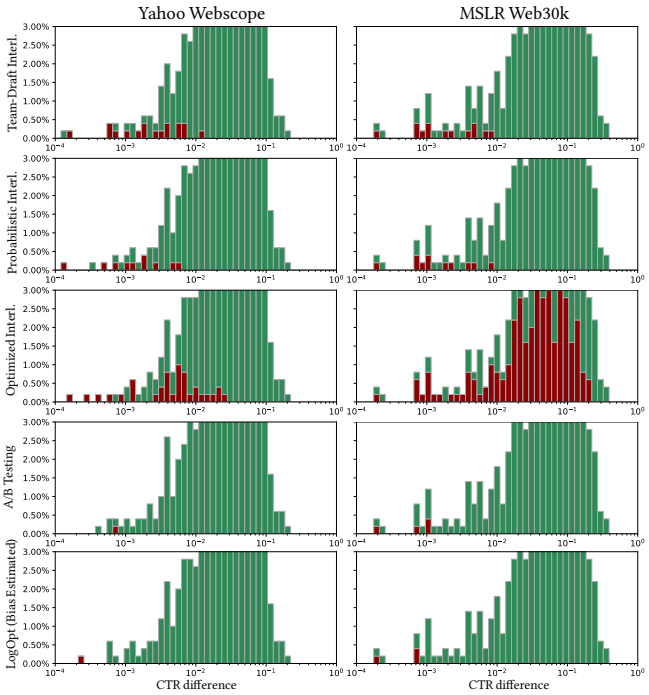

Figure 4: Distribution of errors over the CTR differences of the rankers in the comparison; red indicates a binary error; green indicates a correctly inferred binary preference; results are on estimates based on $3 \cdot 10^6$ sampled queries.

concerned that the claims of unbiasedness in previous interleaving work (see Section 3.2) give practitioners an unwarranted sense of reliability in interleaving.

7 CONCLUSION

In this paper, we have introduced the Logging-Policy Optimization Algorithm (LogOpt): the first method that optimizes a logging policy for minimal variance counterfactual evaluation. Counterfactual evaluation is proven to be unbiased w.r.t. position bias and item-selection bias under a wide range of logging policies. With the introduction of LogOpt, we now have an algorithm that can decide which rankings should be displayed for the fastest convergence. Therefore, we argue that LogOpt turns the existing counterfactual evaluation approach – which is indifferent to the logging policy – into an online approach – which instructs the logging policy.

Our experimental results show that LogOpt can lead to a better data-efficiency than A/B testing, without introducing the bias of interleaving. While our findings are mostly theoretical, they do suggest that future work should further investigate the bias in interleaving methods. Our results suggest that all interleaving methods make systematic errors, in particular when rankers with a similar CTR are compared. Furthermore, to the best of our knowledge, no empirical studies have been performed that could measure such a bias, our findings strongly show that such a study would be highly valuable to the field. Finally, LogOpt shows that in theory an evaluation method that is both unbiased and efficient is possible, if future work finds that these theoretical findings match empirical results with real users, this could be the start of a new line of theoretically-justified online evaluation methods.

Acknowledgements

We want to thank the anonymous reviewers for their feedback. This research was partially supported by the Netherlands Organisation for Scientific Research (NWO) under project nr 612.001.551 and by the Innovation Center for AI (ICAI). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Reproducibility

Our experimental implementation is publicly available at https://github.com/HarrieO/2020ictir-evaluation.

REFERENCES

[1] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. 2019. Estimating Position Bias without Intrusive Interventions. In *WSDM*. ACM, 474–482.
[2] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *SIGIR*. ACM, 385–394.
[3] Olivier Chapelle and Yi Chang. 2011. Yahoo! Learning to Rank Challenge Overview. *Journal of Machine Learning Research* 14 (2011), 1–24.
[4] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-scale Validation and Analysis of Interleaved Search Evaluation. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 1–41.
[5] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.
[6] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *WSDM*. 87–94.
[7] Zhichong Fang, Aman Agarwal, and Thorsten Joachims. 2019. Intervention Harvesting for Context-dependent Examination-bias Estimation. In *SIGIR*. 825–834.
[8] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Foundations and Trends in Information Retrieval* 10, 1 (2016), 1–117.
[9] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. 2011. A Probabilistic Method for Inferring Preferences from Clicks. In *CIKM*. ACM, 249–258.
[10] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. 2013. Fidelity, Soundness, and Efficiency of Interleaved Comparison Methods. *ACM Transactions on Information Systems (TOIS)* 31, 4 (2013), 1–43.
[11] Thorsten Joachims. 2003. Evaluating Retrieval Performance Using Clickthrough Data. In *Text Mining*, J. Franke, G. Nakhaeizadeh, and I. Renz (Eds.). Physica Verlag.
[12] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *SIGIR*. ACM, 154–161.
[13] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *WSDM*. ACM, 781–789.
[14] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of Machine Learning and Data Mining* 7, 8 (2017), 922–929.
[15] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H Chi. 2020. Off-policy Learning in Two-stage Recommender Systems. In *The Web Conference 2020*. ACM, 463–473.
[16] Harrie Oosterhuis and Maarten de Rijke. 2020. Policy-Aware Unbiased Learning to Rank for Top-k Rankings. In *SIGIR*. ACM.
[17] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-rank Systems. In *WWW*. 1863–1873.
[18] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597* (2013).
[19] Filip Radlinski and Nick Craswell. 2013. Optimized Interleaving for Online Retrieval Evaluation. In *WSDM*. ACM, 245–254.
[20] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How Does Clickthrough Data Reflect Retrieval Quality?. In *CIKM*. ACM, 43–52.
[21] Anne Schuth, Katja Hofmann, and Filip Radlinski. 2015. Predicting Search Satisfaction Metrics with Interleaved Comparisons. In *SIGIR*. 463–472.
[22] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to Rank with Selection Bias in Personal Search. In *SIGIR*. ACM, 115–124.
[23] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *WSDM*. ACM, 610–618.
[24] Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2018. The LambdaLoss Framework for Ranking Metric Optimization. In *CIKM*. ACM, 1313–1322.