The State of the Online Learning to Rank Field

Harrie Oosterhuis

December 2, 2019

University of Amsterdam

oosterhuis@uva.nl

Based on the SIGIR 2019 Tutorial:
Unbiased Learning to Rank: Counterfactual and Online Approaches
Harrie Oosterhuis, Rolf Jagerman, Maarten de Rijke

Introduction: Learning to Rank from User Interactions

Learning to Rank from User Interactions

Solves most of the problems of expert-annotations:

- Interactions are **virtually free** if you have users.
- User **behaviour** gives **implicit feedback**.

These methods have to handle:

- **Noise**: Users click for **unexpected reasons**.
- **Biases**: Interactions are affected by **position and selection bias**.

The Golden Triangle

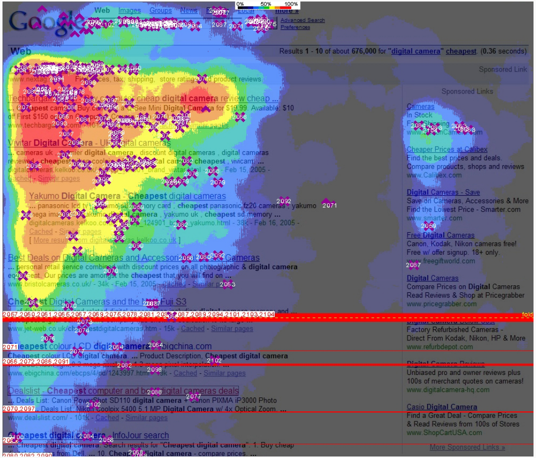


Image source: `http://www.mediative.com/`

Learning from User Interactions: Goal

Goal of unbiased learning to rank:

- Optimize a ranker w.r.t. **relevance preferences** of users from their interactions.
- **Avoid** being **biased by other factors** that influence interactions.

Online Evaluation

Online Evaluation: Interleaving

Often we need to answer the **question**:

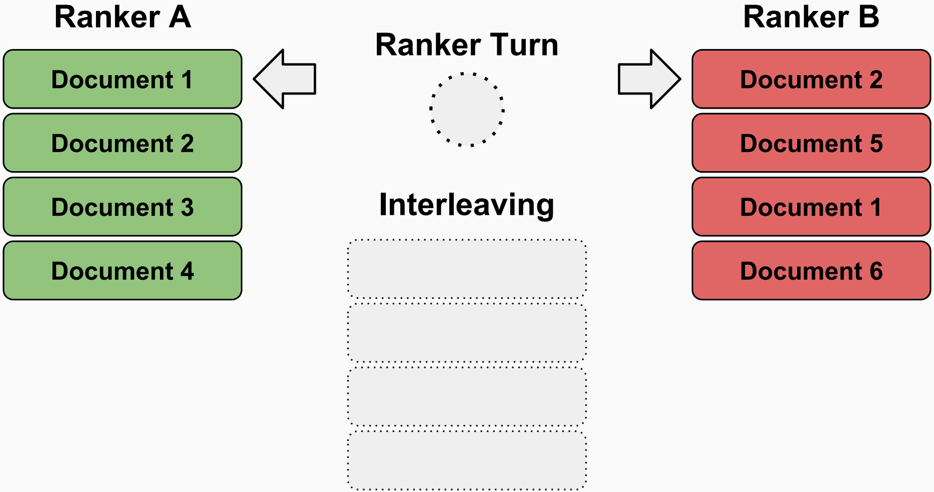
Is ranker A be preferred over ranker B?

Specific aspects of interactions in rankings can be used for **efficient comparisons**.

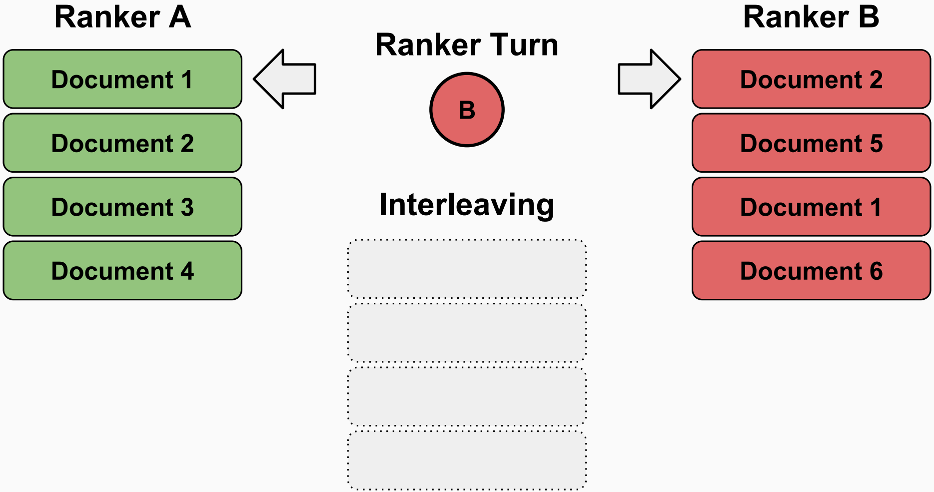
Interleaving (Joachims, 2003):

- Take the two rankings for a query from two rankers .
- Create an **interleaved ranking**, based on both rankings.
- **Clicks** on an interleaved ranking provide **preference signals** between rankers.

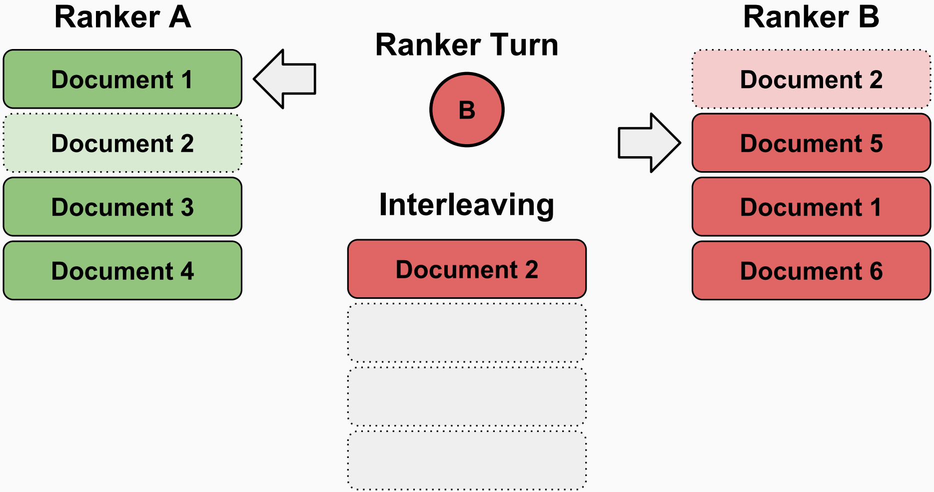
Online Evaluation: Team-Draft Interleaving



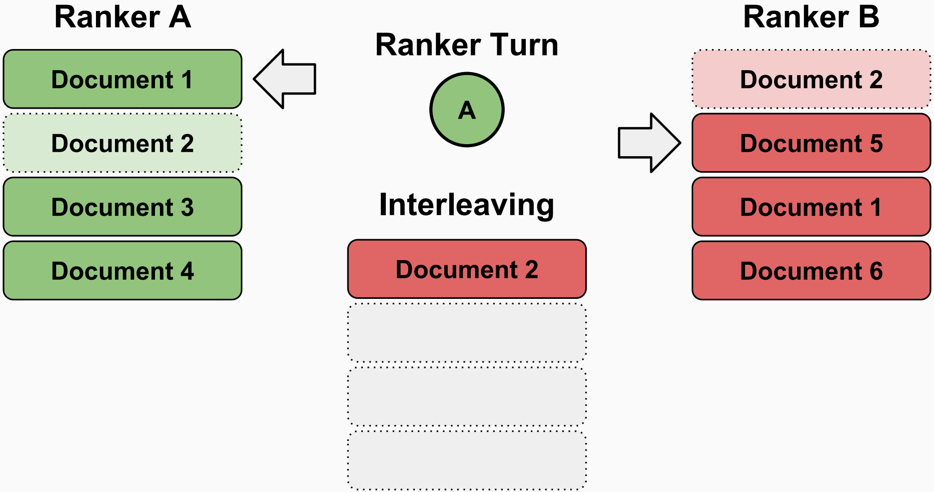
Online Evaluation: Team-Draft Interleaving



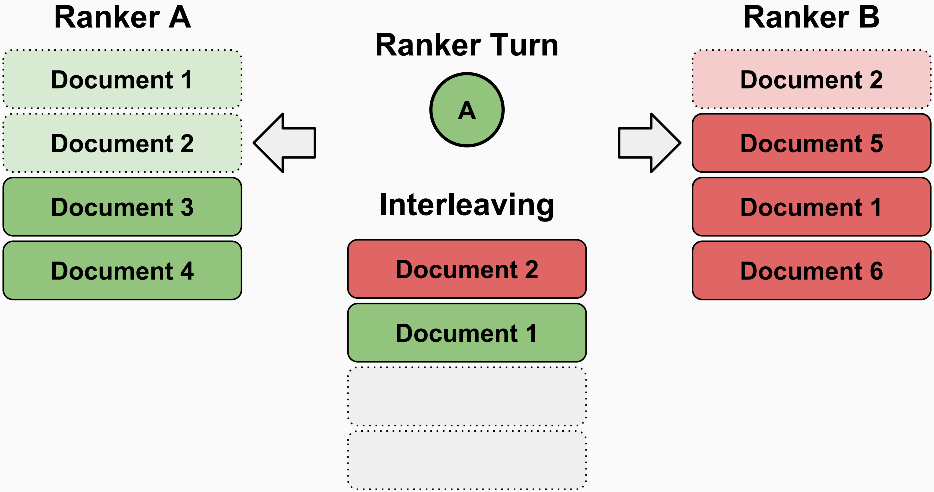
Online Evaluation: Team-Draft Interleaving



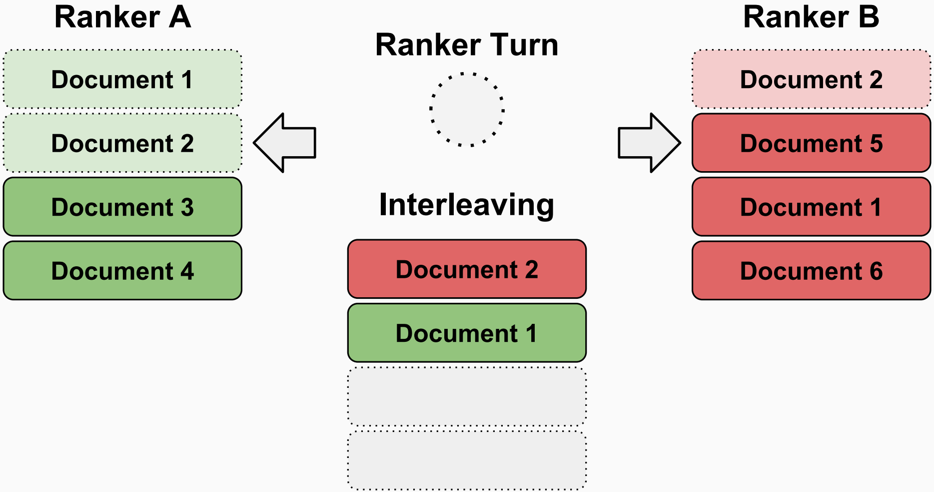
Online Evaluation: Team-Draft Interleaving



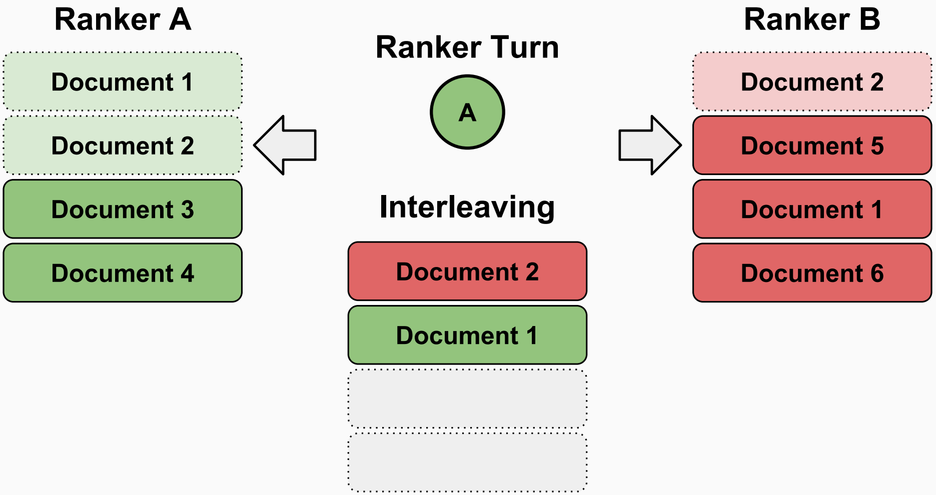
Online Evaluation: Team-Draft Interleaving



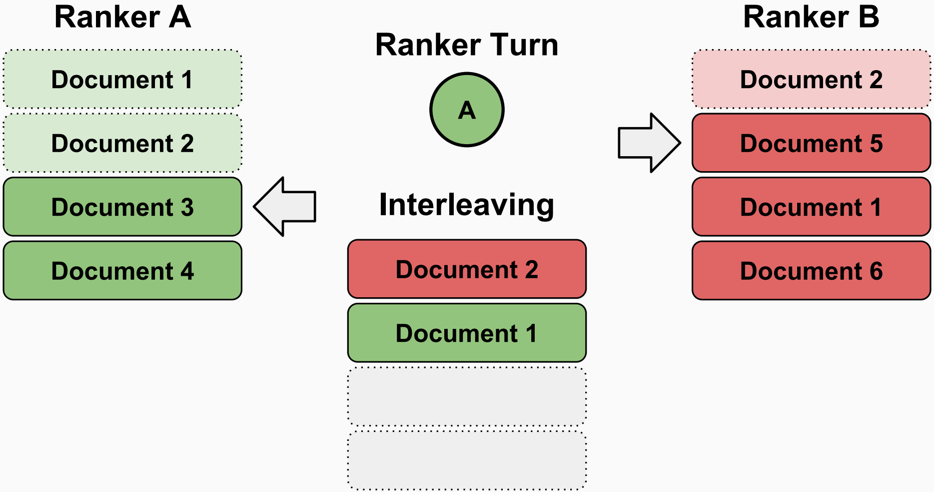
Online Evaluation: Team-Draft Interleaving



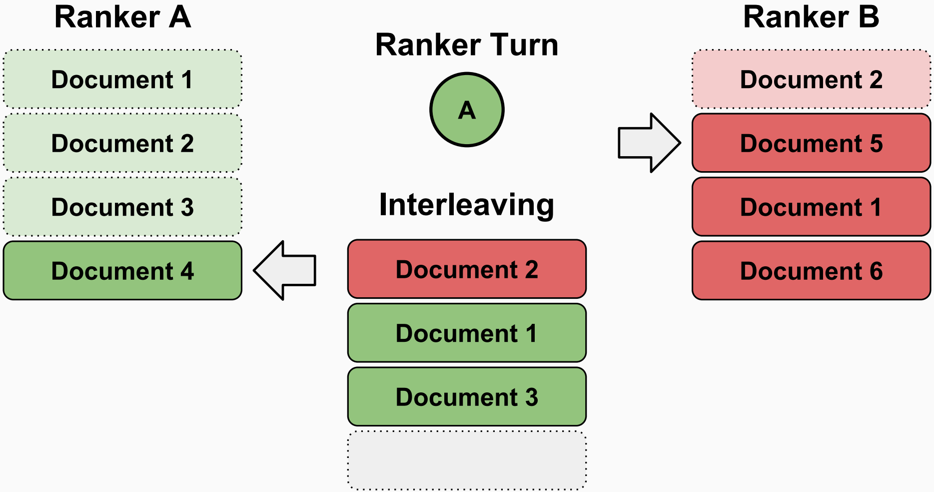
Online Evaluation: Team-Draft Interleaving



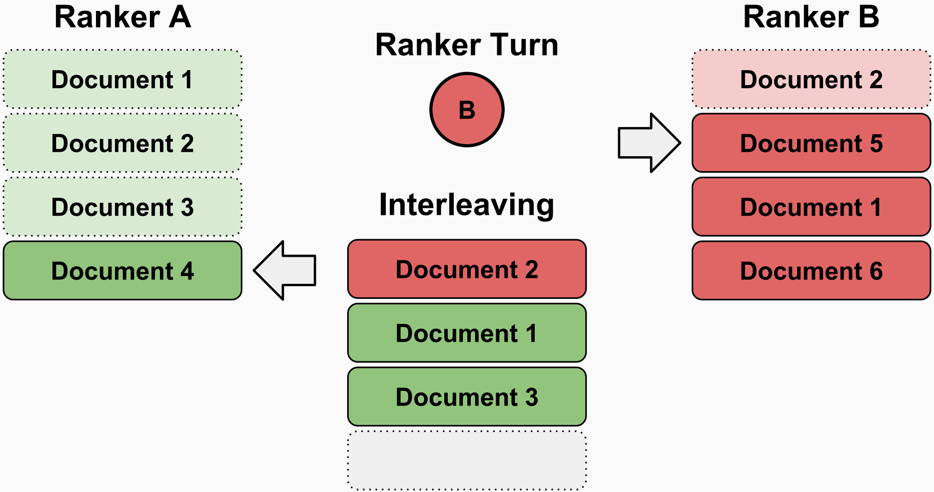
Online Evaluation: Team-Draft Interleaving



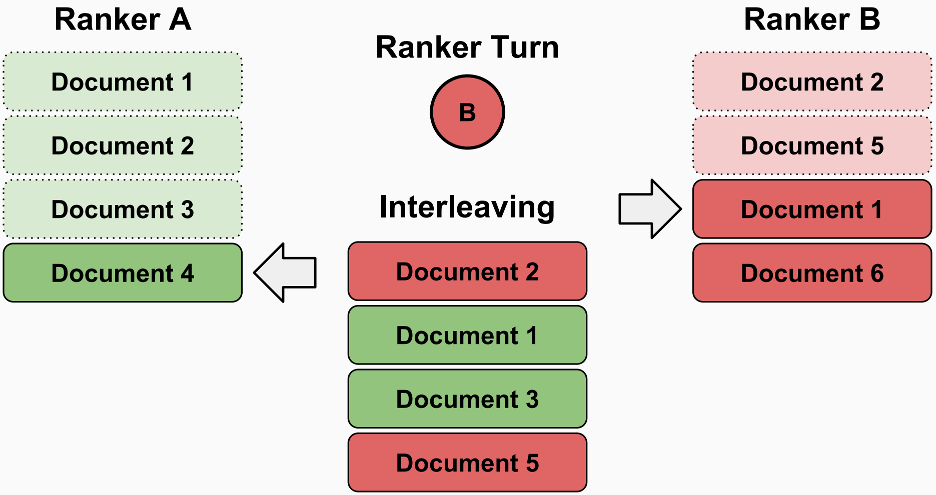
Online Evaluation: Team-Draft Interleaving



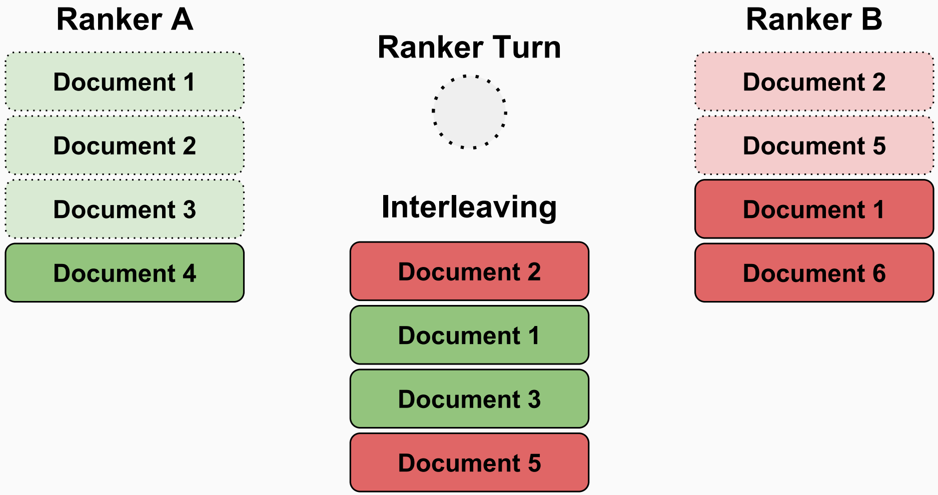
Online Evaluation: Team-Draft Interleaving



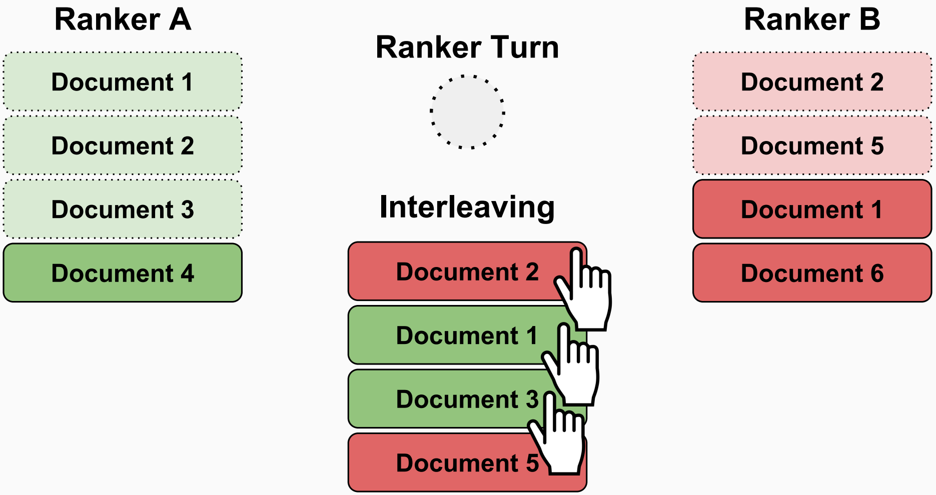
Online Evaluation: Team-Draft Interleaving



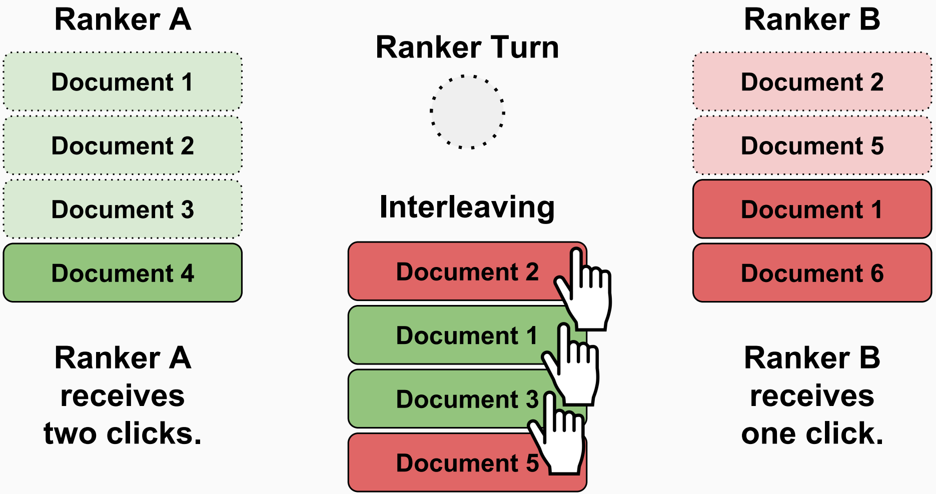
Online Evaluation: Team-Draft Interleaving



Online Evaluation: Team-Draft Interleaving



Online Evaluation: Team-Draft Interleaving



Online Evaluation: Interleaving

The idea behind interleaving:

- **Randomize display positions** of documents to deal with position bias.
- Limit randomization to **maintain user experience**.

Team-Draft Interleaving (Radlinski et al., 2008) is **affected by position bias**:

- Similar rankers can be inferred equal when a preference should be found.

Other interleaving methods are **proven** to be **unbiased**:

- **Probabilistic Interleaving** (Hofmann et al., 2011)
- **Optimized Interleaving** (Radlinski and Craswell, 2013)

Dueling Bandit Gradient Descent

Dueling Bandit Gradient Descent: Introduction

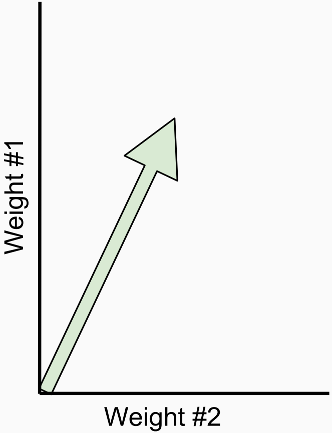
Introduced by Yue and Joachims (2009) as the **first online learning to rank** method.

Intuition:

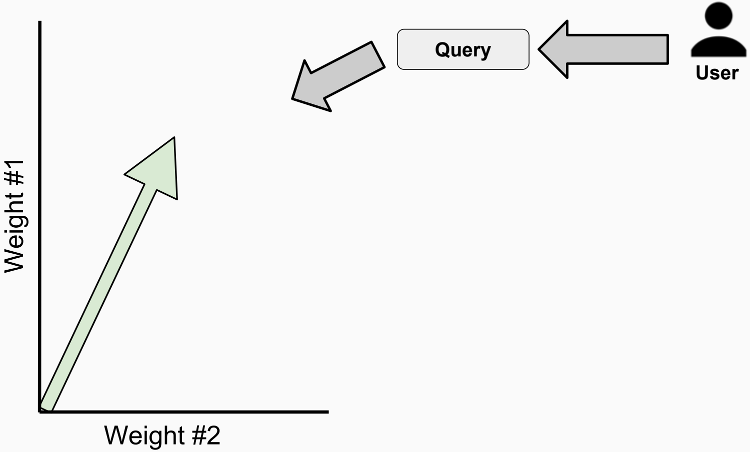
- if **online evaluation** can tell us if a **ranker is better** than another, then we can use it to **find an improvement** of our system.

By **sampling model variants** and **comparing** them with **interleaving**, the *gradient* of a model w.r.t. user satisfaction can be **estimated**.

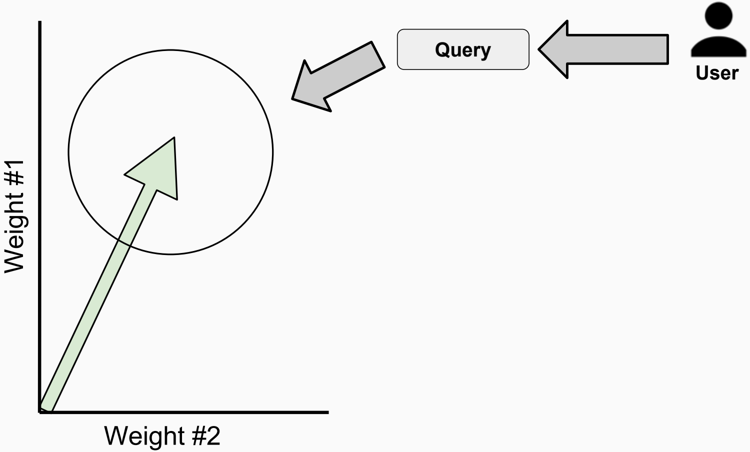
Dueling Bandit Gradient Descent: Visualization



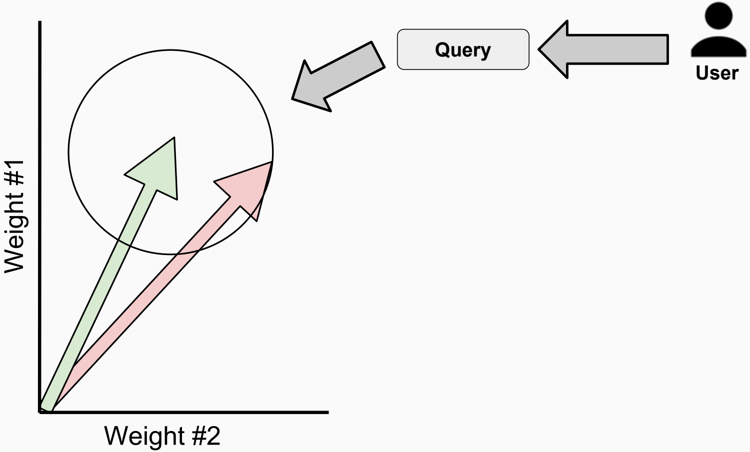
Dueling Bandit Gradient Descent: Visualization



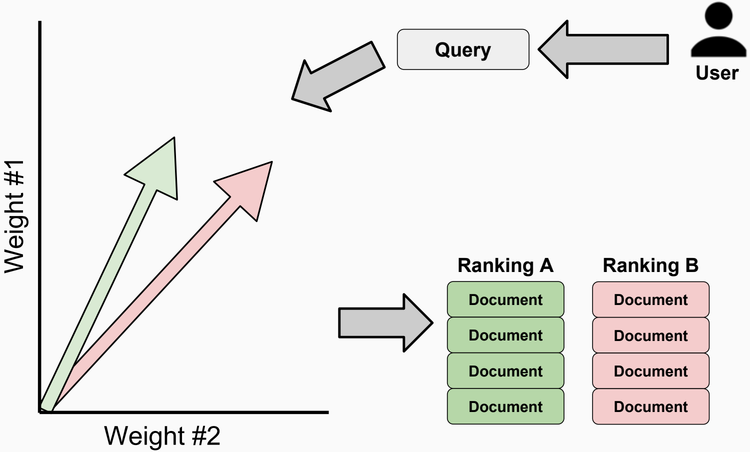
Dueling Bandit Gradient Descent: Visualization



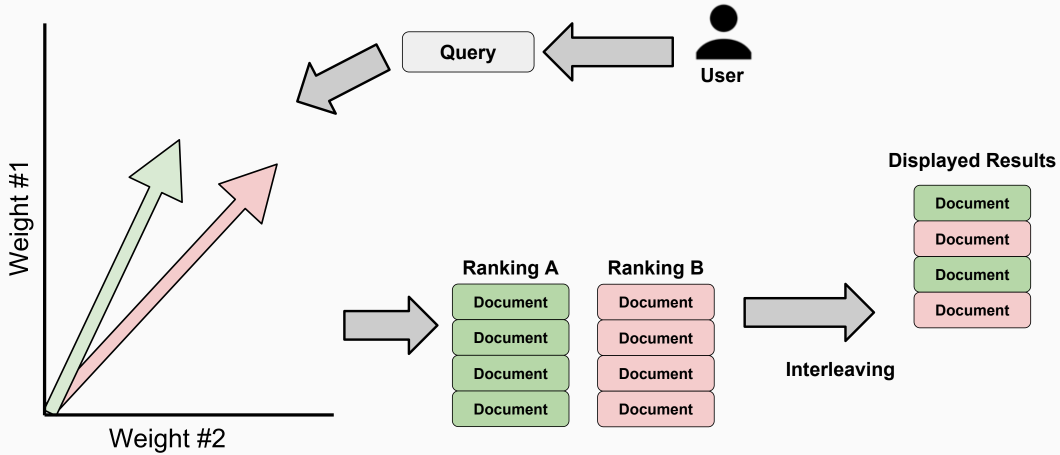
Dueling Bandit Gradient Descent: Visualization



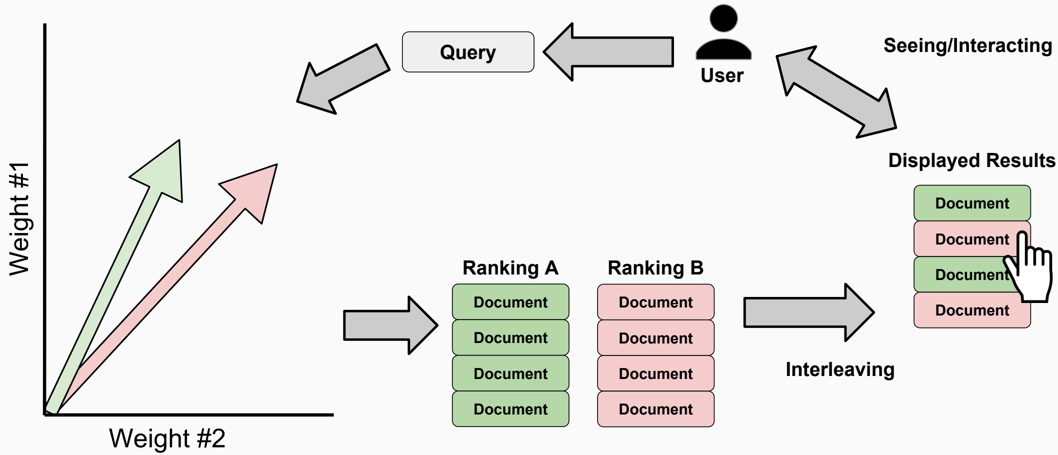
Dueling Bandit Gradient Descent: Visualization



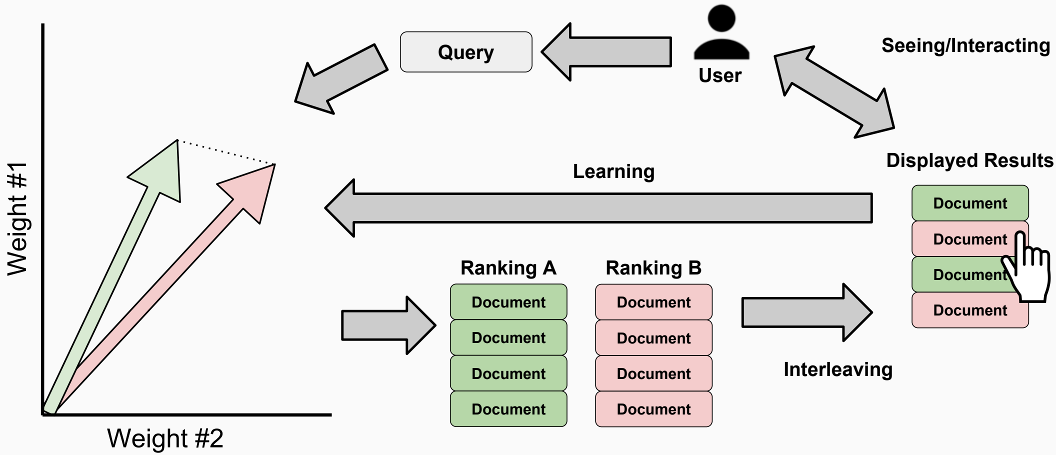
Dueling Bandit Gradient Descent: Visualization



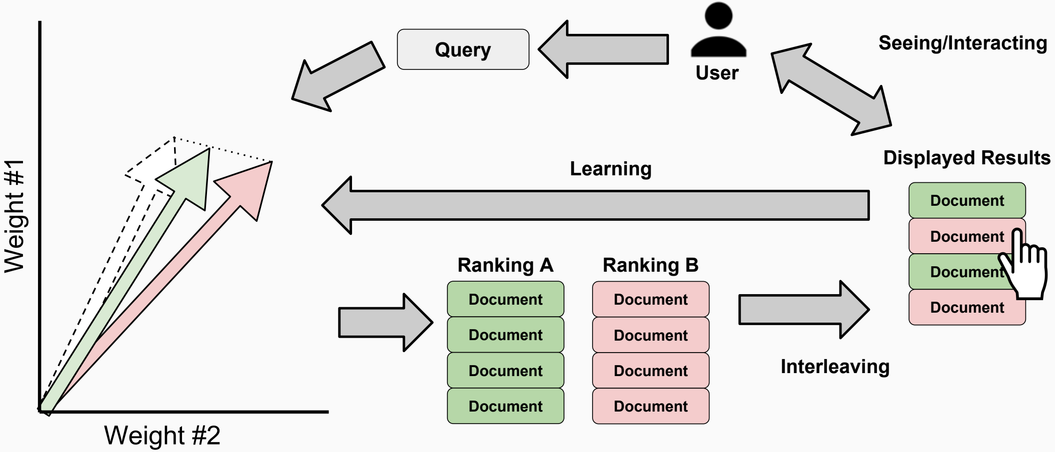
Dueling Bandit Gradient Descent: Visualization



Dueling Bandit Gradient Descent: Visualization



Dueling Bandit Gradient Descent: Visualization



Dueling Bandit Gradient Descent: Properties

Yue and Joachims (2009) prove that under the **assumptions**:

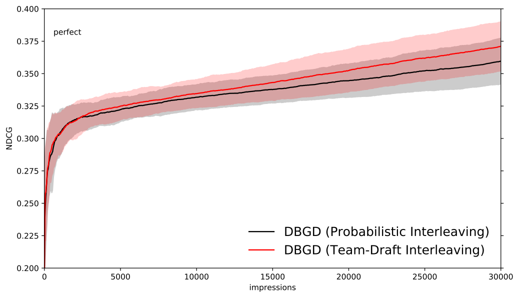
- There is a **single optimal** set of parameters: θ^*.
- The **utility space** w.r.t. θ is **convex** and **smooth**,
  i.e., small changes in θ lead to small changes in user experience.

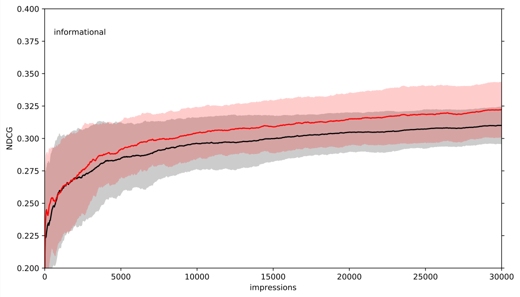
Then Dueling Bandit Gradient Descent is **proven** to have a **sublinear regret**:

- The algorithm will **eventually** approximate the ideal model.
- The duration of time is effected by the number of parameters of the model, the smoothness of the space, the unit chosen, etc.

Dueling Bandit Gradient Descent: Visualization

Simulations based on offline datasets: **user behavior** is based on the **annotations**.
As a result, we can **measure** how close the **model** is getting to their **satisfaction**.





Simulated results on the MSLR-WEB10k dataset,
a perfect user (left) and an informational user (right).

Image credits: (Oosterhuis, 2018).

Reusing Historical Interactions

Reusing Historical Interactions

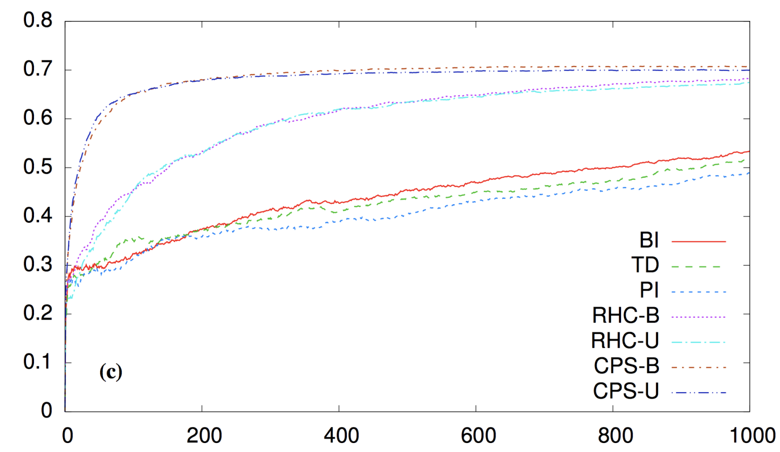
Hofmann et al. (2013) introduced the idea of **guiding exploration** by **reusing previous interactions**.

Intuition: if **previous interactions** showed that a **direction is unfruitful** then we should **avoid it in the future**.

Hofmann et al. (2013) introduced the **Candidate Pre-Selection** method:

- Sample a **large number** of rankers to create a **candidate set**.
- Use historical interactions to select the **most promising candidate** for DBGD.

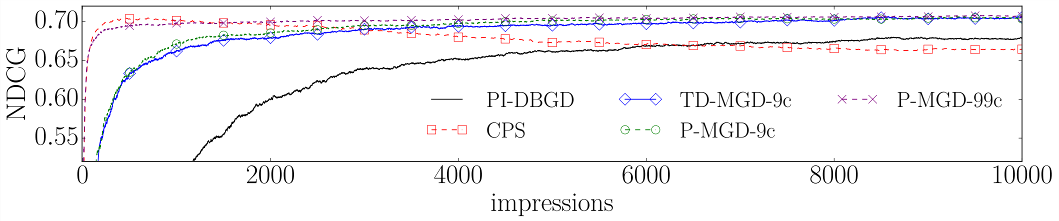
Reusing Historical Interactions: Performance



Simulated results on the NP2003 dataset.

Image credits: graph from (Hofmann et al., 2013).

Reusing Historical Interactions: Long Term Performance



Simulated results on the NP2003 dataset.

Remember, in the online setting the **performance cannot be measured**, thus **early-stopping is unfeasible**.

Image credits: graph from (Oosterhuis et al., 2016).

Reusing Historical Interactions: Other Work

Besides Hofmann et al. (2013) **other work** has also tried **reusing historical interactions** for online learning to rank: (Zhao and King, 2016; Wang et al., 2018).

The problem with these works is that:

- they **do not consider the long-term convergence**.
- they were **not evaluated** on the **largest available industry datasets**.

As a result, it is **still unclear** whether we can **reliably reuse historical interactions** during online learning.

Multileave Gradient Descent

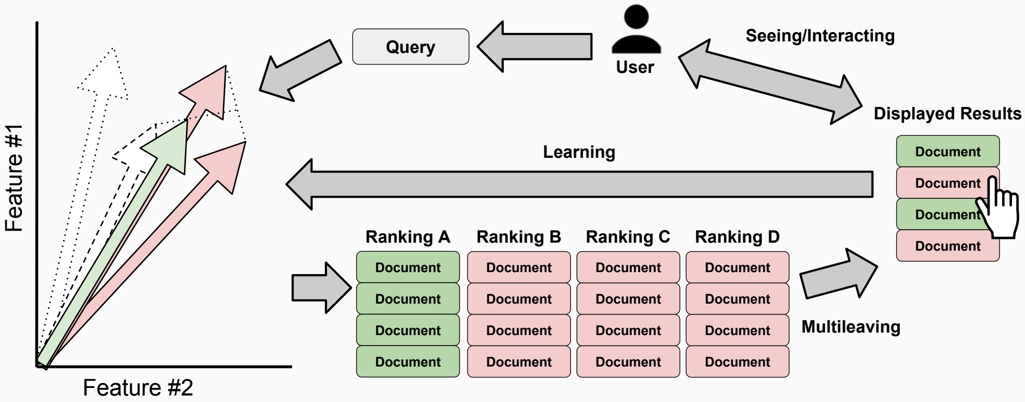
Multileave Gradient Descent

The introduction of **multileaving** in online evaluation allowed for **multiple rankers being compared simultaneously** from a single interaction.

A **natural extension** of Dueling Bandit Gradient Descent is to combine it with multileaving, resulting in **Multileave Gradient Descent** (Schuth et al., 2016).

Multileaving allows comparisons with **multiple candidate rankers**, **increasing** the **chance of finding an improvement**.

Multileave Gradient Descent: Visualization



Multileave Gradient Descent: Results

Results on the MSRL10k dataset under simulated users:

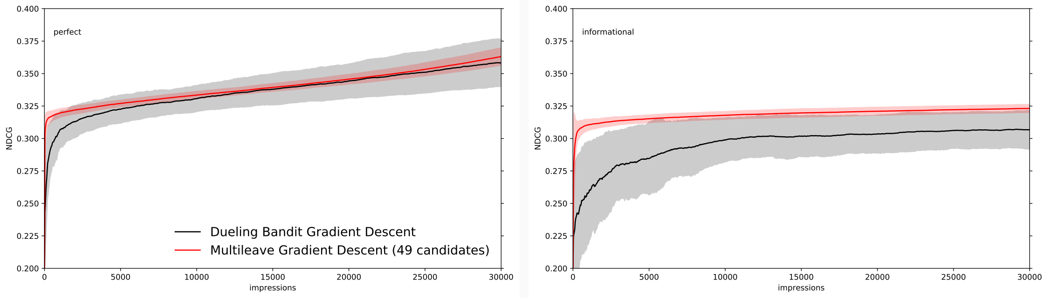


Image credits: (Oosterhuis, 2018).

Multileave Gradient Descent: Conclusion

Properties of Multileave Gradient Descent:

- **Vastly speeds up the learning rate** of Dueling Bandit Gradient Descent.
  - Much better user experience.
- Instead of **limiting (guiding) exploration**, it is done more **efficiently**.
- **Huge computational costs**, large number of rankers have to be applied.

Problems with Dueling Bandit Gradient Descent

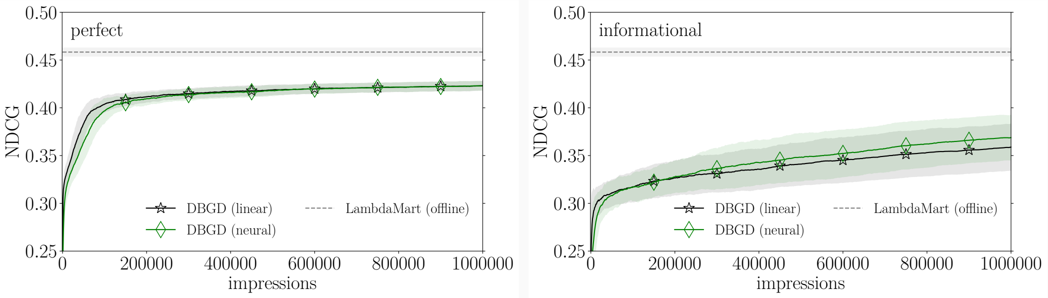
Problems with Dueling Bandit Gradient Descent

A **problem** with Dueling Bandit Gradient Descent and **all its extensions**:

- Their **performance at convergence** is **much worse** than offline approaches, even **under ideal user interactions**.

DBGD problems: Empirical

Results on the MSRL10k dataset under simulated users:



How is this possible, if it has **proven sub-linear regret**?

Image credits: (Oosterhuis, 2018).

Problems with the Dueling Bandit Gradient Descent Bounds

Remember the **regret** of Dueling Bandit Gradient Descent made **two assumptions**:

- There is a **single optimal model**: θ^*.
- The **utility space is smooth** w.r.t. to the model weights θ.

These **assumptions do not hold** for all models that are used in practice (Oosterhuis and de Rijke, 2019).

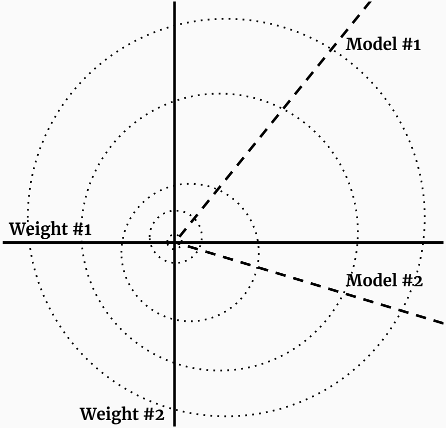
To prove this we use the fact that **the utility u is scale invariant** w.r.t. a ranking function $f_\theta(\cdot)$:

$$\forall\theta, \quad \forall\alpha \in \mathbb{R}_{>0}, \quad u(f_\theta(\cdot)) = u(\alpha f_\theta(\cdot)).$$

DBGD Assumptions: Smoothness Visualization

Intuition behind the **smoothness problem** for linear ranking models:

- **Every model** in a **line** from the origin in any direction is **equivalent**.
- **Any sphere** around the origin contains **every possible ranking model**[a].
- The **distance** between the *best* and the *worst* model becomes **infinitely small** near the origin.



[a] Except for the trivial random model on the origin.

DBGD Problems: Conclusion

Theoretical properties:

- Currently, no **sound regret bounds proven**.

Empirical observations:

- Methods do **not approach optimal performance**.
- Neural models have no advantage over linear models.

Possible solutions:

- Extend the algorithm (the last decade of research) or introduce new model.
- **Find an approach different to the bandit approach.**

Pairwise Differentiable Gradient Descent

Pairwise Differentiable Gradient Descent

We recently introduced **Pairwise Differentiable Gradient Descent** (Oosterhuis and de Rijke, 2018):

- Very different from previous Online Learning to Rank methods, that relied on sampling model variations similar to evolutionary approaches.

Intuition:

- A **pairwise** approach can be made **unbiased**, while being **differentiable**, without relying on online evaluation method or the sampling of models.

Plackett Luce Model

Pairwise Differentiable Gradient Descent optimizes a **Plackett Luce** ranking model, this models a **probabilistic distribution over documents**.

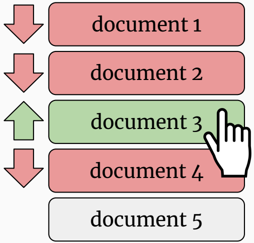
With the ranking scoring model $f_\theta(\mathbf{d})$ the distribution is:

$$P(d|D, \theta) = \frac{\exp^{f_\theta(\mathbf{d})}}{\sum_{d' \in D} \exp^{f_\theta(\mathbf{d}')}}.$$

Confidence is explicitly modelled and **exploration** depends on the **available documents**, thus it **naturally varies per query** and even within the ranking.

Bias in Pairwise Inference

Similar to existing pairwise methods (Oosterhuis and de Rijke, 2017; Joachims, 2002), Pairwise Differentiable Gradient Descent infers **pairwise document preferences from user clicks**:

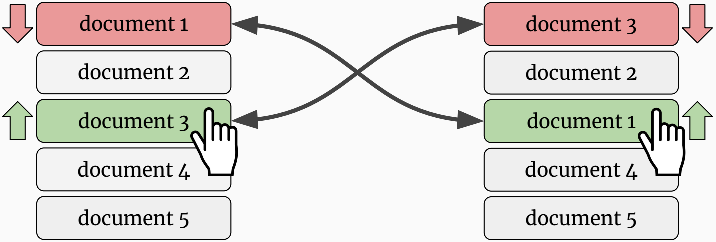


This approach is **biased**:

- Some preferences are **more likely to be inferred** due to **position/selection bias**.

Reversed Pair Rankings

Let $R^*(d_i, d_j, R)$ be R but with the **positions** of d_i and d_j **swapped**:



We assume:

- For a preference $d_i \succ d_j$ inferred from ranking R, if both are **equally relevant** the opposite preference $d_j \succ d_i$ is **equally likely** to be inferred from $R^*(d_i, d_j, R)$.

Then scoring **as if** R and R^* are **equally likely to occur** makes the gradient **unbiased**.

Unbiasing the Pairwise Update

The **ratio** between the probability of the ranking and the reversed pair ranking indicates the **bias between the two directions**:

$$\rho(d_i, d_j, R) = \frac{P(R^*(d_i, d_j, R)|f, D)}{P(R|f, D) + P(R^*(d_i, d_j, R)|f, D)}.$$

We use this ratio to **unbias the gradient estimation**:

$$\nabla f_\theta(\cdot) \approx \sum_{d_i >_{\mathbf{c}} d_j} \rho(d_i, d_j, R) \nabla P(d_i \succ d_j | D, \theta).$$

Unbiasedness of Pairwise Differentiable Gradient Descent

Under the reversed pair ranking assumption, we prove that **the expected estimated gradient** can be written as:

$$E[\nabla f_\theta(\cdot)] = \sum_{d_i, d_j} \alpha_{ij} (f'_\theta(\mathbf{d_i}) - f'_\theta(\mathbf{d_j})).$$

Where the weights α_{ij} will **match the user preferences** in expectation:

$$d_i =_{rel} d_j \Leftrightarrow \alpha_{ij} = 0,$$
$$d_i >_{rel} d_j \Leftrightarrow \alpha_{ij} > 0,$$
$$d_i <_{rel} d_j \Leftrightarrow \alpha_{ij} < 0.$$

Thus the estimated gradient is **unbiased w.r.t. document pair preferences**.

Pairwise Differentiable Gradient Descent: Method

Start with initial model θ_t, then indefinitely:

1. Wait for a user query.
2. **Sample** (without replacement) a **ranking** R from the document distribution:

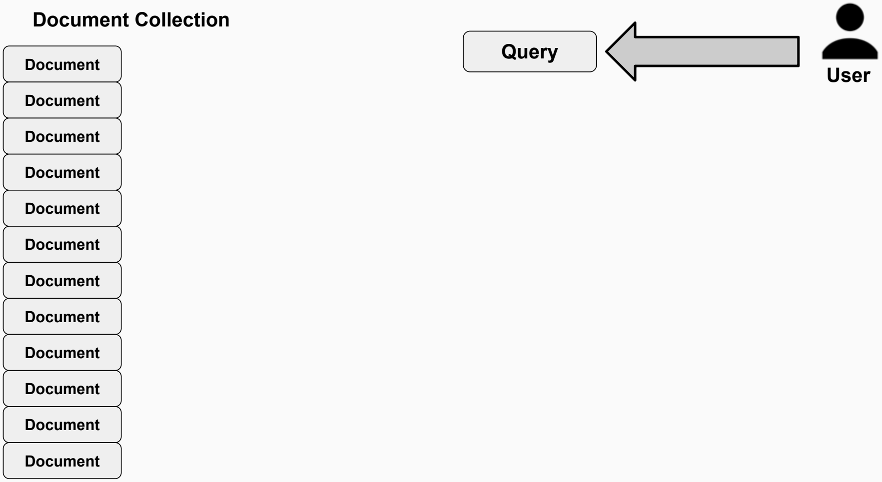
$$P(d|D,\theta_t) = \frac{\exp^{f_{\theta_t}(\mathbf{d})}}{\sum_{d' \in D} \exp^{f_{\theta_t}(\mathbf{d}')}}.$$

3. **Display** the ranking R to the user.
4. **Infer document preferences** from the **user clicks**: $\mathbf{c}$.
5. **Update** model according to the **estimated (unbiased) gradient**:

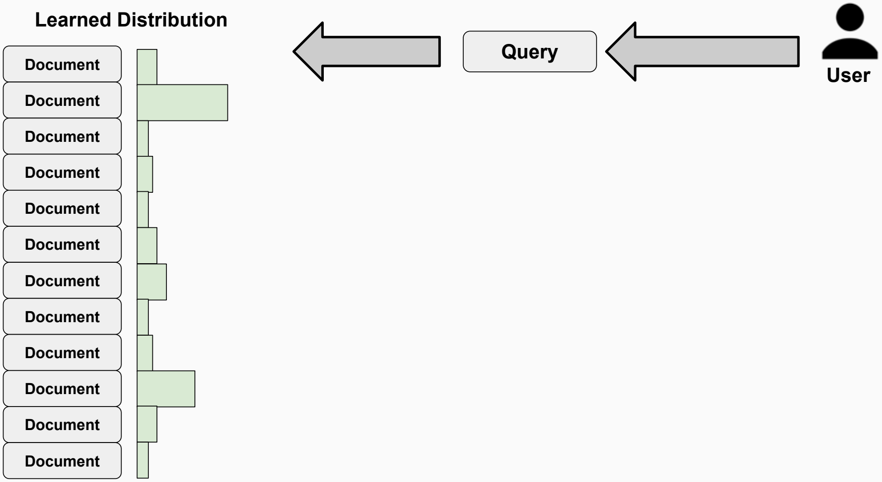
$$\nabla f_{\theta_t}(\cdot) \approx \sum_{d_i >_{\mathbf{c}} d_j} \rho(d_i, d_j, R) \nabla P(d_i \succ d_j | D, \theta_t).$$

Pairwise Differentiable Gradient Descent: Visualization

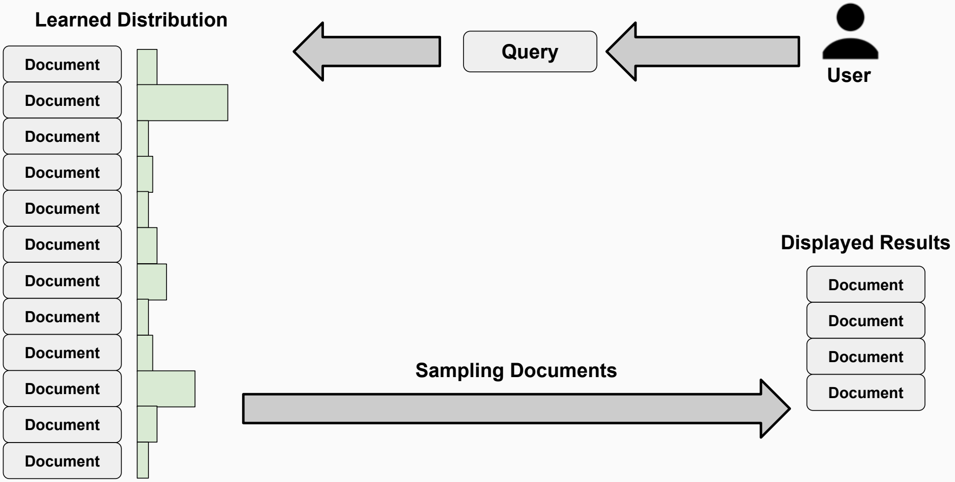
Pairwise Differentiable Gradient Descent: Visualization



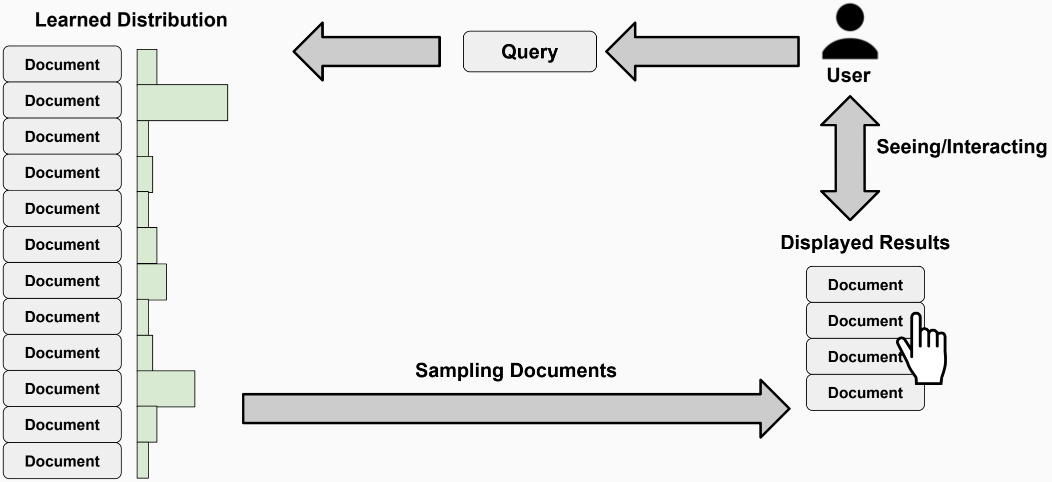
Pairwise Differentiable Gradient Descent: Visualization



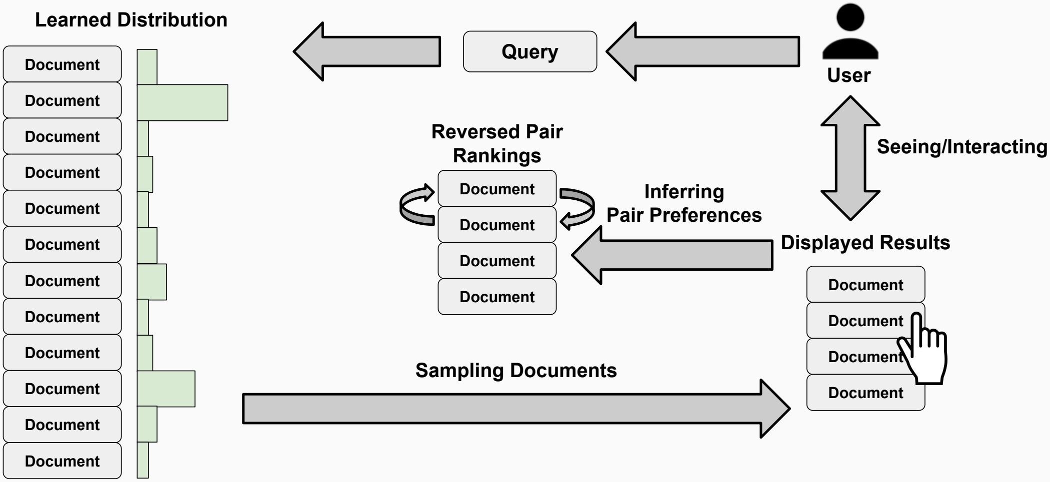
Pairwise Differentiable Gradient Descent: Visualization



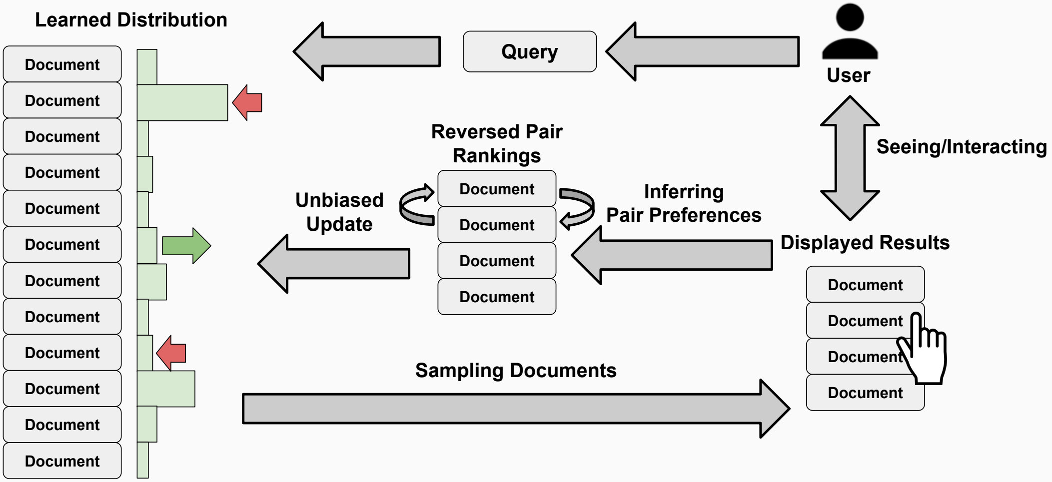
Pairwise Differentiable Gradient Descent: Visualization



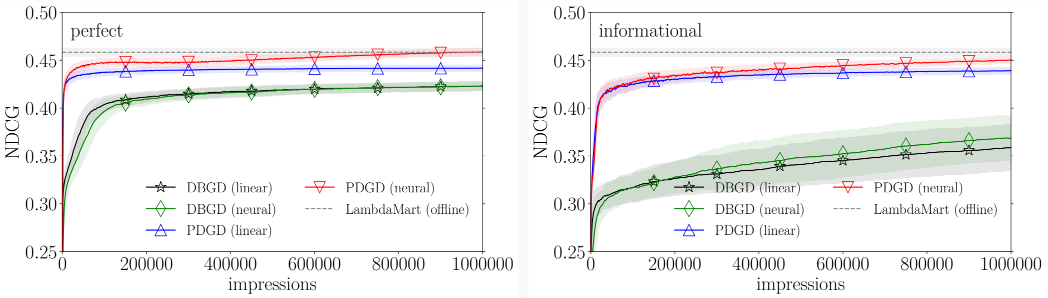
Pairwise Differentiable Gradient Descent: Visualization



Pairwise Differentiable Gradient Descent: Visualization



Pairwise Differentiable Gradient Descent: Results Long Term



Results of simulations on the MSLR-WEB10k dataset, a perfect user (left) and an informational user (right).

Image credits: (Oosterhuis and de Rijke, 2018).

Comparison of Online Methods

Empirical Comparison: Introduction

Recent most generalized comparison so far (Oosterhuis and de Rijke, 2019).

Simulations based on **largest available industry datasets**:

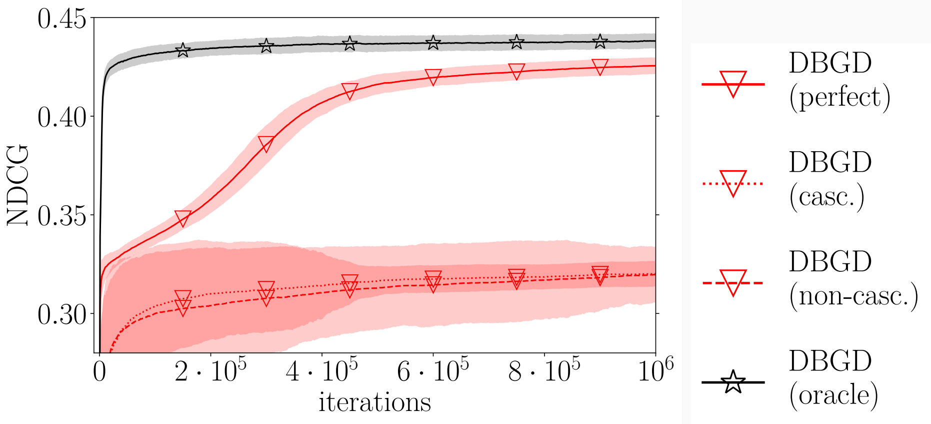
- MSLR-Web10k, Yahoo Webscope, Istella.

Simulated behavior ranging from:

- **ideal:** no noise, no position bias,
- **extremely difficult:** mostly noise, very high position bias.

Dueling Bandit Gradient Descent with an **oracle instead of interleaving**, to see the **maximum potential** of better interleaving methods.

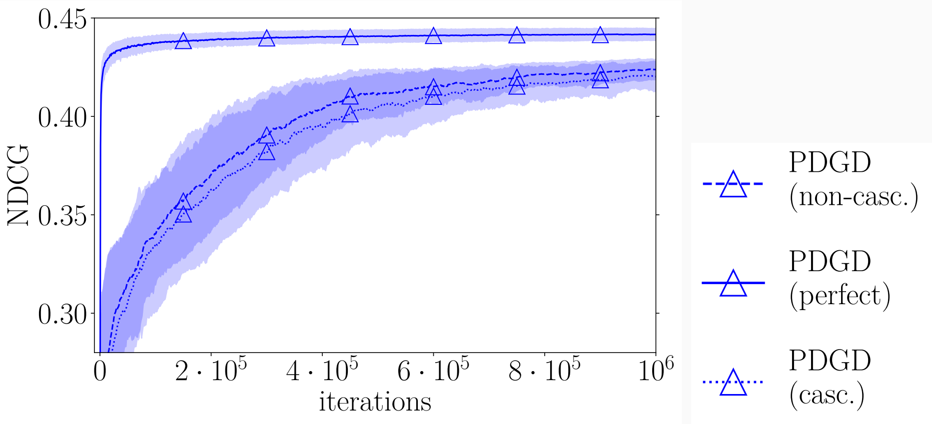
Empirical Comparison: DBGD



Results of simulations on the MSLR-WEB10k dataset.

Image credits: (Oosterhuis and de Rijke, 2019).

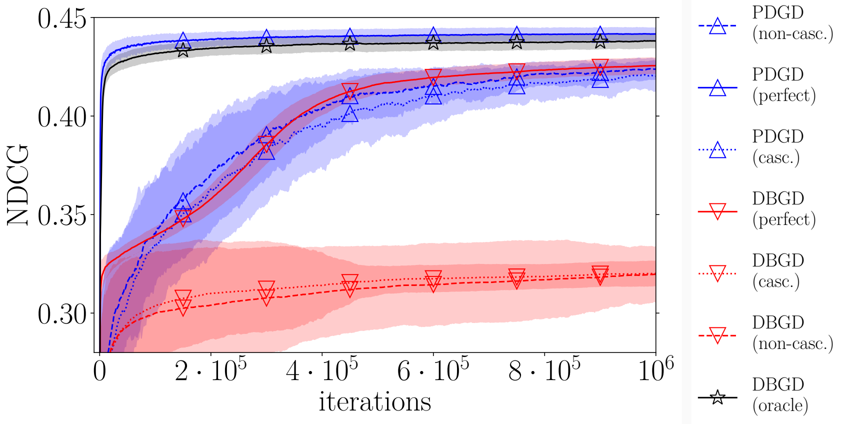
Empirical Comparison: PDGD



Results of simulations on the MSLR-WEB10k dataset.

Image credits: (Oosterhuis and de Rijke, 2019).

Empirical Comparison: All



Results of simulations on the MSLR-WEB10k dataset.

Image credits: (Oosterhuis and de Rijke, 2019).

Empirical Comparison: Conclusion

Dueling Bandit Gradient Descent (DBGD):

- **Unable** to reach **optimal performance** in **ideal** settings.
- Strongly affected by noise and position bias.

Pairwise Differentiable Gradient Descent (PDGD):

- **Capable** of reaching **optimal performance** in ideal settings.
- **Robust** to noise and position bias.
- Considerably **outperforms** DBGD in **all tested experimental settings**.

Theoretical Comparison

Dueling Bandit Based Approaches:

- Sublinear regret bounds proven,
  unsound for ranking problems as commonly applied.
- *Single update steps* are as **unbiased** as its **interleaving method**.

The Differentiable Pairwise Based Approach:

- **No regret bounds** proven.
- *Single update steps* are unbiased w.r.t. **pairwise document preferences**.

For the common ranking problem, neither approach has a theoretical advantage.

The Future for Online Learning to Rank

The **theory** for Online Learning to Rank is **inadequate** and needs **re-evaluation**.

The Dueling Bandit approach appears to be **lacking for optimizing ranking systems**.

Novel alternative approaches have high potential:

- Pairwise Differential Gradient Descent is a clear example.

What about Counterfactual Learning to Rank?

Empirical Comparison

Single empirical comparison so far (Jagerman et al., 2019).

Using the simulated setup common in unbiased learning to rank, we apply both **Inverse Propensity Scoring** and **Pairwise Differentiable Gradient Descent.**

Empirical Comparison: Conclusion

Counterfactual Learning to Rank:

- Slightly higher performance under:
  - no item-selection-bias,
  - little interaction noise.
- Very affected by high interaction noise.

Online Learning to Rank:

- More reliable performance across settings.
- Handles item-selection bias well.
- More robust to noise

Overall the empirical results suggest that **Online Learning to Rank** is more reliable.

Theoretical Comparison: Conclusion

Counterfactual Learning to Rank:

- Explicit position bias model.
- Proven to unbiasedly optimize ranking metrics.
- Can be interactive.
- Applicable to any historical interactions.

Online Learning to Rank:

- No explicit user model.
- Not proven to unbiasedly optimize ranking metrics.
- Only effective when interactive.
- Not applicable to all historical interactions.

In theory **Counterfactual Learning to Rank** has all the advantageous properties.

Conclusion

Conclusion

- **Online approaches** allow for **unbiased** and **responsive** learning to rank:
  - **Immediately adapt** to user behavior.
  - Perform **randomization** at each step, though limited.
- The Online Learning to Rank field seems to be in **trouble**:
  - **Theoretical guarantees** are **unsound** for the standard ranking problem.
  - Dueling Bandit Gradient Descent method is **unable to solve toy-problems**.
- **Comparison** with the **Counterfactual** approach:
  - **Empirically:** Online methods appear to be more reliable.
  - **Theoretically:** Counterfactual methods are much more advantageous.

Future of Online Learning to Rank

Time for Some Fortune Telling

I am now going to attempt to predict the future!

(Based on my personal expectations.)

The Future of Online Learning to Rank

If you are seeing this you missed the talk!

Questions and Answers

Thank you for listening!

Notation

Notation Used in the Slides i

Definition	Notation	Example
Query	q	–
Candidate documents	D	–
Document	$d \in D$	–
Ranking	R	$(R_1, R_2, \dots, R_n)$
Document at rank i	R_i	$R_i = d$
Relevance	$y : D \rightarrow \mathbb{N}$	$y(d) = 2$
Ranker model with weights θ	$f_\theta : D \rightarrow \mathbb{R}$	$f_\theta(d) = 0.75$
Click	$c_i \in \{0, 1\}$	–
Observation	$o_i \in \{0, 1\}$	–
Rank of d when f_θ ranks D	$\mathit{rank}(d \mid f_\theta, D)$	$\mathit{rank}(d \mid f_\theta, D) = 4$

Notation Used in the Slides ii

Differentiable upper bound on $\mathit{rank}(d, \mid f_\theta, D)$	$\overline{\mathit{rank}}(d, \mid f_\theta, D)$	–
Average Relevant Position metric	ARP	–
Discounted Cumulative Gain metric	DCG	–
Precision at k metric	$\mathit{Prec}@k$	–
A performance measure or estimator	Δ	–

Resources i

- Pairwise Differentiable Gradient Descent and Multileave Gradient Descent: `https://github.com/HarrieO/OnlineLearningToRank`
- Data and code for comparing counterfactual and online learning to rank `http://github.com/rjagerman/sigir2019-user-interactions`
- An older online learning to rank framework: Lerot `https://bitbucket.org/ilps/lerot/`

References i

K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 249–258. ACM, 2011.

K. Hofmann, A. Schuth, S. Whiteson, and M. de Rijke. Reusing historical interaction data for faster online learning to rank for ir. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 183–192. ACM, 2013.

R. Jagerman, H. Oosterhuis, and M. de Rijke. To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions. In *42nd International ACM SIGIR Conference on Research & Development in Information Retrieval*, page (to appear). ACM, 2019.

T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

T. Joachims. Evaluating retrieval performance using clickthrough data. In J. Franke, G. Nakhaeizadeh, and I. Renz, editors, *Text Mining*. Physica Verlag, 2003.

H. Oosterhuis. Learning to rank and evaluation in the online setting. 12th Russian Summer School in Information Retrieval (RuSSIR 2018), 2018.

H. Oosterhuis and M. de Rijke. Sensitive and scalable online evaluation with theoretical guarantees. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 77–86. ACM, 2017.

H. Oosterhuis and M. de Rijke. Differentiable unbiased online learning to rank. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1293–1302. ACM, 2018.

H. Oosterhuis and M. de Rijke. Optimizing ranking models in an online setting. In *Advances in Information Retrieval*, pages 382–396, Cham, 2019. Springer International Publishing.

H. Oosterhuis, A. Schuth, and M. de Rijke. Probabilistic multileave gradient descent. In *European Conference on Information Retrieval*, pages 661–668. Springer, 2016.

F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 245–254. ACM, 2013.

References iii

F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52. ACM, 2008.

A. Schuth, H. Oosterhuis, S. Whiteson, and M. de Rijke. Multileave gradient descent for fast online learning to rank. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 457–466. ACM, 2016.

H. Wang, R. Langley, S. Kim, E. McCord-Snook, and H. Wang. Efficient exploration of gradient space for online learning to rank. *arXiv preprint arXiv:1805.07317*, 2018.

Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

T. Zhao and I. King. Constructing reliable gradient exploration for online learning to rank. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1643–1652. ACM, 2016.

Acknowledgments

All content represents the opinion of the author(s), which is not necessarily shared or endorsed by their employers and/or sponsors.