Counterfactual Evaluation and Learning for Interactive Systems: Foundations, Implementations, and Recent Advances

Yuta Saito
Cornell University
Ithaca, NY, USA
ys552@cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY, USA
tj@cs.cornell.edu

ABSTRACT

Counterfactual estimators enable the use of existing log data to estimate how some new target policy would have performed, if it had been used instead of the policy that logged the data. We say that those estimators work "off-policy", since the policy that logged the data is different from the target policy. In this way, counterfactual estimators enable *Off-policy Evaluation* (OPE) akin to an unbiased offline A/B test, as well as learning new decision-making policies through *Off-policy Learning* (OPL). The goal of this tutorial is to summarize *Foundations, Implementations, and Recent Advances* of OPE and OPL (OPE/OPL), with applications in recommendation, search, and an ever growing range of interactive systems. Specifically, we will introduce the fundamentals of OPE/OPL and provide theoretical and empirical comparisons of conventional methods. Then, we will cover emerging practical challenges such as how to handle large action spaces, distributional shift, and hyper-parameter tuning. We will then present *Open Bandit Pipeline*, an open-source Python software for OPE/OPL to better enable new research and applications. We will conclude the tutorial with future directions.

CCS CONCEPTS

• **Computing methodologies → Batch learning**; • **Theory of computation → Sequential decision making**.

KEYWORDS

counterfactual inference, off-policy evaluation and learning, contextual bandits, reinforcement learning, recommender systems

ACM Reference Format:
Yuta Saito and Thorsten Joachims. 2022. Counterfactual Evaluation and Learning for Interactive Systems: Foundations, Implementations, and Recent Advances. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542601

1 TUTORIAL OUTLINE

This tutorial consists of the following contents (total 3 hours).

(1) **Introduction to OPE/OPL (Thorstem Joachims; 30min)**: We will introduce conventional formulation of OPE and how it helps improve interactive systems quickly and safely. We also introduce basic estimators in OPE including Direct Method (DM) and Inverse Propensity Score (IPS) weighting with some empirical illustrations to highlight their bias-variance trade-off.

(2) **Bias-Variance Control (Yuta Saito; 40min)**
This section summarizes a wide range of existing estimators in OPE including Self-Normalized IPS [17], Doubly Robust [2], Switch [20], and Doubly Robust with Shrinkage [14]. These estimators aim at achieving a better bais-variance trade-off compared to DM and IPS. We will provide comprehensive comparisons of these estimators from both theoretical and empirical perspectives.

(3) **Recent Advances (Yuta Saito; 40min)**
This section will cover recent related methods to handle emerging practical challenges such as OPE of ranking policies [4, 5, 7, 9], large-scale applications [6, 12], deficient support [8], multiple loggers [1, 3], and hyper-parameter tuning for OPE [13–15, 19]. These challenges are closely related to real-world applications such as recommender and retrieval systems where the estimators have to deal with many number of actions and non-stationary dynamics.

(4) **Off-Policy Learning (Thorsten Joachims; 40min)**
This section will cover the fundamental methods for OPL [16, 17] where we aim at training a new decision-making policy using only the logged bandit data.

(5) **Implementations (Yuta Saito; 20min)**
This section will introduce *Open Bandit Pipeline*[1] [10], an open-source Python package for OPE/OPL, and demonstrate how it helps us implement OPE/OPL for both research and practical purposes.

(6) **Conclusions and QAs (Both Presenters; 10min)**
This section will conclude the tutorial by summarizing the previous sections and presenting remaining research challenges of the area. There will also be a live QA session.

The learning outcomes are to enable the participants:

(1) to know fundamental concepts and methods of OPE/OPL
(2) to be familiar with recent advances to address practical challenges such as large action spaces and parameter tuning
(3) to understand how to implement OPE/OPL
(4) to be aware of remaining challenges and opportunities in the relevant field

Note that all materials, including slides and demo code, will be available during and after the tutorial on our tutorial website.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542601

[1] https://github.com/st-tech/zr-obp

2 TARGETED AUDIENCE

This tutorial is aimed at an audience with intermediate experience in machine learning, data mining, or recommender systems who are interested in using OPE/OPL methods in their research and applications. Participants are expected to have basic knowledge of machine learning, probability theory, and statistics.

3 RELATED TUTORIALS

We presented a similar tutorial at RecSys 2021 in Amsterdam, Netherlands titled "Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances" [11].[2]. Our new tutorial is based on this previous version, but additionally highlights the emerging topic of OPE/OPL for large-scale applications.

We also want to highlight a related tutorial presented remotely at KDD 2021 titled "Causal Inference and Machine Learning in Practice with EconML and CausalML" [18][3], which focuses on recent advances and real-world use cases of treatment effect prediction. The technical aspect of this tutorial is closely related to ours, but our focus is rather on evaluating and training decision-making policies using only logged bandit data, which is a goal substantially different from predicting the heterogeneous causal effect of often binary treatments.

4 PRESENTER BIO

Yuta Saito *(ys552@cornell.edu).* is a Ph.D. student in the Department of Computer Science at Cornell University, advised by Prof. Thorsten Joachims. His current research focuses on OPE of bandit algorithms and fairness in ranking. Some of his recent work has been published at top-tier conferences, including ICML, NeurIPS, KDD, RecSys, and WSDM. He has also co-lectured a tutorial related to counterfactual inference at RecSys 2021.

Thorsten Joachims *(tj@cs.cornell.edu).* is a Professor in the Department of Computer Science and in the Department of Information Science at Cornell University, and he is an Amazon Scholar. His research interests center on the synthesis of theory and system building in machine learning, with applications in information retrieval and recommendation. His past research focused on support vector machines, learning to rank, learning with preferences, and learning from implicit feedback, text classification, and structured output prediction. Working with his students and collaborators, his papers won 10 Best Paper Awards and 4 Test-of-Time Awards. He is also an ACM Fellow, AAAI Fellow, KDD Innovations Award recipient, and member of the SIGIR Academy.

ACKNOWLEDGMENTS

This research was supported in part by NSF Awards IIS-1901168 and IIS-2008139. Yuta Saito was supported by Funai Overseas Scholarship. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

[2] https://sites.google.com/cornell.edu/recsys2021tutorial
[3] https://causal-machine-learning.github.io/kdd2021-tutorial/

REFERENCES

[1] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. 2017. Effective Evaluation using Logged Bandit Feedback from Multiple Loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 687–696.

[2] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. 2014. Doubly Robust Policy Evaluation and Optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[3] Nathan Kallus, Yuta Saito, and Masatoshi Uehara. 2021. Optimal Off-Policy Evaluation from Multiple Logging Policies. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, 5247–5256.

[4] Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. 2022. Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 487–497.

[5] Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, S Muthukrishnan, Vishwa Vinay, and Zheng Wen. 2018. Offline Evaluation of Ranking Policies with Click Models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 1685–1694.

[6] 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 *Proceedings of The Web Conference 2020*. 463–473.

[7] James McInerney, Brian Brost, Praveen Chandar, Rishabh Mehrotra, and Benjamin Carterette. 2020. Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 1779–1788.

[8] Noveen Sachdeva, Yi Su, and Thorsten Joachims. 2020. Off-policy Bandits with Deficient Support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 965–975.

[9] Yuta Saito. 2020. Doubly Robust Estimator for Ranking Metrics with Post-Click Conversions. In *Fourteenth ACM Conference on Recommender Systems*. Association for Computing Machinery, 92–100.

[10] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2020. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. *arXiv preprint arXiv:2008.07146* (2020).

[11] Yuta Saito and Thorsten Joachims. 2021. Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances. In *Fifteenth ACM Conference on Recommender Systems*. 828–830.

[12] Yuta Saito and Thorsten Joachims. 2022. Off-Policy Evaluation for Large Action Spaces via Embeddings. *arXiv preprint arXiv:2202.06317* (2022).

[13] Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. 2021. Evaluating the Robustness of Off-Policy Evaluation. In *Fifteenth ACM Conference on Recommender Systems*. 114–123.

[14] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. 2020. Doubly Robust Off-Policy Evaluation with Shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, 9167–9176.

[15] Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. 2020. Adaptive Estimator Selection for Off-Policy Evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, 9196–9205.

[16] Adith Swaminathan and Thorsten Joachims. 2015. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *The Journal of Machine Learning Research* 16, 1 (2015), 1731–1755.

[17] Adith Swaminathan and Thorsten Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, Vol. 28. 3231–3239.

[18] Vasilis Syrgkanis, Greg Lewis, Miruna Oprescu, Maggie Hei, Keith Battocchi, Eleanor Dillon, Jing Pan, Yifeng Wu, Paul Lo, Huigang Chen, et al. 2021. Causal Inference and Machine Learning in Practice with Econml and Causalml: Industrial Use Cases at Microsoft, TripAdvisor, Uber. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4072–4073.

[19] George Tucker and Jonathan Lee. 2021. Improved Estimator Selection for Off-Policy Evaluation. *Workshop on Reinforcement Learning Theory at the 38th International Conference on Machine Learning* (2021).

[20] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. 2017. Optimal and Adaptive Off-Policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, 3589–3597.