A Large-scale Open Dataset for Bandit Algorithms

Yuta Saito saito.y.bj@m.titech.ac.jp
Tokyo Institute of Technology

Shunsuke Aihara shunsuke.aihara@zozo.com
ZOZO Technologies, Inc.

Megumi Matsutani megumi.matsutani@zozo.com
ZOZO Technologies, Inc.

Yusuke Narita yusuke.narita@yale.edu
Yale University

Abstract

We build and publicize the *Open Bandit Dataset and Pipeline* to facilitate scalable and reproducible research on bandit algorithms. They are especially suitable for *off-policy evaluation* (OPE), which attempts to predict the performance of hypothetical algorithms using data generated by a different algorithm in use. We construct the dataset based on experiments and implementations on a large-scale fashion e-commerce platform, ZOZOTOWN. The data contain the ground-truth about the performance of several bandit policies and enable the fair comparisons of different OPE estimators. We also provide a pipeline to make its implementation easy and consistent. As a proof of concept, we use the dataset and pipeline to evaluate and compare the performance of OPE estimators. Our pipeline and sample data are available at `https://github.com/st-tech/zr-obp`.

1. Introduction

Interactive bandit and reinforcement learning systems produce log data valuable for evaluating and redesigning the systems. For example, the logs of a news recommendation system record which news article was presented and whether the user read it, giving the system designer a chance to make its recommendation more relevant. Exploiting log data is, however, more difficult than conventional supervised machine learning: the result is only observed for the action chosen by the system but not for all the other actions the system could have taken. The logs are also biased in that the logs over-represent the actions favored by the system.

A potential solution to this problem is an A/B test that compares the performance of counterfactual systems in an online environment. However, A/B testing counterfactual systems is often difficult, since deploying a new policy is time- and money-consuming, and entails a risk of failure.

This leads us to the problem of *off-policy evaluation* (OPE), which aims to estimate the performance of a counterfactual policy using only log data collected by a past (or behavior) policy. Such an evaluation allows us to compare the performance of candidate counterfactual policies to decide which policy should be deployed. This alternative approach thus solves the above problem with the A/B test approach. Applications range from contextual bandits (Bottou et al., 2013; Li et al., 2010, 2011; Narita et al., 2019; Strehl et al., 2010; Swaminathan

and Joachims, 2015a,b) and reinforcement learning in the web industry (Farajtabar et al., 2018; Irpan et al., 2019; Jiang and Li, 2016; Kallus and Uehara, 2019; Liu et al., 2018; Narita et al., 2020; Thomas and Brunskill, 2016) to other social domains such as healthcare (Murphy et al., 2001) and education (Mandel et al., 2014).

Issues with current experimental procedures. While the research community has produced theoretical breakthroughs, the experimental evaluation of OPE remains primitive. Specifically, it lacks a public benchmark dataset for comparing the performance of different methods. Researchers often validate their methods using synthetic simulation environments (Kallus and Uehara, 2019; Kato et al., 2020; Liu et al., 2018; Voloshin et al., 2019; Xie et al., 2019). A version of the synthetic approach is to modify multi-class classification datasets and treat supervised machine learning methods as bandit policies to evaluate off-policy estimators (Dudík et al., 2014; Farajtabar et al., 2018; Vlassis et al., 2019; Wang et al., 2017). An obvious problem with these studies is that there is no guarantee that their simulation environment is similar to real-world settings. To solve this issue, (Gilotte et al., 2018; Gruson et al., 2019; Narita et al., 2019, 2020) use proprietary real-world datasets. Since these datasets are not public, however, it remains challenging to reproduce the results, and compare their methods with new ideas in a fair manner.

Contributions. Our goal is to implement and evaluate OPE of bandit algorithms in realistic and reproducible ways. We release the *Open Bandit Dataset*, a logged bandit feedback collected on a large-scale fashion e-commerce platform, ZOZOTOWN.[1] ZOZOTOWN is the largest fashion EC platform in Japan with over 3 billion USD annual Gross Merchandise Value. When the platform produced the data, it used Bernoulli Thompson Sampling (Bernoulli TS) and Random policies to recommend fashion items to users. The dataset includes an A/B test of these policies and collected over 26 million records of users' clicks and the ground-truth about the performance of Bernoulli TS and Random. To streamline and standardize the analysis of the Open Bandit Dataset, we also provide the *Open Bandit Pipeline*, a series of implementations of dataset preprocessing, behavior bandit policy simulators, and OPE estimators.

To illustrate how to use the dataset and pipeline, we compare and evaluate state-of-the-art OPE estimators. Specifically, for each estimator, we use the log data of one of the behavior policies to predict the click through rates (CTR) of the other policy. We then assess the accuracy of the prediction by comparing it with the ground truth contained in the data. We compare the three estimators by their prediction performance. This exercise shows the following:

> **Empirical Result 1.** Inverse Probability Weighting (IPW; (Strehl et al., 2010)) and Doubly Robust (DR; (Dudík et al., 2014)) well predict performance of counterfactual policies (prediction errors being lower than 8.5%), while Direct Method (DM; (Beygelzimer and Langford, 2009)) produces much larger prediction errors.

This result is presented in Figure 1, where IPW and DR predict the ground-truth policy values of Bernoulli TS and Random well. In contrast, DM exhibits poor predictions. This experiment

1. https://corp.zozo.com/en/service/

suggests that a well-established estimator like DM may fail to predict the performance of a counterfactual policy. It is therefore essential to use an appropriate method, such as IPW and DR in this case.

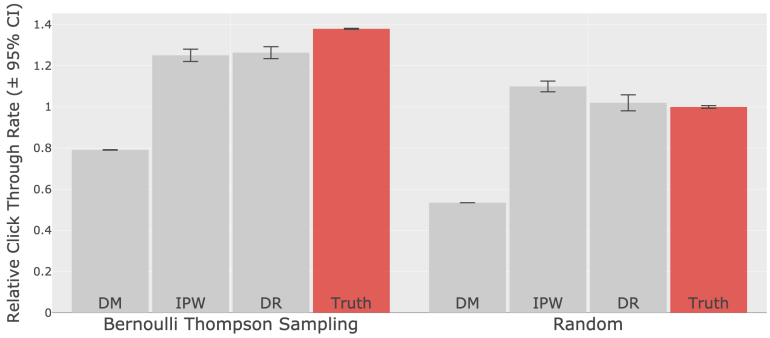

Figure 1: Comparing Off-Policy Evaluations with the Ground Truth

Notes: We report the estimated performances of Bernoulli Thompson Sampling and Random policies and their 95% confidence intervals (CI) in one of the three campaigns. IPW and DR predict the ground-truths well, while DM exhibits poor predictions.

We believe that our open data and pipeline help researchers evaluate the empirical performance of their methods, thereby advancing the future OPE research. Our case study showcases how to use our data to compare different estimators and use an appropriate one to improve the bandit systems.

2. Off-Policy Evaluation

2.1 Setup

We consider a general multi-armed contextual bandit setting. Let $\mathcal{A} = \{0, ..., m\}$ be a finite set of $m + 1$ *actions*, that the decision maker can choose from. Let $Y(\cdot) : \mathcal{A} \to \mathbb{R}$ denote a potential reward function that maps actions into rewards or outcomes, where $Y(a)$ is the reward when action a is chosen. Let X denote *context* vector that the decision maker observes when picking an action. We think of $(Y(\cdot), X)$ as a random vector with unknown distribution G. Given a vector of $(Y(\cdot), X)$, we define the mean reward function $\mu : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as $\mu(x, a) = \mathbb{E}[Y(a)|X = x]$.

We call a function $\pi : \mathcal{X} \to \mathcal{A}$ a *policy*, which maps each context $x \in \mathcal{X}$ into a distribution over actions, where $\pi(a|x)$ is the probability of taking action a given a context vector x. Let $\{(Y_t, X_t, D_t)\}_{t=1}^T$ be historical logged bandit feedback with T rounds of observations. $D_t := (D_{t0}, ..., D_{tm})'$ where D_{ta} is a binary variable indicating whether action a is chosen in round t. $Y_t := \sum_{a=0}^m D_{ta} Y_t(a)$ and X_t denote the reward and the context observed in round t, respectively. We assume that a logged bandit feedback is generated by a *behavior policy* π_b as follows: (i) In each round $t = 1, ..., T$, $(Y_t(\cdot), X_t)$ is i.i.d. drawn from distribution G., (ii) Given X_t, an action is randomly chosen based on $\pi_b(\cdot|X_t)$, creating the action choice D_t and the associated reward Y_t.

We are interested in using the historical logged bandit data to estimate the following *policy value* of any given *counterfactual policy* π which might be different from π_b:

$$V^{\pi} \coloneqq \mathbb{E}_{(Y(\cdot),X)\sim G}[\sum_{a=0}^{m} Y(a)\pi(a|X)] = \mathbb{E}_{(Y(\cdot),X)\sim G,\ D\sim\pi_b}[\sum_{a=0}^{m} Y(a)D_a\frac{\pi(a|X)}{\pi_b(a|X)}] \quad (1)$$

where the last equality uses the independence of D and $Y(\cdot)$ conditional on X and the definition of $\pi_b(\cdot|X)$. We allow the counterfactual policy π to be degenerate, i.e., it may choose a particular action with probability 1. Estimating V^{π} before implementing π in an online environment is valuable because π may perform poorly and damage user satisfaction. Additionally, it is possible to select a counterfactual policy that maximizes the policy value by comparing their estimated performances.

2.2 Benchmark OPE Estimators

Direct Method (DM). There are several approaches to estimate the value of the counterfactual policy. A widely-used method, DM (Beygelzimer and Langford, 2009), first learns a supervised machine learning model, such as ridge regression and gradient boosting, to predict the mean reward function. DM then uses it to estimate the policy value as

$$\hat{V}^{\pi}_{DM} = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m}\pi(a|X_t)\hat{\mu}(a|X_t).$$

where $\hat{\mu}(a|x)$ is the estimated reward function. If $\hat{\mu}(a|x)$ is a good approximation to the mean reward function, this estimator accurately predicts the policy value of the counterfactual policy V^{π}. If $\hat{\mu}(a|x)$ fails to approximate the mean reward function well, however, the final estimator is no longer consistent.

Inverse Probability Weighting (IPW). To alleviate the issue with DM, researchers often use another estimator called IPW (Precup et al., 2000; Strehl et al., 2010). IPW re-weights the rewards by the ratio of the counterfactual policy and behavior policy as

$$\hat{V}^{\pi}_{IPW} = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m} Y_t D_{ta}\frac{\pi(a|X_t)}{\pi_b(a|X_t)}.$$

When the behavior policy is known, the IPW estimator is unbiased and consistent for the policy value. However, it can have a large variance, especially when the counterfactual policy significantly deviates from the behavior policy.

Doubly Robust (DR). The final approach is DR (Dudík et al., 2014), which combines the above two estimators as

$$\hat{V}^{\pi}_{DR} = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m}\left\{(Y_t - \hat{\mu}(a|X_t))D_{ta}\frac{\pi(a|X_t)}{\pi_b(a|X_t)} + \pi(a|X_t)\hat{\mu}(a|X_t)\right\}.$$

DR mimics IPW to use a weighted version of rewards, but DR also uses the estimated mean reward function as a control variate to decrease the variance. It preserves the consistency of IPW if either the importance weight or the mean reward estimator is accurate (a property called *double robustness*).

3. Open Bandit Dataset and Pipeline

We apply and evaluate the above methods by using real-world data. Our data is logged bandit feedback data we call the *Open Bandit Dataset*. The dataset is provided by ZOZO Inc., the largest Japanese fashion e-commerce company. The company recently started using context-free multi-armed bandit algorithms to recommend fashion items to users in their large-scale fashion e-commerce platform called ZOZOTOWN.

We collected the data in a 7-days experiment in late November 2019 on three "campaigns," corresponding to "all", "men's", and "women's" items, respectively. Each campaign randomly uses either the Random algorithm or the Bernoulli Thompson Sampling (Bernoulli TS) algorithm for each user impression. In the notation of our bandit setups, action a is one of the possible fashion items, while reward Y is a click indicator.

To facilitate the usage of the Open Bandit Dataset, we also build a toolkit called the *Open Bandit Pipeline*. Our pipeline contains implementations of dataset preprocessing, behavior policy simulators, and evaluation of OPE estimators. This pipeline allows researchers to focus on building their OPE estimator and easily compare it with other methods in realistic and reproducible ways. To our knowledge, our real-world dataset and pipeline are the first to include multiple behavior policies, their implementations used in production, and their ground-truth policy values. These features enable the evaluation of OPE for the first time.

We provide detailed descriptions of the Open Bandit Dataset and usage of Open Bandit Pipeline at `https://github.com/st-tech/zr-obp`.

4. Experiments

We empirically evaluate and compare DM, IPW, and DR using our open data as follows (for each campaign separately):

1. For each of the Random and Bernoulli TS policies, randomly split the data collected by that policy into training (70%) and test (30%) sets.

2. Estimate the ground-truth value of each policy π by the empirical mean of clicks in the test set collected by that policy: $V^{\pi} = (T^{\pi}_{test})^{-1} \sum_{t=1}^{T^{\pi}_{test}} Y_t$, where T^{π}_{test} is the size of the test set of policy π.

3. Estimate the policy value of each policy by DM, IPW, and DR with the training set collected by the other policy.

4. Repeat the above process $K = 15$ times by sampling different training sets.

5. Compare the ground-truth and policy value estimated by the bagging prediction (Breiman, 1996).

We measure each estimator's performance with the *Relative-Estimation Error* defined below: *Relative-Estimation Error* of $\hat{V}^{\pi} = \left| \frac{(K^{-1} \sum_{k=1}^{K} \hat{V}^{\pi}_{k}) - V^{\pi}}{V^{\pi}} \right|$. where V^{π} is a ground-truth policy value of π in a test set. We ensured that the ground-truth estimation for each pair of behavior policies and campaigns has small confidence intervals and thus is accurate.

Table 1: Comparing Relative-Estimation Errors of Alternative Off-policy Estimators

	Campaigns and Behavior Policies (Prediction Target)					
	All		**Men's**		**Women's**	
Methods	**Bernoulli TS**	**Random**	**Bernoulli TS**	**Random**	**Bernoulli TS**	**Random**
DM	0.64162	0.08482	0.42645	0.46560	0.62527	0.02357
IPW	**0.04556**	**0.07532**	**0.09352**	**0.09940**	**0.06473**	**0.00942**
DR	**0.04512**	**0.02063**	**0.08410**	**0.01997**	**0.06538**	**0.00744**

Notes: The relative-estimation errors of the three estimators are reported. The **red fonts** and **blue fonts** represent the best and the second best OPE estimators for each prediction target.

$K^{-1}\sum_{k=1}^{K}\hat{V}_k^{\pi}$ is a bagging prediction where $\hat{V}_k^{\pi}$ is an estimated policy value with the k-th bootstrapped samples. $K = 15$ is the number of folds.

For IPW and DR, we compute the true behavior policy by Monte Carlo simulation of the beta distribution used in Bernoulli TS. For DM and DR, we need to obtain a reward estimator $\hat{\mu}$. We do so by using LightGBM (Ke et al., 2017) implemented in *scikit-learn* and training it with the whole training set.

The results of off-policy estimator selection are given in Table 1 and Figure 1. First, DM fails to predict the policy values in all settings. The failure of DM likely comes from the bias of the regression model. We observe that the prediction by LightGBM does not improve upon a naive prediction using the mean CTR for every prediction. Specifically, the improvements of the regression model over the naive prediction are only 1.51%-7.04% in the binary cross-entropy measure.

The problem with DM makes us expect that IPW and DR may perform better, because the two methods do not rely on the correct specification of the regression model. We confirm this expectation in Table 1, where IPW and DR drastically outperform DM. In particular, DR performs best in five out of the six scenarios. A possible reason for the best performance of DR is that DR is robust to the bias of the nuisance estimation, which is known as the *Neyman orthogonality* (Narita et al., 2020).

As we show in this section, researchers can evaluate and compare their methods with others easily by using our open data and pipeline.

5. Conclusion

To enable realistic and reproducible evaluation of off-policy evaluation of bandit algorithms, we have publicized the Open Bandit Dataset–a benchmark logged bandit dataset collected on a large-scale fashion e-commerce platform. The data comes with the Open Bandit Pipeline, a collection of implementations that makes it easy to evaluate and compare different OPE estimators. We have also presented a case evaluation of OPE estimators by using the dataset and pipeline. We believe that our open data and pipeline will allow researchers and practitioners to easily evaluate and compare their bandit algorithms and OPE estimators with others in a large, real-world setting.

References

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138, 2009.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29:485–511, 2014.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1447–1456, 2018.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 198–206, 2018.

Alois Gruson, Praveen Chandar, Christophe Charbuillet, James McInerney, Samantha Hansen, Damien Tardieu, and Ben Carterette. Offline evaluation to make decisions about playlistrecommendation algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 420–428, 2019.

Alex Irpan, Kanishka Rao, Konstantinos Bousmalis, Chris Harris, Julian Ibarz, and Sergey Levine. Off-policy evaluation via off-policy classification. In *Advances in Neural Information Processing Systems*, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 652–661, 2016.

Nathan Kallus and Masatoshi Uehara. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

Masahiro Kato, Masatoshi Uehara, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *arXiv preprint arXiv:2002.11642*, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A Contextual-bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 297–306, 2011.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2644–2653, 2018.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084, 2014.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Yusuke Narita, Shota Yasui, and Kohei Yata. Efficient counterfactual learning from bandit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4634–4641, 2019.

Yusuke Narita, Shota Yasui, and Kohei Yata. Safe counterfactual reinforcement learning. *arXiv preprint arXiv:2002.08536*, 2020.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from Logged Implicit Exploration Data. In *Advances in Neural Information Processing Systems*, pages 2217–2225, 2010.

Adith Swaminathan and Thorsten Joachims. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research*, 16: 1731–1755, 2015a.

Adith Swaminathan and Thorsten Joachims. The Self-normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, pages 3231–3239, 2015b.

Philip Thomas and Emma Brunskill. Data-efficient Off-policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2139–2148, 2016.

Nikos Vlassis, Aurelien Bibaut, Maria Dimakopoulou, and Tony Jebara. On the design of estimators for bandit off-policy evaluation. In *International Conference on Machine Learning*, pages 6468–6476, 2019.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3589–3597, 2017.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 9665–9675, 2019.