A Simulation Framework for Off-Policy Evaluation of Ranking Models

Dan Qiao[1], Mohsen Malmir[2], Houssam Nassif[2], and Murat Bayir[2]

[1]University of California, Santa Barbara, danqiao@ucsb.edu
[2]Meta, {mohsenm,houssamn,mbayir}@meta.com

1 Introduction

Ranking policies are central to information retrieval and recommender systems, yet evaluating new policies is challenging because online experiments are costly and risky. Small changes to ranking models can significantly affect user experience and business outcomes, making reliable offline (off-policy) evaluation a critical component of system development. Off-policy evaluation (OPE) methods estimate the performance of counterfactual policies using data logged under a production policy, thereby avoiding direct user exposure [2].

Researchers proposed a variety of estimators for OPE, including inverse propensity scoring (IPS) [5], self-normalized IPS (SNIPS) [6], and doubly robust (DR) methods [3]. Beyond these estimators, system-level frameworks such as Genie [1] have also been developed, using open-box simulation and log replay to improve counterfactual evaluation in sponsored search. Complementary to these directions, *Causal Transfer Random Forest* (CTRF) combines limited randomized-experiment data with large-scale logs to learn invariant decision structures and improve robustness under covariate shift, enabling scalable offline policy estimation in sponsored search [7].

In our prior work on counterfactual evaluation [4], we proposed a domain-adapted reward model that leverages importance weighting to reduce bias when training on source-domain logs, and demonstrated that this approach improves upon vanilla IPS and simple direct methods. Building on that foundation, this paper shifts focus toward a controlled simulation environment. By generating synthetic ads, policies, and user responses, we can systematically evaluate the role of domain adaptation and compare multiple estimators—including Direct Method, SNIPS, and Doubly Robust—against ground-truth values. Our goal is to provide a comprehensive study of how domain adaptation influences off-policy evaluation of ranking policies in a controlled simulation.

2 Simulation Environment

We build a controlled simulation that mimics a production ads-ranking stack while retaining access to ground truth. A catalog of 10,000 ads is generated with feature vectors and scored by an oracle reward function: a randomly initialized, then frozen, multilayer perceptron (MLP) with a sigmoid output that maps each ad a to its true conversion probability $r^*(a) \in [0, 1]$. This fixed oracle defines the data-generating mechanism and lets us compute the true value of any policy for benchmarking.

To emulate production, we first train a logging policy as a pointwise ranking MLP. Training data are created by uniformly sampling ads and drawing $r \sim \text{Bernoulli}(r^*(a))$ using the oracle probabilities. The logging policy is trained with binary cross-entropy (BCE) loss and, at serving time, ranks ads by its scores, thereby inducing the source distribution and propensities $P_{\pi_S}(a \mid C)$.

We then introduce a shifted "domain" policy—

another MLP, potentially with different feature masks or architecture—trained on interaction data collected while serving the logging policy (ads selected by the logger, labels supplied by the oracle), again with a pointwise BCE objective. This mirrors candidate models trained on biased logs from an existing production system.

Our goal is to evaluate the domain policy without target-domain labels. We therefore construct two datasets: (i) a labeled source dataset by serving the logging policy (features, 0/1 labels), and (ii) an unlabeled target dataset by serving the domain policy. From the source dataset, we perform off-policy evaluation using propensity-based estimators such as Vanilla IPS. We reweight the observed outcomes by the target-to-logging propensity ratio to correct for the fact that the target policy would have selected a different set of actions than the logging policy.

We train reward models (MLPs) on the labeled source dataset and adapt them to the target distribution using importance weighting during training to emphasize source examples that are more likely under the domain policy. We then apply a Direct Method (DM) on the unlabeled target dataset: the adapted reward model predicts per-impression conversion probabilities that we aggregate under the domain policy to obtain an offline value estimate. Because the oracle is known, we can compute ground-truth policy values and quantify the error of IPS, DM (with/without adaptation), and related variants under controlled distribution shifts. This provides a concise yet rigorous testbed for studying off-policy evaluation and domain adaptation in ads ranking.

In the next section, we provide a comprehensive explanation of the evaluation methodology for ranking policies, including propensity-based estimators (e.g., IPS and its variants), reward-modeling approaches such as Direct Method with domain adaptation, and the metrics used to assess estimator accuracy.

3 Policy Evaluation

We now describe our framework for evaluating ranking policies in the simulation environment. Our goal is to estimate the expected value of a target policy π_T, defined as the expected user conversion rate under the distribution of ads that π_T would select.

3.1 Datasets and Notation

We evaluate ranking policies in a single-slot setting where each ad request t produces a random set of ten ad candidates $C_t = \{a_t^1, \ldots, a_t^5\}$. A policy π assigns a score $s_\pi(a)$ to each candidate ad $a \in C_t$. The final ad to be presented to the user is then selected greedily, using maximum score to fill in the slot.

The *source domain* consists of logged impressions collected by serving the production policy π_S with observed outcomes:

$$\mathcal{D}_S = \{(a_i, r_i, C_i)\}_{i=1}^N, \quad C_i = \{a_i^1, \ldots, a_i^5\}, \\ a_i \sim P_{\pi_S}(\cdot \mid C_i), \quad r_i \in \{0, 1\}. \quad (1)$$

The *target domain* mirrors this construction for the target policy π_T but without user labels; instead, it carries reward-model predictions for the selected ad:

$$\mathcal{D}_T = \{(a_j, \hat{r}_j, C_j)\}_{j=1}^M, \quad C_j = \{a_j^1, \ldots, a_j^5\}, \\ a_j \sim P_{\pi_T}(\cdot \mid C_j), \quad \hat{r}_j = \hat{r}(a_j) \in [0, 1]. \quad (2)$$

3.2 Evaluation Methods

We consider four evaluation methods:

Ground Truth. In simulation, we can compute the true value of the target policy by combining the oracle conversion probabilities with the target policy's selection distribution:

$$V_{\mathrm{GT}}(\pi_T) = \sum_a P_T(a)\, r^*(a).$$

Direct Method (DM). The DM estimator uses the reward model trained on the source domain to predict user responses across all ads:

$$\hat{V}_{\mathrm{DM}}(\pi_T) = \sum_a P_T(a)\, \hat{r}(a).$$

Self-Normalized IPS (SNIPS). The SNIPS estimator reuses the logged source-domain impressions and reweights them to match the target policy:

$$\hat{V}_{\text{SNIPS}}(\pi_T) = \frac{\sum_{i=1}^{n_S} w_i r_i}{\sum_{i=1}^{n_S} w_i}, \quad w_i = \frac{P_T(a_i)}{P_S(a_i)}.$$

Doubly Robust with Self-Normalized IPS (DR-SNIPS). We consider the doubly robust estimator where the IPS correction term is self-normalized. This approach augments the low-variance Direct Method (DM) estimate with a correction term based on logged data, using self-normalized importance weights:

$$\hat{V}_{\text{DR-SNIPS}}(\pi_T) = \sum_a P_T(a)\,\hat{r}(a) + \frac{1}{\sum_{i=1}^{n_S} w_i} \sum_{i=1}^{n_S} w_i \left(r_i - \hat{r}(a_i)\right).$$

All estimators target the same underlying quantity $V(\pi_T)$, the expected value of the target policy. In simulation, V_{GT} serves as the benchmark. DM relies solely on the reward model, SNIPS relies solely on logged outcomes, and DR-SNIPS balances the two by combining model predictions with logged corrections using self-normalized weights.

4 Experiments

We evaluate our approach in a controlled simulation without user contexts. The simulator generates a pool of 10,000 ads $\{a_i\}_{i=1}^{10000}$, each drawn i.i.d. from $\mathcal{N}(0, I_{10})$ where I_{10} represents the 10-dimensional identity covariance matrix. The oracle reward function f_θ is a 3-layer ReLU MLP with sigmoid output, defining ground-truth conversion probabilities $P(r = 1|a) = f_\theta(a)$. Rewards are sampled from Bernoulli($f_\theta(a)$).

Policy setup. Ads are served through a ranking-based delivery policy: from a random subset of 5 ads, the ad with the highest score is selected. The source (production) policy is a neural network similar to the oracle but restricted to a subset of features. It is trained with cross-entropy loss on randomly sampled ads and their oracle-based labels. Using this trained source policy, we collect a logged dataset $\mathcal{D}_S$. Target policies are then trained on $\mathcal{D}_S$, each with a different architecture or feature mask. For each target policy we also sample an unlabeled dataset $\mathcal{D}_T$ of impressions.

Reward models. For each target policy, we train a reward model on $\mathcal{D}_S$. We compare two weighting strategies:

1. **No domain adaptation:** Unweighted strategy.
2. **Domain adaptation:** Importance weighting with $\widehat{w}(a) = \frac{n(a)/n_T}{\widetilde{n}(a)/n_S}$ where $n_S = |\mathcal{D}_S|$, $n_T = |\mathcal{D}_T|$, $n(a)$ and $\widetilde{n}(a)$ are ad counts in $\mathcal{D}_T$ and $\mathcal{D}_S$.

All models minimize the weighted cross-entropy loss.

5 Results

Figure 1 reports the estimation error of each method, defined as

$$\text{error} = \frac{|\, V_{\text{GT}} - \widehat{V}\,|}{V_{\text{GT}}}.$$

Doubly robust (DR) estimator with domain adaptation achieves the lowest error overall, confirming its robustness to distribution shift. In contrast, the **direct method (DM)** on both the source and target domains performs comparably, while **SNIPS** improves upon DM but does not reach the accuracy of DR. Importantly, domain adaptation consistently improves on unweighted DR, yielding the most accurate estimates across all settings.

6 Conclusions

We present a simulation framework to systematically study offline evaluation of ranking policies under distribution shift. Our experiments demonstrate that domain-adaptive reward models effectively reduce estimation bias. The doubly robust estimator with domain adaptation consistently achieves the lowest estimation error, outperforming both direct methods

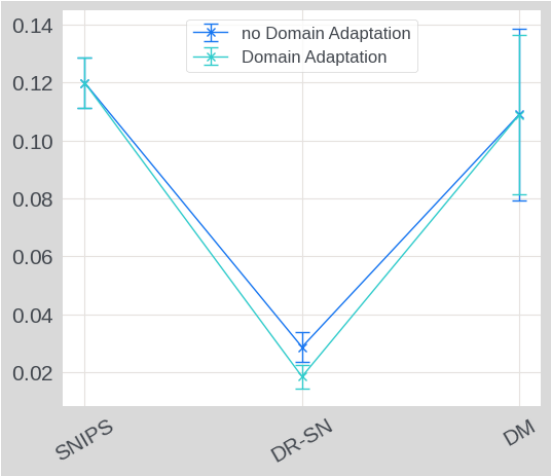

Figure 1: Comparison of evaluation methods with and without domain adaptation. DM: direct method, SNIPS: self normalized IPS, DR-SN: doubly robust with self-normalized weights.

and importance sampling baselines. While the Direct Method (DM) shows minimal sensitivity to domain adaptation, it performs slightly worse than the doubly robust approach. SNIPS demonstrates the poorest performance among all methods tested, performing worse than both DM and doubly robust estimators.

References

[1] M. A. Bayir, M. Xu, Y. Zhu, and Y. Shi. Genie: An open box counterfactual policy estimator for optimizing sponsored search marketplace. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*, pages 465–473, 2019.

[2] A. Biswas, T. T. Pham, M. Vogelsong, B. Snyder, and H. Nassif. Seeker: Real-time interactive search. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2867–2875, 2019.

[3] M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

[4] M. A. Radwan, Q. Lanners, J. Zhang, S. Karakulak, H. Nassif, and M. A. Bayir. Counterfactual evaluation of ads ranking models through domain adaptation. In *Proceedings of the Workshops at the 18th ACM Conference on Recommender Systems (RecSys 2024)*, 2024.

[5] A. L. Strehl, J. Langford, S. M. Kakade, and L. Li. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23 (NeurIPS 2010)*, pages 2217–2225, 2010.

[6] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pages 3231–3239, 2015.

[7] S. Zeng, M. A. Bayir, J. J. P. III, D. Charles, and E. Kıcıman. Causal transfer random forest: Combining logged data and randomized experiments for robust prediction. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*, pages 758–767, 2021.