With probability δ, the learner may either guess incorrectly or not terminate. Therefore, it is difficult so say anything about the **expectation** of the complexity.

SAMPLE COMPLEXITY

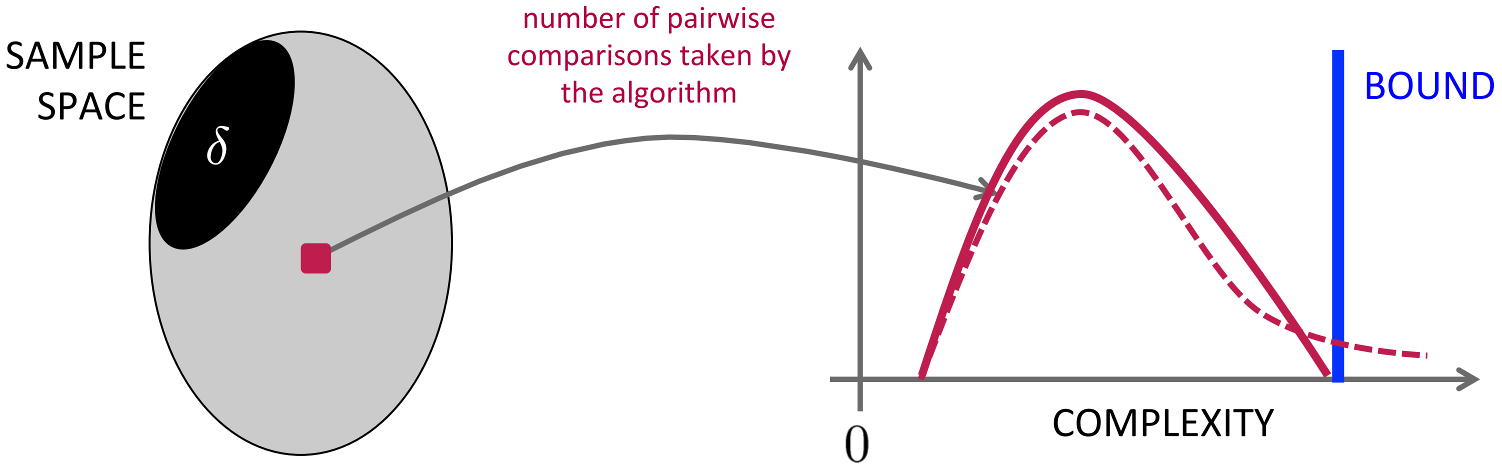


SAMPLE SPACE

number of pairwise comparisons taken by the algorithm

0

COMPLEXITY



SAMPLE SPACE

δ

number of pairwise comparisons taken by the algorithm

BOUND

0

COMPLEXITY

PAC ALGORITHMS

- Gaining efficiency at the cost of optimality!
- An algorithm is called (ϵ, δ)-PAC preference-based MAB algorithm with a **sample complexity**

  $$B(\mathbf{Q}, K, \epsilon, \delta)$$

  if it terminates and returns an ϵ-optimal recommendation with probability at least $1 - \delta$, and the number of comparisons is at most $B(\mathbf{Q}, K, \epsilon, \delta)$.
- Depending on the type of recommendation, the definition of ϵ-optimality is not necessarily straightforward.

THE NOTION OF REGRET

- Suboptimality of decision making is typically measured in terms of **expected regret** (cost of ignorance).

- If options have an inherent (expected) value μ_i, and actions correspond to selecting single options, a natural notion of (expected) regret is

$$r_t = \mu^* - \mu_{i(t)} = \max_{j \in [K]} \mu_j - \mu_{i(t)} \quad .$$

- The cumulative regret at time $T \in \mathbb{T}$ is

$$R^T = \sum_{t=0}^{T} r_t \quad .$$

THE NOTION OF REGRET

- How to penalize the selection of **pair of options** in the qualitative setting?

- Yue and Joachims (2009) proposed

$$r_t = f\big(\Delta_{i^*,i(t)}, \Delta_{i^*,j(t)}\big)$$

with $f(a,b) = \max(a,b)$, $f(a,b) = \min(a,b)$, $f(a,b) = (a+b)/2$.

- The regret is 0 if the best arm is compared to itself.

- Note that this definition presupposes the existence of a unique best arm a_{i^*} (in the form of a Condorcet winner).

THE NOTION OF REGRET

- The regret accumulated by an algorithm is a random variable that depends on the stochastic nature of the data-generating process (and maybe randomized decisions of the learner).
- An **expected regret bound** is of the form

$$\mathbf{E}\left[R^T\right] \le B(\mathbf{Q}, K, T) \ .$$

- A **high-probability regret bound** is of the form

$$\mathbf{P}\Big(R^T < B(\mathbf{Q}, K, T, \delta) \Big) \ge 1 - \delta \ .$$

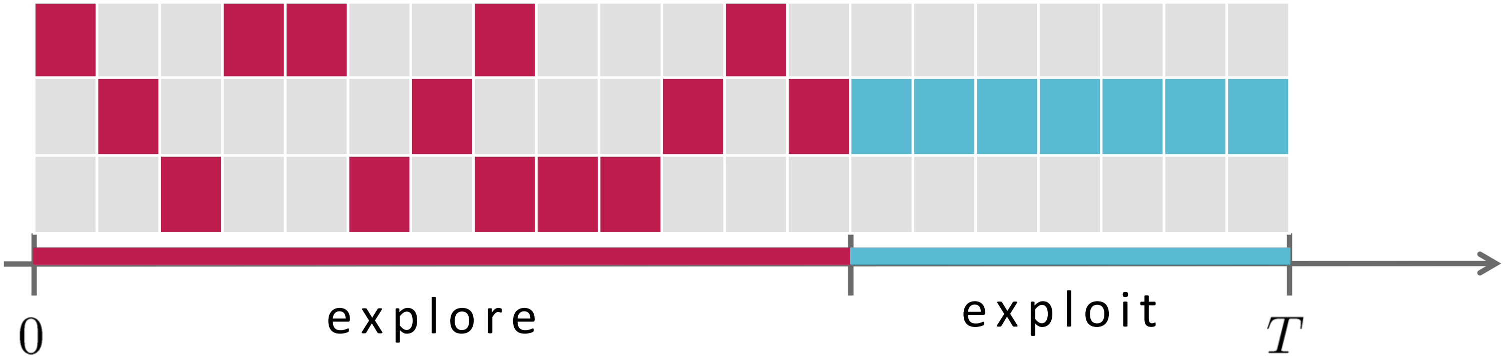
We say the regret of the learner is $\mathcal{O}(B(\mathbf{Q}, K, T, \delta))$ with high probability.

COMPARISON OF THE SETTINGS

	pure exploration	explore/exploit (finite time horizon)	expore/exploit (infinite horizon)
actions	pairwise comparison or termination	pairwise comparison	pairwise comparison
recommendation	top-1, top-k, ranking, ...	top-1	---
cost of action	unit cost	suboptimality of selected pair (regret)	suboptimality of selected pair (regret)
evaluation	sample complexity	simple regret or cumulative regret	cumulative regret
type of analysis	bounds, PAC bounds in the case of approximation	bounds on regret, expectation or high probability	bounds on expected regret, high probability bounds

EXPLORE-THEN-EXPLOIT

- **Explore-then-exploit** algorithms first try to identify the best arm with high probability, and then fully commit to this arm for the rest of the time horizon T (which is fixed and known beforehand).



What is a good tradeoff, i.e., how much time should be devoted to exploration?

EXPLORE-THEN-EXPLOIT

- **Explore-then-exploit** algorithms first try to identify the best arm with high probability, and then fully commit to this arm for the rest of the time horizon T (which is fixed and known beforehand).

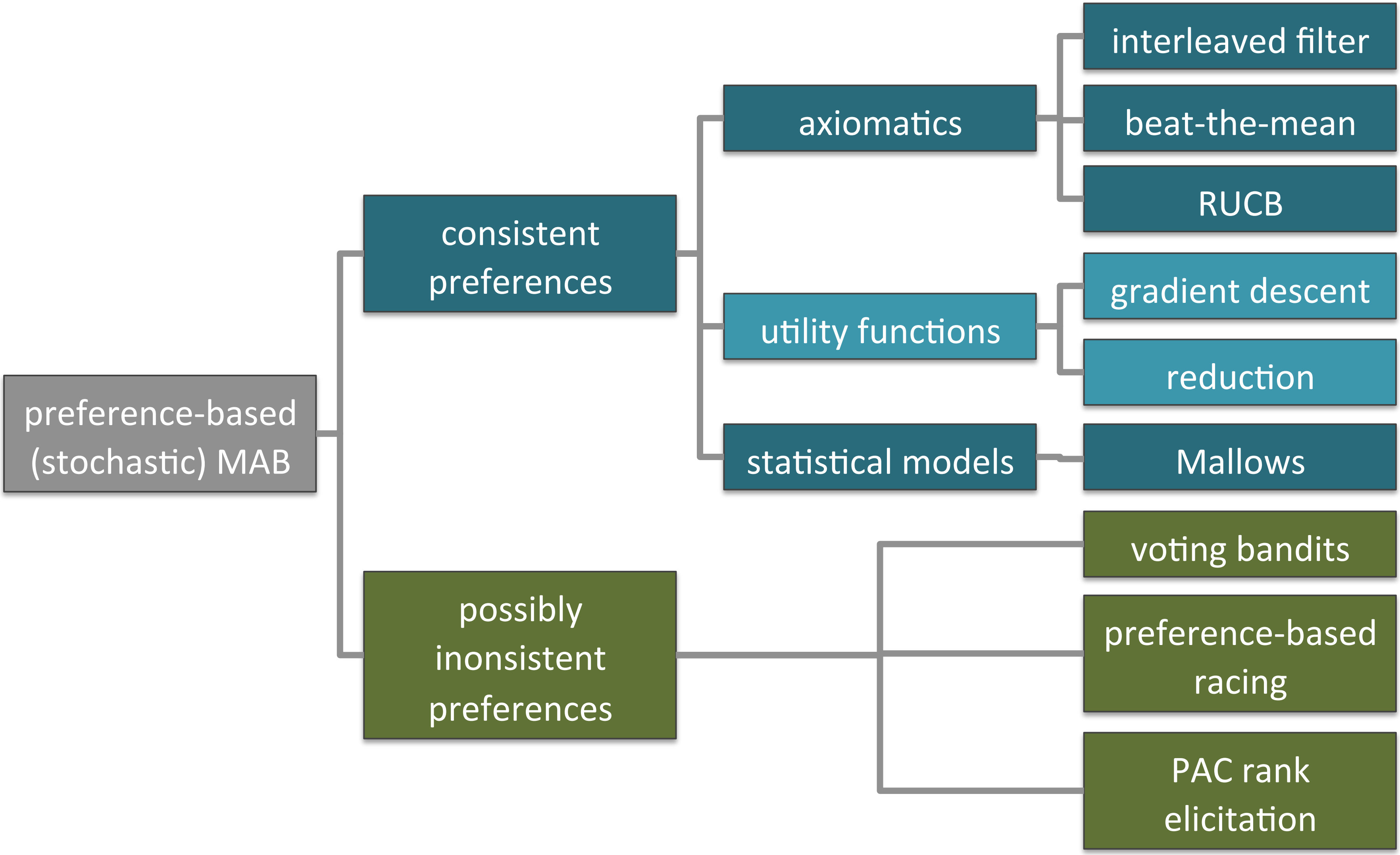
- Suppose **exploratory algorithm** A identifes a_{i^*} with probability $\geq 1 - \delta$. Then, with $\delta = 1/T$, the expected regret of an explore-then-exploit algorithm is

$$\mathbf{E}[R^T] \leq (1 - 1/T)\,\mathbf{E}[R^T_A] + (1/T)\,\mathcal{O}(T) = \mathcal{O}\left(\mathbf{E}[R^T_A] + 1\right) \quad .$$

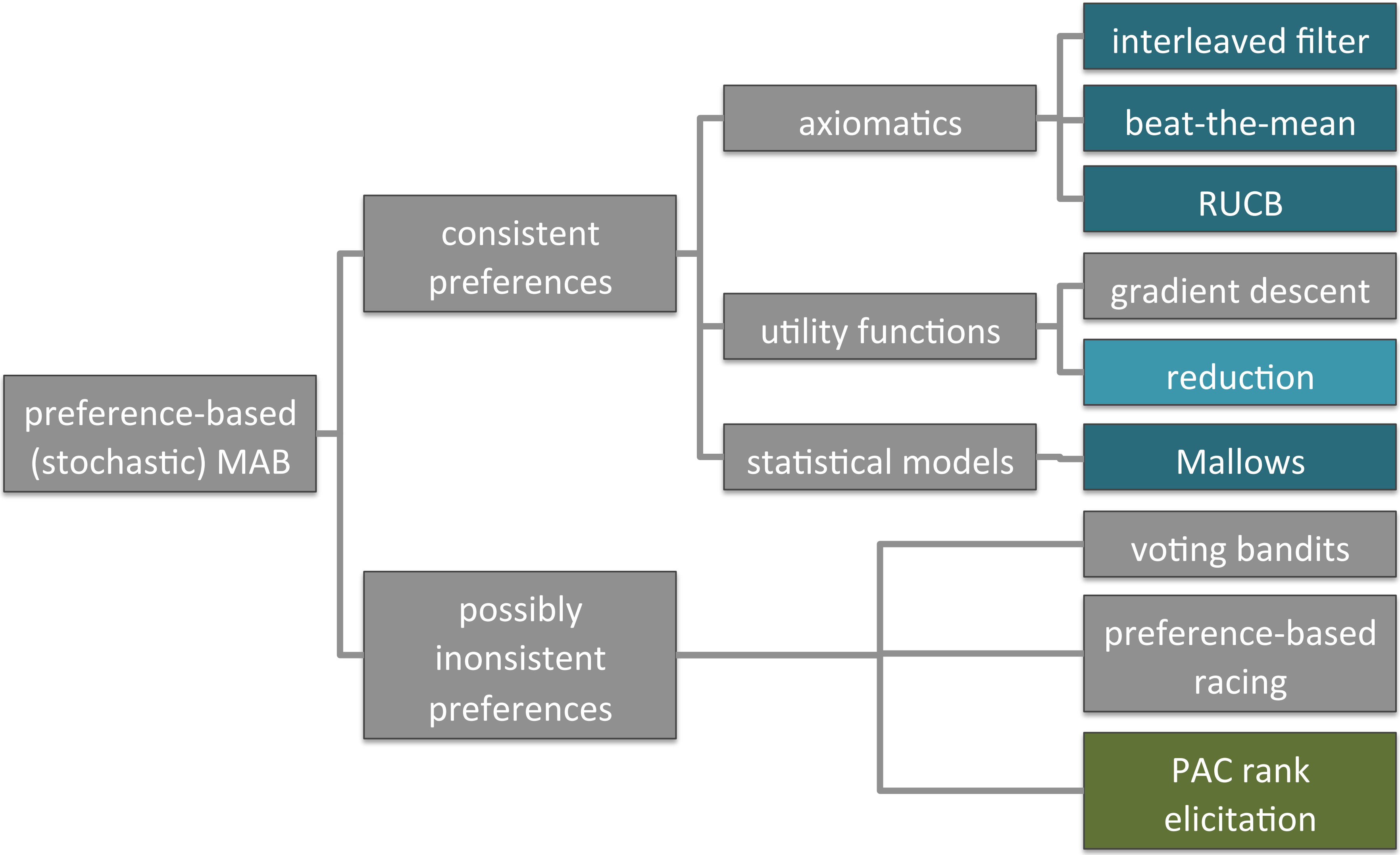
- Since the per round regret is at most one, the **sample complexity** of A upper-bounds the expected regret.

- Explore-then-exploit algorithms somehow blur the distinction between the two settings.

OVERVIEW OF METHODS



OVERVIEW OF METHODS



INTERLEAVED FILTER

Regularity assumptions on $\mathbf{Q}$ [Yue et al., 2012]:

- **Total order over arms**: there exists a total order $\succ$ on $\mathcal{A}$, such that $a_i \succ a_j$ implies $\Delta_{i,j} > 0$.

- **Strong stochastic transitivity**: for any triplet of arms such that $a_i \succ a_j \succ a_k$, the pairwise probabilities satisfy

$$\Delta_{i,k} \geq \max\left(\Delta_{i,j}, \Delta_{j,k}\right).$$

- **Stochastic triangle inequality**: for any triplet of arms such that $a_i \succ a_j \succ a_k$, the pairwise probabilities satisfy

$$\Delta_{i,k} \leq \Delta_{i,j} + \Delta_{j,k}.$$

INTERLEAVED FILTER

- Yue and Joachims (2009) proposed the first explore-then-exploit algorithm (with time horizon T given in advance).

- Exploration consists of a sequential elimination strategy called **Interleaved Filtering** (IF), which identifies the best arm with probability at least $1 - \delta$.

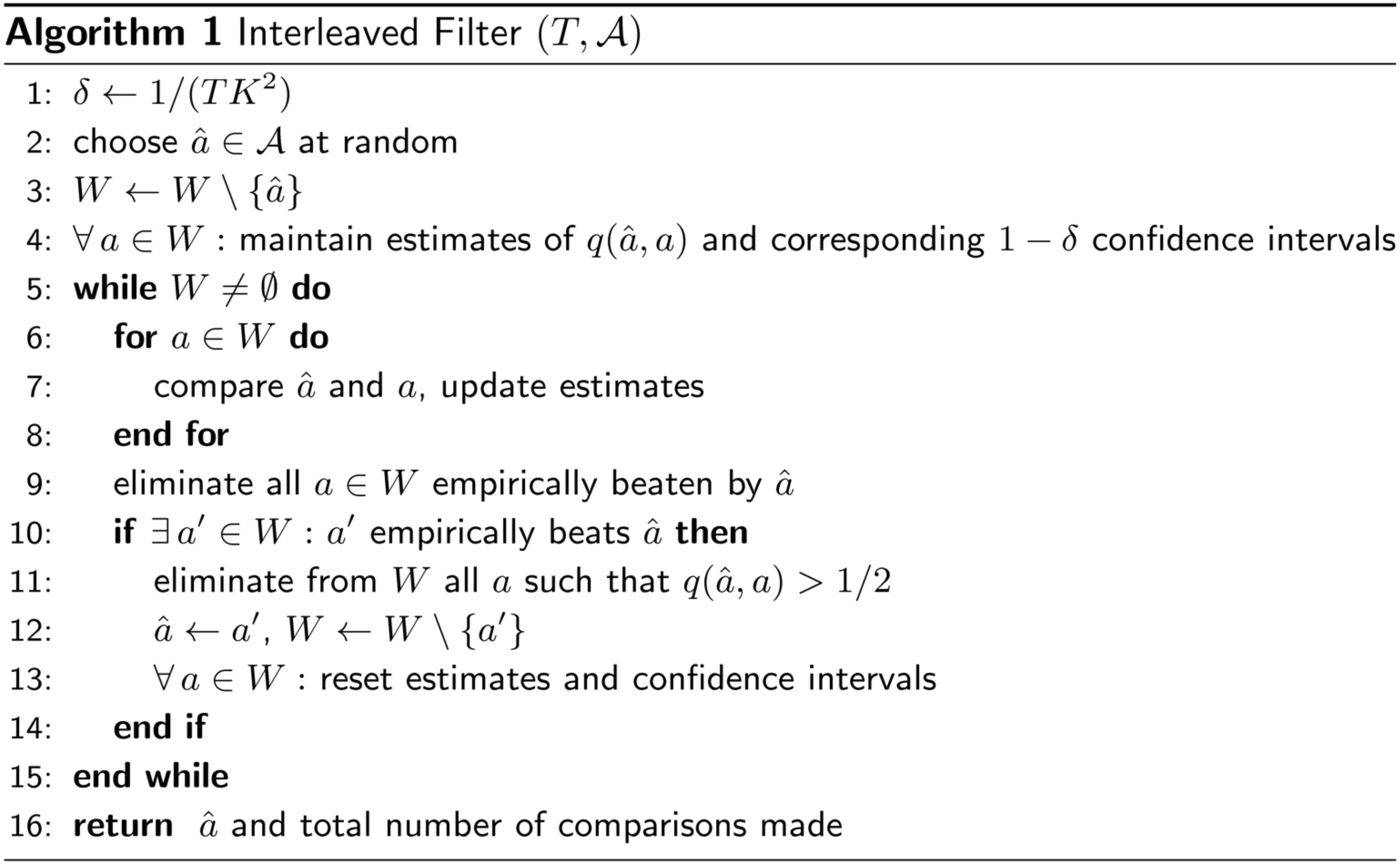
- The currently **selected arm** a_i is compared to the rest of the **active arms**. If a_j beats a_i ($\widehat{q}_{i,j} + c_{i,j} < 1/2$), then a_i is eliminated and a_j selected.

- **Pruning**: if $\widehat{q}_{i,j} - c_{i,j} > 1/2$ for an arm a_j, then a_j is eliminated.

- Assuming the horizon T to be finite and known in advance, IF incurs an **expected regret**

$$\mathbf{E}\left[R_{IF}^{T}\right] = \mathcal{O}\left(\frac{K}{\min_{j \neq i^*} \Delta_{i^*,j}} \log T\right) \quad .$$

INTERLEAVED FILTER



Algorithm 1 Interleaved Filter $(T, \mathcal{A})$

```
$\delta \leftarrow 1/(TK^2)$
choose $\hat{a} \in \mathcal{A}$ at random
$W \leftarrow W \setminus \{\hat{a}\}$
$\forall\, a \in W$ : maintain estimates of $q(\hat{a}, a)$ and corresponding $1 - \delta$ confidence intervals
while $W \neq \emptyset$ do
    for $a \in W$ do
        compare $\hat{a}$ and $a$, update estimates
    end for
    eliminate all $a \in W$ empirically beaten by $\hat{a}$
    if $\exists\, a' \in W$ : $a'$ empirically beats $\hat{a}$ then
        eliminate from $W$ all $a$ such that $q(\hat{a}, a) > 1/2$
        $\hat{a} \leftarrow a'$, $W \leftarrow W \setminus \{a'\}$
        $\forall\, a \in W$ : reset estimates and confidence intervals
    end if
end while
return $\hat{a}$ and total number of comparisons made
```

STRONG STOCHASTIC TRANSITIVITY

	a_1	a_2	a_3	a_4
a_1	--	0.6	0.6	0.6
a_2	0.4	--	0.8	0.9
a_3	0.4	0.2	--	0.6
a_4	0.4	0.1	0.4	--

violation of strong stochastic transitivity

BEAT THE MEAN

- Yue and Joachims (2011) only require **relaxed stochastic transitivity**: There is a $\gamma \geq 1$ such that, for any triplet of arms such that $a_{i^*} \succ a_i \succ a_j$ with respect to the total order $\succ$,

$$\gamma \, \Delta_{i^*,j} \geq \max\{\Delta_{i^*,i}, \Delta_{i,j}\} \quad .$$

- **Beat-The-Mean** (BTM) is an elimination strategy resembling IF. However, it follows a different strategy for pairing arms.

- Like for IF, the time horizon is fixed in advance (more correctly, one may denote algorithms IF$_T$ and BTM$_T$).

BEAT THE MEAN

- BTM maintains a set of **active arms**. It always selects an arm with the fewest comparisons so far and pairs it with a randomly chosen competitor.

- Based on the pairwise comparisons, a **score** b_i is assigned to each a_i, which is an empirical estimate of the probability that a_i is winning in a "random" pairwise comparison.

- The idea is that comparing an arm a_i to the "mean" arm, which beats half of the arms, is equivalent to comparing a_i to a random arm.

- A **confidence interval** for each of the b_i scores is derived.

- An arm is eliminated from the set of active arms as soon as there is another arm with a significantly higher score.

- The process ends when there is only a single active arm left (or the number of comparisons is large enough to guarantee ϵ-approximation).

BEAT THE MEAN

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**0.6**
a_2	0.4	--	0.8	0.9	**0.7**
a_3	0.4	0.2	--	0.6	**0.4**
a_4	0.4	0.1	0.4	--	**0.3**

The best arm does not necessarily have the best expected performance (highest probability of winning), though it cannot be the worst either.

BEAT THE MEAN

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**0.6**
a_2	0.4	--	0.8	0.9	**0.7**
a_3	0.4	0.2	--	0.6	**0.4**
~~a_4~~	~~0.4~~	~~0.1~~	~~0.4~~	~~--~~	~~**0.3**~~

	a_1	a_2	a_3	
a_1	--	0.6	0.6	**0.6**
a_2	0.4	--	0.8	**0.6**
~~a_3~~	~~0.4~~	~~0.2~~	~~--~~	~~**0.3**~~

	a_1	a_2	
a_1	--	0.6	**0.6**
~~a_2~~	~~0.4~~	~~--~~	~~**0.4**~~

BEAT THE MEAN

- **Online setting** ("explote-then-exploit"): For a finite time horizon, the regret accumulated by BTM is

$$\mathcal{O}\left(\frac{\gamma^7 K}{\min_{j\neq i^*} \Delta_{i^*,j}} \log T\right)$$

with high probability.

- **PAC setting**: Moreover, BTM is an (ϵ, δ)-PAC preference-based learner, the sample complexity of which is

$$\mathcal{O}\left(\frac{\gamma^6 K}{\epsilon^2} \log \frac{K\gamma \log(K/\delta)}{\delta\epsilon}\right)$$

for large enough N (maximum number of comparisons).

RUCB

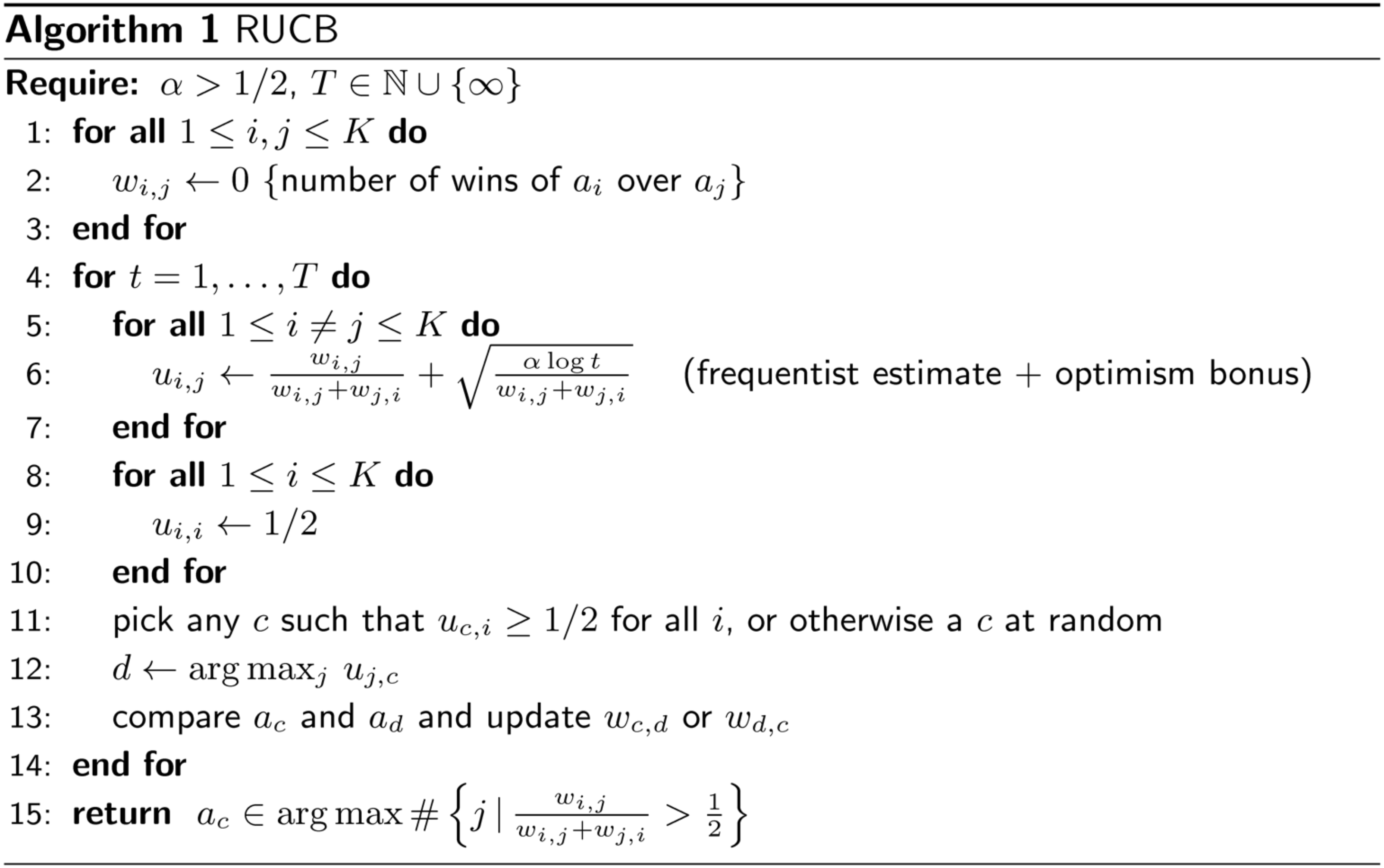
- Zoghi et al. (2014) proposed **Relative UCB**, which only assumes the existence of a Condorcet winner.
- Like the well-known UCB, this algorithm is based on the "optimism in the face of uncertainty" principle: the arms to be compared next are selected based on the upper boundaries of their confidence intervals

$$\left[\, \widehat{q}_{i,j} - c_{i,j} \, , \, \widehat{q}_{i,j} + c_{i,j} \,\right] \; .$$

- In an iteration, RUCB selects a_c randomly from the set of potential Condorcet winners ($\widehat{q}_{c,j} + c_{c,j} > 1/2$ for all j).
- Finally, regular UCB is performed relative to a_c, namely, a_c is compared to the arm a_d supposed to yield the smallest regret:

$$d = \arg\max_{\ell \neq c} \; \widehat{q}_{\ell,d} + c_{\ell,d}$$

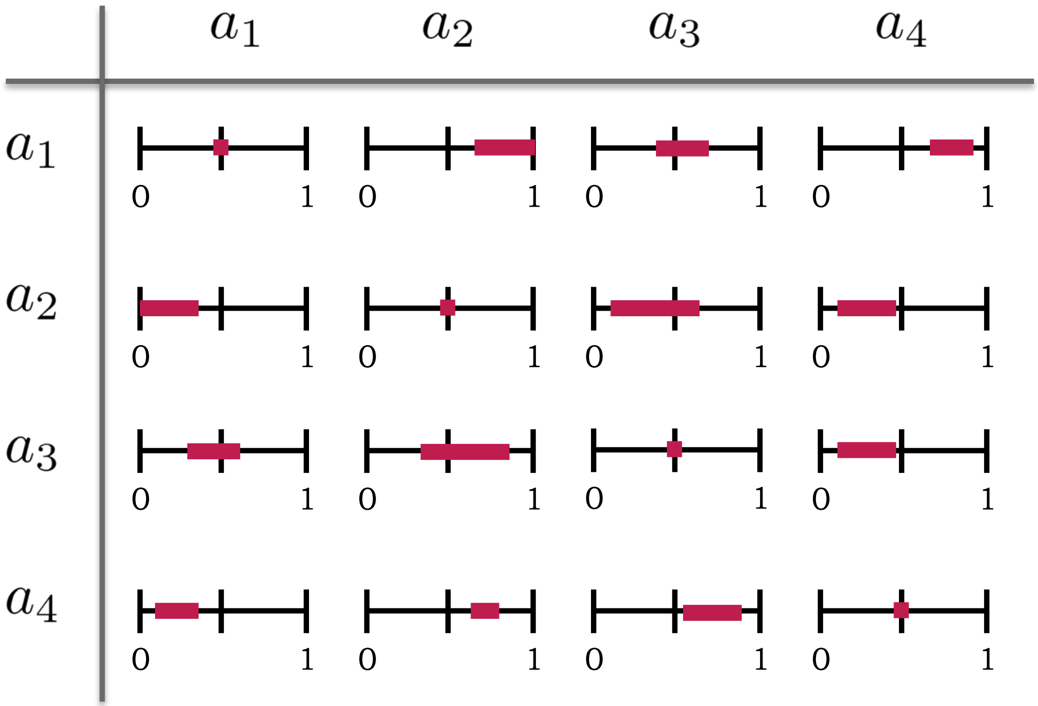
RUCB



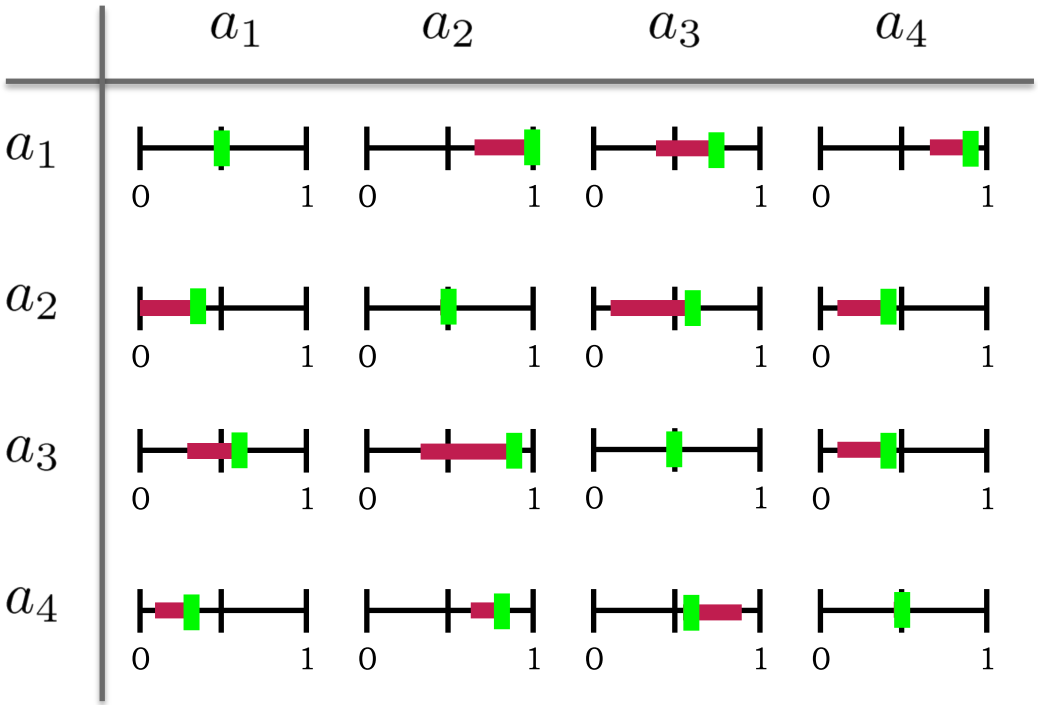
Algorithm 1 RUCB

```
Require: $\alpha > 1/2$, $T \in \mathbb{N} \cup \{\infty\}$
for all $1 \leq i, j \leq K$ do
    $w_{i,j} \leftarrow 0$ {number of wins of $a_i$ over $a_j$}
end for
for $t = 1, \ldots, T$ do
    for all $1 \leq i \neq j \leq K$ do
        $u_{i,j} \leftarrow \frac{w_{i,j}}{w_{i,j}+w_{j,i}} + \sqrt{\frac{\alpha \log t}{w_{i,j}+w_{j,i}}}$    (frequentist estimate + optimism bonus)
    end for
    for all $1 \leq i \leq K$ do
        $u_{i,i} \leftarrow 1/2$
    end for
    pick any $c$ such that $u_{c,i} \geq 1/2$ for all $i$, or otherwise a $c$ at random
    $d \leftarrow \arg\max_j u_{j,c}$
    compare $a_c$ and $a_d$ and update $w_{c,d}$ or $w_{d,c}$
end for
return $a_c \in \arg\max \# \left\{ j \mid \frac{w_{i,j}}{w_{i,j}+w_{j,i}} > \frac{1}{2} \right\}$
```

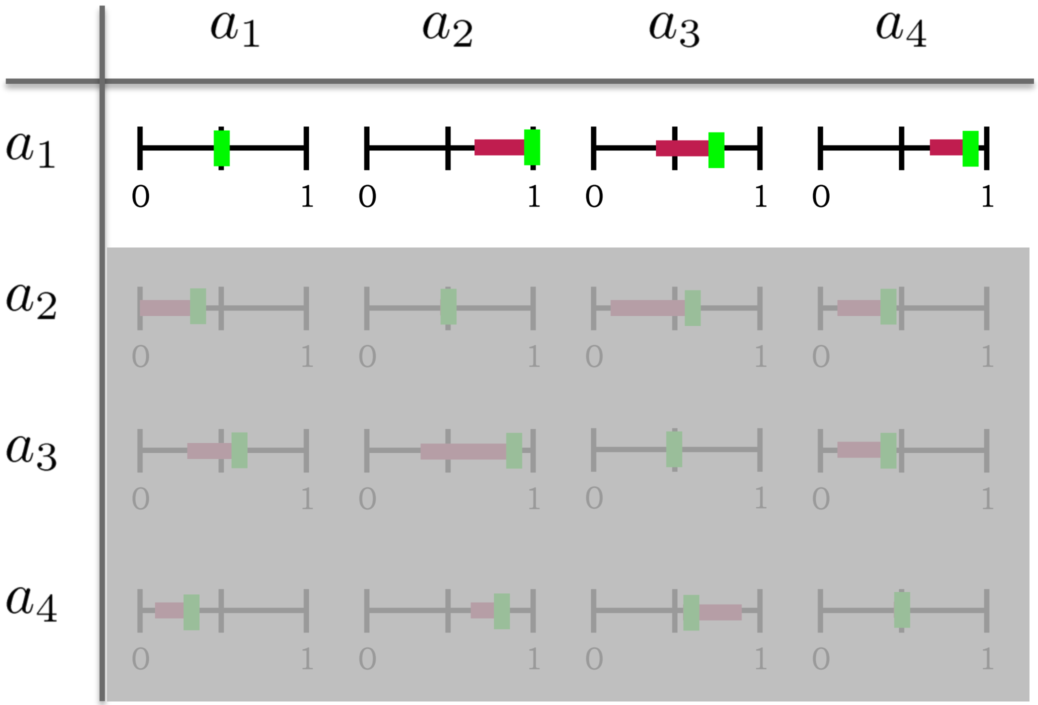
RUCB



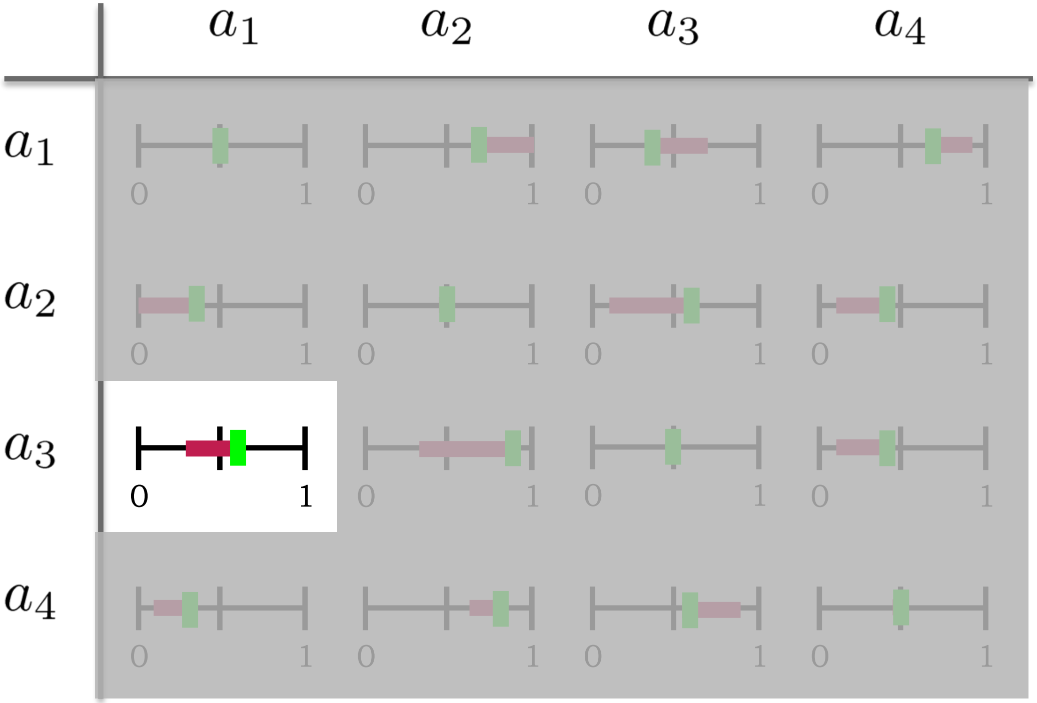
RUCB



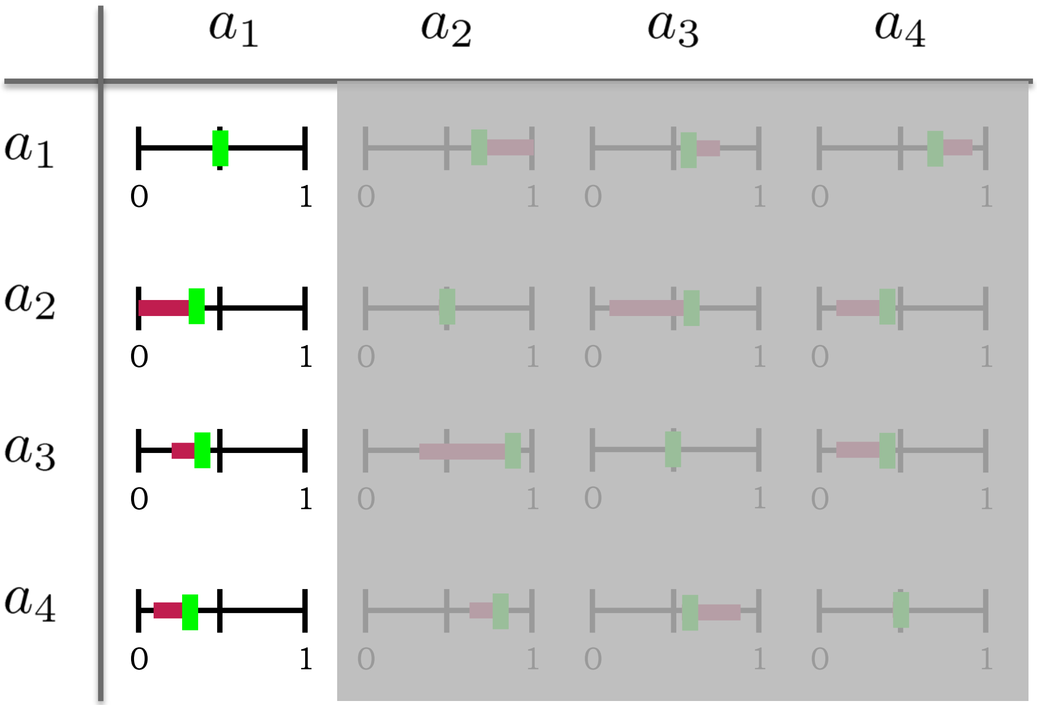
RUCB



RUCB



RUCB



first arm will be compared to itself

RUCB

- **Horizonless regret bounds** are provided, both for the expected and high probability case. Unlike the bounds for **IF** and **BTM**, they are valid for each time step (no need to specify/guess the exploration horizon T beforehand).

- Both bounds are of order $\mathcal{O}(K \log T)$.

- The constants again depend on the $\Delta_{i,j}$ values, however, they are not directly comparable to those of **IF** and **BTM**.

- Empirically, RUCB seems to outperform both IF and BTM.

RUCB: MAIN RESULT

Theorem (Zoghi et al., 2014): Suppose a_1 to be the best arm (Condorcet winner). Given $\delta > 0$ and $\alpha > 1/2$, define

$$C(\delta) = \left(\frac{(4\alpha - 1)K^2}{(2\alpha - 1)\delta} \right)^{\frac{1}{2\alpha - 1}} \text{ and } D_{i,j} = \frac{4\alpha}{\min(\Delta_i^2, \Delta_j^2)}$$

for all $1 \leq i \neq j \leq K$, where $\Delta_i = 1/2 - q_{i,1}$. Moreover, let $D_{i,i} = 0$ for all i. Then, for any pair $(i, j) \neq (1, 1)$, the number of comparisons $N_{i,j}(t)$ between a_i and a_j taken by RUCB up to time t satisfies

$$\mathbf{P}\Big(\forall\, t : \; N_{i,j}(t) \leq \max\left\{ C(\delta), D_{i,j} \log t \right\} \Big) > 1 - \delta.$$

Moreover, the following high probability bound holds for the regret:

$$\mathbf{P}\Big(\forall\, t : \; R^t \leq C(\delta)\Delta^* + \sum_{i>j} D_{i,j} \Delta_{i,j} \log t \Big) > 1 - \delta,$$

where $\Delta^* = \max_i \Delta_i$ and $\Delta_{i,j} = \frac{\Delta_i + \Delta_j}{2}$.

RUCB: SKETCH OF PROOF

Lemma: Suppose a_1 to be the best arm (Condorcet winner), i.e., $q_{1,j} > 1/2$ for all $j > 1$. For any $\alpha > 0$, let

$$u_{i,j} = \frac{w_{i,j}(t)}{w_{i,j}(t) + w_{j,i}(t)} + \sqrt{\frac{\alpha \log t}{w_{i,j}(t) + w_{j,i}(t)}}$$

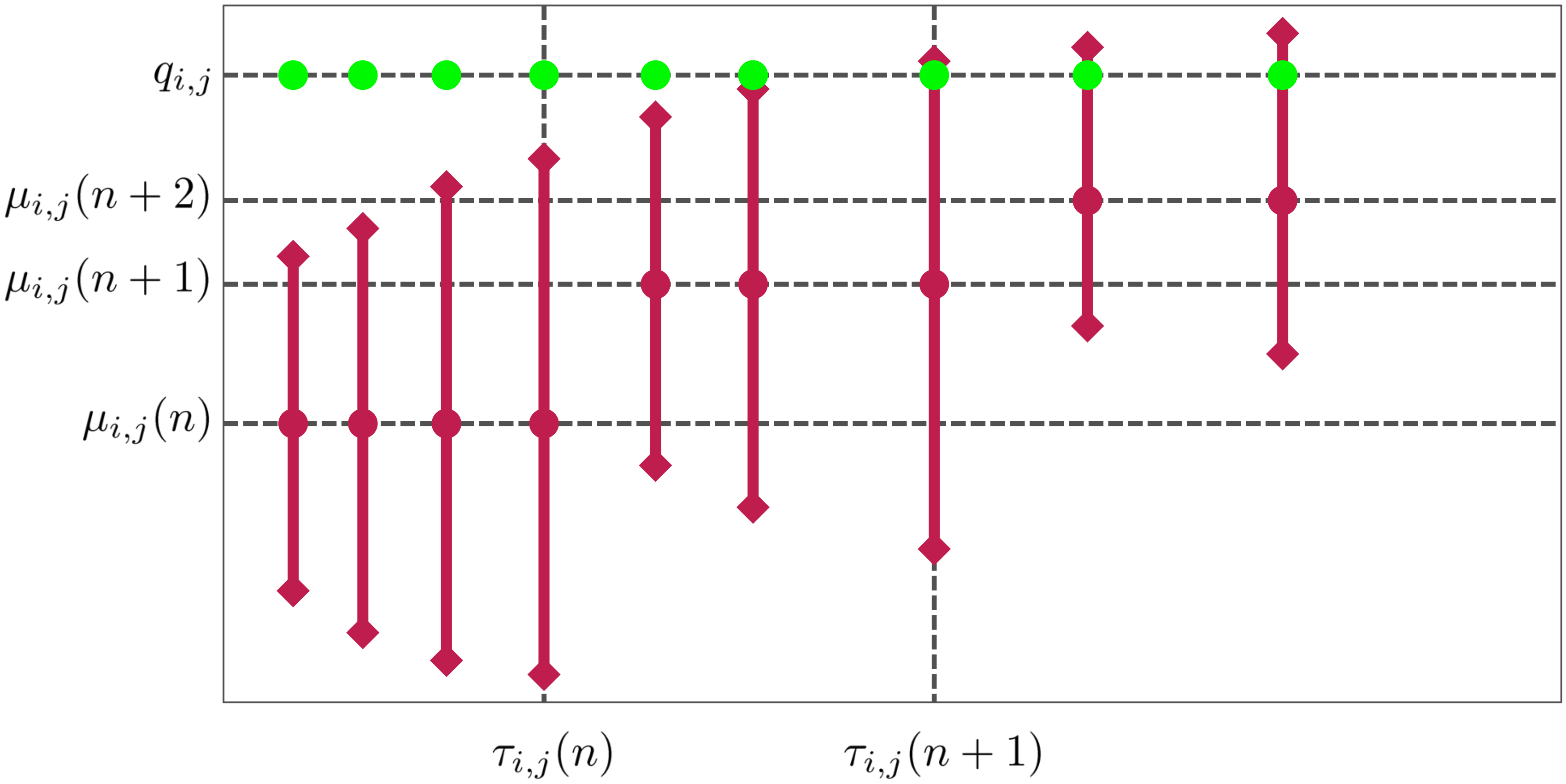
and $l_{i,j}(t) = 1 - u_{j,i}(t)$. Moreover, for any $\delta > 0$, let

$$C(\delta) = \left(\frac{(4\alpha - 1)K^2}{(2\alpha - 1)\delta} \right)^{\frac{1}{2\alpha - 1}} \quad .$$

Then,

$$\mathbf{P}\Big(\forall\, t > C(\delta), i, j : q_{i,j} \in [l_{i,j}(t), u_{i,j}(t)]\Big) > 1 - \delta \ .$$

RUCB: SKETCH OF PROOF



Confidence intervals $[l_{i,j}(t), u_{i,j}(t)]$ grow logarithmically unless a_i and a_j are compared. In that case, the interval shrinks and moves toward the ground truth $q_{i,j}$.

RUCB: SKETCH OF PROOF

- The first part of the theorem is obtained by showing that, for $(i,j) \neq (1,1)$, $N_{i,j}(t) > \max\{C(\delta), D_{i,j} \log t\}$ leads to a contradition (w.h.p).

- This is done by showing that $u_{i,j}(t) - l_{i,j}(t)$ must be "small" and "large" at the same time, which, together with the correctness of the confidence intervals (due to $t > C(\alpha)$ and the lemma), is not possible.

- The width of the interval $[u_{i,j}(t), l_{i,j}(t)]$ must be small because of the comparatively large number of comparisons $N_{i,j}(t)$.

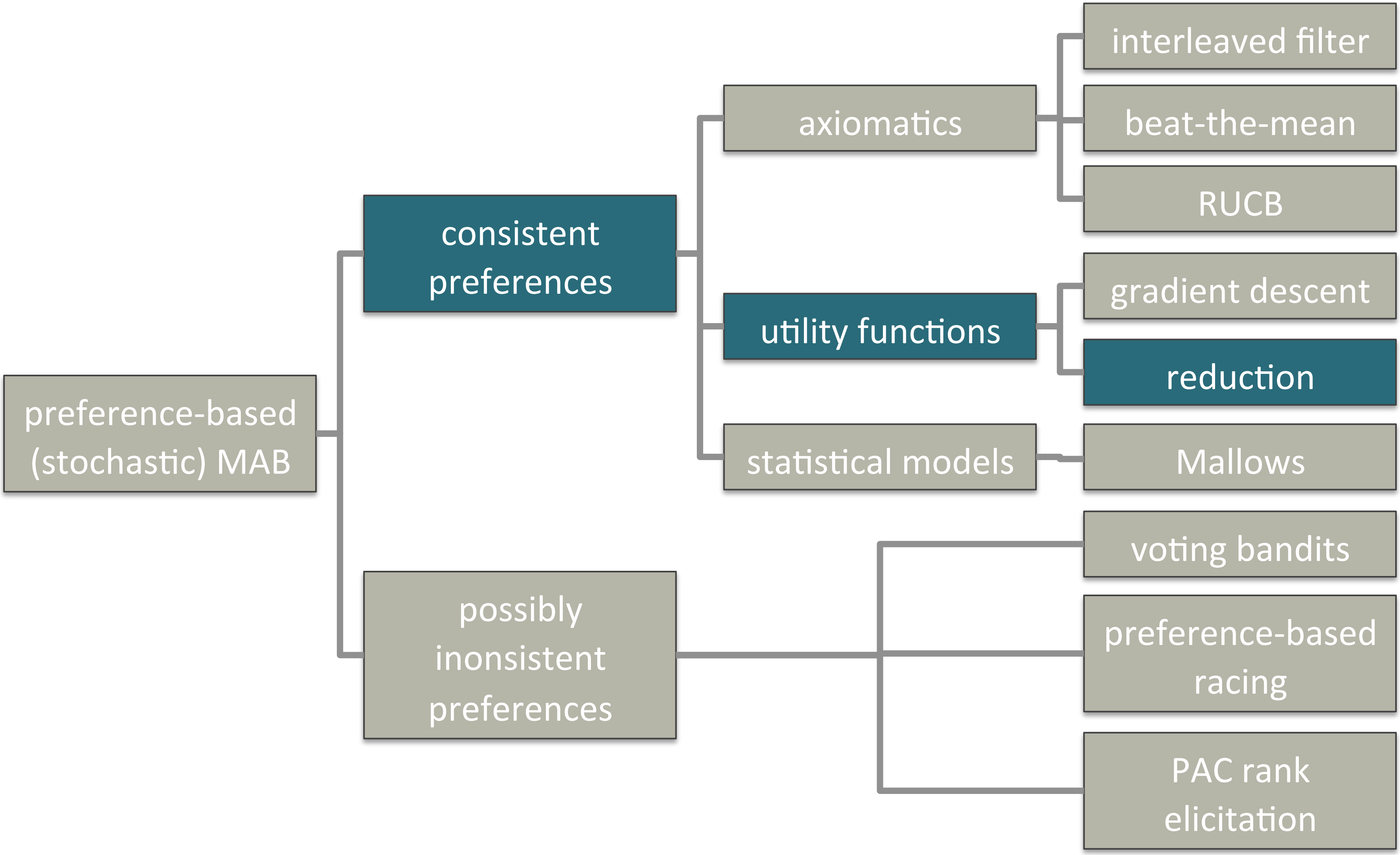
- Moreover, for (i,j) to be picked by RUCB for comparison, both $u_{i,j}$ and $u_{j,i}$ must be large, and hence $l_{i,j} = 1 - u_{j,i}$ small:

  $u_{i,j} > 1/2$, since otherwise i would not be picked as index c.

  $u_{j,i} \geq u_{1,i} \geq q_{1,i}$, since otherwise j would not be picked as index d.

- The second part of the theorem is shown by assuming the largest regret, Δ^*, to occur in the first $C(\delta)$ steps and, moreover, adding the regret for $D_{i,j} \log t$ comparisons of a_i and a_j.

OVERVIEW OF METHODS



LATENT UTILITY FUNCTIONS

Suppose each arm a_i is associated with a **latent utility** $\mu_i = \mu(a_i)$, and the pairwise probabilities $q_{i,j}$ are connected to these utilities via a **link function** ϕ such that

$$q_{i,j} \; = \; \phi\big(\mu_i, \mu_j\big)$$

An example is the **Bradley-Terry model**, which is well-known in discrete choice theory:

$$q_{i,j} = \mathbf{P}(a_i \succ a_j) \; = \; \frac{\mu_i}{\mu_i + \mu_j}$$

Ailon et al. (2014) make use of a **linear link function** (assuming utilities in $[0,1]$):

$$q_{i,j} = \mathbf{P}(a_i \succ a_j) \; = \; \frac{1 + \mu_i - \mu_j}{2}$$

LATENT UTILITY FUNCTIONS

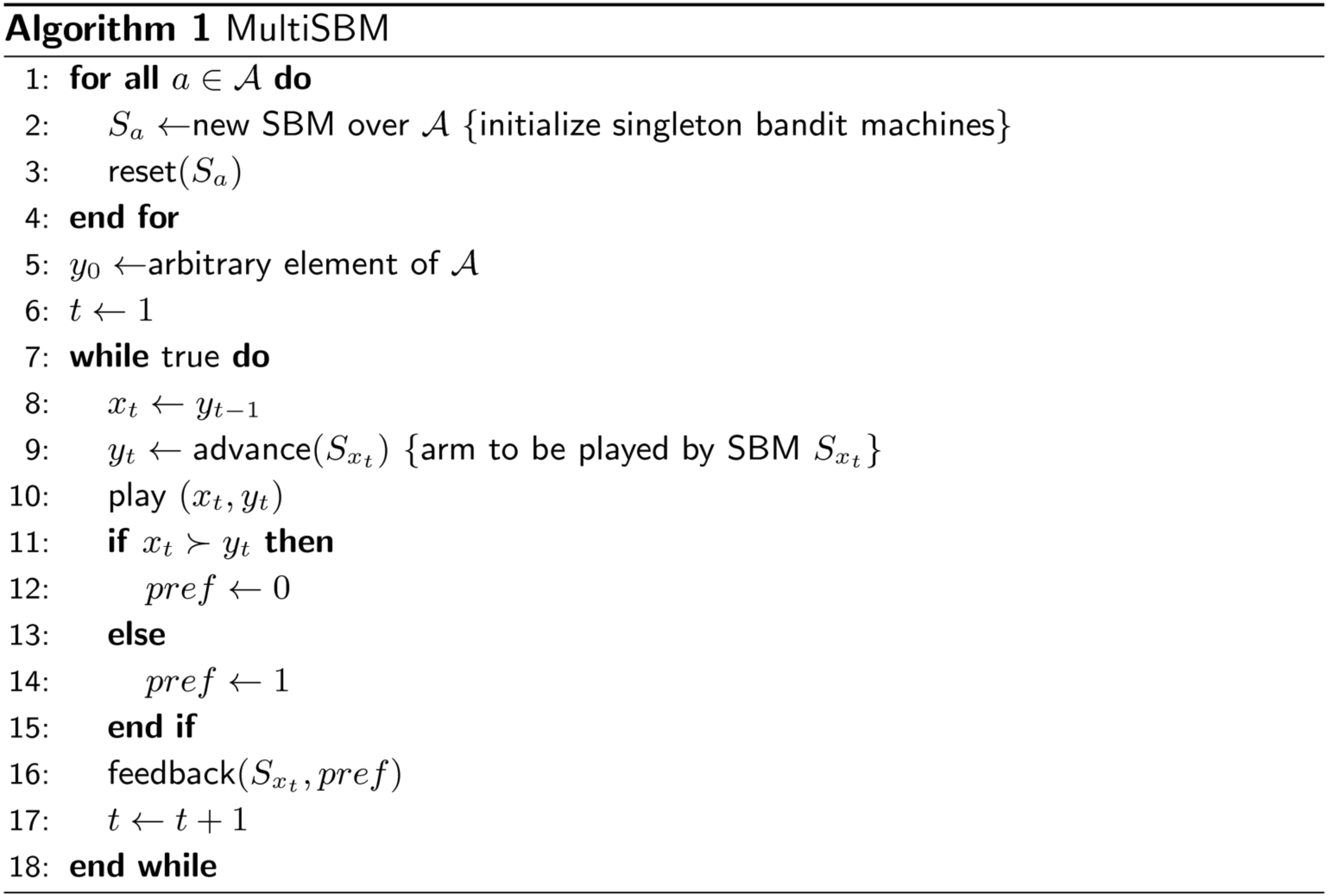
Note that properties such as total order, stochastic transitivity, Condorcet-winner, etc. will normally be implied.

Utilities μ_i (or noisy versions thereof) are not observed directly, however, the observed preferences do at least provide **hints** at these utilities.

Since this setting is somwhat in-between preference-based and the standard (cardinal) MAB problems, an obvious idea is to somehow **reduce the former to the latter**.

This idea has been realized by Ailon et al. (2014), both for the case of a finite and an infinite number of bandits.

MultiSBM



Algorithm 1 MultiSBM

```
for all $a \in \mathcal{A}$ do
    $S_a \leftarrow$new SBM over $\mathcal{A}$ {initialize singleton bandit machines}
    reset($S_a$)
end for
$y_0 \leftarrow$arbitrary element of $\mathcal{A}$
$t \leftarrow 1$
while true do
    $x_t \leftarrow y_{t-1}$
    $y_t \leftarrow$ advance($S_{x_t}$) {arm to be played by SBM $S_{x_t}$}
    play $(x_t, y_t)$
    if $x_t \succ y_t$ then
        $pref \leftarrow 0$
    else
        $pref \leftarrow 1$
    end if
    feedback($S_{x_t}, pref$)
    $t \leftarrow t + 1$
end while
```

MultiSBM

MultiSBM runs K singleton bandit machines in parallel, one for each arm.

The (unobserved) regret is defined as follows:

$$r_t = \max_{a \in \mathcal{A}} \mu(a) - \frac{\mu(x_t) + \mu(y_t)}{2}$$

The feedback for the SBM is

$$pref = \begin{cases} 0 & \text{if } x_t \succ y_t \\ 1 & \text{if } y_t \succ x_t \end{cases}$$

MultiSBM

INTELLIGENT SYSTEMS

right arm

left arm

	a_1	a_2	a_3	a_4	$\mu(a_i)$
a_1	0	1/6	2/6	3/6	1
a_2	1/6	2/6	3/6	4/6	2/3
a_3	2/6	3/6	4/6	5/6	1/3
a_4	3/6	4/6	5/6	1	0

regret $r_t = 1 - \frac{\mu(x_t)+\mu(y_t)}{2}$

MultiSBM

right arm

left arm		a_1	a_2	a_3	a_4	$\mu(a_i)$
	a_1	1/2	2/6	1/6	0	1
	a_2	4/6	1/2	2/6	1/6	2/3
	a_3	5/6	4/6	1/2	2/6	1/3
	a_4	1	5/6	4/6	1/2	0

$$\mathbf{P}(pref = 1) = \mathbf{P}(y_t \succ x_t)$$

MultiSBM

right arm

left arm

	a_1	a_2	a_3	a_4	$\mu(a_i)$
a_1	1/2	2/6	1/6	0	1
a_2	4/6	1/2	2/6	1/6	2/3
a_3	5/6	4/6	1/2	2/6	1/3
a_4	1	5/6	4/6	1/2	0

$$\mathbf{P}(pref = 1) = \mathbf{P}(b_t \succ a_t)$$

LATENT UTILITY FUNCTIONS

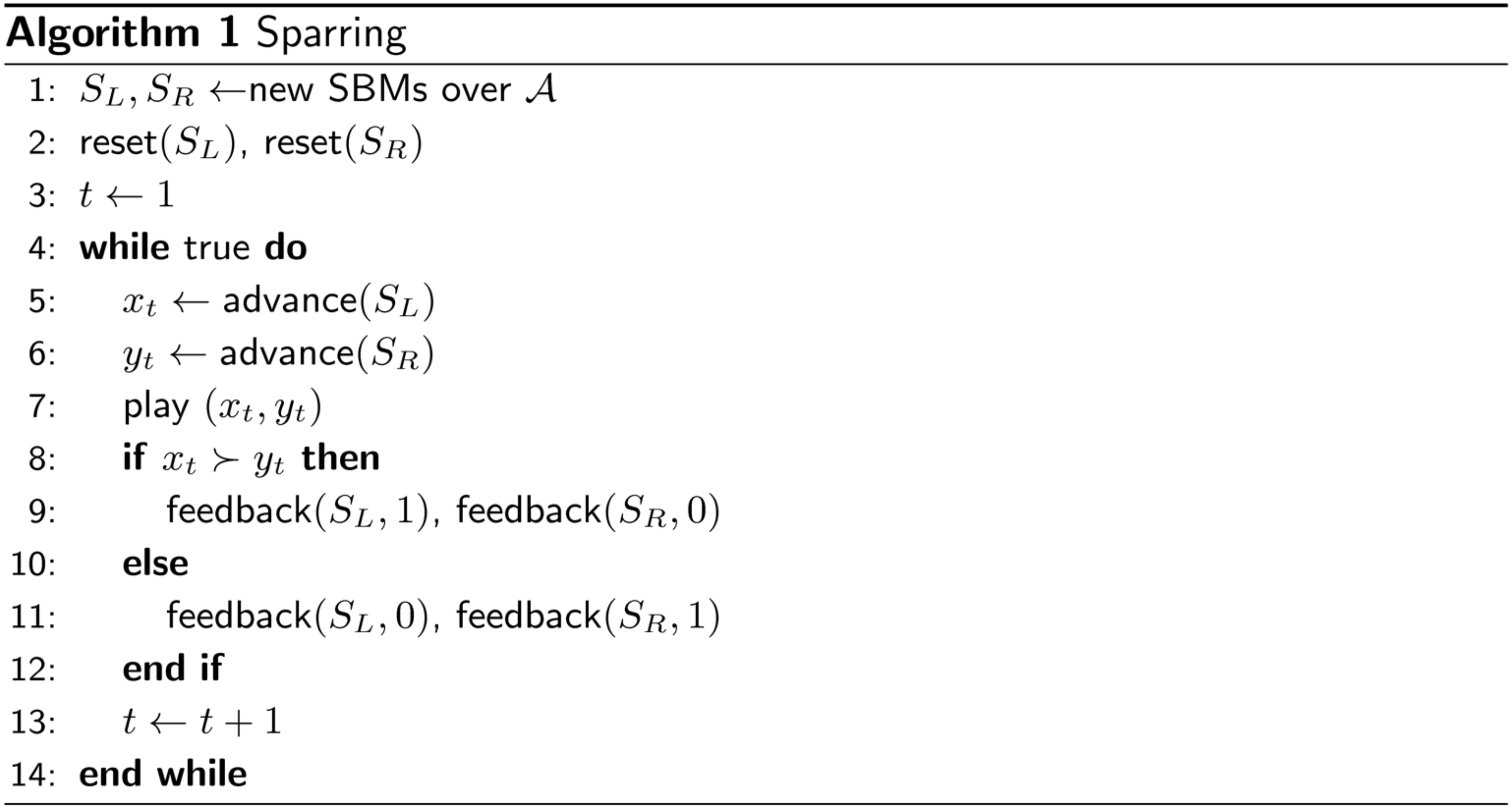
The **expected cumulative regret** of MultiSBM can be shown to be asymptotically competitive to that of a single SBM (namely the one corresponding to the best arm), provided this bound holds for the SBM itself and, moreover, the latter obeys a certain robustness property.

(Inverse polynomial tail distribution for the regret: $\mathbf{P}(T_x > s) < 2/\alpha(s/2)^{-\alpha}$, where T_x is the number of times a suboptimal arm x is played).

Proof is based on showing a "positive feedback loop":

- If the expected regret incurred by the right arm y_t is low, then there is a higher chance that x^* will be played as the left arm at the next time step.
- Conversely, if any fixed arm (in particular x^*) is played very often as the left arm, then the expected regret incurred by the right arm decreases quickly.

SPARRING

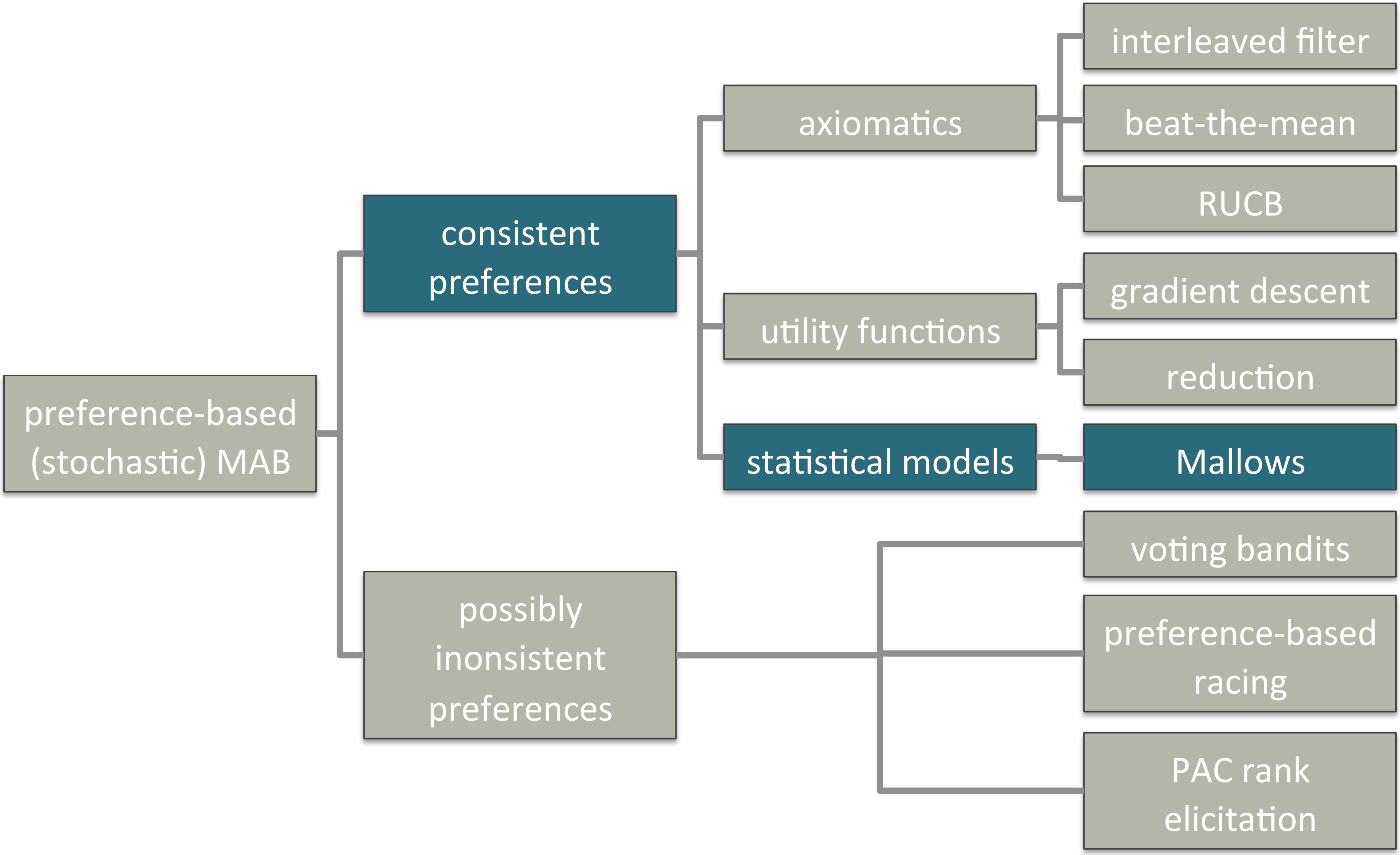


Algorithm 1 Sparring

```
$S_L, S_R \leftarrow$new SBMs over $\mathcal{A}$
reset$(S_L)$, reset$(S_R)$
$t \leftarrow 1$
while true do
    $x_t \leftarrow$ advance$(S_L)$
    $y_t \leftarrow$ advance$(S_R)$
    play $(x_t, y_t)$
    if $x_t \succ y_t$ then
        feedback$(S_L, 1)$, feedback$(S_R, 0)$
    else
        feedback$(S_L, 0)$, feedback$(S_R, 1)$
    end if
    $t \leftarrow t + 1$
end while
```

Sparring (Ailon et al., 2014) performs very well in practice but is difficult to analyze formally (due to non-stochastic nature of feedback).

OVERVIEW OF METHODS



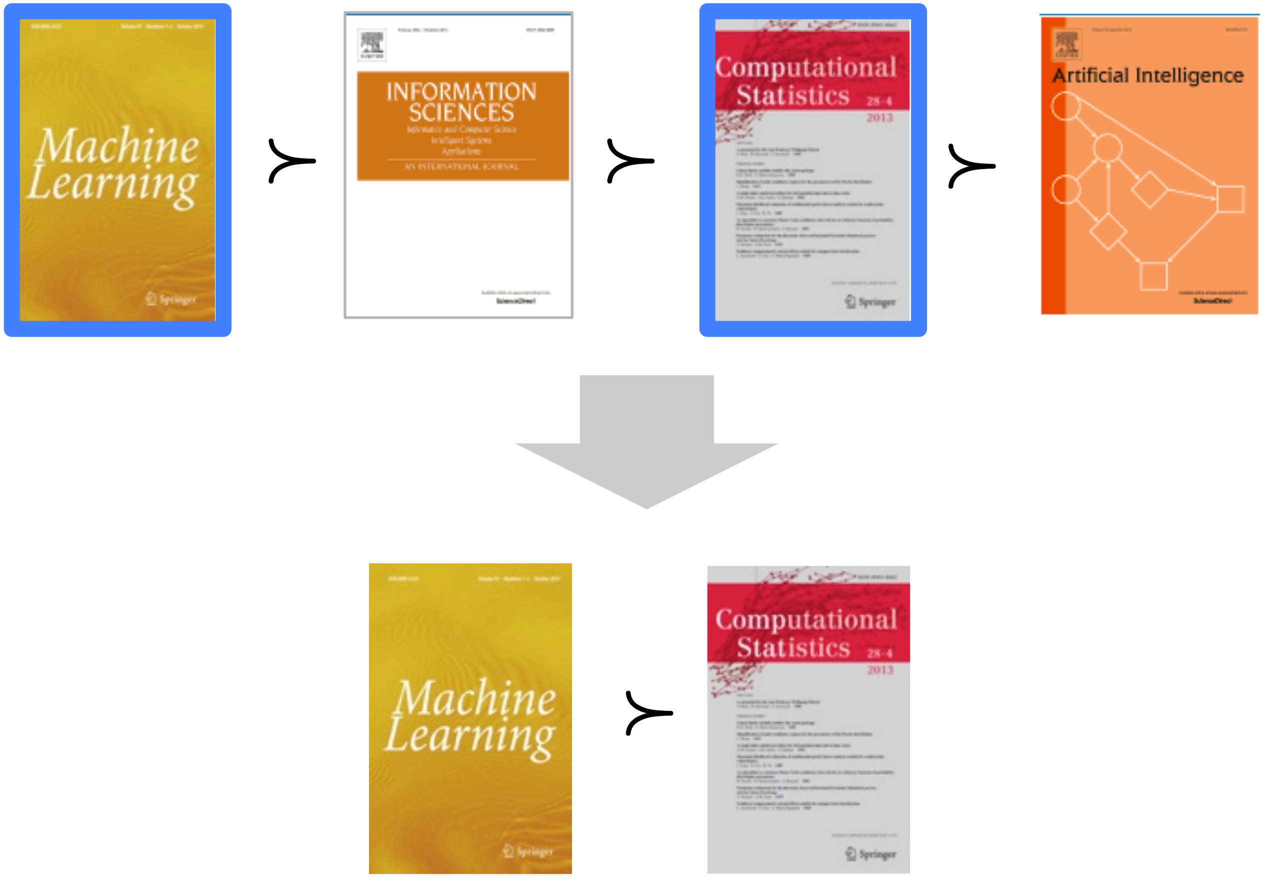
STATISTICAL MODELS

- We might not only be interested in the single best (top-1) alternative, but perhaps in a **complete ranking** of arms, or at least larger parts of it (for example, a top-k ranking).

- In that case, it appears natural to refer to **statistical models of rank data**, which have been studied in statistics (and applied fields such as voting theory, social choice, etc.) for a long time.

STATISTICAL MODELS

Pairwise marginals of a permutation-valued random variable:



PROBABILITIES ON RANKINGS

A ≻ B ≻ C	p_1
A ≻ C ≻ B	p_2
B ≻ A ≻ C	p_3
B ≻ C ≻ A	p_4
C ≻ A ≻ B	p_5
C ≻ B ≻ A	p_6

3 items { A, B, C }

4 items { A, B, C, D }

A ≻ B ≻ C ≻ D	p_1
A ≻ B ≻ D ≻ C	p_2
A ≻ C ≻ B ≻ D	p_3
A ≻ C ≻ D ≻ B	p_4
A ≻ D ≻ B ≻ C	p_5
A ≻ D ≻ C ≻ B	p_6
B ≻ A ≻ C ≻ D	p_7
B ≻ A ≻ D ≻ C	p_8
B ≻ C ≻ A ≻ D	p_9
B ≻ C ≻ D ≻ A	p_{10}
B ≻ D ≻ A ≻ C	p_{11}
B ≻ D ≻ C ≻ A	p_{12}
C ≻ A ≻ B ≻ D	p_{13}
C ≻ A ≻ D ≻ B	p_{14}
C ≻ B ≻ A ≻ D	p_{15}
C ≻ B ≻ D ≻ A	p_{16}
C ≻ D ≻ A ≻ B	p_{17}
C ≻ D ≻ B ≻ A	p_{18}
D ≻ A ≻ B ≻ C	p_{19}
D ≻ A ≻ C ≻ B	p_{20}
D ≻ B ≻ A ≻ C	p_{21}
D ≻ B ≻ C ≻ A	p_{22}
D ≻ C ≻ A ≻ B	p_{23}
D ≻ C ≻ B ≻ A	p_{24}

Need a parameterized family of distributions on the permutation space!

PROBABILITIES ON RANKINGS

item	A	B	C	D
rank	2	3	4	1

$\longleftrightarrow \quad \mathsf{D} \succ \mathsf{A} \succ \mathsf{C} \succ \mathsf{B}$

- Rankings can be represented by permutations $\pi : \{1, \ldots, K\} \to \{1, \ldots, K\}$.
- $\pi(i)$ is the rank of the i-th item.
- The set of all permutations is the symmetric group of order K, denoted $\mathcal{S}_K$.

THE MALLOWS MODELL

… is a **distance-based** probability distribution $\mathbf{P} : \mathcal{S}_K \rightarrow [0,1]$, which belongs to the exponential family:

reference ranking

spread parameter

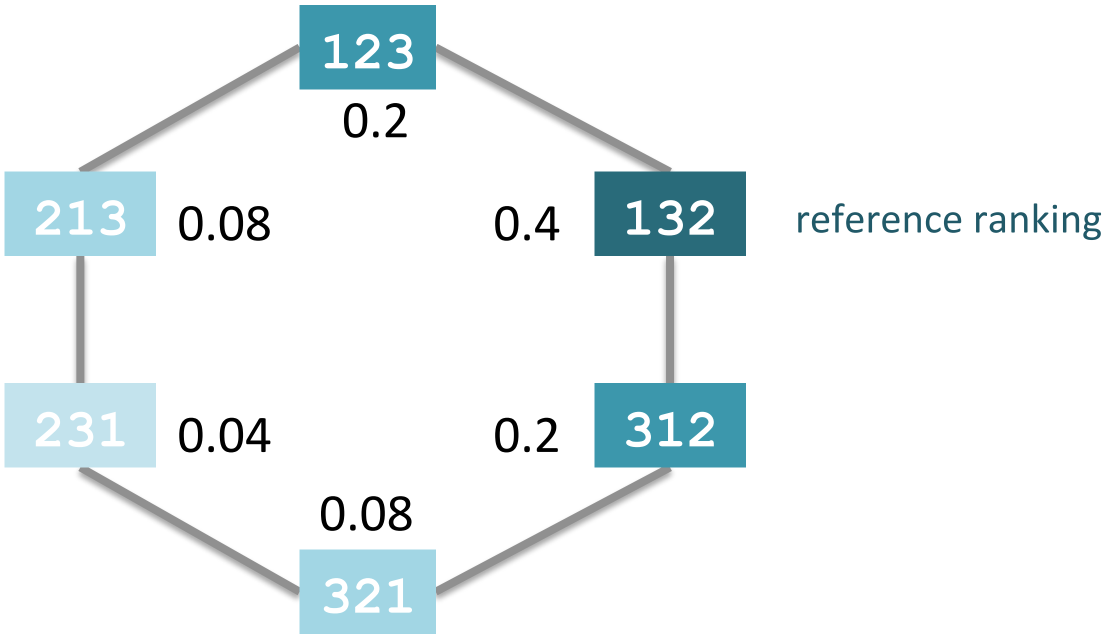
$$\mathbf{P}(\pi \,|\, \pi_0, \theta) \;=\; \frac{\exp\big(-\theta\Delta(\pi,\pi_0)\big)}{\phi(\pi_0,\theta)}$$

normalization constant

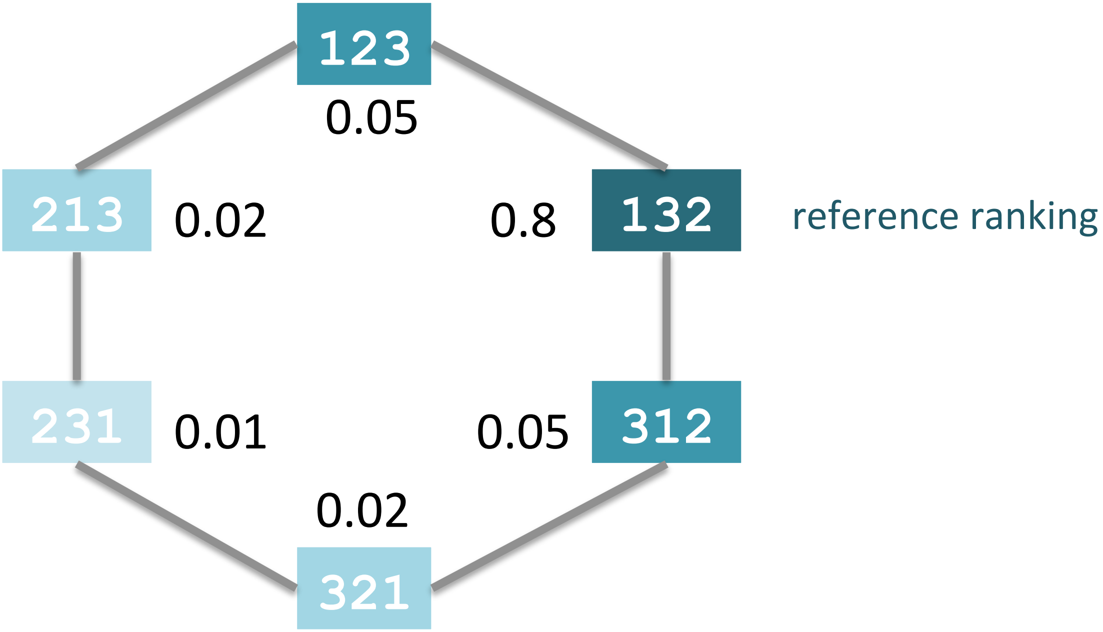
where Δ is the Kendall distance on permutations (number of item pairs differently ordered):

$$\Delta(\pi,\pi_0) \;=\; \#\big\{1 \leq i < j \leq K \,|\, (\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\,\big\}$$

THE MALLOWS MODELL



THE MALLOWS MODELL



PROBABILITY OF INCOMPLETE RANKINGS

Observations are not complete rankings such as

$$\pi : \mathsf{B} \succ \mathsf{C} \succ \mathsf{A} \succ \mathsf{D}$$

but **pairwise preferences** like

$$\sigma : \mathsf{D} \succ \mathsf{C}$$

or **incomplete rankings** like

$$\sigma : \mathsf{B} \succ \mathsf{D} \succ \mathsf{A} \quad .$$

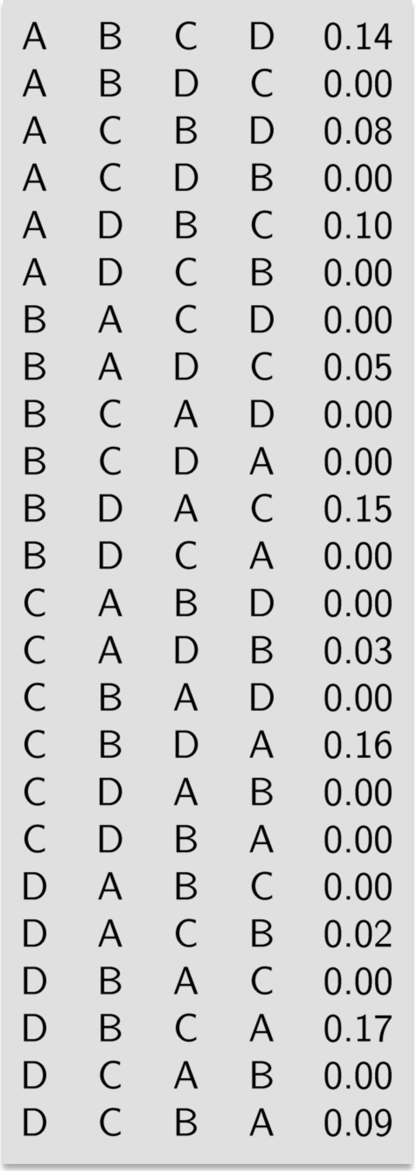
PROBABILITY OF INCOMPLETE RANKINGS

Given a probability $\mathbf{P}(\cdot)$ on $\mathcal{S}_K$, the probability of an **incomplete ranking** σ is given by the probability of its linear extensions:

$$\mathbf{P}(\sigma) = \mathbf{P}(E(\sigma)) = \sum_{\pi \in E(\sigma)} \mathbf{P}(\pi)$$

linear extensions

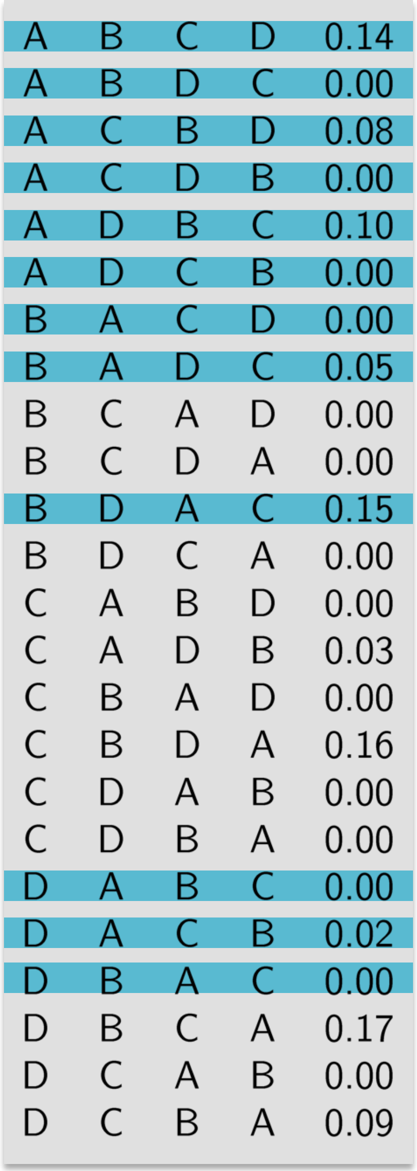
PROBABILITY OF INCOMPLETE RANKINGS



A	B	C	D	0.14
A	B	D	C	0.00
A	C	B	D	0.08
A	C	D	B	0.00
A	D	B	C	0.10
A	D	C	B	0.00
B	A	C	D	0.00
B	A	D	C	0.05
B	C	A	D	0.00
B	C	D	A	0.00
B	D	A	C	0.15
B	D	C	A	0.00
C	A	B	D	0.00
C	A	D	B	0.03
C	B	A	D	0.00
C	B	D	A	0.16
C	D	A	B	0.00
C	D	B	A	0.00
D	A	B	C	0.00
D	A	C	B	0.02
D	B	A	C	0.00
D	B	C	A	0.17
D	C	A	B	0.00
D	C	B	A	0.09

$$\mathbf{P}(\mathsf{A} \succ \mathsf{C}) =$$

PROBABILITY OF INCOMPLETE RANKINGS



A	B	C	D	0.14
A	B	D	C	0.00
A	C	B	D	0.08
A	C	D	B	0.00
A	D	B	C	0.10
A	D	C	B	0.00
B	A	C	D	0.00
B	A	D	C	0.05
B	C	A	D	0.00
B	C	D	A	0.00
B	D	A	C	0.15
B	D	C	A	0.00
C	A	B	D	0.00
C	A	D	B	0.03
C	B	A	D	0.00
C	B	D	A	0.16
C	D	A	B	0.00
C	D	B	A	0.00
D	A	B	C	0.00
D	A	C	B	0.02
D	B	A	C	0.00
D	B	C	A	0.17
D	C	A	B	0.00
D	C	B	A	0.09

$$\mathbf{P}(\mathsf{A} \succ \mathsf{C}) = 0.54$$

THE MALLOWS MODEL

In the case of Mallows, the induced pairwise marginals are

$$q_{i,j} = \mathbf{P}(\, a_i \succ a_j \,) \;=\; \sum_{\pi \colon \pi(i) < \pi(j)} \mathbf{P}(\pi \,|\, \pi_0, \theta)$$

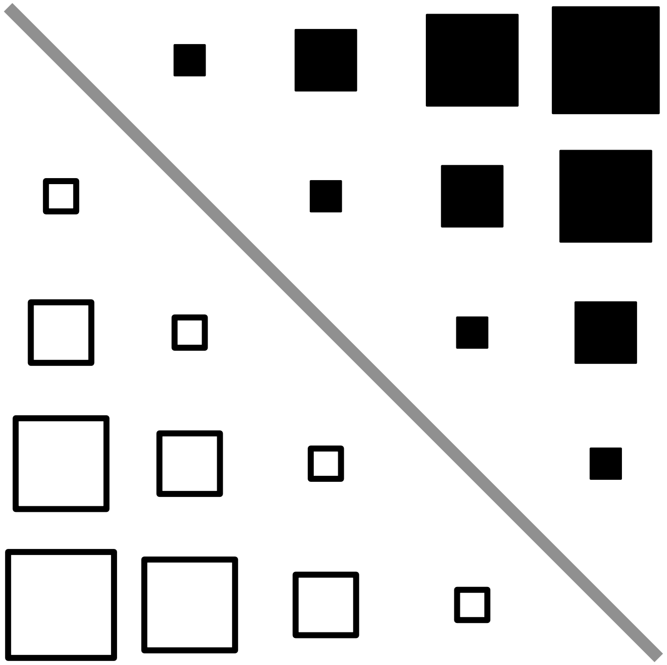
$$= \frac{1}{\phi(\pi_0, \theta)} \sum_{\pi \colon \pi(i) < \pi(j)} \exp\big(- \theta \Delta(\pi, \pi_0)\big)$$

THE MALLOWS MODEL

Important observation: With π_0 the identity, the matrix $\mathbf{Q} = (q_{i,j})$ has a Toeplitz structure:

$$q_{i,j} = h(j-i+1,\theta) - h(j-i,\theta) \ ,$$

with $h(k,\theta) = k/(1-\exp(-k\theta))$.

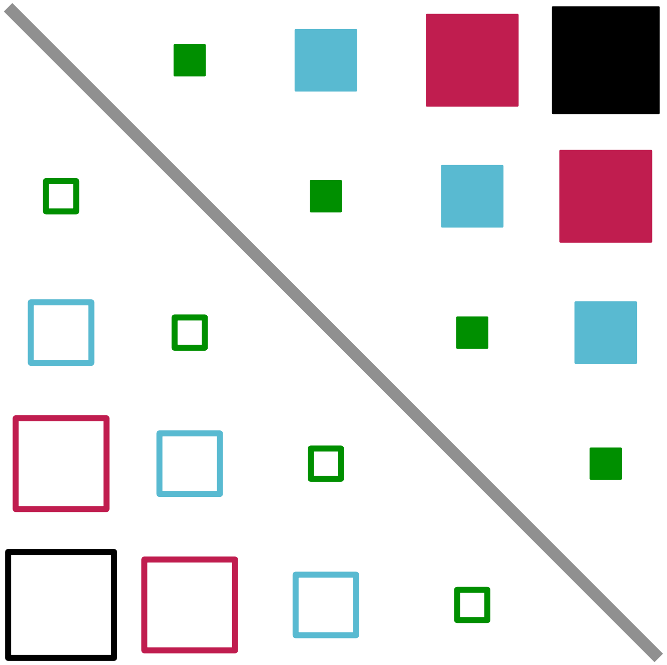


THE MALLOWS MODEL

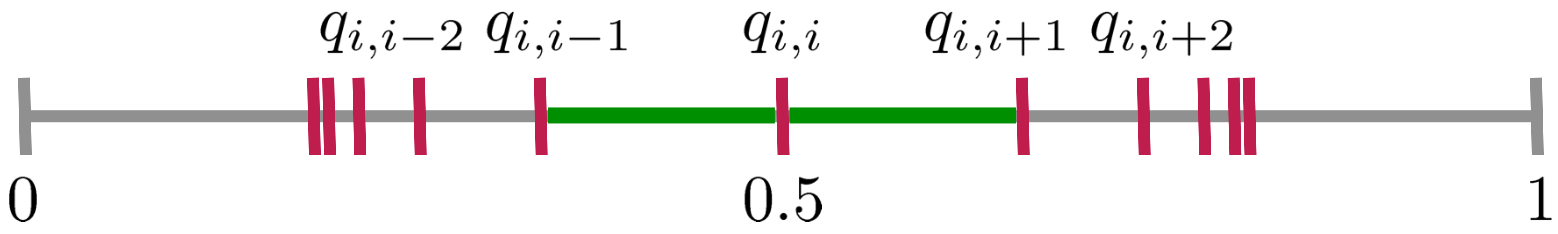
Important observation: With π_0 the identity, the matrix $\mathbf{Q} = (q_{i,j})$ has a Toeplitz structure:

$$q_{i,j} = h(j - i + 1, \theta) - h(j - i, \theta) \;\; ,$$

with $h(k, \theta) = k/(1 - \exp(-k\theta))$.



THE MALLOWS MODEL



- Compared to weaker model assumptions, Mallows induces a **highly regular structure** on the pairwise marginals.

- These are coherent with the target ranking in the sense that $\pi_0(i) < \pi_0(j)$ implies $q_{i,j} > 1/2$ and $\pi_0(i) < \pi_0(j) < \pi_0(k)$ implies $q_{i,j} < q_{i,k}$. (Yet, stochastic triangle inequality does not hold.)

- Most importantly, Mallows assures a **minimum separation** ρ between neighbored options, which depends on θ.

- This allows for establishing a connection to (noisy) **sorting**.

PREFERENCE-BASED RANK ELICITATION

- Busa-Fekete et al. (2014) propose a sampling strategy called **MallowsMPR**, which is based on the **merge sort** algorithm for selecting the arms to be compared.

- However, two arms a_i and a_j are not only compared once but until

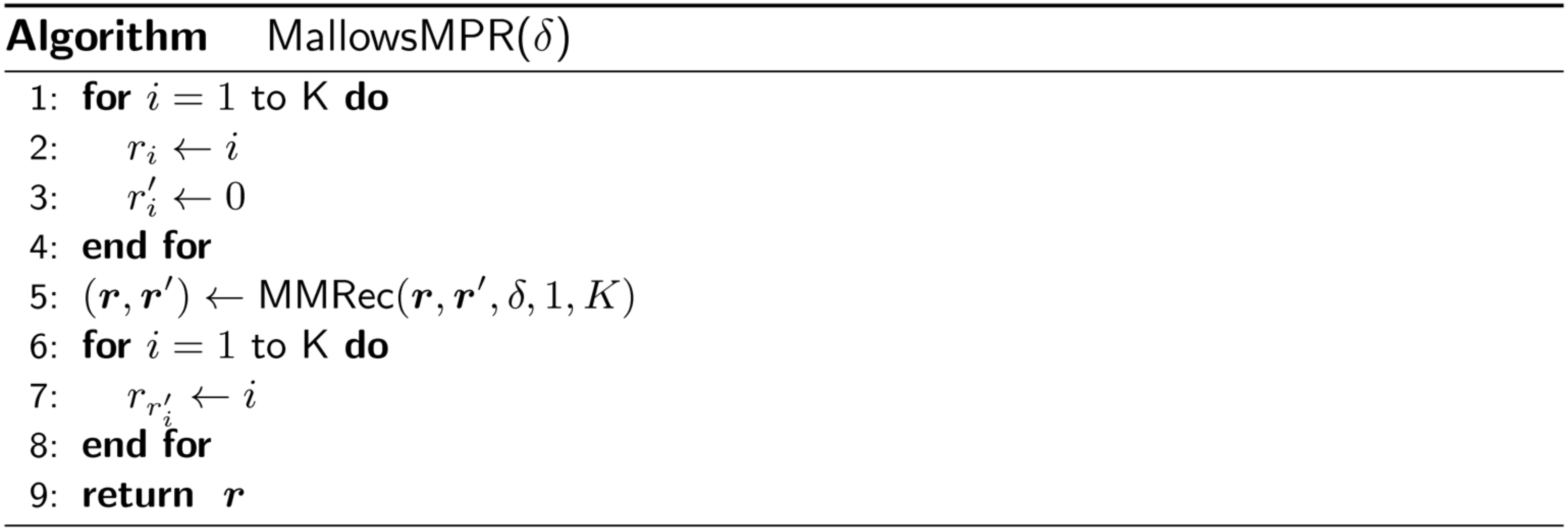
$$1/2 \notin \left[\, \widehat{q}_{i,j} - c_{i,j}, \widehat{q}_{i,j} + c_{i,j} \,\right] \ .$$

- **Theorem:** For any $0 < \delta < 1$, MallowsMPR outputs the reference ranking π_0 with probability at least $1 - \delta$, and the number of pairwise comparisons taken by the algorithm is

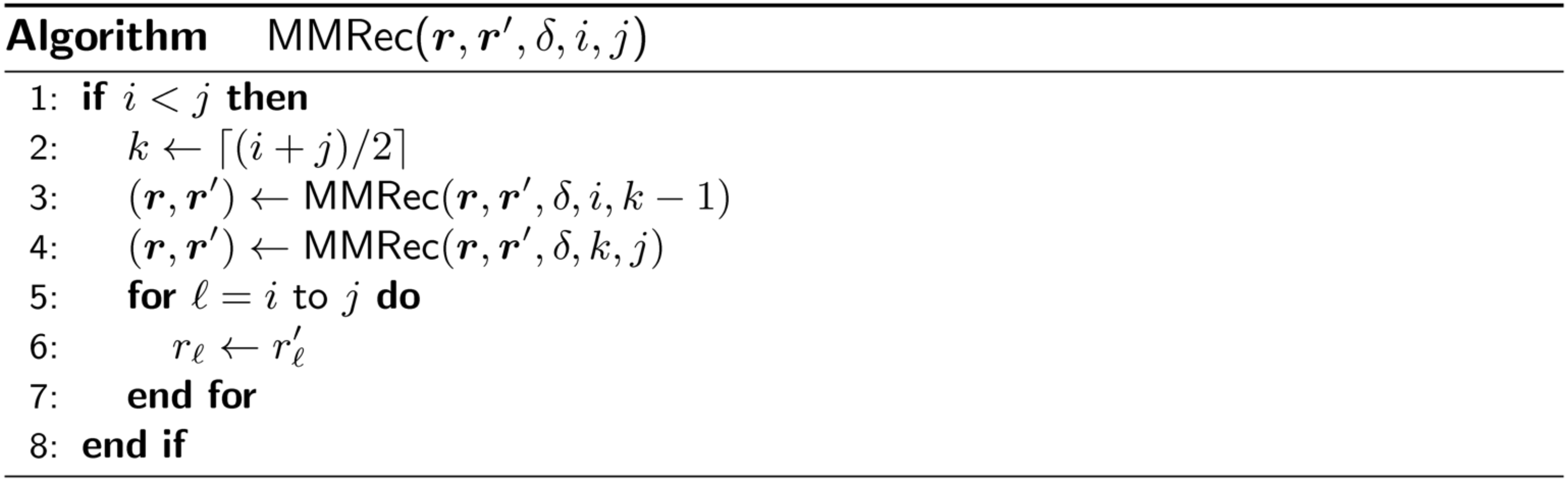
$$\mathcal{O}\left(\frac{K \log_2 K}{\rho^2} \log \frac{K \log_2 K}{\delta \rho} \right) \ ,$$

where $\rho = \frac{1-\phi}{1+\phi}$, $\phi = \exp(-\theta)$.

PREFERENCE-BASED RANK ELICITATION

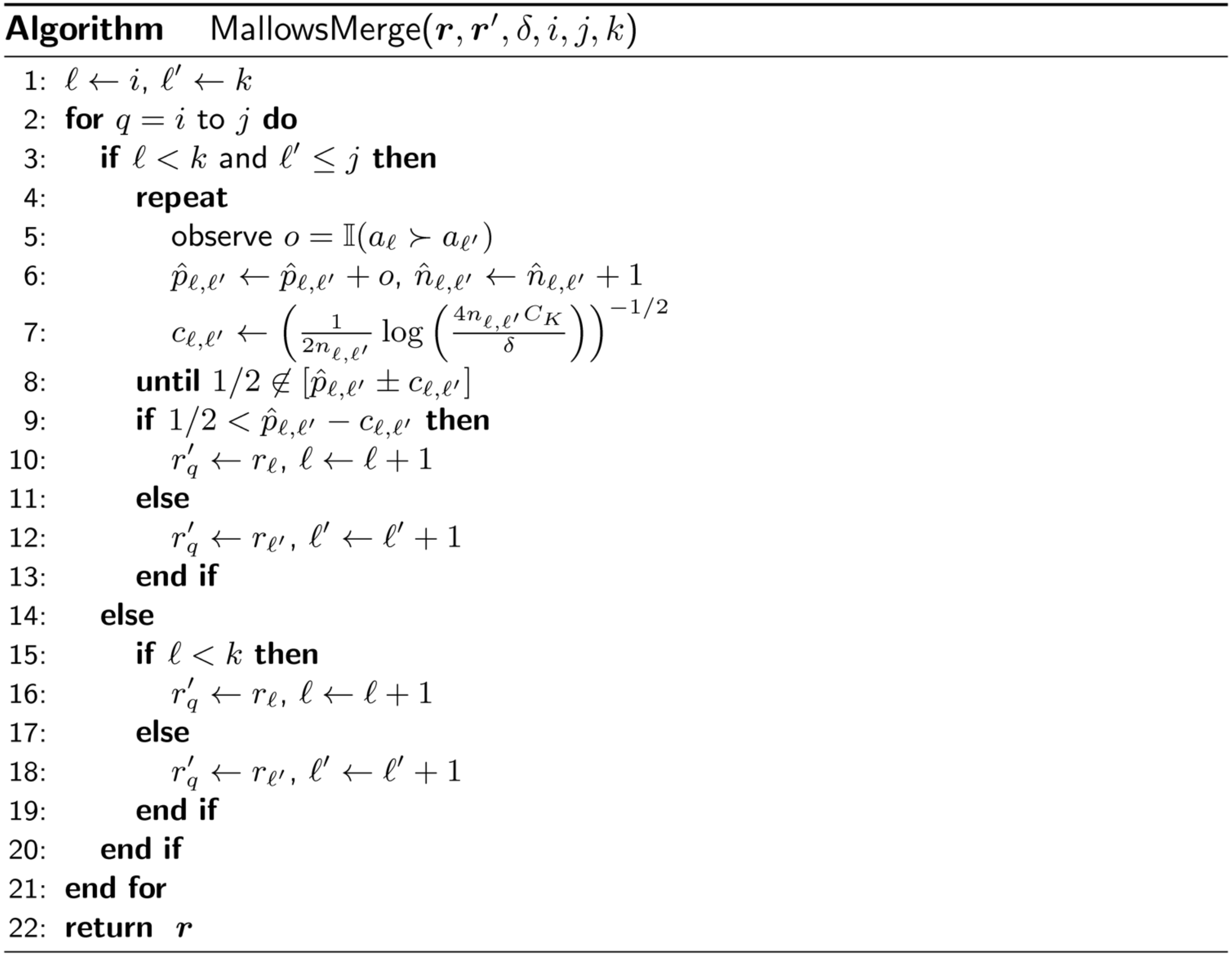


```
Algorithm    MallowsMPR($\delta$)
for $i = 1$ to K do
    $r_i \leftarrow i$
    $r'_i \leftarrow 0$
end for
$(\boldsymbol{r}, \boldsymbol{r}') \leftarrow$ MMRec($\boldsymbol{r}, \boldsymbol{r}', \delta, 1, K$)
for $i = 1$ to K do
    $r_{r'_i} \leftarrow i$
end for
return $\boldsymbol{r}$
```



```
Algorithm    MMRec($\boldsymbol{r}, \boldsymbol{r}', \delta, i, j$)
if $i < j$ then
    $k \leftarrow \lceil (i + j)/2 \rceil$
    $(\boldsymbol{r}, \boldsymbol{r}') \leftarrow$ MMRec($\boldsymbol{r}, \boldsymbol{r}', \delta, i, k - 1$)
    $(\boldsymbol{r}, \boldsymbol{r}') \leftarrow$ MMRec($\boldsymbol{r}, \boldsymbol{r}', \delta, k, j$)
    for $\ell = i$ to $j$ do
        $r_\ell \leftarrow r'_\ell$
    end for
end if
```

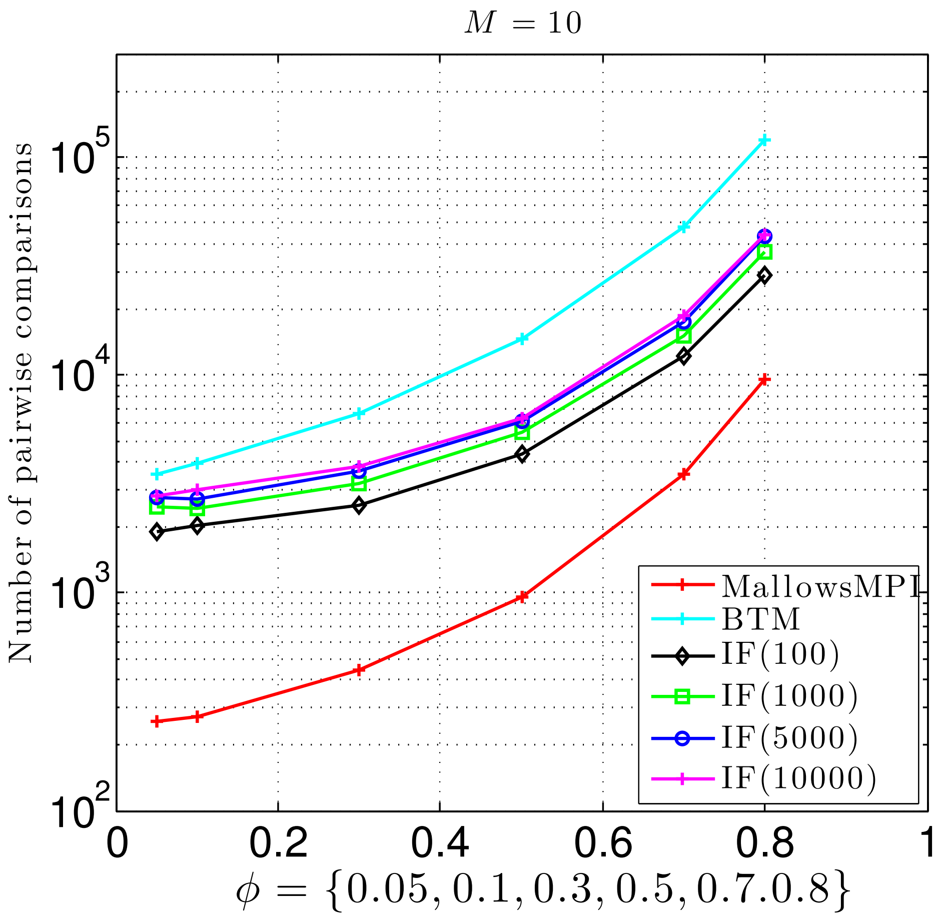
PREFERENCE-BASED RANK ELICITATION



Algorithm MallowsMerge$(\boldsymbol{r}, \boldsymbol{r}', \delta, i, j, k)$

```
$\ell \leftarrow i$, $\ell' \leftarrow k$
for $q = i$ to $j$ do
    if $\ell < k$ and $\ell' \leq j$ then
        repeat
            observe $o = \mathbb{I}(a_\ell \succ a_{\ell'})$
            $\hat{p}_{\ell,\ell'} \leftarrow \hat{p}_{\ell,\ell'} + o$, $\hat{n}_{\ell,\ell'} \leftarrow \hat{n}_{\ell,\ell'} + 1$
            $c_{\ell,\ell'} \leftarrow \left( \frac{1}{2 n_{\ell,\ell'}} \log \left( \frac{4 n_{\ell,\ell'} C_K}{\delta} \right) \right)^{-1/2}$
        until $1/2 \notin [\hat{p}_{\ell,\ell'} \pm c_{\ell,\ell'}]$
        if $1/2 < \hat{p}_{\ell,\ell'} - c_{\ell,\ell'}$ then
            $r'_q \leftarrow r_\ell$, $\ell \leftarrow \ell + 1$
        else
            $r'_q \leftarrow r_{\ell'}$, $\ell' \leftarrow \ell' + 1$
        end if
    else
        if $\ell < k$ then
            $r'_q \leftarrow r_\ell$, $\ell \leftarrow \ell + 1$
        else
            $r'_q \leftarrow r_{\ell'}$, $\ell' \leftarrow \ell' + 1$
        end if
    end if
end for
return $\boldsymbol{r}$
```

EMPIRICAL VALIDATION



Sample complexity for K=10, δ = 0.05 and different values of ϕ.

PREFERENCE-BASED RANK ELICITATION

- For the problem of **finding the best arm**, Busa-Fekete et al. (2014) devise an algorithm that is similar to the one used for finding the largest element in an array.

- Again, two arms a_i and a_j are compared until significance is achieved.

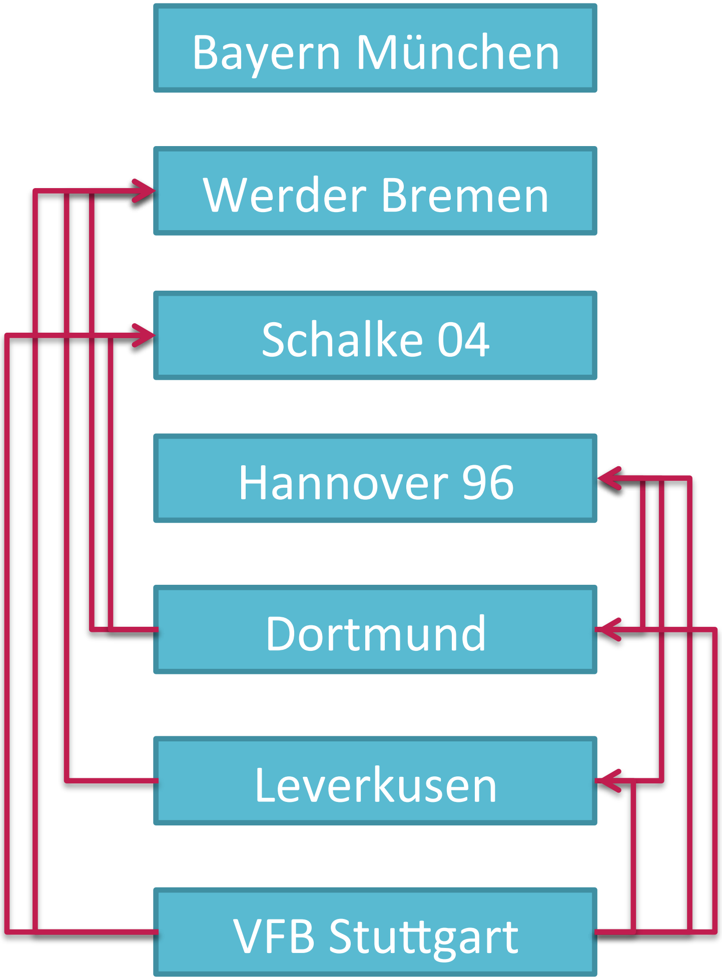
- **Theorem:** MallowsMPI finds the most preferred arm with probability at least $1 - \delta$ for a sample complexity that is of the form

$$\mathcal{O}\left(\frac{K}{\rho^2} \log \frac{K}{\delta\rho}\right) \quad ,$$

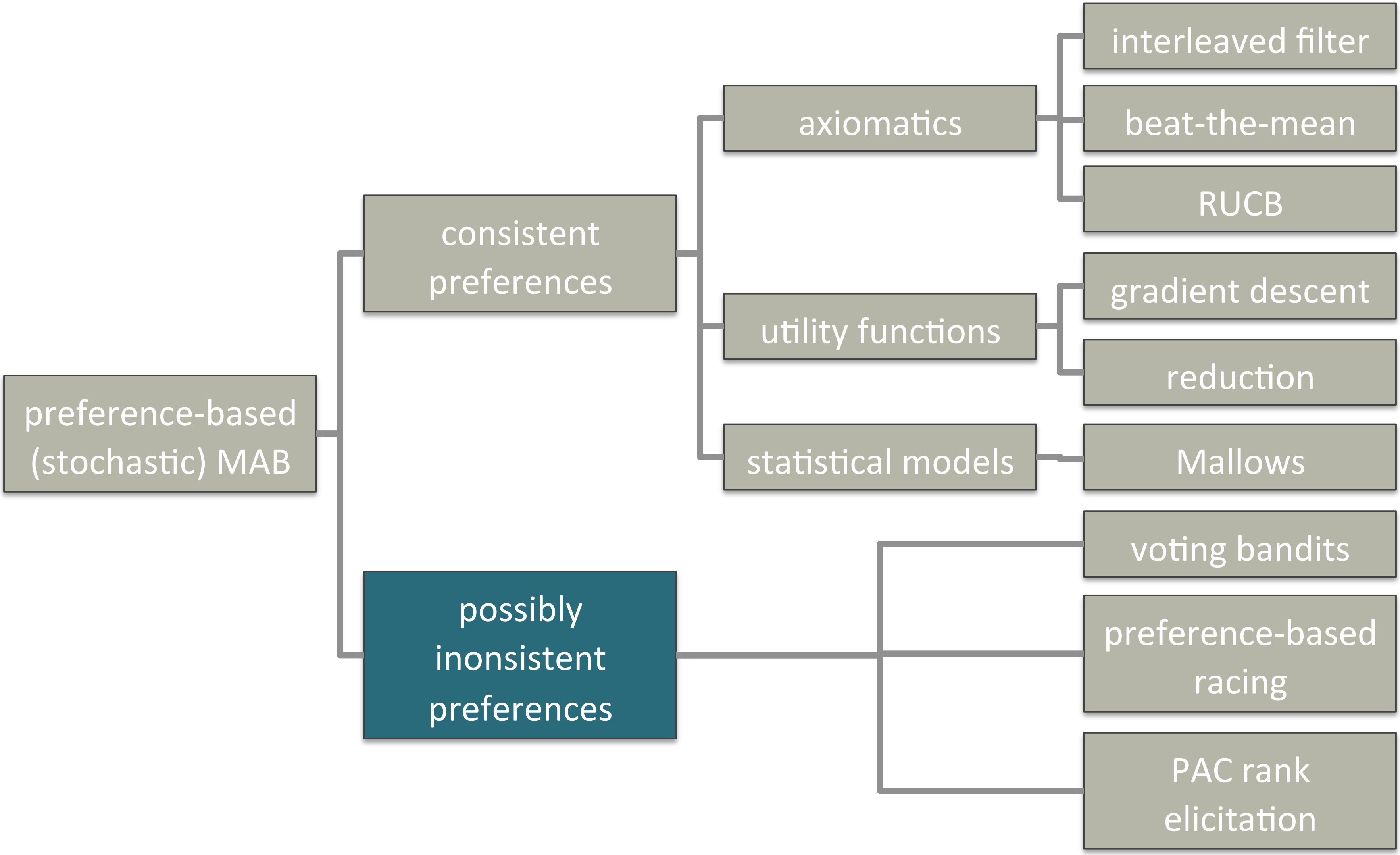
  where $\rho = \frac{1-\phi}{1+\phi}$.

EMPIRICAL VALIDATION

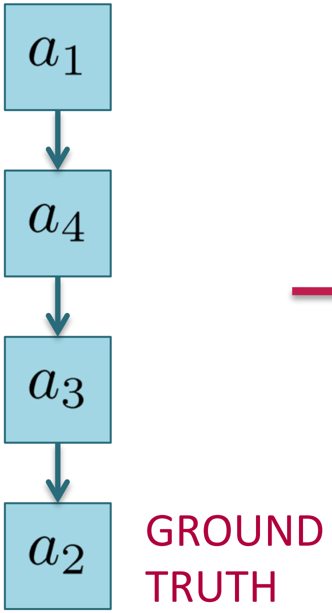
- In general, the approach performs quite well compared to baselines.
- However, it may fail if the underlying data is not enough „Mallowsian" ...



OVERVIEW OF METHODS



PREFERENCE-BASED RANK ELICITATION



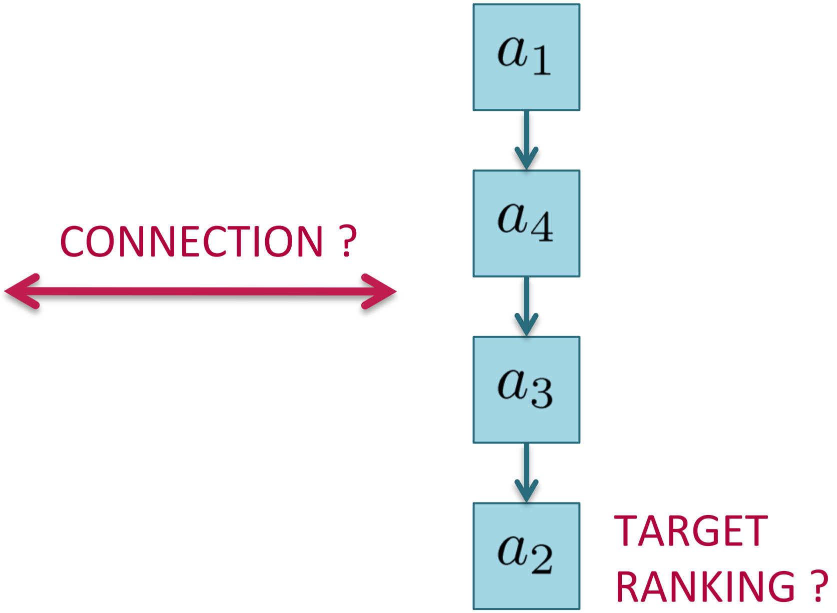
$$\xrightarrow{\text{COHERENCE}} \quad \mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K,1} & q_{K,2} & \cdots & q_{K,K} \end{bmatrix}$$

... the preference relation is derived from, or at least strongly restricted by the target (ranking or best arm)!

PREFERENCE-BASED RANK ELICITATION

Now, take any preference relation as a point of departure ...

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K,1} & q_{K,2} & \cdots & q_{K,K} \end{bmatrix}$$



CONNECTION ?

TARGET RANKING ?

*The target can then be defined by means of a **ranking procedure**!*

RANKING PROCEDURES

Copeland (number of wins, binary voting):

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**3**
a_2	0.4	--	0.8	0.9	**2**
a_3	0.4	0.2	--	0.6	**1**
a_4	0.4	0.1	0.4	--	**0**

$$a_1 \succ a_2 \succ a_3 \succ a_4$$

RANKING PROCEDURES

Borda (weighted voting, sum of expectations):

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**1.8**
a_2	0.4	--	0.8	0.9	**2.1**
a_3	0.4	0.2	--	0.6	**1.2**
a_4	0.4	0.1	0.4	--	**0.9**

$a_2 \succ a_1 \succ a_3 \succ a_4$

Easy reduction for the case of Borda:

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**1.8**
a_2	0.4	--	0.8	0.9	**2.1**
a_3	0.4	0.2	--	0.6	**1.2**
a_4	0.4	0.1	0.4	--	**0.9**

Choosing an arm = pairing it with a randomly chosen alternative:

	a_1	a_2	a_3	a_4
reward 0	0.4	0.3	0.6	0.7
reward 1	0.6	0.7	0.4	0.3

RANKING PROCEDURES

- **Copeland's ranking** (CO) : $a_i \preceq^{\mathsf{CO}} a_j$ if and only if

  $$d_i = \#\{k \in [K] \,|\, 1/2 < q_{i,k}\} \leq \#\{k \in [K] \,|\, 1/2 < q_{i,k}\} = d_j$$

  Problem: Minimal changes of $q_{i,j}$ (around $1/2$) may strongly influence $\preceq^{\mathsf{CO}}$.

- The ϵ-**insensitive CO relation**: $a_i \preceq^{\mathsf{CO}_\epsilon} a_j$ if and only if

  $$d_i^* + s_i^* \leq d_j^*,$$

  with

  $$d_i^* = \#\left\{k \,:\, 1/2 + \epsilon < q_{i,k}, i \neq k\right\},$$
  $$s_i^* = \#\left\{k : |1/2 - q_{i,k}| \leq \epsilon,\ i \neq k\right\}.$$

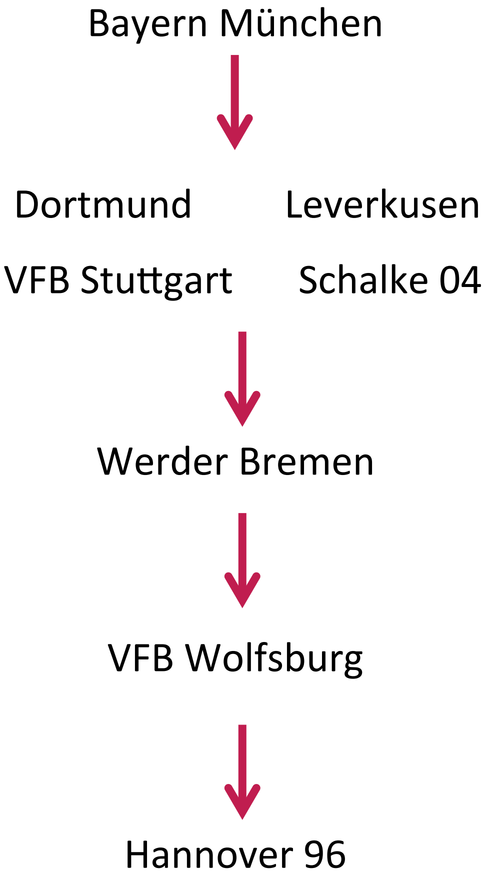
RANKING PROCEDURES

	Bayern München	Dortmund	Leverkusen	VFB Stuttgart	Schalke 04	Werder Bremen	VFB Wolfsburg	Hannover 96
Bayern München		0.7	0.55	0.575	0.75	0.55	0.775	0.7
Dortmund	0.3		0.55	0.475	0.425	0.525	0.6	0.675
Leverkusen	0.45	0.45		0.425	0.55	0.55	0.65	0.6
VFB Stuttgart	0.425	0.525	0.575		0.4	0.6	0.5	0.65
Schalke 04	0.25	0.575	0.45	0.6		0.45	0.65	0.675
Werder Bremen	0.45	0.475	0.45	0.5	0.35		0.5	0.675
VFB Wolfsburg	0.225	0.4	0.35	0.5	0.35	0.45		0.675
Hannover 96	0.3	0.325	0.4	0.35	0.325	0.35	0.325	

German Bundesliga data

RANKING PROCEDURES

	ϵ = 0.02	ϵ = 0.1
Bayern München	[7,7]	[4,7]
Dortmund	[4,4]	[1,6]
Leverkusen	[4,4]	[1,7]
VFB Stuttgart	[4,5]	[1,7]
Schalke 04	[4,4]	[2,6]
Werder Bremen	[3,3]	[1,7]
VFB Wolfsburg	[1,2]	[1,4]
Hannover 96	[0,0]	[0,1]



Bayern München

Dortmund Leverkusen

VFB Stuttgart Schalke 04

Werder Bremen

VFB Wolfsburg

Hannover 96

FORMALIZING CLOSENESS

Distance measures that compare a (predicted) permutation τ with a (target) preorder $\preceq$:

- The **number of discordant pairs** (NDP)

$$d_K(\tau, \preceq) = \sum_{i=1}^{K} \sum_{j \neq i} \mathbb{I}(\tau_j < \tau_i)\mathbb{I}(a_i \prec a_j)$$

- The **maximum rank difference** (MRD)

$$d_M(\tau, \preceq) = \min_{\tau' \in \mathcal{L}^{\preceq}} \max_{1 \leq i \leq K} |\tau_i - \tau'_i|,$$

with $\mathcal{L}^{\preceq}$ the set of linear extensions of $\preceq$.

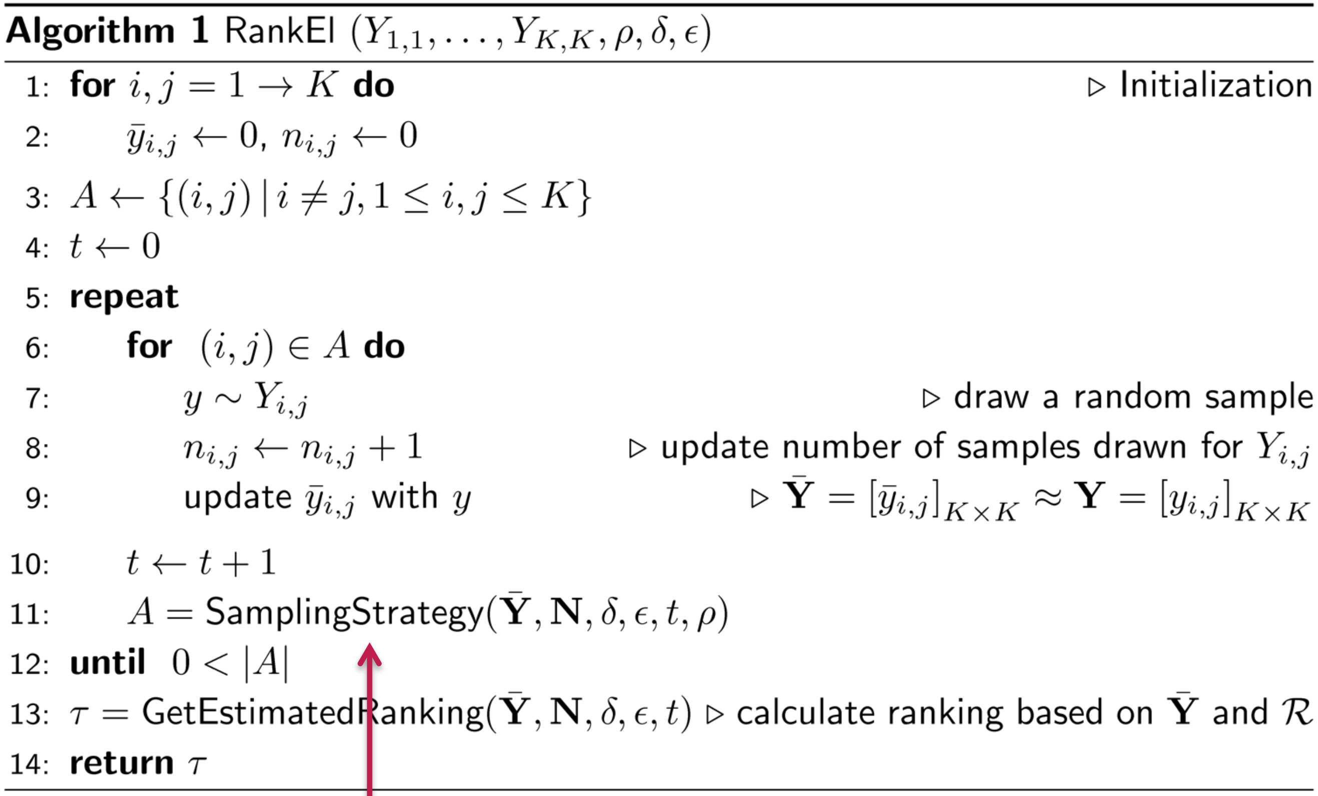
PAC RANK ELICITATION

An algorithm $\mathbf{A}$ is a (ρ, δ)-**PAC rank elicitation algorithm** with respect to a ranking procedure $\mathcal{R}$ and rank distance d, if it returns a ranking τ for which

$$d(\tau, \preceq^{\mathcal{R}}) < \rho$$

with probability at least $1 - \delta$.

PAC RANK ELICITATION



Algorithm 1 RankEl $(Y_{1,1}, \ldots, Y_{K,K}, \rho, \delta, \epsilon)$

```
for $i, j = 1 \rightarrow K$ do                                   ▷ Initialization
    $\bar{y}_{i,j} \leftarrow 0$, $n_{i,j} \leftarrow 0$
$A \leftarrow \{(i,j) \,|\, i \neq j, 1 \leq i,j \leq K\}$
$t \leftarrow 0$
repeat
    for $(i,j) \in A$ do
        $y \sim Y_{i,j}$                                           ▷ draw a random sample
        $n_{i,j} \leftarrow n_{i,j} + 1$                           ▷ update number of samples drawn for $Y_{i,j}$
        update $\bar{y}_{i,j}$ with $y$                            ▷ $\bar{\mathbf{Y}} = [\bar{y}_{i,j}]_{K \times K} \approx \mathbf{Y} = [y_{i,j}]_{K \times K}$
    $t \leftarrow t + 1$
    $A$ = SamplingStrategy($\bar{\mathbf{Y}}, \mathbf{N}, \delta, \epsilon, t, \rho$)
until $0 < |A|$
$\tau$ = GetEstimatedRanking($\bar{\mathbf{Y}}, \mathbf{N}, \delta, \epsilon, t$)   ▷ calculate ranking based on $\bar{\mathbf{Y}}$ and $\mathcal{R}$
return $\tau$
```

decides which pairwise preferences
still need to be sampled

PAC RANK ELICITATION

Define a ranking τ^t over the arms a_i by sorting them in decreasing order according to

$$d_i^t = \#\left\{ j \,|\, 1/2 - \epsilon < \bar{y}^t_{i,j} - c^t_{i,j}, j \neq i \right\} \quad ,$$

sure wins

and in case of a tie ($d_i^t = d_j^t$) according to

$$u_i^t = \#\left\{ j \,|\, [1/2 - \epsilon, 1/2 + \epsilon] \subseteq [\widehat{p}^t_{i,j} - c^t_{i,j}, \widehat{p}^t_{i,j} + c^t_{i,j}], j \neq i \right\} \quad .$$

Moreover, let

$$\mathbb{I}^t_{i,j} = [\![(d_i^t < d_j^t + u_j^t) \wedge (d_j^t < d_i^t + u_i^t)]\!]$$

possible wins

for all $1 \leq i \neq j \leq K$. Then, for any time step t, and for any sampling strategy,

$$d_K(\tau^t, \prec^{\mathrm{CO}_\epsilon}) \leq \frac{1}{2} \sum_{i=1}^{K} \sum_{j \neq i} \mathbb{I}^t_{i,j} \quad \text{and} \quad d_M(\tau^t, \prec^{\mathrm{CO}_\epsilon}) \leq \max_{i \neq j} |\tau^t_i - \tau^t_j| \cdot \mathbb{I}^t_{i,j}$$

with probability at least $1 - \delta$.

Choose a pair, for which the upper bound on the prediction loss promises to decrease the most!

SAMPLE COMPLEXITY

Theorem (Busa-Fekete et al., 2014): The expected sample complexity for $\mathrm{RankEl}^{\mathrm{CO}_\epsilon}_{d_M}$ is

$$\mathcal{O}\left(K^2\, R_1 \log\left(\frac{R_1}{\delta}\right)\right),$$

with

$$R_1 = \sum_{r=1}^{K^2 - r_1} \left(\Delta_{(r)} + \epsilon\right)^{-2},$$

where $\Delta_{(r)}$ denotes the r-th smallest among the values $\Delta_{i,j} = |1/2 - y_{i,j}|$ for all distinct $i, j \in [K]$, and

$$r_1 = 2 \arg\max \left\{ r \in [K^2] \,|\, v^{\mathrm{CO}_\epsilon}_{d_M}(r) < \rho \right\},$$

$$v^{\mathrm{CO}_\epsilon}_{d_M}(r, \mathbf{Y}) = \max_{\tilde{\mathbf{Y}} \in (\mathbf{Y})_r} \; \min_{\tau \in \mathcal{L}^{\preceq^{\mathrm{CO}_\epsilon}_{\tilde{\mathbf{Y}}}}} \; \max_{\mathbf{Y}' \in (\mathbf{Y})_r} d_M(\tau, \preceq^{\mathrm{CO}_\epsilon}_{Y'}).$$

Empirically, significant gains are reported in comparison to random sampling.

SUMMARY & CONCLUSION

Preference learning is

- **methodologically** interesting,
- **theoretically** challenging,
- and **practically** useful, with many potential **applications**;
- more **general** than could be shown in this talk („preferences“ in the broad sense, standard ML problems as special cases, ...); in fact, a flexible machine learning framework for learning from **weak supervision**;
- **interdisciplinary** (connections to operations research, decision sciences, economics, social choice, recommender systems, information retrieval, ...).

SUMMARY & CONCLUSION

Preference-based online learning with multi-armed bandits (PB-MAB):

- **emerging** research topic,
- no complete and **coherent framework** so far,
- many **open questions and problems** (e.g., necessary assumptions on preference relation to guarantee certain bounds on regret or sample complexity, lower bounds, statistical tests for verifying model assumptions, generalizations to large (structured) set of arms, contextual bandits, adversarial setting, etc., practical applications, ...)

SELECTED LITERATURE

- N. Ailon, K. Hatano, and E. Takimoto. Bandit online optimization over the permutahedron. CoRR, abs/1312.1530, 2014.
- N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. ICML 2014.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47:235-256, 2002.
- R. Busa-Fekete, E. Hüllermeier, and B. Szorenyi. Preference-based rank elicitation using statistical models: The case of Mallows. ICML 2014.
- R. Busa-Fekete, B. Szorenyi, and E. Hüllermeier. PAC rank elicitation through adaptive sampling of stochastic pairwise preferences. AAAI 2014.
- R. Busa-Fekete, B. Szorenyi, P. Weng, W. Cheng, and E. Hüllermeier. Top-k selection based on adaptive sampling of noisy preferences. ICML 2013.
- W.W. Cohen, R.E. Schapire and Y. Singer. Learning to order things. Journal of Artificial Intelligence Research, 10:243–270, 1999.
- J. Duchi, L. Mackey, and M. Jordan. On the consistency of ranking algorithms. ICML 2010.
- J. Fürnkranz and E. Hüllermeier, editors. Preference Learning. Springer-Verlag, 2011.
- E. Hüllermeier, J. Fürnkranz, W. Cheng and K. Brinker. Label ranking by learning pairwise preferences. Artificial Intelligence, 172, 2008.
- F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? CIKM 2008.
- T. Urvoy, F. Clerot, R. Feraud, and S. Naamane. Generic exploration and k-armed voting bandits. ICML 2013.
- Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The K-armed dueling bandits problem. Journal of Computer and System Sciences, 78(5):1538-1556, 2012.
- Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. ICML 2009.
- Y. Yue and T. Joachims. Beat the mean bandit. ICML 2011.
- M. Zoghi, S. Whiteson, R. Munos, and M. de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. ICML 2014.