Bayesian Perspectives on Offline Evaluation for Recommender Systems

Michael Benigni
Politecnico di Milano
Milan, Italy
michael.benigni@polimi.it

Abstract

Offline evaluation is a fundamental component in the deployment and development of better recommender systems. In recent years, the contextual bandit framework has emerged as a valuable approach for offline and counterfactual evaluation, leading to the increasing interest in estimators based on inverse propensity scoring (IPS), direct methods (DM), and doubly robust (DR) techniques. However, nearly all existing methods rely on frequentist statistics, limiting their ability to capture model uncertainty and reflecting it in evaluation outcomes.

This work explores the novel research direction of Bayesian statistics for Off-Policy Evaluation in recommendation tasks, motivated by the need for reliable estimators that are more robust to distribution shift, data sparsity, and model misspecification. Three underexplored research directions are identified in this work: (i) using posterior uncertainty from Bayesian reward models to design adaptive hybrid estimators, (ii) explicitly modeling all components of the OPE problem (contexts, actions, and rewards) using a joint probabilistic framework, and (iii) quantifying epistemic uncertainty over policy value estimates via posterior inference.

By leveraging the Bayesian framework, the aim is to improve the reliability, interpretability, and safety of offline evaluation protocols, offering a new perspective on one of the most persistent challenges in recommender systems research. This perspective is especially relevant in data-scarce or high-stakes settings, where understanding uncertainty is essential for trustworthy decision-making.

CCS Concepts

• **Computing methodologies** → **Learning from implicit feedback**; • **Mathematics of computing** → **Bayesian computation**.

Keywords

Offline Evaluation, Recommender Systems, Bayesian Statistics

ACM Reference Format:
Michael Benigni. 2025. Bayesian Perspectives on Offline Evaluation for Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3705328.3748762

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748762

1 Introduction

Nowadays, online platforms such as social networks, entertainment services (e.g., music and video streaming), e-commerce are pervasive in everyday life [9, 21]. These services offer a wide variety of products to users, aiming to fulfill a broad range of demands and needs. However, from a user's perspective, navigating an extensive catalog to find the most suitable item or service can be tedious and frustrating. A recommender system plays a crucial role in simplifying this process by suggesting items that may be of interest to the user. Technically, a recommender system is an algorithm, often based on machine learning, that seeks to recommend relevant items to a user, based on previously collected data about users' past interactions, contextual information about these interactions, and metadata about users and items in the catalog.

A critical challenge in recommender systems is their evaluation. The most reliable evaluation methodology is online testing, which involves deploying an algorithm in a live system and measuring user engagement metrics, often through A/B testing. A/B testing is an online evaluation procedure for recommender systems that consists of dividing the user traffic into two (or more) portions, providing recommendations using a different strategy for each user group. After collecting interactions, evaluation metrics can be computed. Such evaluations are considered highly reliable but are typically feasible only for large technology companies. In fact, relying solely on online methods is impractical, as testing multiple algorithms requires a vast user base. Moreover, deploying a poorly performing algorithm may lead to poor recommendations, and consequently to user dissatisfaction and revenue loss, making the use of online evaluations techniques limited also for companies with large user traffic.

For these reasons, offline evaluation is often the only viable alternative. Offline evaluation is designed to mimic online evaluation but is conducted on historical data. Historical data is usually split into two parts: one portion (training data) is used for recommendation algorithm design, while the remaining portion (testing data) is employed to evaluate the algorithm, using the same metrics used in online evaluation settings.

However, it has been shown that offline evaluation is often unreliable, making a robust approach a valuable achievement. Indeed, several industry studies [8, 10] have shown a significant discrepancy between online and offline metrics. This discrepancy arises due to methodological differences between online and offline evaluation:

- Exposure bias: offline evaluation relies on historical interactions collected when a different algorithm, compared to the one being evaluated, was providing recommendations. In an online setting instead, the evaluation metrics are computed

when the algorithm to evaluate is providing recommendations. This difference likely influences the outcome of the evaluation, as users tend to interact only with items that were recommended to them, skewing the data collection. This phenomenon is known as exposure bias. Due to this shift in the evaluation settings, there is a risk that new recommender systems are primarily evaluated not on their ability to model user preferences, but rather on how well they mimic the previous model that mediated the data collection.
- Non-stationary preferences: user preferences evolve over time, so historical interactions may not accurately reflect current user interests, leading to distribution shifts [11].
- Catalog modifications: The item catalog may change, meaning that historical interactions do not contain information about newly introduced items that could influence evaluation outcomes.

Moreover, in large-scale systems, recommenders suggest combinations of items rather than single items, making the recommendation space combinatorial and further complicating offline evaluation [1, 22, 23].

To address the evaluation issues in offline evaluation, recent studies leverage reinforcement learning and causal inference, in particular using the paradigm of contextual bandit (CB) [14, 38]. This family of approaches models the recommendation system as an active agent: it observes the user's *context*, takes an *action* that consists of recommending items according to a recommendation strategy (named *logging* policy, modeled as a distribution over possible items), and receives a *reward* if the user interacts somehow with one of the items recommended. This paradigm is employed in the context of recommender systems to solve the problem of offline evaluation, which can be seen as a problem of estimating the quality of a new policy, termed *evaluation* policy, which is different from the logging one. The historical data are assumed to be collected following the logging policy, while no interaction data are collected under the evaluation policy. In reinforcement learning this problem is usually called offline Off-Policy Evaluation (OPE) [25].

In the next sections it will be explained how this paradigm allows to model the distribution shift between online and offline evaluation scenarios, potentially providing more accurate estimations of the effectiveness of recommendation algorithm even in an offline setting.

However, despite some advances, no single estimator has proven universally reliable [37], making this a hot topic in the field of recommendation systems evaluation.

This work aims to investigate a new possible research direction of counterfactual offline evaluation methods that consists in shifting the paradigm to a Bayesian statistical perspective. Traditionally, common approaches employ frequentist statistics to address distribution shifts and exposure bias. Bayesian statistical techniques are instead less explored. The primary advantage of a Bayesian framework is its intrinsic ability to quantify uncertainty, which can help understanding the reliability of evaluation metrics, essential for conducting meaningful offline evaluations in real-world scenarios.

This work presents several key insights. First, last advances in the field of offline evaluation are examined, highlighting the differences with a bayesian perspective (see Section 2), then the counterfactual contextual bandit paradigm is explored, demonstrating its relevance and practicality for recommendation system use cases (Section 3). Next, the potential of bayesian modeling in this setting is outlined, placing special emphasis on the advantages that this formulation could provide (Section 4). Lastly, the discussion is wrapped up in Section 5.

2 Related Work

A substantial body of work has addressed the challenges of offline evaluation through techniques adapted from counterfactual learning and causal inference. Inverse Propensity Scoring (IPS) [18, 28] remains a foundational method in this space, providing an unbiased estimate of a target policy's performance by reweighting observed outcomes via importance sampling[1]. However, IPS suffers from high variance, especially when the logging policy probability assigned to actions chosen by the target policy is low [19, 24]. Several variants have been proposed to mitigate this issue, including Self-Normalized IPS [32], which normalize estimates with the sum of propensities, and clipped IPS [27], which introduces bias to reduce variance by clipping small propensities.

At the opposite end of the spectrum lies the Direct Method [3] (DM), which models expected user feedback by learning a reward predictor from historical data. While DM can drastically reduce variance, it introduces bias when the model is misspecified. To overcome the respective weaknesses of IPS and DM, hybrid estimators such as the Doubly Robust (DR) approach have been proposed [5, 12, 33]. DR estimators maintain unbiasedness if either the propensity model (if the real propensities are not available) or the reward model is correctly specified and typically achieves lower variance than IPS. Several refinements of the DR estimator have been developed [15, 16, 27, 29, 39], with different characteristics.

Despite these advances, offline evaluation remains intrinsically difficult in practice. One fundamental obstacle is the support mismatch problem [7], which arises when the target policy recommends items not sufficiently explored by the logging policy. In such cases, the estimators break down due to the lack of reliable observations. This issue is exacerbated in large or combinatorial action spaces (e.g., top-k recommendation or slate ranking), where even mild shifts in policy can lead to severe extrapolation [24]. Proposed solutions include restricting evaluation to the support set of the logging policy, introducing clipping thresholds to control importance weights [29], or imputing outcomes for unobserved actions using model-based or semi-parametric techniques.

A further barrier is the limited availability of datasets with logged propensities, which are necessary for accurate IPS-based evaluation. Most public datasets only record user interactions (e.g., clicks or ratings) without the full exposure context, thereby precluding proper counterfactual evaluation. Recent works [20] have emphasized the importance of impression datasets, where the full list of recommended items per session is retained, enabling a potentially more accurate modeling of selection bias and exposure mechanisms. New benchmarks such as the Open Bandit Dataset and synthetic simulators [20] are being developed to facilitate progress in this direction, making the propensity scores directly available.

[1]Assuming known or accurately estimated propensity scores [7], i.e., the probability of selecting an item under the logging policy.

A less often discussed, yet increasingly important issue is that of estimator selection [4, 6, 30, 34, 35]. No estimator, be it IPS, DM, DR, or one of their variants is universally reliable. Their performance depends on a variety of factors, including model misspecification, data sparsity, and the magnitude of distributional shift between the logging and target policies [37]. Selecting an appropriate estimator for a given evaluation scenario is therefore a challenging task. This has motivated research into adaptive selection strategies, ensemble combinations, and meta-learned policies that aim to optimize evaluation reliability across varying regimes. For instance, methods like OPERA [17] dynamically combine multiple OPE estimators by minimizing prediction error on a held-out validation set, offering a robust aggregation mechanism that adapts to the underlying data characteristics. Complementary work by Sun et al. [31] explores the role of offline reward modeling in supporting estimator selection, highlighting how accurate reward predictions can guide the choice or weighting of estimators, especially in complex recommendation scenarios. These contributions underscore the growing recognition that estimator selection is not merely a model selection problem, but a core component of the offline evaluation pipeline requiring dedicated algorithmic and theoretical attention.

Surprisingly, Bayesian statistical methods remain underexplored in this context, despite their conceptual compatibility with offline evaluation goals. Bayesian inference provides a natural framework for modeling uncertainty over both propensities and reward functions, enabling posterior predictive estimators that integrate uncertainty rather than relying on point estimates. Such formulations have the potential to improve robustness, especially in data-scarce settings or when estimator selection is ambiguous. This work aims to highlight the potential of Bayesian approaches to offline evaluation, both as a principled alternative to standard estimators and as a unifying perspective that can guide estimator selection.

Bayesian approaches to off-policy evaluation (OPE) have recently begun to gain traction as a principled alternative to traditional frequentist methods, particularly in settings characterized by uncertainty, sparse data, and large action spaces. A few works have proposed Bayesian estimators that explicitly model the posterior distribution over expected rewards or policy values, enabling more calibrated evaluation and decision-making under uncertainty. For instance, Aouali et al. [1] propose a scalable Bayesian framework for DM estimator that leverages correlations among actions via structured priors, improving sample efficiency in OPE and off-policy learning. Similarly, Vlassis et al. [36] analyze OPE through the lens of Bayes risk minimization, proposing a new control variate estimator for slate settings that reduces variance in proportion to slot-level policy divergences. Sakhi et al. [26] adopt a PAC-Bayesian perspective, deriving generalization bounds for policy value estimation and proposing tractable optimization objectives with provable guarantees. In parallel, empirical likelihood techniques have been adapted to the contextual bandit setting by Karampatziakis et al. [13], yielding convex formulations for uncertainty-aware confidence intervals and robust policy optimization. Finally, Aouali et al. [2] present a probabilistic model for slate recommendation that combines ranking and reward signals to enable efficient and scalable off-policy estimation. While these contributions highlight growing interest in Bayesian OPE, this work contributes to this emerging line by proposing Bayesian formulations not only for modeling rewards and policies, but also for reasoning about estimator uncertainty and guiding estimator design.

3 Background

This work addresses the problem of offline evaluation of recommendation policies using the contextual bandit paradigm, a widely adopted abstraction for modeling user-item interactions in recommender systems. In the CB setting, the system operates in discrete time. At each timestep t, it observes a user context $x_t \in \mathcal{X}$, drawn from a distribution $p(x_t)$, and selects an item $a_t \in \mathcal{A}$, where $\mathcal{A} \subseteq \mathbb{N}$ denotes a finite set of candidate items. The item is selected according to a stochastic *logging policy* π_b, i.e.,

$$a_t \sim \pi_b(a_t \mid x_t).$$

After the item is recommended, a binary *reward* $r_t \in \{0, 1\}$ is observed, drawn from a conditional distribution $p(r_t \mid x_t, a_t)$. This signal captures whether the user engaged with the recommended item, through a click, watch, like, or other interaction event, and serves as a proxy for user satisfaction.

The goal of offline policy evaluation is to estimate the performance of a new or target policy π_e, without deploying it online. The performance is formalized as the *policy value*, defined as the expected reward under the target policy:

$$V(\pi_e) \coloneqq \mathbb{E}_{x_t}\left[\mathbb{E}_{a_t}\left[\mathbb{E}_{r_t}[r_t]\right]\right].$$

Only historical data collected under the logging policy π_b is available, forming the dataset:

$$\mathcal{D}_b = \left\{\left(x_t, a_t, r_t, \{\pi_b(a_t \mid x_t)\}_{a_t \in \mathcal{A}}\right)\right\}_{t=1}^{n_b},$$

where n_b is the number of logged interactions. An estimator $\hat{V}(\pi_e)$ aims to infer $V(\pi_e)$ using $\mathcal{D}_b$ and the definition of π_e. Estimator quality is typically measured via its *mean squared error* (MSE):

$$\mathrm{MSE}(\hat{V}(\pi_e)) = \mathbb{E}\left[\left(\hat{V}(\pi_e) - V(\pi_e)\right)^2\right],$$

where the expectation is over the stochasticity of sampling of contexts, selection of actions, and generation of rewards.

While this binary-reward formulation is commonly used for theoretical analysis and algorithm development, real-world recommender systems often involve ranking or set-based objectives, where the goal is to evaluate top-k recommendation lists using metrics such as *nDCG*, *Recall*, or *MAP*, that depend on the relative order of items and their relevance.

Extending OPE techniques to support these structured metrics can introduce additional technical challenges. Nonetheless, the foundational principles established in the binary reward setting, such as importance weighting, reward modeling, and doubly robust estimation, remain applicable and provide a valuable starting point for these generalizations.

4 Research Questions

While the vast majority of off-policy evaluation methods in recommendation systems are rooted in frequentist statistics [12, 18, 28, 33, 39], the application of Bayesian approaches remains largely unexplored. This gap represents a significant missed opportunity, as Bayesian inference offers a natural framework for reasoning under

uncertainty, model averaging, and integrating prior knowledge, capabilities that are particularly well-suited to the high-variance and data-scarce scenarios typical of offline evaluation. In this section, we propose three research directions that aim to bridge this gap by investigating the advantages of Bayesian statistics for offline evaluation of recommendation policies.

4.1 Bayesian Uncertainty for Hybrid Estimators

The trade-off between bias and variance has long motivated the development of hybrid estimators in OPE, such as the Doubly Robust (DR) estimator [5, 12, 29, 33, 39]. These methods typically combine an outcome model (Direct Method) with importance weighting (IPS) to balance robustness against model misspecification. However, existing hybrid approaches treat the outcome model as a point estimate, ignoring epistemic uncertainty in its predictions. This is particularly limiting when using expressive reward models like neural networks, where overconfidence on out-of-distribution inputs can lead to catastrophic bias.

The first proposal of this work is to extend DR-style estimators by integrating Bayesian reward models, such as Bayesian neural networks or ensembles with approximate posterior distributions. The idea is to use the predictive uncertainty of the reward model to adaptively weight the IPS and DM components, down-weighting the outcome model when uncertainty is high. This opens the door to a class of uncertainty-aware hybrid estimators, where the confidence of the model directly informs the estimator's structure. This approach would allow a more principled way to interpolate between importance-weighted and model-based estimates, potentially improving estimator robustness in low-support regions.

4.2 Probabilistic Modeling of OPE Components

Most OPE estimators focus solely on estimating expected rewards under a target policy [12, 18, 28, 33, 39], without modeling the underlying generative mechanisms. However, Bayesian methods allow us to explicitly model the full stochastic process that generates logged data, including the context distribution, the logging and evaluation policies, and the reward distribution. A fully Bayesian approach would treat all unknown components as latent variables, define priors over their distributions, and infer posterior predictive distributions given the data.

The second research direction proposed by this work is to build a unified Bayesian generative model for recommendation logs, treating user contexts, action policies, and outcomes as random variables with learnable distributions. This probabilistic modeling approach enables posterior sampling of counterfactual user responses, thereby producing predictive distributions over the policy value rather than just point estimates. Such models could incorporate structure e.g., temporal priors, hierarchical user/item dependencies, or side information—and could also be used to regularize estimators resulting in more accurate estimates.

In addition to enabling principled OPE, this perspective aligns naturally with the downstream goal of counterfactual learning: a Bayesian generative model provides a coherent framework for both evaluation and learning under uncertainty, facilitating risk-sensitive decision-making.

4.3 Modeling Estimator Uncertainty

A well-known limitation of current OPE methods is the absence of calibrated confidence intervals or uncertainty estimates. While some frequentist techniques provide asymptotic variance bounds [26], these are often intractable or loose in high-dimensional settings. In contrast, Bayesian statistics provides a natural mechanism for uncertainty quantification: posterior distributions over parameters directly induce posterior distributions over estimated policy values.

The last proposal is to explicitly model the uncertainty of the OPE estimator by treating the policy value as a random variable and deriving its posterior distribution given the logged data. This can be achieved either by sampling from the posterior of the underlying models (e.g., reward and propensity models) or via fully Bayesian decision-theoretic formulations. Such distributions can then be used to construct credible intervals, hypothesis tests, or even Bayesian model selection procedures that account for estimator risk.

Moreover, this line of work can be connected with Bayesian experimental design: once a posterior over policy values is obtained, we can ask which additional data (or user queries) would most reduce uncertainty, paving the way toward active OPE or safe policy deployment strategies.

5 Conclusions

Offline policy evaluation is a critical yet inherently difficult task in the development of recommender systems. While the field has made significant progress using frequentist estimators such as IPS, DM, and DR, current approaches face challenges related to model misspecification, high variance, deficient support, and the lack of principled uncertainty quantification. Bayesian statistics, though mostly overlooked in this domain, offers a promising and underexplored toolkit for addressing these limitations.

In this work, three concrete and novel research directions are proposed. First, the use of Bayesian uncertainty in reward models can improve hybrid estimators by dynamically adapting to regions of high uncertainty. Second, a fully Bayesian treatment of OPE, explicitly modeling the distributions of users, policies, and rewards, would enable posterior predictive evaluations and structured regularization. Third, modeling the epistemic uncertainty of OPE estimators allows for credible intervals and Bayesian decision-making, opening new pathways for safe deployment and risk-sensitive evaluation.

Together, these contributions could ground OPE methods on a more coherent probabilistic foundation, enabling more reliable, interpretable, and data-efficient evaluation protocols for modern recommender systems.

Acknowledgments

I sincerely thank Maurizio Ferrari Dacrema and Nicolò Felicioni for their mentorship and support during the development of this work.

References

[1] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. 2024. Bayesian Off-Policy Evaluation and Learning for Large Action Spaces. doi:10.48550/arXiv.2402.14664 arXiv:2402.14664 [cs] version: 1.

[2] Imad Aouali, Achraf Ait Sidi Hammou, Otmane Sakhi, David Rohde, and Flavian Vasile. 2024. Probabilistic Rank and Reward: A Scalable Model for Slate Recommendation. doi:10.48550/arXiv.2208.06263 arXiv:2208.06263 [cs].

[3] Alina Beygelzimer and John Langford. 2009. The offset tree for learning with partial labels. In *Proceedings of the ACM international conference on Knowledge discovery and data mining (SIGKDD)*. 129–138.

[4] Matej Cief, Branislav Kveton, and Michal Kompan. 2025. Cross-Validated Off-Policy Evaluation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 16073–16081. doi:10.1609/AAAI.V39I15.33765

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

[6] Nicolò Felicioni, Michael Benigni, and Maurizio Ferrari Dacrema. 2024. AutoOPE: Automated Off-Policy Estimator Selection. *CoRR* abs/2406.18022 (2024). doi:10.48550/ARXIV.2406.18022 arXiv:2406.18022

[7] Nicolò Felicioni, Maurizio Ferrari Dacrema, Marcello Restelli, and Paolo Cremonesi. 2022. Off-Policy Evaluation with Deficient Support Using Side Information. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren (Eds.). ACM, 169–176. doi:10.1145/2645710.2645745

[9] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 198–206.

[10] Carlos Alberto Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manag. Inf. Syst.* 6, 4 (2016), 13:1–13:19. doi:10.1145/2843948

[11] Balázs Hidasi and Ádám Tibor Czapp. 2023. Widespread Flaws in Offline Evaluation of Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 848–855. doi:10.1145/3604915.3608839

[12] Nan Jiang and Lihong Li. 2016. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 48. 652–661.

[13] Nikos Karampatziakis, John Langford, and Paul Mineiro. 2020. Empirical Likelihood for Contextual Bandits. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9597–9607. https://proceedings.neurips.cc/paper/2020/hash/6d34d468ac8876333c4d7173b85efed9-Abstract.html

[14] John Langford and Tong Zhang. 2007. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *Advances in Neural Information Processing Systems (NeurIPS)*. 817–824.

[15] Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. 2020. Importance Sampling Techniques for Policy Optimization. *J. Mach. Learn. Res.* 21 (2020), 141:1–141:75.

[16] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. 2021. Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 8119–8132.

[17] Allen Nie, Yash Chandak, Christina J Yuan, Anirudhan Badrinath, Yannis Flet-Berliac, and Emma Brunskil. 2024. OPERA: Automatic Offline Policy Evaluation with Re-weighted Aggregates of Multiple Estimators. *arXiv preprint arXiv:2405.17708* (2024).

[18] Doina Precup, Richard S. Sutton, and Satinder P. Singh. 2000. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the International Conference on Machine Learning (ICML)*. 759–766.

[19] Noveen Sachdeva, Lequn Wang, Dawen Liang, Nathan Kallus, and Julian McAuley. 2024. Off-policy evaluation for large action spaces via policy convolution. In *Proceedings of the ACM on Web Conference (WWW)*. 3576–3585.

[20] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2021. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*.

[21] Yuta Saito and Thorsten Joachims. 2021. Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*. 828–830.

[22] Yuta Saito and Thorsten Joachims. 2022. Off-Policy Evaluation for Large Action Spaces via Embeddings. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 19089–19122. https://proceedings.mlr.press/v162/saito22a.html

[23] Yuta Saito, Qingyang Ren, and Thorsten Joachims. 2023. Off-Policy Evaluation for Large Action Spaces via Conjunct Effect Modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 29734–29759. https://proceedings.mlr.press/v202/saito23b.html

[24] Yuta Saito, Qingyang Ren, and Thorsten Joachims. 2023. Off-policy evaluation for large action spaces via conjunct effect modeling. In *Proceedings of the International Conference on Machine Learning (ICML)*. 29734–29759.

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

[26] Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. [n. d.]. PAC-Bayesian Offline Contextual Bandits With Guarantees. ([n. d.]).

[27] Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. 2024. Logarithmic Smoothing for Pessimistic Off-Policy Evaluation, Selection and Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37. 80706–80755.

[28] Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. 2010. Learning from Logged Implicit Exploration Data. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2217–2225.

[29] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. 2020. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 119. 9167–9176.

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

[31] Hao Sun, Alex James Chan, Nabeel Seedat, Alihan Hüyük, and Mihaela van der Schaar. 2024. When is Off-Policy Evaluation (Reward Modeling) Useful in Contextual Bandits? A Data-Centric Perspective. *Journal of Data-centric Machine Learning Research* (2024).

[32] Adith Swaminathan and Thorsten Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3231–3239.

[33] Philip S. Thomas and Emma Brunskill. 2016. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 48. 2139–2148.

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

[35] Takuma Udagawa, Haruka Kiyohara, Yusuke Narita, Yuta Saito, and Kei Tateno. 2023. Policy-Adaptive Estimator Selection for Off-Policy Evaluation. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*. 10025–10033.

[36] Nikos Vlassis, Fernando Amat Gil, and Ashok Chandrashekar. 2021. Off-Policy Evaluation of Slate Policies under Bayes Risk. doi:10.48550/arXiv.2101.02553 arXiv:2101.02553 [cs].

[37] Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. 2021. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*.

[38] Chih-Chun Wang, Sanjeev R Kulkarni, and H Vincent Poor. 2005. Bandit problems with side observations. *IEEE Trans. Automat. Control* 50, 3 (2005), 338–355.

[39] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. 2017. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 70. 3589–3597.