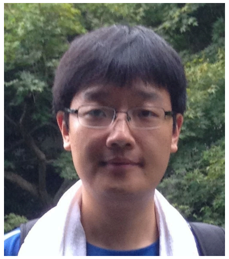
dolcit

Caltech

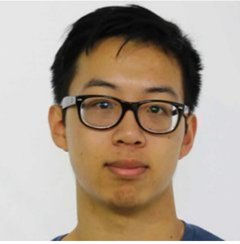
The Dueling Bandits Problem

Yisong Yue

Collaborators



Yanan
Sui



Vincent
Zhuang

Josef
Broder

Joel
Burdick

Thorsten
Joachims

Bobby
Kleinberg

Outline

- **Brief Overview of Multi-Armed Bandits**
  - Sequential Experimental Design

- **Dueling Bandits**
  - Mathematical properties
  - Connections to other problems

- **Recent Results & Ongoing Research**

Multi-Armed Bandit Problem
(stochastic version)

- K actions (aka arms or bandits)
- Each action has an average reward: μ_k
  - Unknown to us
  - Assume WLOG that u_1 is largest

- For t = 1…T
  - Algorithm chooses action a(t)
  - Receives random reward y(t)
    - Expectation $\mu_{a(t)}$

Algorithm only receives feedback on chosen action

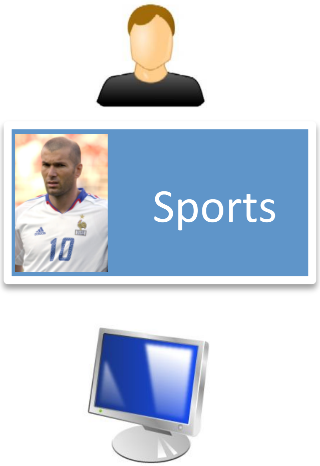
- **Goal:** minimize $Tu_1 - (\mu_{a(1)} + \mu_{a(2)} + \ldots + \mu_{a(T)})$

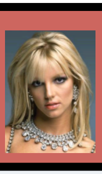
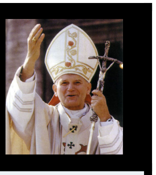
"Regret"

If we had perfect information to start

Expected Reward of Algorithm

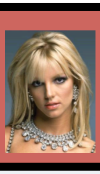
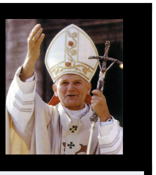
Example:
Interactive Personalization



					
Average Likes	--	--	--	--	--
# Shown	0	0	0	1	0

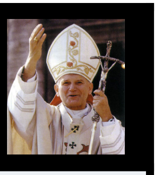

: 0

Example:
Interactive Personalization

					
Average Likes	--	--	--	0	--
# Shown	0	0	0	1	0

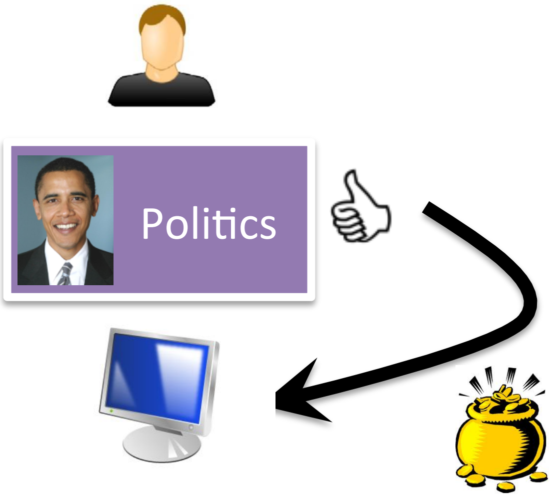

: 0

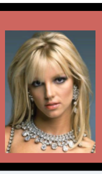
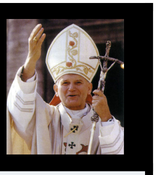
Example:
Interactive Personalization

					
Average Likes	--	--	--	0	--
# Shown	0	0	1	1	0

💰 : 0

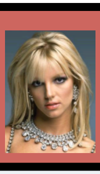
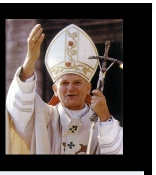
Example:
Interactive Personalization



					
Average Likes	--	--	1	0	--
# Shown	0	0	1	1	0

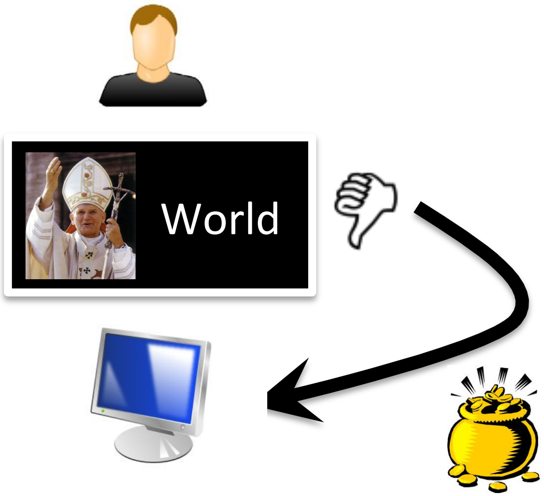
💰 : 1

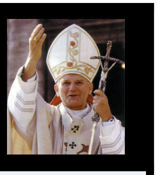
Example:
Interactive Personalization

		$			
Average Likes	--	--	1	0	--
# Shown	0	0	1	1	1

💰 : 1

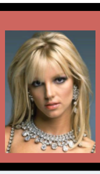
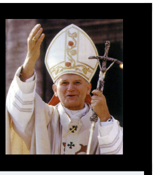
Example:
Interactive Personalization



					
Average Likes	--	--	1	0	0
# Shown	0	0	1	1	1

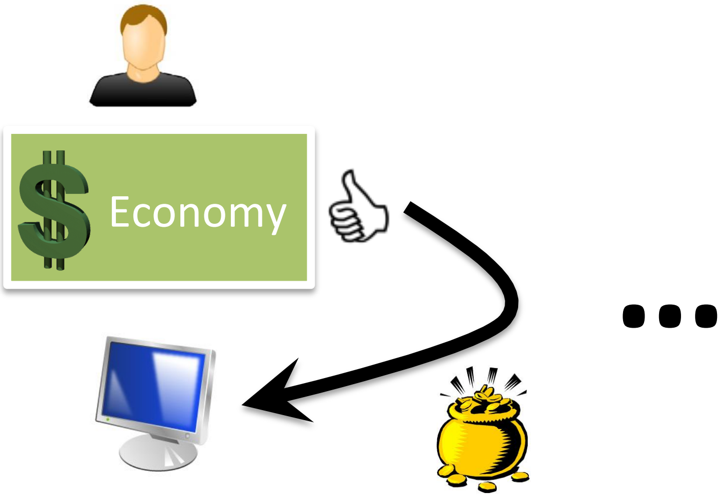

: 1

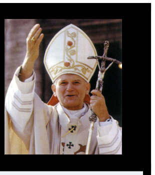
Example:
Interactive Personalization

					
Average Likes	--	--	1	0	0
# Shown	0	1	1	1	1


: 1

Example:
Interactive Personalization



					
Average Likes	--	1	1	0	0
# Shown	0	1	1	1	1

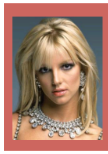
 : 2

What Should Algorithm Recommend?

Exploit:

Economy

Explore:



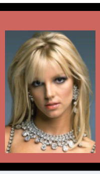
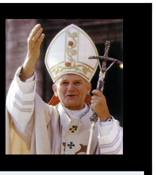
Celebrity

Best:

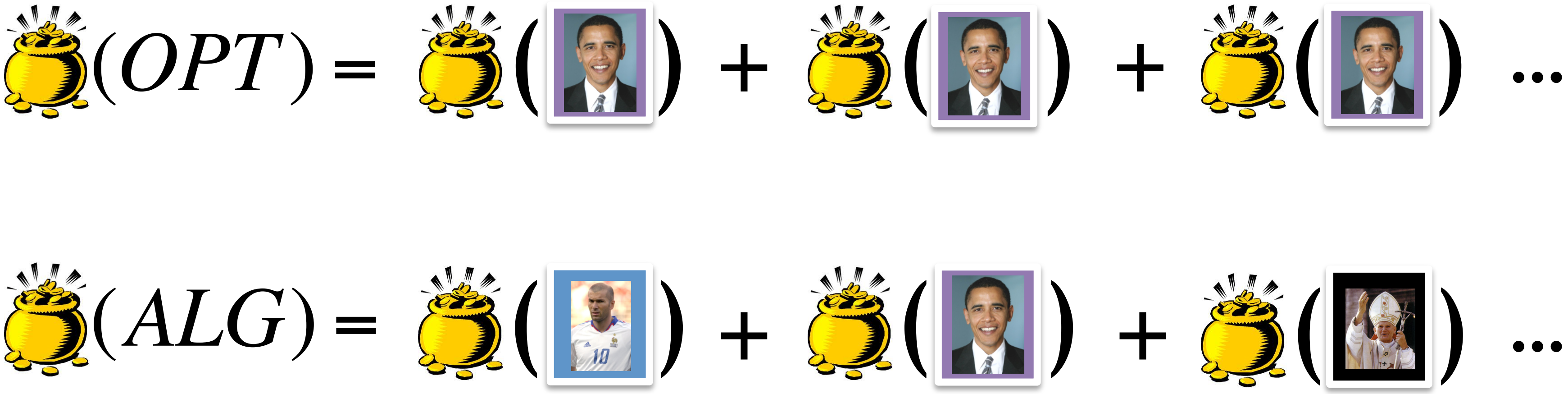
Politics

How to Optimally Balance Explore/Exploit Tradeoff?

Characterized by the Multi-Armed Bandit Problem

					
Average Likes	--	0.44	0.4	0.33	0.2
# Shown	0	25	10	15	20

: 24



Time Horizon

Regret: $R(T) =$ 💰$(OPT) -$ 💰(ALG)

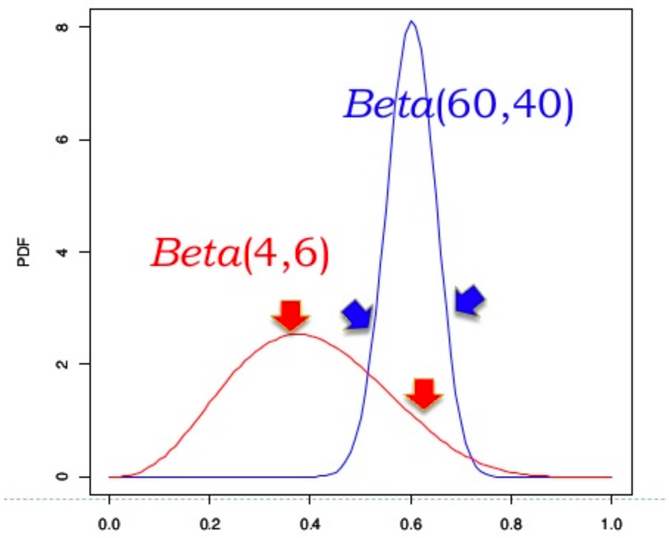
- Opportunity cost of not knowing preferences
- “**no-regret**” if R(T)/T ➔ 0
  - Efficiency measured by convergence rate

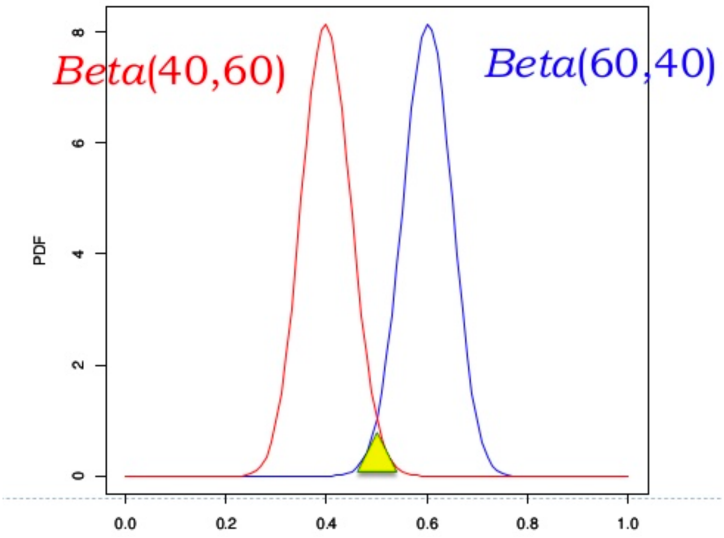
Thompson Sampling

- Maintain distribution over rewards
  - $P(\mu_1 ,\ldots\mu_K | Y)$

- Every round:
  - Sample $\mu_1 ,\ldots\mu_K$
  - Play arm with highest μ_a
  - Incorporate feedback into Y

Incentivizing Exploration





Arms

$$O(K/\varepsilon \log(T))$$

Regret Bound:

Time horizon

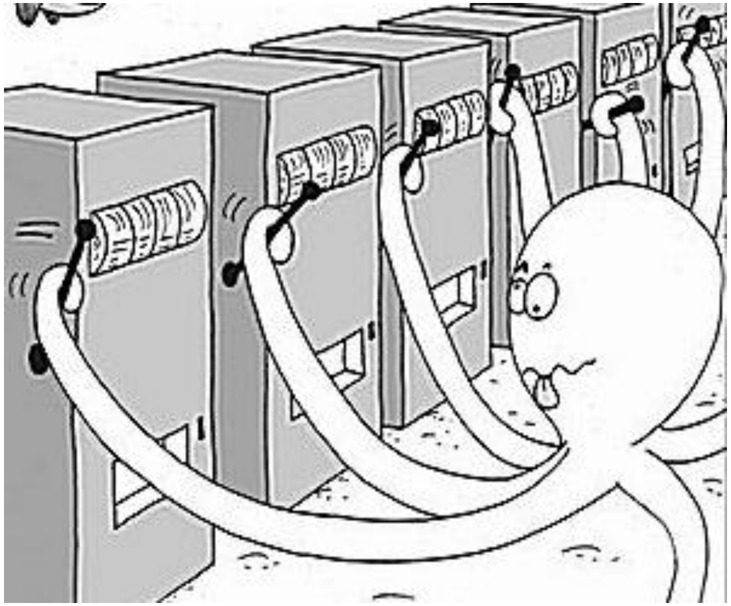
Gap between best & 2nd best

[Agrawal & Goyal; COLT 2012]

Images from Chu-Cheng Hsieh

The Motivating Problem

- Slot Machine = One-Armed Bandit



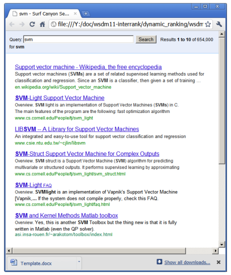
Each Arm Has Different Payoff

- **Goal:** Minimize regret From pulling suboptimal arms

Image source: http://research.microsoft.com/en-us/projects/bandits/

Many Applications

Online Advertising



Query: svm

Results **1 to 10** of 654,000 for **svm**

Support vector machine - Wikipedia, the free encyclopedia
Support vector machines (**SVMs**) are a set of related supervised learning methods used for classification and regression. Since an **SVM** is a classifier, then given a set of training ...
en.wikipedia.org/wiki/Support_vector_machine

SVM-Light Support Vector Machine
Overview. **SVM** light is an implementation of Support Vector Machines (**SVMs**) in C. The main features of the program are the following: fast optimization algorithm
www.cs.cornell.edu/People/tj/svm_light

LIB**SVM** -- A Library for Support Vector Machines
An integrated and easy-to-use tool for support vector classification and regression
www.csie.ntu.edu.tw/~cjlin/libsvm

SVM-Struct Support Vector Machine for Complex Outputs
Overview. **SVM** struct is a Support Vector Machine (**SVM**) algorithm for predicting multivariate or structured outputs. It performs supervised learning by approximating
www.cs.cornell.edu/People/tj/svm_light/svm_struct.html

SVM-Light FAQ
Overview. **SVMlight** is an implementation of Vapnik's Support Vector Machine [Vapnik,**...** If the system does not compile properly, check this FAQ.
www.cs.cornell.edu/People/tj/svm_light/faq.html

SVM and Kernel Methods Matlab toolbox
Overview. Yes, this is another **SVM** Toolbox but the thing new is that it is fully written in Matlab (even the QP solver).
asi.insa-rouen.fr/~arakotom/toolbox/index.html

Search Engines

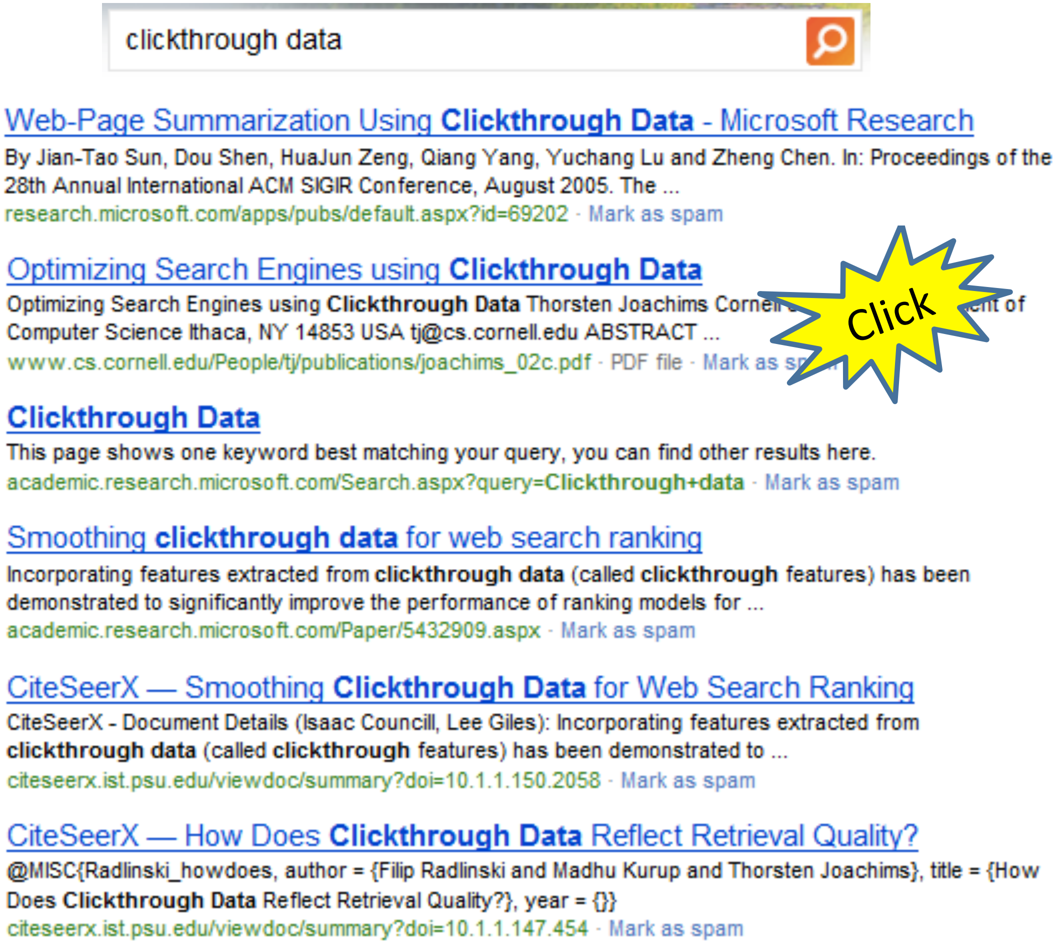
Recommender Systems

Personalized Clinical Treatment

Sequential Experimental Design

What if Rewards aren’t Directly Measureable?

Evaluating using Click Data



clickthrough data

Web-Page Summarization Using **Clickthrough Data** - Microsoft Research
By Jian-Tao Sun, Dou Shen, HuaJun Zeng, Qiang Yang, Yuchang Lu and Zheng Chen. In: Proceedings of the 28th Annual International ACM SIGIR Conference, August 2005. The ...
research.microsoft.com/apps/pubs/default.aspx?id=69202 - Mark as spam

Optimizing Search Engines using **Clickthrough Data**
Optimizing Search Engines using **Clickthrough Data** Thorsten Joachims Cornell ent of Computer Science Ithaca, NY 14853 USA tj@cs.cornell.edu ABSTRACT ...
www.cs.cornell.edu/People/tj/publications/joachims_02c.pdf - PDF file - Mark as sp

Clickthrough Data
This page shows one keyword best matching your query, you can find other results here.
academic.research.microsoft.com/Search.aspx?query=**Clickthrough+data** - Mark as spam

Smoothing **clickthrough data** for web search ranking
Incorporating features extracted from **clickthrough data** (called **clickthrough** features) has been demonstrated to significantly improve the performance of ranking models for ...
academic.research.microsoft.com/Paper/5432909.aspx - Mark as spam

CiteSeerX — Smoothing **Clickthrough Data** for Web Search Ranking
CiteSeerX - Document Details (Isaac Councill, Lee Giles): Incorporating features extracted from **clickthrough data** (called **clickthrough** features) has been demonstrated to ...
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.2058 - Mark as spam

CiteSeerX — How Does **Clickthrough Data** Reflect Retrieval Quality?
@MISC{Radlinski_howdoes, author = {Filip Radlinski and Madhu Kurup and Thorsten Joachims}, title = {How Does **Clickthrough Data** Reflect Retrieval Quality?}, year = {}}
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.454 - Mark as spam

Interpretation 1:
Result #2 is good.
(Absolute)

Interpretation 2:
Result #2 is better than Result #1.
(Relative / Preference)

Evaluating using Click Data

Retrieval Function A

Retrieval Function B

Personalized Search

Personalized Search ▸ Personalized Web Search Personalized Web ▸ Data Integration in Web Data Extraction System Personalized Web Search J I -R ONG ...

research.microsoft.com/pubs/79334/publishedversion.pdf · PDF file

A personalized search research based on vocabulary semantic net

Along with the fast developing of network technology, the number of Web page and us search become very enormous. In order to solve the problem of ...

portal.acm.org/citation.cfm?id=1794768

Zakta – Personalized Social Search E earch, firefox

Zakta, unlike other social search engines, Personal Researc with its ability to dig deeper to get the required inf

techpp.com/2009/10/15/zakta-personalized-s gine

Related Searches for personalized search research

Ontology-based Personalized S...
Bing Personalized Search
Personalized Search Engines

Disable Personalized Search
Personalized Search Results
Personalization Business

Personalized search - Wikipedia, the free encyclopedia

Personalized search refers to search experiences ... specific groups of people, personalized search depends on a user profile that is unique to the individual. Research ...

en.wikipedia.org/wiki/Personalized_Search

Research from Microsoft: Personalized rmining a Query ...

The other day I posted about a paper presente rence a few week's ago. Apparently, that got Findory CEO, Greg Linden, looking fo

blog.searchenginewatch.com/blog/050826-12 40

Adapting SEO for Personalized Search

Ok, but seriously, the last round of personalized search research we did here on it seems to suggest that a lot of the personalization, in relatively new query ...

www.searchenginejournal.com/adapting-seo-for-personalized-search/22207

ACM SIGIR Special Interest Group on Information Retriev ge

Welcome to the ACM SIGIR Web site. ACM SIGIR addresses issues user demands in the application of computers to the acquisition, organizati

www.sigir.org

Personalizi arch via Automated Analysis of Interests and ...

Automated Analysis of Interests and Activities Jaime Teevan MIT, CSAIL 32 ridge, MA 02138 USA tee van@csail. mit. edu Susan T ...

/um/people/sdumais/SIGIR2005-PersonalizedSearch.pdf · PDF file

for personalized search

We propose a alized search framework to utilize folksonomy for ... SIGIR '08 Proceedings of the 31st annual in tional ACM SIGIR conference on Research ...

portal.acm.org itation.cfm?id=1390363

Susan Dumais Homepage

Research Activities: I am interested ... issues, including: personal information management, web search ... and prospective. SIGIR 2010 Desktop Search Workshop ...

research.microsoft.com/en-us/um/people/sdumais

Personalized search - Wikipedia, the free encyclopedia

Personalized search refers to search experiences ... Research systems that personalize search results model their users in ... to personalize global Web search". SIGIR: 287 ...

en.wikipedia.org/wiki/Personalized_Search

Xuehua's Publications

Proceedings of 2003 ACM Conference on Research and Development on Information Retrieval (SIGIR'2003), pages 377-378. pdf ppt; Demos. UCAIR Toolbar: A Personalized Search ...

sifaka.cs.uiuc.edu/xshen/publication.html

Event: IR

SIGIR is the major international forum for the presentation of new research results and for the demonstration of ... summarization, task models, personalized search ...

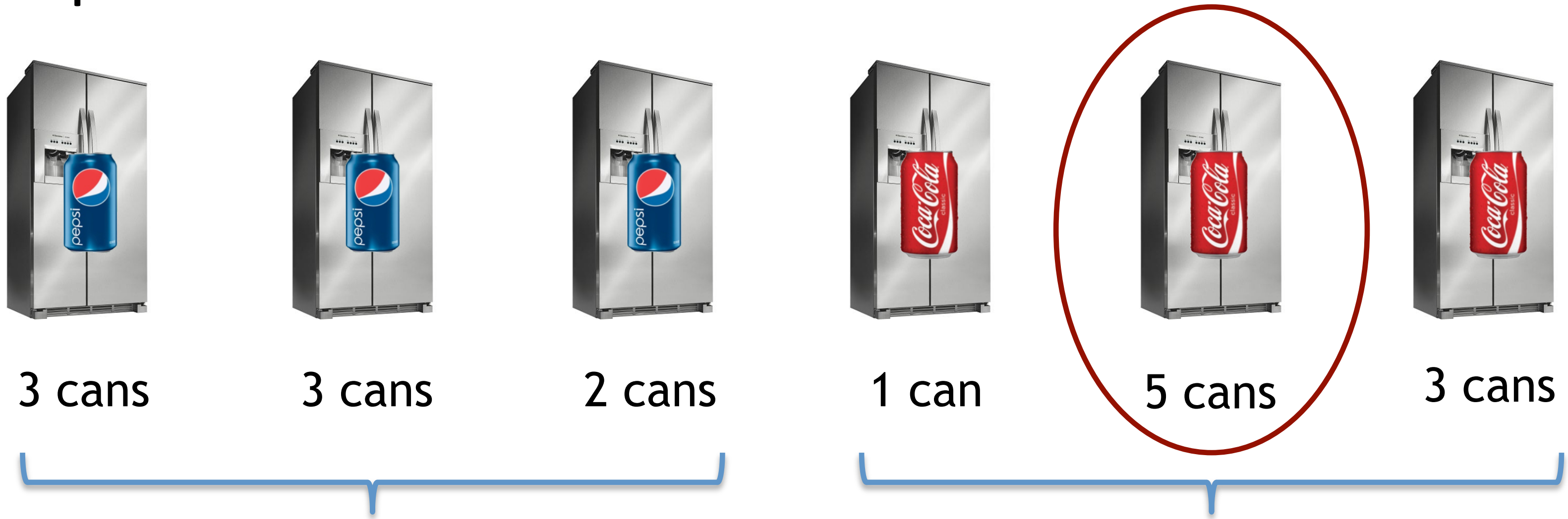
portal.acm.org/browse_dl.cfm?linked=1&part=series&idx=SERIES278&coll=ACM&dl=ACM

Analogy to Sensory Testing

- (Hypothetical) taste experiment:
  - Natural usage context

- Experiment 1: **Absolute Metrics**

Very Thirsty!



3 cans

3 cans

2 cans

1 can

5 cans

3 cans

Total: 8 cans

Total: 9 cans

Analogy to Sensory Testing

- (Hypothetical) taste experiment:
  - Natural usage context


vs


- Experiment 1: **Relative Metrics**

2 - 1 3 - 0 2 - 0 1 - 0 4 - 1 2 - 1

All 6 prefer Pepsi

Interleaving (Taste Test in Search)

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.ivebeenthere.co.u
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

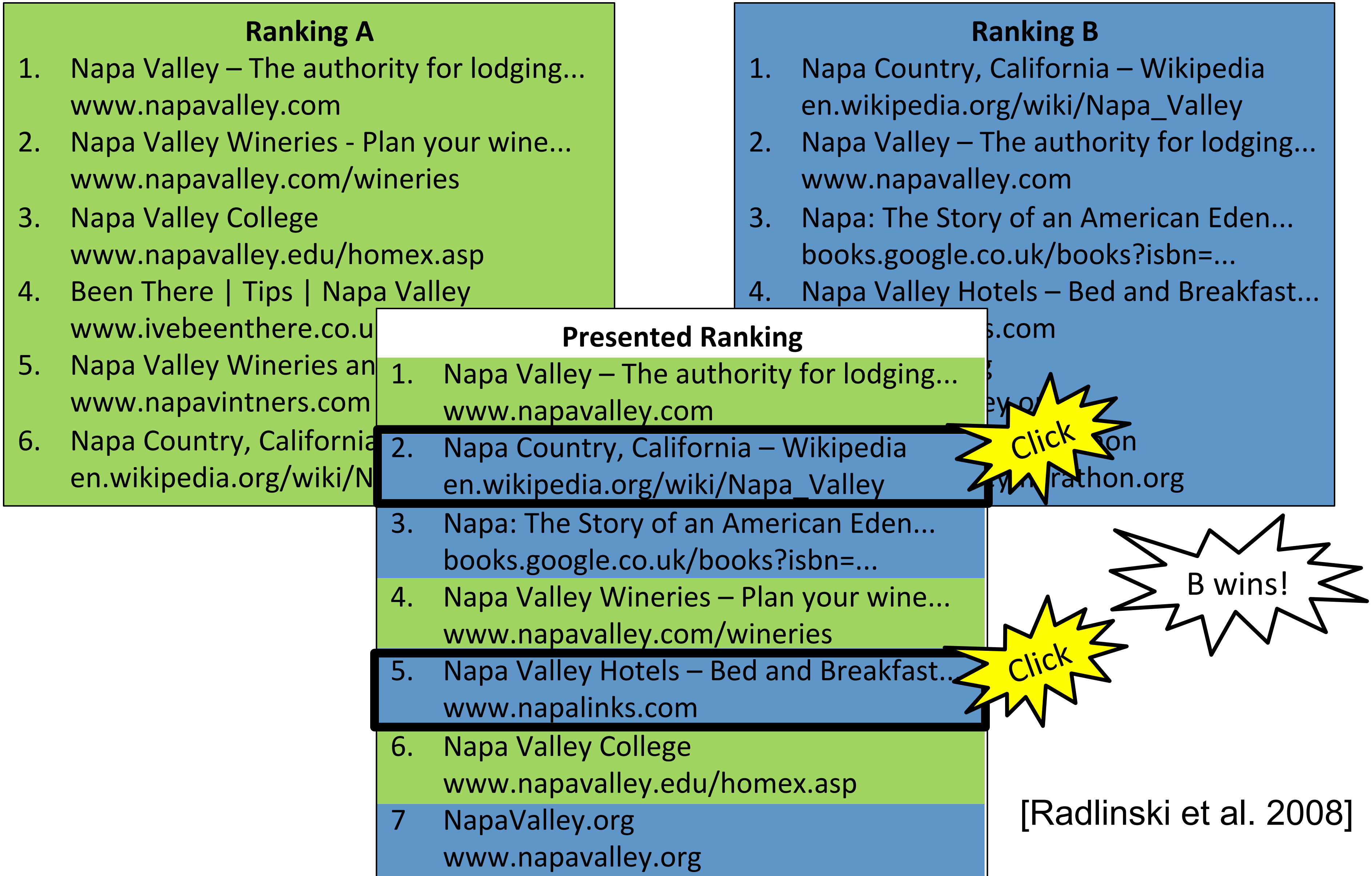
1. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...
   s.com
   g
   ey.org
   y Marathon
   eymarathon.org

Presented Ranking

A B

[Radlinski et al. 2008]

Interleaving (Taste Test in Search)



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.ivebeenthere.co.u
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

1. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...
   s.com

   ey.o

   on

   rathon.org

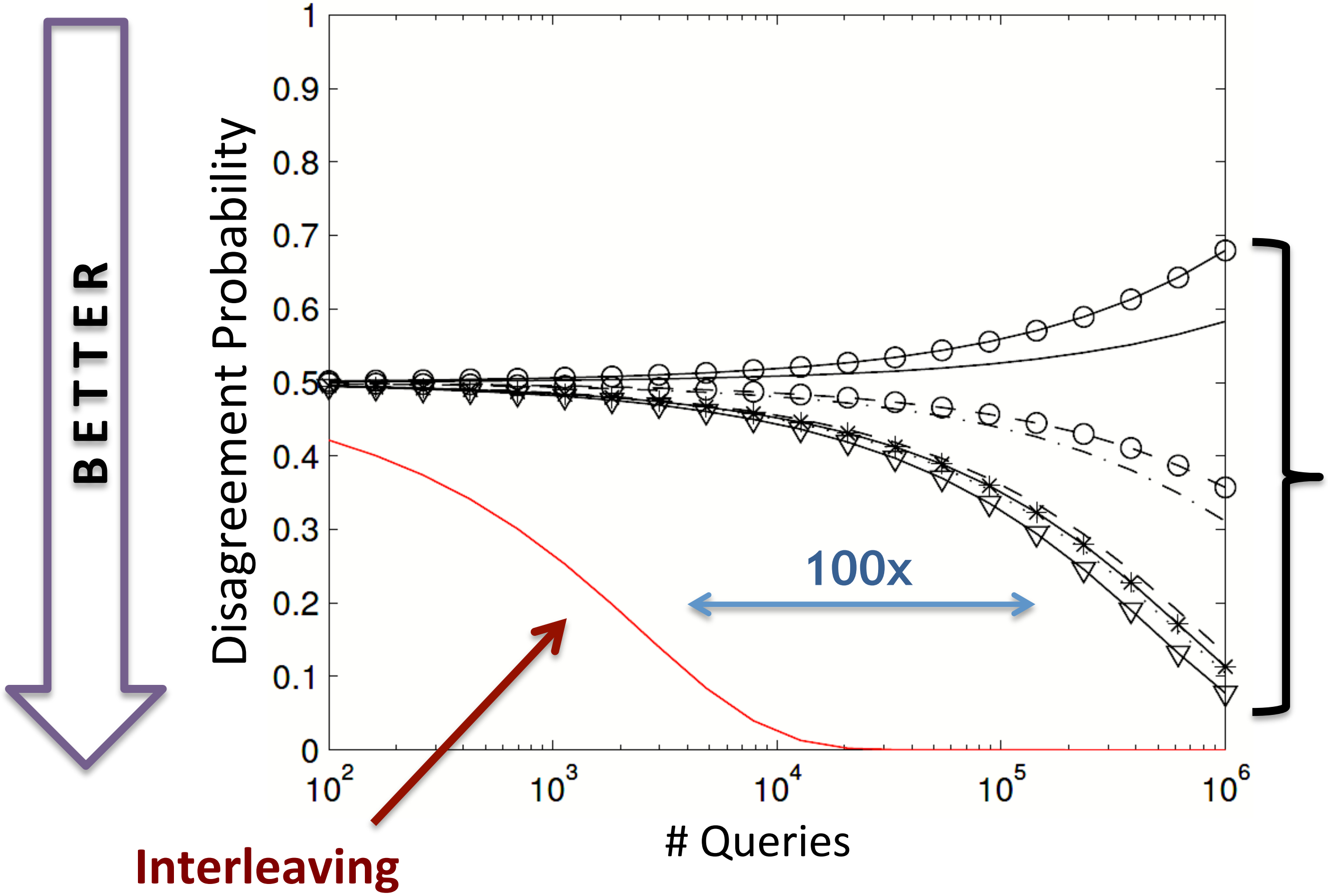
Presented Ranking

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 Valley College
   www.napavalley.edu/homex.asp
7 NapaValley.org
   www.napavalley.org

[Radlinski et al. 2008]

Deployment on Yahoo! Search Engine

Comparing Two Ranking Functions



BETTER

100x

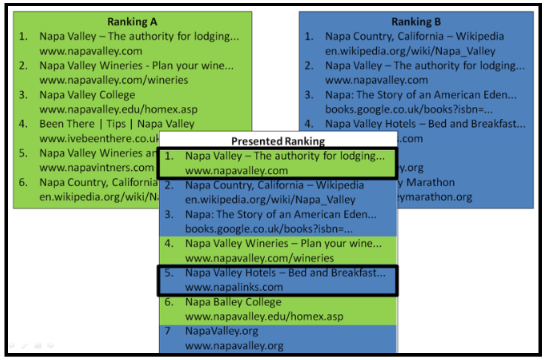
Interleaving

Absolute Metrics
E.g., #Clicks@1,
Total #Clicks, etc.

Each ranking function
receives 50% traffic

- Interleaving is more **sensitive** and more **reliable**

[Chapelle, Joachims, Radlinski & Yue, TOIS 2012]



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.ivebeenthere.co.uk
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

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

y.org
y Marathon
eymarathon.org

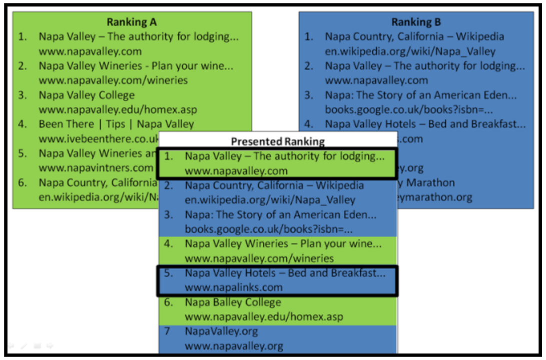
Presented Ranking

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
   www.napavalley.edu/homex.asp
7 NapaValley.org
   www.napavalley.org

Interleave A vs B

. . .

	Left wins	Right wins
A vs B	0	1
A vs C	0	0
B vs C	0	0



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.ivebeenthere.co.uk
5. Napa Valley Wineries a
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

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

y.org

y Marathon
eymarathon.org

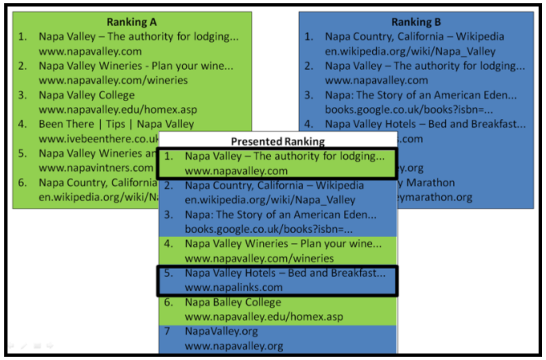
Presented Ranking

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
   www.napavalley.edu/homex.asp

7 NapaValley.org
www.napavalley.org

Interleave A vs C

	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	0



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.ivebeenthere.co.uk
5. Napa Valley Wineries ar
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

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

y.org

y Marathon
eymarathon.org

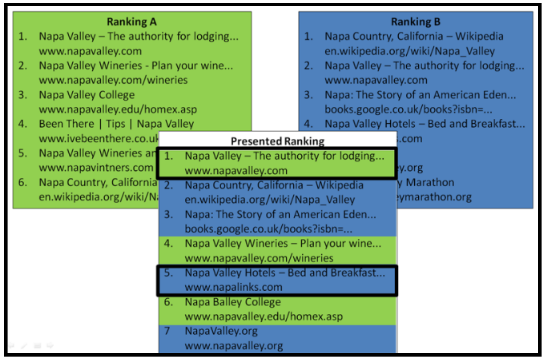
Presented Ranking

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
   www.napavalley.edu/homex.asp

7 NapaValley.org
www.napavalley.org

Interleave B vs C

	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	1



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.ivebeenthere.co.uk
5. Napa Valley Wineries ar
www.napavintners.com
6. Napa Country, California
en.wikipedia.org/wiki/Na

Ranking B

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

y.org
y Marathon
eymarathon.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
www.napavalley.edu/homex.asp
7 NapaValley.org
www.napavalley.org

Interleave A vs C

. . .

	Left wins	Right wins
A vs B	0	1
A vs C	1	1
B vs C	0	1

Dueling Bandits Problem

Goal: Maximize total user utility

Exploit: run C
(interleave C with itself)

Explore: interleave A vs B

Best: A
(interleave A with itself)

How to interact optimally?

	Left wins	Right wins
A vs B	0	1
A vs C	1	1
B vs C	0	1

Example Pairwise Preferences

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

- **Utility function may not exist**
- **How to define regret?**

Values are Pr(row > col) – 0.5

Example Pairwise Preferences

	A	B	C	D	E	F
A	**0**	**0.03**	**0.04**	**0.06**	**0.10**	**0.11**
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

- **Utility function may not exist**
- **How to define regret?**
- **Compare against best bandit!**

Values are Pr(row > col) – 0.5

Dueling Bandits Problem

(with Josef Broder, Robert Kleinberg and Thorsten Joachims)

- K bandits $b_1, \ldots, b_K$
- Each iteration: compare (duel) two bandits
  - Observe (noisy) outcome

Requires Dueling Mechanism

- Cost function (regret):

$$R_T = \sum_{t=1}^{T} P(b^* > b_t) + P(b^* > b_t') - 1$$

- (b_t, b_t') are the two bandits chosen
- b^* is the overall best one
- (How much human user preferred b^* over chosen bandits)

[Yue, Broder, Kleinberg & Joachims, COLT 2009]

Dueling Bandits Problem

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	**0.10**	**0.11**
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	**0.03**
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Observe

$$R_T = \sum_{t=1}^{T} P(b^* > b_t) + P(b^* > b_t') - 1$$

Compare E & F:

- P(A > E) = 0.60
- P(A > F) = 0.61
- **Incurred Regret = 0.21**

Values are Pr(row > col) – 0.5

Dueling Bandits Problem

	A	B	C	D	E	F
A	0	**0.03**	**0.04**	0.06	0.10	0.11
B	-0.03	0	**0.03**	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Observe

$$R_T = \sum_{t=1}^{T} P(b^* > b_t) + P(b^* > b_t') - 1$$

Compare B & C:
- P(A > B) = 0.53
- P(A > C) = 0.54
- **Incurred Regret = 0.07**

Values are Pr(row > col) – 0.5

Dueling Bandits Problem

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Observe

$$R_T = \sum_{t=1}^{T} P(b^* > b_t) + P(b^* > b_t') - 1$$

Values are Pr(row > col) – 0.5

Compare A & A:

- P(A > A) = 0.50
- P(A > A) = 0.50
- **Incurred Regret = 0.00**

Basic Modeling Assumptions

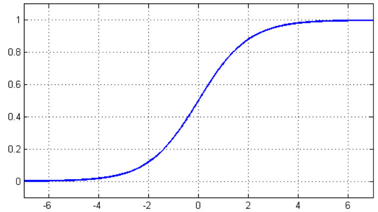
- $P(b_i > b_j) = ½ + \varepsilon_{ij}$ (distinguishability)
- **Strong Stochastic Transitivity**

$$\varepsilon_{ik} \geq \max\left\{\varepsilon_{ij}, \varepsilon_{jk}\right\}$$

  - For three bandits $b_i > b_j > b_k$:
  - Monotonicity property
- **Stochastic Triangle Inequality**

$$\varepsilon_{ik} \leq \varepsilon_{ij} + \varepsilon_{jk}$$

  - For three bandits $b_i > b_j > b_k$:
  - Diminishing returns property
- Satisfied by many standard models
  - E.g., Logistic / Bradley-Terry



[Yue, Broder, Kleinberg & Joachims, COLT 2009]

Strong Stochastic Transitivity

(Assumes Condorcet Winner)

$$\varepsilon_{ik} \geq \max\left\{\varepsilon_{ij}, \varepsilon_{jk}\right\}$$

Monotonic

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Monotonic

Values are Pr(row > col) – 0.5

Stochastic Triangle Inequality

(Assumes Condorcet Winner)

$$\varepsilon_{ik} \leq \varepsilon_{ij} + \varepsilon_{jk}$$

Red ≤ Blue + Green

	A	B	C	D	E	F
A	0	0.03	**0.04**	**0.06**	0.10	0.11
B	-0.03	0	**0.03**	**0.05**	0.08	0.11
C	-0.04	-0.03	0	**0.04**	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Values are Pr(row > col) – 0.5

Stochastic Triangle Inequality

(Assumes Condorcet Winner)

$$\varepsilon_{ik} \leq \varepsilon_{ij} + \varepsilon_{jk}$$

Red ≤ Blue + Green

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

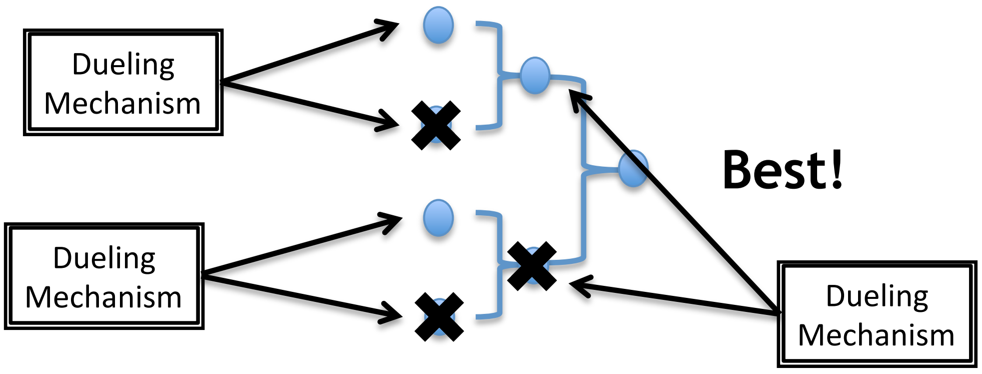
Values are Pr(row > col) – 0.5

Other Modeling Assumptions

- Approximate Linearity $\varepsilon\downarrow ik - \varepsilon\downarrow jk \geq \gamma\varepsilon\downarrow ij$
- Other Solution Concepts
  - Borda Winner [Jamieson et al., 2015]
  - Copeland Winner [Zoghi et al., 2015]
  - Von Neuman Winner [Dudik et al., 2015]
  - General Tournament Solutions [Ramamohan et al., 2016]
- Conditioning on Context [Dudik et al., 2015]
- Adversarial Setting [Gajane et al., 2015]
- Continuous Convex Setting [Yue & Joachims, 2009]

Connection to Tournaments

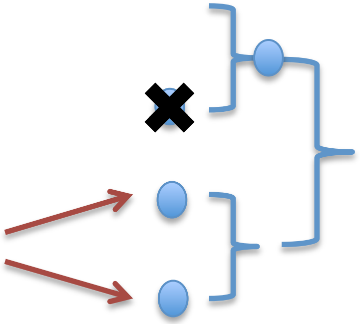
- Each pair "duels" until statistical significance



- Aka Noisy Tournament [Feige et al., 1994]
  - Guarantees finding best bandit w.h.p.
  - **Can we use as explore algorithm?**

Tournament is Bad

- Each pair "duels" until statistical significance



Problem: two
Equally bad bandits

- **Analogy:** Hypothetical Soccer Tournament
  - A team wins when it has a 3-goal lead
  - Audience prefers good teams play **(regret)**
  - **Two (nearly) equally bad teams will play for a long time**

Many Algorithms

- Interleaved Filter [Yue et al., 2009]
- Beat the Mean [Yue & Joachims, 2011]
- SAVAGE [Urvoy et al., 2013]
- RMED [Komiyama et al., 2015]
- RUCB [Zoghi et al., 2014; 2015]
- Double Thompson Sampling [Wu & Liu, 2016]
- Sparring [Ailon et al., 2014]
- SelfSparring (under review)
- …

Many Algorithms

- Interleaved Filter [Yue et al., 2009]
- Beat the Mean [Yue & Joachims, 2011]
- SAVAGE [Urvoy et al., 2013]
- RMED [Komiyama et al., 2015]
- RUCB [Zoghi et al., 2014; 2015]
- Double Thompson Sampling [Wu & Liu, 2016]
- **Sparring [Ailon et al., 2014]**
- **SelfSparring (under review)**

+ Extensions!

- …

Outline

- Algorithms & Theory
  - Sparring [Ailon et al., 2014]
  - Challenges in Regret Analysis
  - SelfSparring
  - Theoretical Results
- Experiments
- Extensions
  - Application to Personalized Clinical Treatment

Dueling Bandits ≈ Zero-Sum Game

Player 1

Player 2

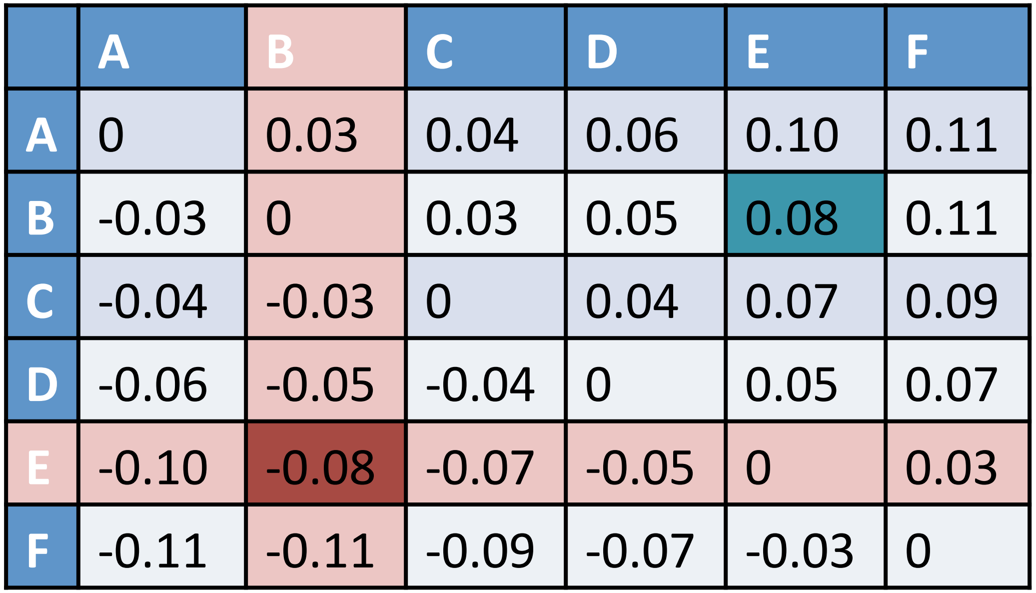
	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Basic Setting: Single Dominant Strategy

Regret = Opportunity Cost to Social Welfare

- Values are Pr(row > col) – 0.5

Dueling Bandits ≈ Zero-Sum Game



Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Basic Setting: Single Dominant Strategy

Regret = Opportunity Cost to Social Welfare

- Values are Pr(row > col) – 0.5

Dueling Bandits ≈ Zero-Sum Game

Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Basic Setting: Single Dominant Strategy

Regret = Opportunity Cost to Social Welfare

- Values are Pr(row > col) – 0.5

Dueling Bandits ≈ Zero-Sum Game

Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

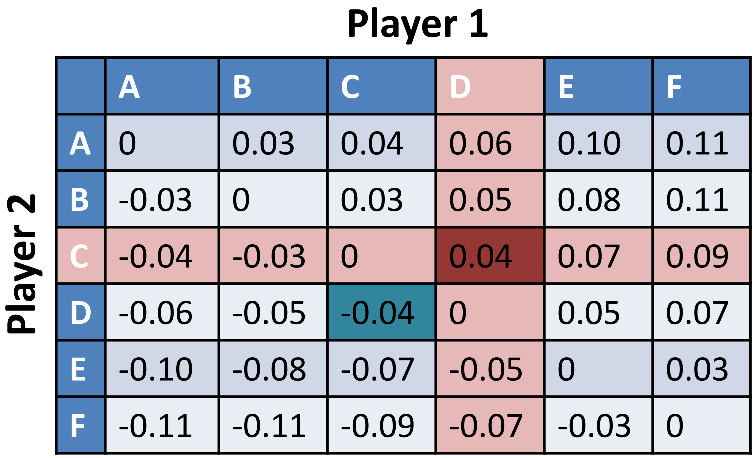
Basic Setting: Single Dominant Strategy

Regret = Opportunity Cost to Social Welfare

- Values are Pr(row > col) – 0.5

Sparring

- Instantiate 2 MAB algorithms: P_1 & P_2
- For t = 1, …
  - – P_1 chooses a_1
  - – P_2 chooses a_2
  - – Duel a_1 vs a_2
  - – Provide feedback



Player 1

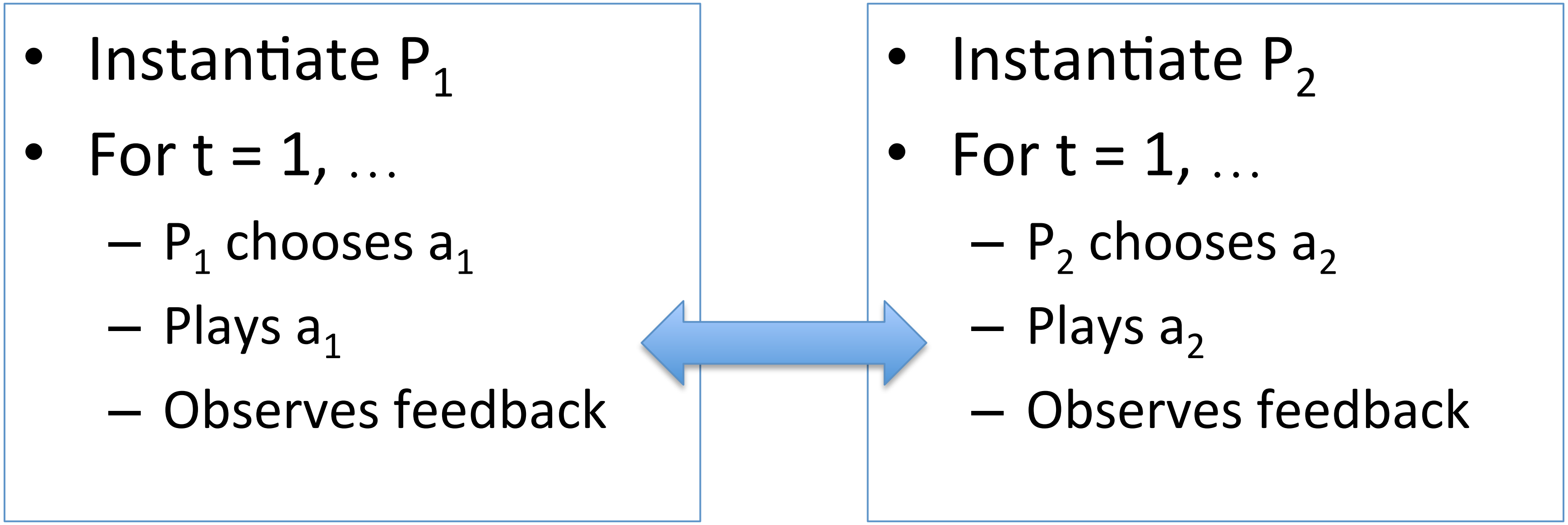
Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Reducing Dueling Bandits to Cardinal Bandits
Ailon, Karnin & Joachims, ICML 2014

Intuition

- Reduction to standard MAB settings
  - Each player selfishly maximizes own reward

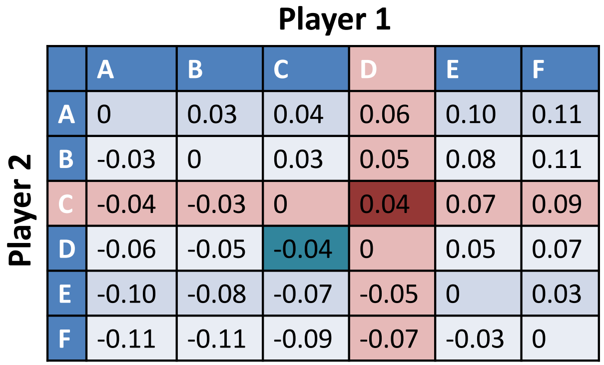


- Instantiate P_1
- For t = 1, …
  - P_1 chooses a_1
  - Plays a_1
  - Observes feedback

- Instantiate P_2
- For t = 1, …
  - P_2 chooses a_2
  - Plays a_2
  - Observes feedback

Drifting Reward Distributions

- Playing against a changing environment
  - – Rewards depend on other player
- Players learn over time
  - – Environment drifts over time



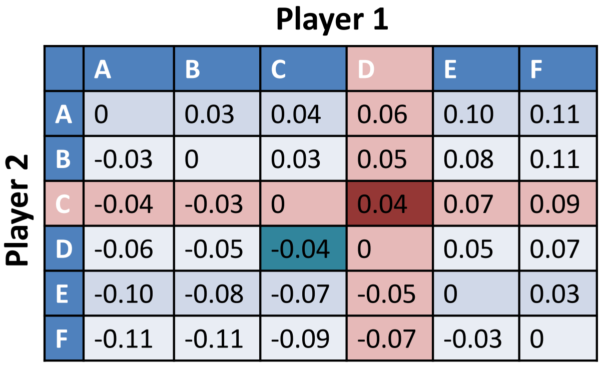
Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Stochastic vs Adversarial

- **Stochastic:** Reward of each arm fixed
  - E.g., UCB1 & Thompson Sampling
  - No guarantees within Sparring
- **Adversarial:** Rewards chosen adversarially
  - E.g., EXP3
  - Very slow in practice
- **Not fully adversarial!**



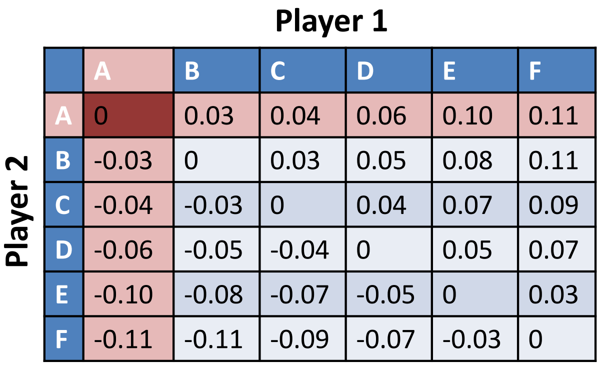
Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Thought Experiment

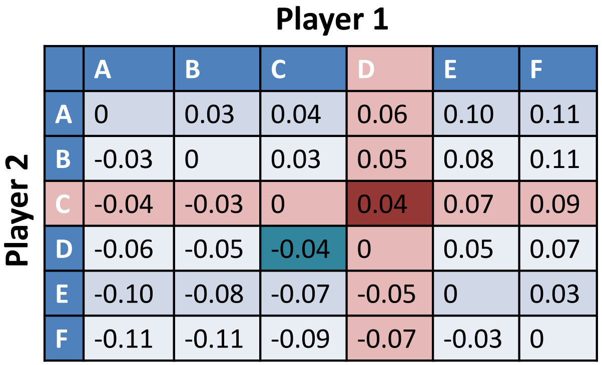
- If one player has converged
  - – Then other player is playing stochastic MAB!
- Both players implement learning algorithms
  - – Slowly drifts to fixed distribution



Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0



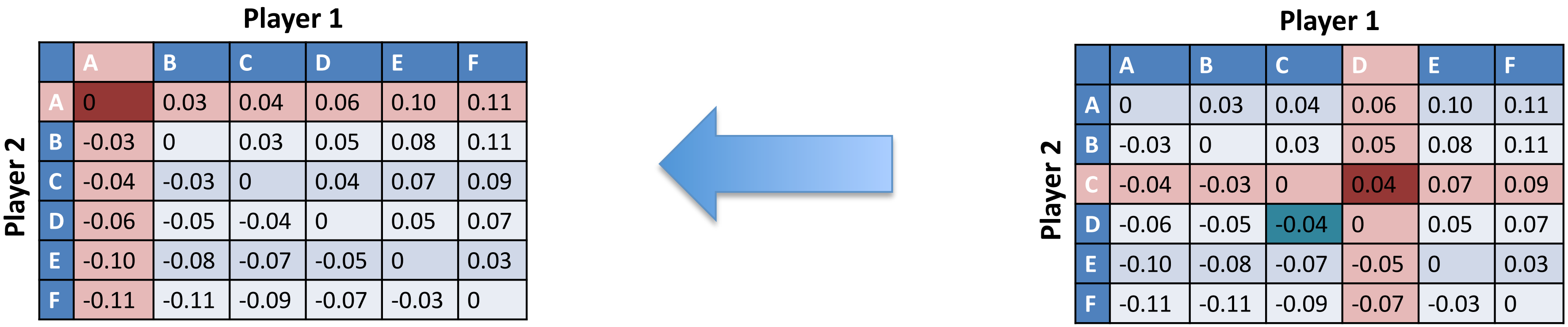
Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Chicken & Egg Problem

- If one player has converged
  - Can prove other player is converging
- If one player is converging
  - Can prove other is converging (slower)



Player 1

Player 2

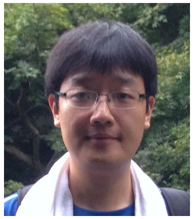
	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

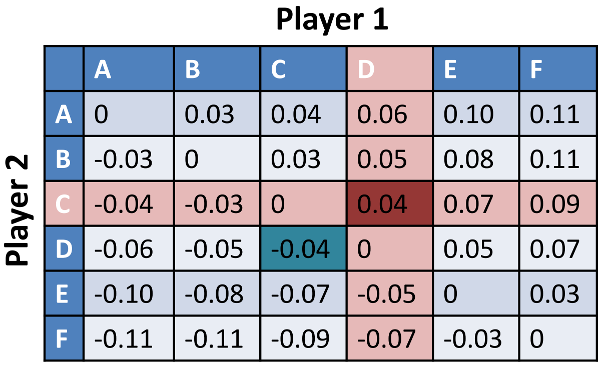
SelfSparring



Yanan
Sui

- Instantiate 1 MAB algorithm P
- For t = 1, …
  - – P chooses a_1
  - – P chooses a_2
  - – Duel a_1 vs a_2
  - – Provide feedback

Probabilistic Bandit Algorithm (Thompson Sampling)



Player 1

Player 2

	A	B	C	D	E	F
A	0	0.03	0.04	0.06	0.10	0.11
B	-0.03	0	0.03	0.05	0.08	0.11
C	-0.04	-0.03	0	0.04	0.07	0.09
D	-0.06	-0.05	-0.04	0	0.05	0.07
E	-0.10	-0.08	-0.07	-0.05	0	0.03
F	-0.11	-0.11	-0.09	-0.07	-0.03	0

Multi-dueling Bandits with Dependent Arms
Sui, Zhuang, Burdick & Yue, (under review)

Theoretical Insights
(SelfSparring)

- Each player playing against itself
  - Can tightly couple convergence of both players

- Once converged enough
  - Can prove optimal regret bound (asymptotic)

\# Arms

$$O(K/\varepsilon \log(T))$$

Time horizon

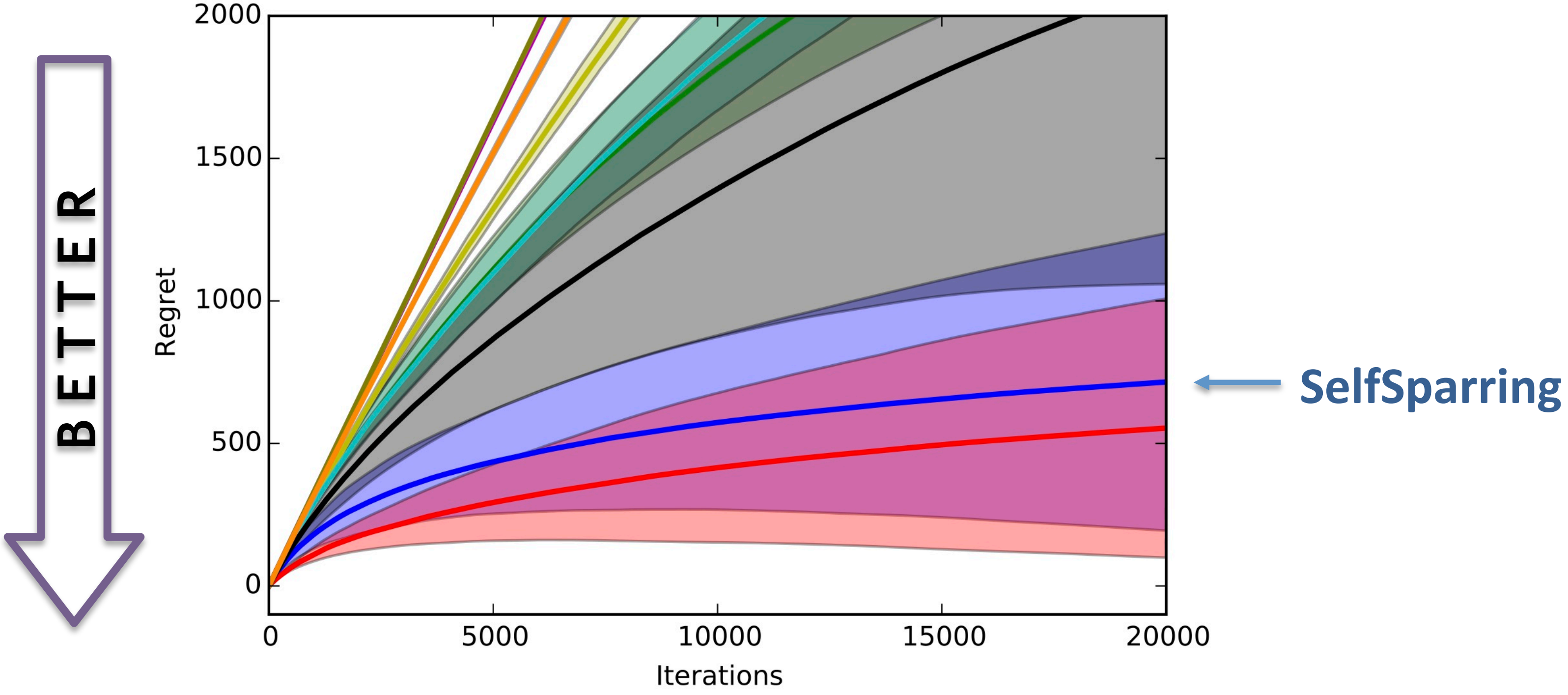
Gap between
best & 2nd best

Multi-dueling Bandits with Dependent Arms
Sui, Zhuang, Burdick & Yue, (under review)

SelfSparring

- Optimal asymptotic regret bound
- Performs very well in practice
- Easily extendable to new settings

Basic Experiments



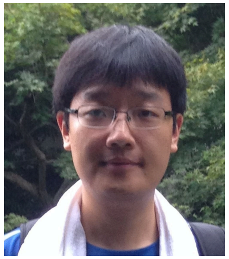
BETTER

SelfSparring

Multi-dueling Bandits with Dependent Arms

Sui, Zhuang, Burdick & Yue, (under review)

Ongoing Work:
Personalized Clinical Treatment



Yanan Sui

Medtronic
human
array

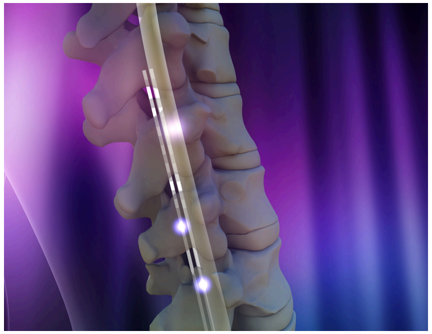


Image source:
williamcapicottomd.com

SCI Patient

Each patient is unique

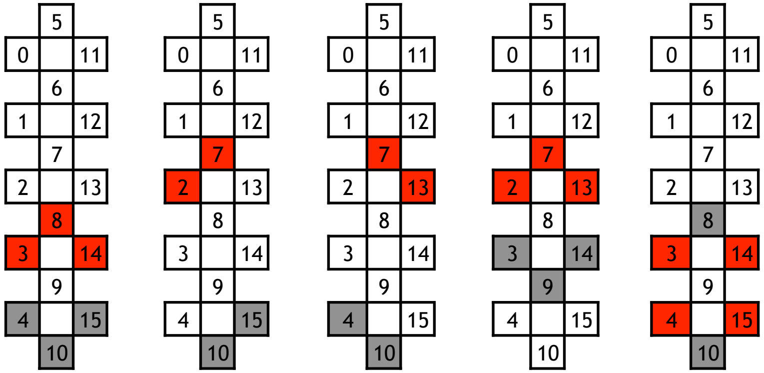
10^6 possible configurations!

Challenges

- Many arms
  - K = 10^6

$$O(K/\varepsilon \log(T))$$

- Duel more than 2 arms

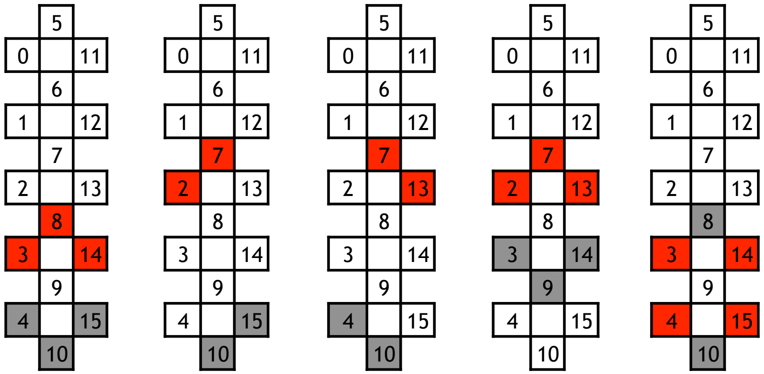


Challenges

- Many arms
  - K = 10^6

$O(K/\varepsilon \log(T))$

- **Duel more than 2 arms**

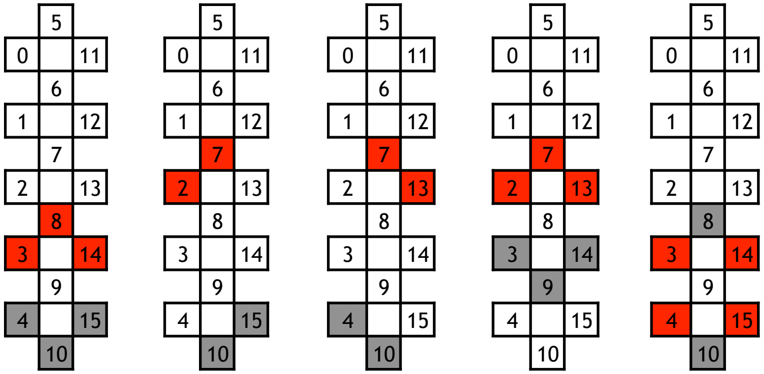


Multi-Dueling Bandits

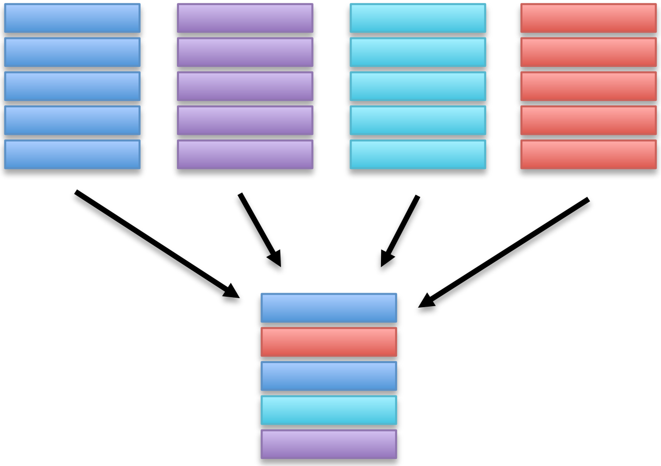
- For t = 1, …
  - Choose M arms
  - Duel M arms
  - Observe outcomes

All Pairs
Winner takes all
Random set of pairs

Comparing Multiple Stimuli



Probabilistic Multi-Leaving



Multi-Dueling SelfSparring

- SelfSparring generalizes trivially!
  - Just sample M times!
  - (Sparring requires M separate bandit algorithms)

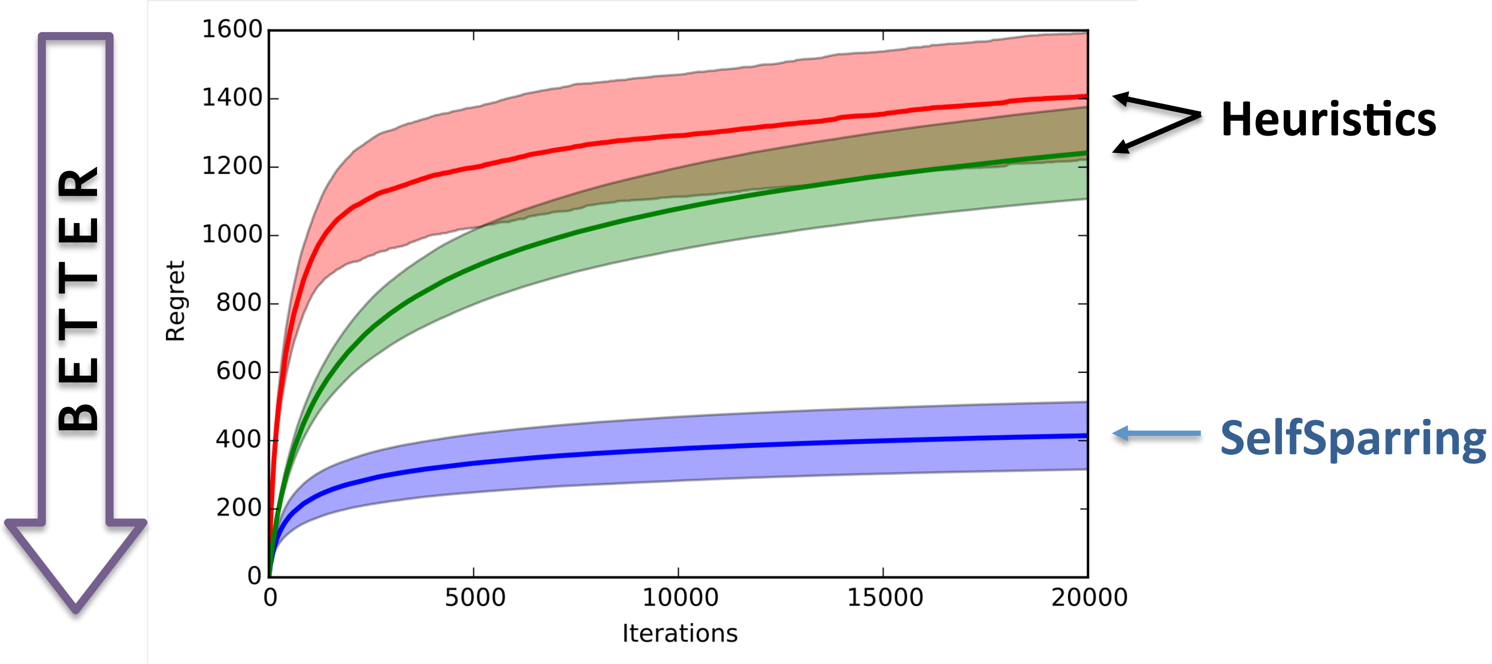
- Can prove same regret bound

$$O(K/\varepsilon \log(T))$$

Constant depends on dueling mechanism

Multi-dueling Bandits with Dependent Arms
Sui, Zhuang, Burdick & Yue, (under review)

Multi-Dueling Experiments



Heuristics

SelfSparring

Sparring not displayed due to very poor scaling
Most DB algorithms not applicable

Multi-dueling Bandits with Dependent Arms
Sui, Zhuang, Burdick & Yue, (under review)

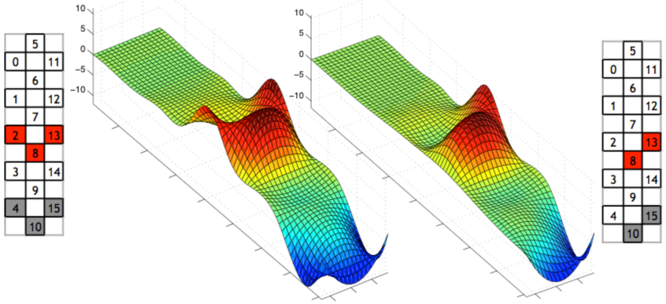
Dueling Bandits w/ Dependent Arms

- Suppose K is very large (possibly infinite)
  - But arms have dependency structure
  - E.g., P(a>b) ≈ P(a'>b) if a similar to a'
  - Measure similarity using kernel

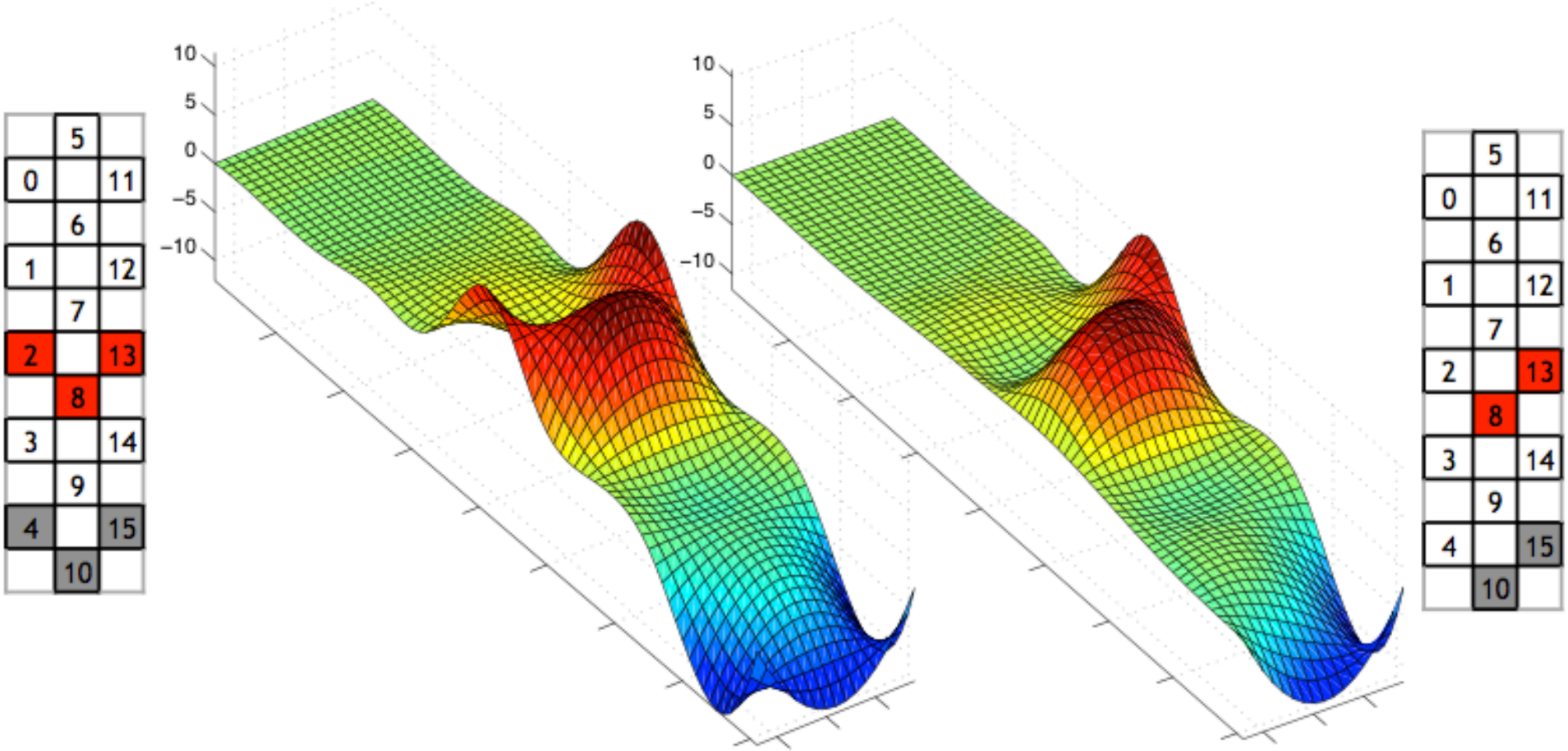
Dimensionality of Kernel

- Want convergence to depend on D
  - And not K!

Multi-dueling Bandits with Dependent Arms
Sui, Zhuang, Burdick & Yue, (under review)



Visualizing Electrical Potentials

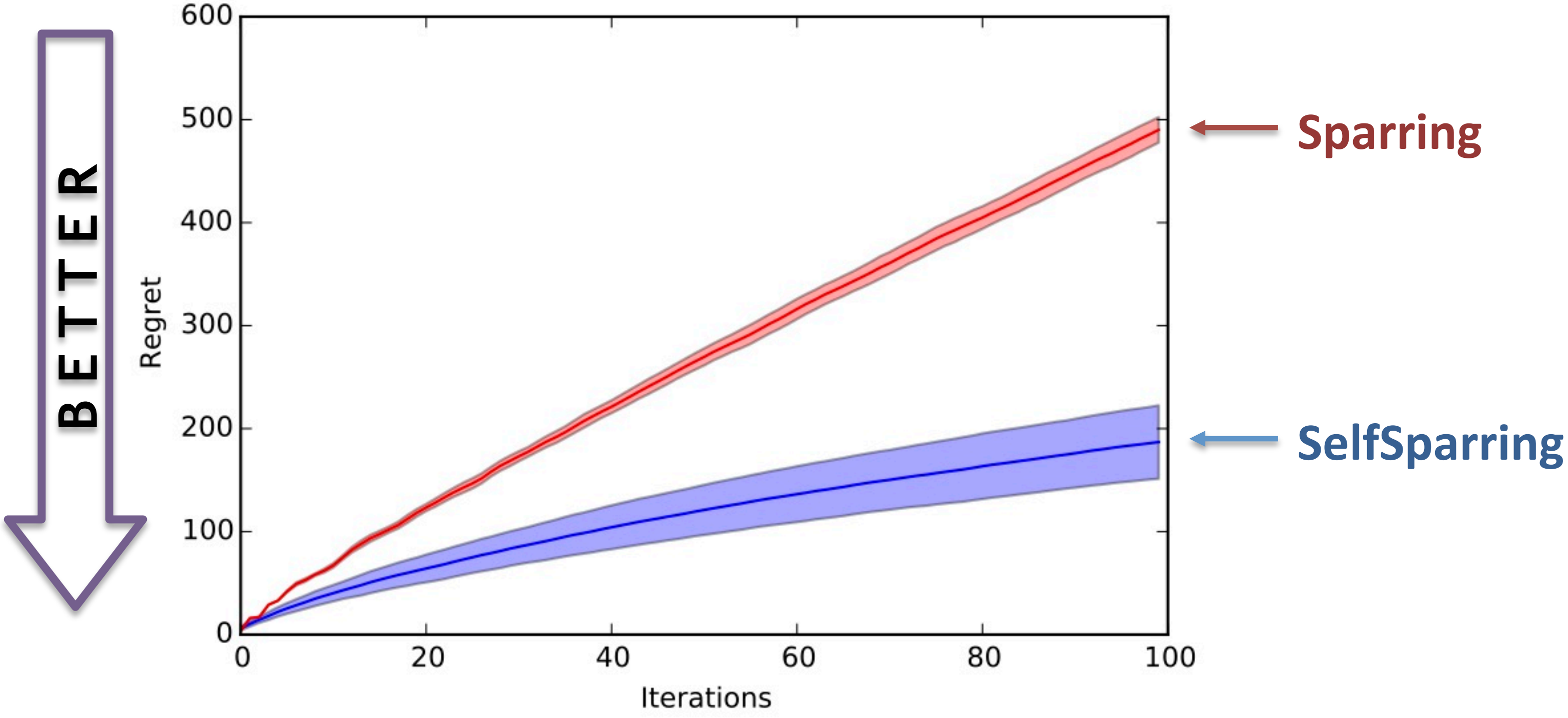


SelfSparring w/ Gaussian Processes

- Maintain Gaussian process prior
  - $f \sim GP(Y)$
  - $f(a)$ = probability arm a beats current distribution

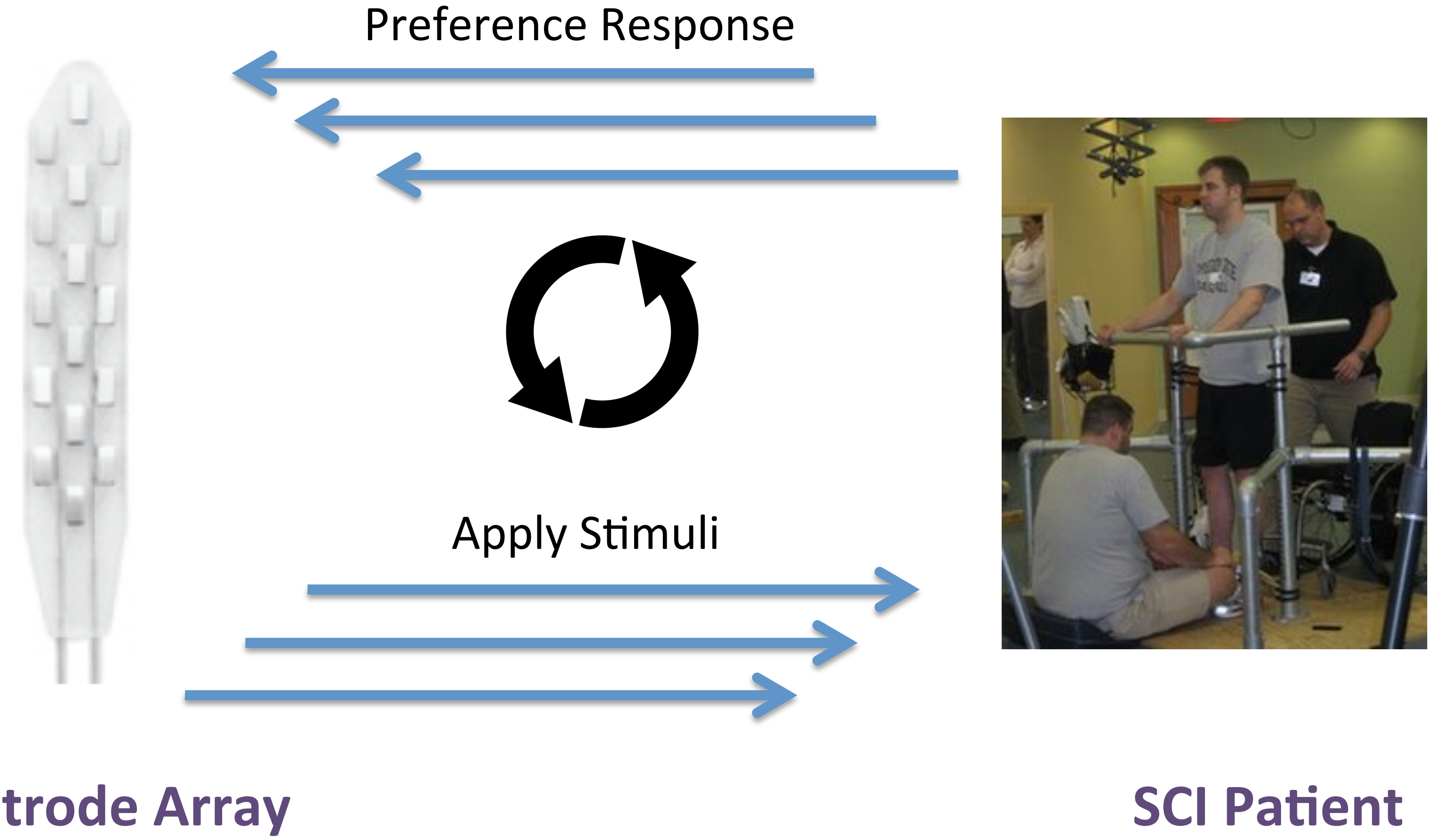
- Each time step:
  - Sample f_1 ,…, f_M
  - Choose a_1 ,…,a_M
  - Duel arms, incorporate feedback into Y

Kernel Multi-Dueling Experiments



Multi-dueling Bandits with Dependent Arms
Sui, Zhuang, Burdick & Yue, (under review)

Back to Motivating Application



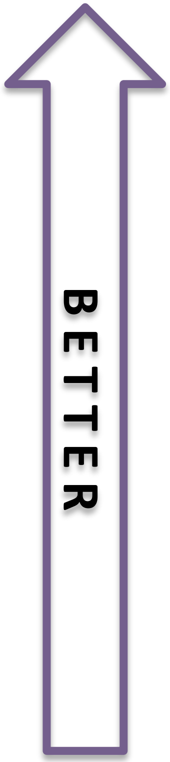
Preference Response

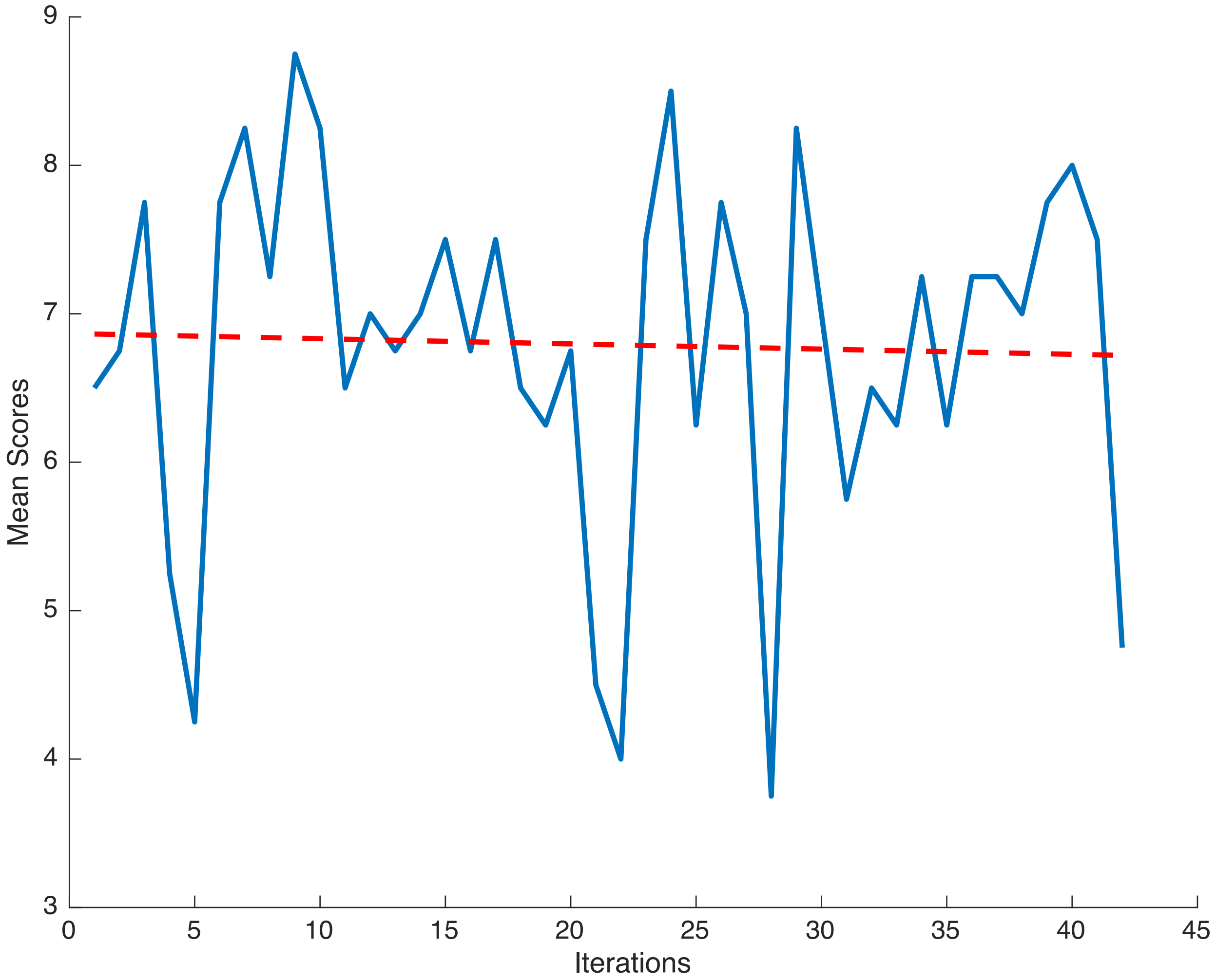
Apply Stimuli

Electrode Array

SCI Patient

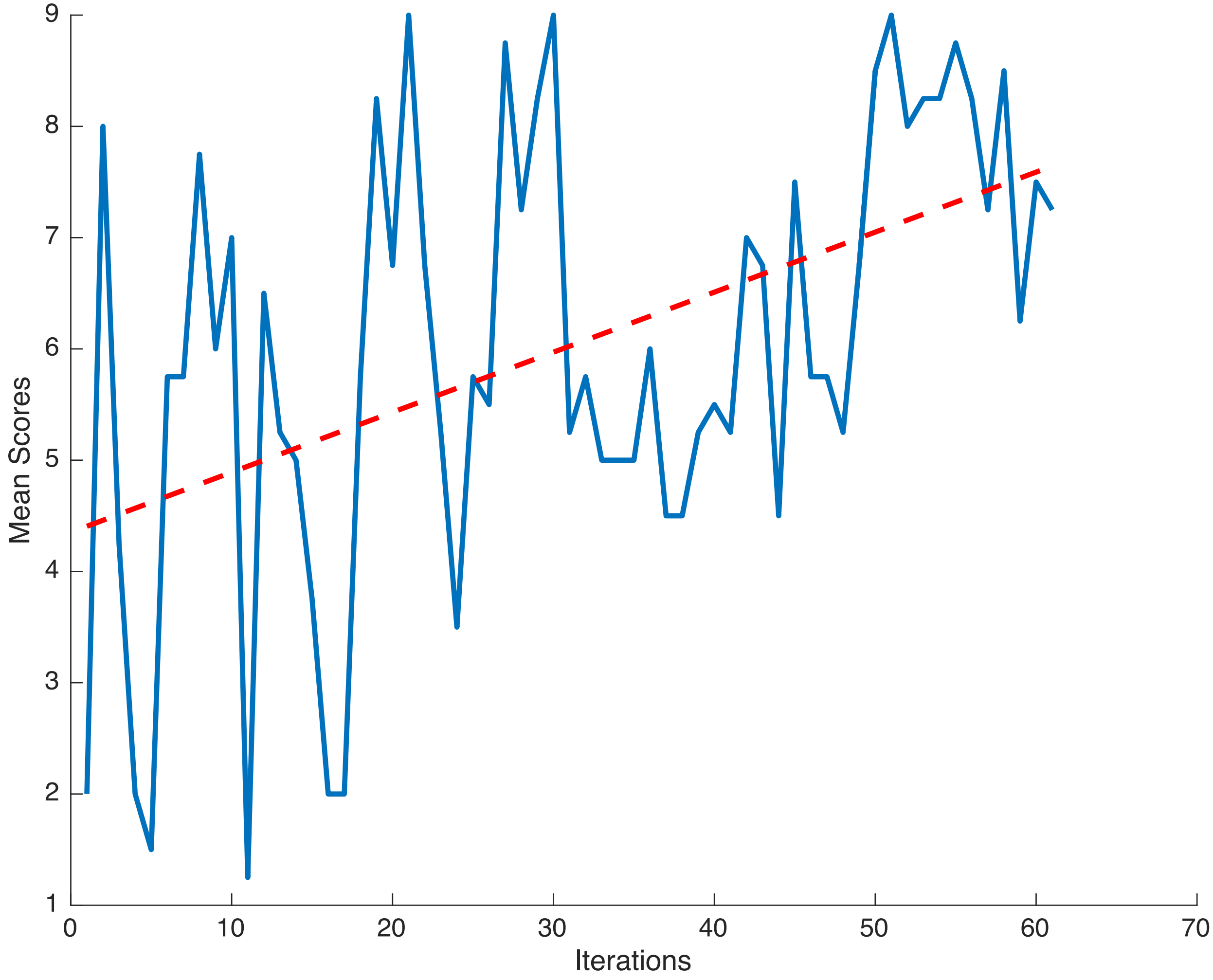
Preliminary Clinical Results: Human





Preliminary Clinical Results: DB Algorithm

BETTER



Summary: Dueling Bandits Problem

- Elicits preference feedback
  - Motivated by human-centric personalization
  - Characterizes explore/exploit tradeoff

- Ongoing research
  - Personalized clinical treatment
  - Dependent arms (regret bound?)
  - Complex dueling mechanisms

The K-armed Dueling Bandits Problem, Yisong Yue, Josef Broder, Robert Kleinberg and Thorsten Joachims, COLT 2009

Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem, Yisong Yue and Thorsten Joachims, ICML 2009

Beat the Mean Bandit, by Yisong Yue and Thorsten Joachims, ICML 2011

Large-Scale Validation and Analysis of Interleaved Search Evaluation, Olivier Chapelle, Thorsten Joachims, Filip Radlinski, Yisong Yue, TOIS 2012

Probabilistic Multileave for Online Retrieval Evaluation, Anne Schuth et al., SIGIR 2015

Reusing Historical Interaction Data for Faster Online Learning to Rank for IR, Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke, WSDM 2013

Generic Exploration and K-armed Voting Bandits, Tanguy Urvoy, Fabrice Clerot, Raphael Feraud and Sami Naamane, ICML 2013

Reducing Dueling Bandits to Cardinal Bandits, Nir Ailon, Zohar Karnin and Thorsten Joachims, ICML 2014

Relative Upper Confidence Bound for the K-armed Dueling Bandit Problem, Masrour Zoghi, Shimon Whiteson, Remi Munos and Maarten de Rijke, ICML 2014

Clinical Online Recommendation with Subgroup Rank Feedback, Yanan Sui and Joel Burdick, RecSys 2014

Sparse Dueling Bandits, Kevin Jamieson, Sumeet Katariya, Atul Deshpande and Robert Nowak, AISTATS 2015

Contextual Dueling Bandits, Miro Dudik, Robert Schapire and Alex Slivkins, COLT 2015

A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits, Pratik Gajane, Tanguy Urvoy and Fabrice Clerot, ICML 2015

Copeland Dueling Bandits, Masrour Zoghi, Zohar Karnin, Shimon Whiteson and Maarten de Rijke, NIPS 2015

Online Rank Elicitation for Plackett-Luce: A Dueling Bandits Approach, Balazs Szorenyi, Robert Busa-Fekete, Adil Paul and Eyke Hullermeier, NIPS 2015

Copeland Dueling Bandit Problem: Regret Lower Bound, Optimal Algorithm, and Computationally Efficient Algorithm, Junpei Komiyama, Junya Honda, Hiroshi Nakagawa, ICML 2016

Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions, Siddartha Ramamohan, Arun Rajkumar, Shivani Agarwal, NIPS 2016

Double Thompson Sampling for Dueling Bandits, Huasen Wu, Xin Liu, NIPS 2016

Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions, Siddartha Ramamohan, Arun Rajkumar, Shivani Agrawal, NIPS 2016

Multi-dueling Bandits with Dependent Arms, Yanan Sui, Vincent Zhuang, Joel Burdick, Yisong Yue, (under review)