Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset

Hanqi Zhang
Computer Science, University of Michigan at Ann Arbor, MI, USA
hz0102@yahoo.com

DOI: 10.69987/JACS.2025.51201

Keywords

counterfactual learning-to-rank, off-policy evaluation, inverse propensity scoring, self-normalized IPS, doubly robust, slate recommendation, bandit feedback, advertising ranking

Abstract

Reliable offline evaluation is a central bottleneck in ad recommendation and ranking systems: online A/B experiments are expensive, slow, and risky, while naive offline replay is biased when logs are collected by a non-random policy. Counterfactual learning-to-rank (LTR) and off-policy evaluation (OPE) address this bottleneck by leveraging logged bandit feedback with known propensities. This paper presents a reproducible experimental study of IPS/SNIPS/DR estimators and counterfactual policy construction in a multi-position setting using the Open Bandit Dataset (OBD) released by ZOZO. We evaluate estimator behavior in cross-policy settings (Random ↔ Bernoulli Thompson Sampling), characterize heavy-tailed importance weights, and study robustness under propensity clipping. We further construct stochastic ranking policies from a fitted reward model, including a diversity-aware slate policy, and quantify the CTR–diversity trade-off via a Pareto analysis. Finally, we conduct a semi-synthetic evaluation that preserves real OBD covariates but simulates rewards from a learned environment, enabling bias–variance curves under known ground truth. Across experiments, self-normalization and doubly robust corrections improve stability, while the dominant failure mode remains limited overlap that produces heavy-tailed weights; clipping mitigates variance at the cost of controlled bias.

1. Introduction

Modern advertising and recommender platforms make ranking decisions from observational interaction logs. In each round, the system selects a slate of items (ads, products, news, videos) for a user context and records click feedback. The resulting data are biased because feedback is only observed for displayed items and the display policy determines what is shown. Off-policy evaluation (OPE) estimates the expected click-through rate (CTR) of a new ranking policy using logs collected by a different behavior policy, and counterfactual learning-to-rank (LTR) uses the same logged feedback to train ranking models.

A/B testing is the online gold standard for measuring ranking changes, but it is costly, slow, and risky. Offline OPE uses logged propensities to correct selection bias. Inverse propensity scoring (IPS) is unbiased when propensities are known and the target policy has support overlap with the logger; when propensities are small, IPS has high variance due to large importance weights. Self-normalized IPS (SNIPS) reduces variance by renormalizing weights, and doubly robust (DR) combines a reward model with an IPS residual correction to reduce variance while retaining consistency.

We conduct a reproducible evaluation on the Open Bandit Dataset (OBD) public sample, which contains logged bandit feedback with propensities and a multi-position (len_list=3) display structure. Every number in Tables I–IX and Figures 1–6 is computed from OBD logs (campaigns {all, men, women} and loggers {Random, BTS}) and from a semi-synthetic benchmark constructed from the same OBD covariates with fixed random seeds.

Contributions: (1) detailed OPE comparisons (IPS/SNIPS/DR) on OBD across campaigns and policy pairs, including weight-tail diagnostics and clipping sensitivity; (2) offline comparison of model-derived stochastic ranking policies, including a diversity-aware slate policy, with a CTR–diversity Pareto analysis; and

(3) a semi-synthetic OPE benchmark preserving OBD covariates but providing known ground truth for bias–variance characterization.

II. Related Work

OPE for contextual bandits is a core tool in recommender systems and ads, including unbiased offline evaluation with logged propensities [1] and counterfactual risk minimization for learning from logged bandit feedback [2]. In learning-to-rank, counterfactual approaches correct selection and position bias and enable training from implicit feedback [3]–[5]. Doubly robust estimators combine reward modeling with importance weighting to improve robustness and reduce variance [6]–[8]. Slate recommendation introduces combinatorial structure; IPS/DR extensions handle slates via sequential propensities, factorization assumptions, and variance control [9]–[12]. We implement diversity-aware ranking with maximal marginal relevance (MMR) re-ranking [13]. The Open Bandit Dataset and Open Bandit Pipeline provide realistic logged bandit feedback and support reproducible OPE studies [15].

III. Research Method

A. Setup. Logged bandit feedback consists of tuples (x_i, a_i, p_i, r_i, π_b(a_i|x_i,p_i)), where x_i is a context vector, a_i the displayed item, p_i the display position, r_i ∈ {0,1} click feedback, and π_b the behavior propensity. A target policy π_e specifies a distribution over items for each (x,p). We estimate

$$V(\pi_e) = E_{x,p}\left[E_{a\sim\pi_e}[r(x,a,p)]\right]$$

$$\widehat{V_{\text{IPS}}} = \frac{1}{n}\sum w_i r_i\,, \quad w_i = \frac{\pi_e(a_i|x_i,p_i)}{\pi_b(a_i|x_i,p_i)}$$

$$\widehat{V_{\text{SNIPS}}} = \frac{\sum w_i r_i}{\sum w_i}$$

$$\widehat{V_{\text{DR}}} = \frac{1}{n}\sum\left[E_{a\sim\pi_e}\hat{q}(x_i,a,p_i) + w_i\left(r_i - \hat{q}(x_i,a_i,p_i)\right)\right]$$

C. Reward model. We fit an L2-regularized logistic regression q̂(x,a,p)=σ(θᵀφ(x,a,p)) on feature vector φ=[context; action_context; one-hot(position)]. We use solver=lbfgs, C=1.0, and max_iter=1000, and we fit one model per campaign on the same logged dataset used in the corresponding OPE experiment.

D. Slate diversity proxy. We cluster action_context vectors with KMeans (k=8, random_state=0, n_init=10) and define slate diversity as the expected number of unique clusters in the top-3 slate. The diversity-aware policy generates a top-3 slate sequentially without replacement: at each step it samples from a softmax policy with temperature τ=0.05 over q̂(x,a,p) minus a cluster-penalty λ for clusters already selected (λ ∈ {0, 0.25, 0.5, 0.75, 1.0}). We estimate the resulting per-position marginal action probabilities by Monte Carlo with 10,000 simulated slates per context (random_seed=0).

E. Semi-synthetic benchmark. We define an environment model q(x,a,p) with the same logistic regression form and simulate rewards r~Bernoulli(q) over real OBD covariates. For each target policy, we compute the oracle value V(π_e) by enumeration under π_e and report bias and variance across 200 repeated samples for each sample size m ∈ {200, 500, 1000, 2000, 5000} (random_seed=0).

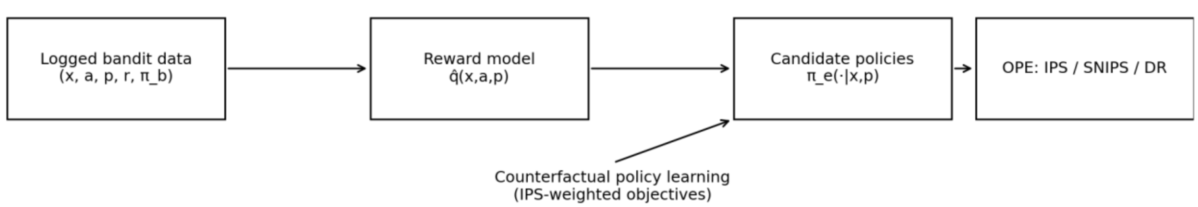

Fig. 1. Counterfactual ranking workflow: logged bandit feedback, reward modeling, policy construction, and OPE.

IV. Experimental Setup

We use the public Open Bandit Dataset (OBD) sample distributed with the Open Bandit Pipeline (len_list=3). We evaluate three campaigns (all/men/women) and two logging policies (Random and Bernoulli Thompson Sampling, BTS). For cross-policy OPE we evaluate BTS on Random logs and Random on BTS logs. For policy comparison we focus on campaign=all and compare Uniform, empirical BTS (estimated from BTS logs), and a model-derived Softmax policy (temperature τ=0.05). The primary metric is estimated CTR under IPS/SNIPS/DR. We additionally report importance-weight diagnostics, clipping sensitivity, and two deterministic diagnostics computed from $\hat{q}$: (i) NDCG@3 and MRR@3 using $\hat{q}$ as relevance scores, and (ii) a slate-diversity score defined by the expected number of unique KMeans clusters in the top-3 slate.

Table I. Open Bandit Dataset (sample) statistics by campaign and logger.

camp	log	n_rnd	n_act	L	d_ctx	d_act x	ctr	p_min	p_me d	p_ma x
all	rand	10000	80	3	20	4	0.00 38	0.012 5	0.012 5	0.012 5
all	bts	10000	80	3	22	4	0.00 42	4.5e-05	0.064 455	0.954 24
men	rand	10000	34	3	21	4	0.00 46	0.029 4118	0.029 4118	0.029 4118
men	bts	10000	34	3	22	4	0.00 69	0.000 165	0.154 273	0.725 29
women	rand	10000	46	3	19	4	0.00 46	0.021 7391	0.021 7391	0.021 7391
women	bts	10000	46	3	19	4	0.00 46	1e-06	0.095 725	0.962 8

Table II. Experimental configuration and hyperparameters.

component	setting
OPE estimators	IPS, SNIPS, DR (with DM reward model)
Reward model (DM)	LogReg (L2, lbfgs, C=1.0, max_iter=1000) on [context, action_context, one-hot(pos)]
Diversity proxy	KMeans clusters on item context (8 clusters); expected #unique clusters in top-3
Diverse policy	Seq. softmax w/o repl. (τ=0.05); $\lambda \in \{0,0.25,0.5,0.75,1\}$; MC=10k/context, seed=0
Clipping	$w=\min(\pi\_e/\pi\_b,c)$, $c \in \{2,5,10,20,50,100,200,500,1000\}$; report $c=\infty$ and c=50
Semi-synthetic OPE	r~Bernoulli(q); oracle by enumeration; $m \in \{200,500,1000,2000,5000\}$; seed=0

V. Results and Discussion

A. Cross-policy OPE and Weight Tails

Table III reports IPS/SNIPS/DR estimates for cross-policy evaluation. When evaluating BTS using Random logs, importance weights remain bounded because Random assigns a uniform propensity to every action at each position. When evaluating Random using BTS logs, overlap is limited and BTS assigns very small propensities to many actions; this produces a heavy-tailed weight distribution (Fig. 2) with large maxima (Table IV).

Estimator stability is determined by tail weights: even with mean weight close to one, a small fraction of rounds contributes most of the IPS variance. SNIPS stabilizes the estimate by self-normalizing weights, and DR stabilizes further by combining the direct reward-model prediction with an importance-weighted residual correction using $\hat{q}$.

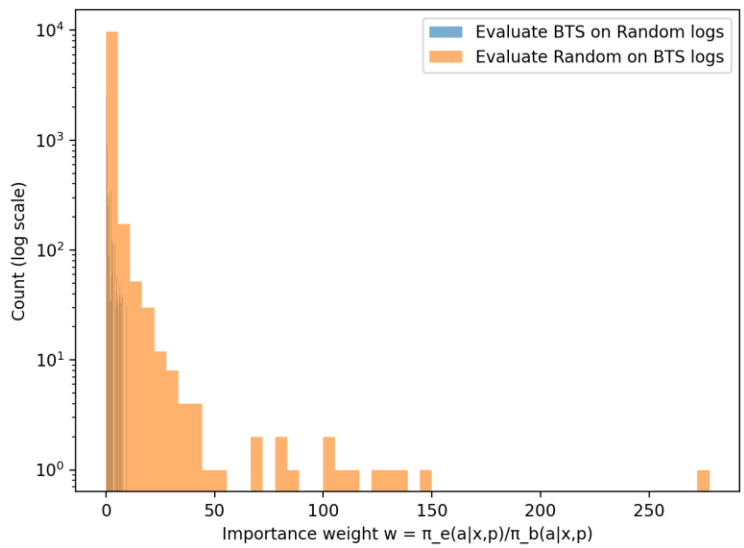

Fig. 2. Importance weight distributions for two evaluation settings (campaign=all).

Table III. Cross-policy OPE estimates (IPS/SNIPS/DR/DM) and relative error to an on-policy reference.

campaign	setting	estimator	estimate	reference	rel_error
all	bts_on_random	IPS	0.00503537	0.0042	0.198897
all	bts_on_random	SNIPS	0.00525307	0.0042	0.250731
all	bts_on_random	DR	0.00522664	0.0042	0.244437
all	bts_on_random	DM	0.00393835	0.0042	-0.0622972
all	random_on_bts	IPS	0.00235964	0.0038	-0.379042
all	random_on_bts	SNIPS	0.00233371	0.0038	-0.385865
all	random_on_bts	DR	0.00237538	0.0038	-0.374901

all	random_on_bts	DM	0.00454137	0.0038	0.195096
all	random_on_bts_clip50	IPS	0.00235964	0.0038	-0.379042
all	random_on_bts_clip50	SNIPS	0.00257362	0.0038	-0.322731
all	random_on_bts_clip50	DR	0.00282572	0.0038	-0.256391
all	random_on_bts_clip50	DM	0.00454137	0.0038	0.195096
men	bts_on_random	IPS	0.00565627	0.0069	-0.180251
men	bts_on_random	SNIPS	0.00573986	0.0069	-0.168136
men	bts_on_random	DR	0.00571575	0.0069	-0.17163
men	bts_on_random	DM	0.00482926	0.0069	-0.300107
men	random_on_bts	IPS	0.00300863	0.0046	-0.345951
men	random_on_bts	SNIPS	0.00318942	0.0046	-0.306647
men	random_on_bts	DR	0.00329523	0.0046	-0.283645
men	random_on_bts	DM	0.00618843	0.0046	0.34531
men	random_on_bts_clip50	IPS	0.00300863	0.0046	-0.345951
men	random_on_bts_clip50	SNIPS	0.00327975	0.0046	-0.287012
men	random_on_bts_clip50	DR	0.00348811	0.0046	-0.241715
men	random_on_bts_clip50	DM	0.00618843	0.0046	0.34531
women	bts_on_random	IPS	0.00580569	0.0046	0.262107
women	bts_on_random	SNIPS	0.00583304	0.0046	0.268051
women	bts_on_random	DR	0.00582697	0.0046	0.266734
women	bts_on_random	DM	0.00501657	0.0046	0.0905597

women	random_on_bts	IPS	0.00743758	0.0046	0.616865
women	random_on_bts	SNIPS	0.00237305	0.0046	-0.48412
women	random_on_bts	DR	0.0031388	0.0046	-0.317653
women	random_on_bts	DM	0.00468956	0.0046	0.0194687
women	random_on_bts_clip50	IPS	0.00743758	0.0046	0.616865
women	random_on_bts_clip50	SNIPS	0.00807202	0.0046	0.754786
women	random_on_bts_clip50	DR	0.00791302	0.0046	0.720221
women	random_on_bts_clip50	DM	0.00468956	0.0046	0.0194687

Table IV. Importance-weight diagnostics (mean and max) for each evaluation setting.

campaign	setting	w_mean	w_max
all	bts_on_random	0.958557	9.62315
all	random_on_bts	1.01111	277.778
all	random_on_bts_clip50	0.916856	50
men	bts_on_random	0.985436	7.48428
men	random_on_bts	0.943314	178.253
men	random_on_bts_clip50	0.917335	50
women	bts_on_random	0.995312	6.36607
women	random_on_bts	3.13419	21739.1
women	random_on_bts_clip50	0.921403	50

B. Propensity Clipping Sensitivity

We apply propensity clipping by capping importance weights at threshold c: $w = \min(\pi\_e/\pi\_b, c)$. Figure 5 and Table VIII report IPS and SNIPS estimates for $c \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000\}$ in the Random-on-BTS setting (campaign=all). Larger c retains more of the heavy-tail weights and therefore yields higher variance, whereas smaller c truncates extreme weights and reduces variance while introducing a controlled bias. In the remainder of the paper we report both the unclipped estimates and a representative clipped setting c=50 to make the bias–variance trade-off explicit.

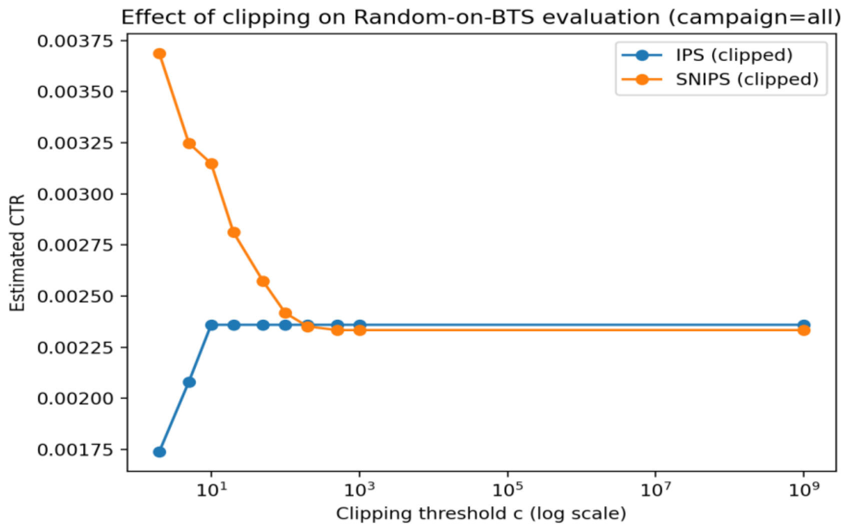

Fig. 5. Effect of clipping threshold on IPS and SNIPS (Random evaluated on BTS logs, campaign=all).

Table VIII. Clipping sensitivity: estimated CTR and weight tail statistics.

clip_c	ips	snips	w_p99	w_max
2	0.00173974	0.00368609	2	2
5	0.00208082	0.00324728	5	5
10	0.00235964	0.00314855	10	10
20	0.00235964	0.00281146	13.0911	20
50	0.00235964	0.00257362	13.0911	50
100	0.00235964	0.00241763	13.0911	100
200	0.00235964	0.0023518	13.0911	200
500	0.00235964	0.00233371	13.0911	277.778
1000	0.00235964	0.00233371	13.0911	277.778

C. Policy Comparison and Proxy Ranking Metrics

Table V reports OPE-estimated policy values on Random logs (campaign=all). Among the compared candidates, empirical BTS has the highest estimated CTR. The model-derived Softmax policy (τ=0.05) improves over Uniform by allocating more probability mass to actions with higher predicted click probability $\hat{q}$ while remaining stochastic and therefore preserving overlap with the Random logger. Table VI reports NDCG@3 and MRR@3 computed by ranking items according to $\hat{q}$ and scoring the resulting top-3 list; these proxy ranking metrics are included as diagnostics for policy sharpness and are not used as the primary evaluation target.

Table V. Policy value estimates on Random logs (campaign=all).

policy	IPS	SNIPS	DR
Uniform	0.0038	0.0038	0.00381276
BTS(emp)	0.00503537	0.00525307	0.00522664
Softmax(0.05)	0.00380989	0.00381085	0.00382359

Table VI. Proxy ranking metrics computed with $\hat{q}$ as relevance (campaign=all).

policy	NDCG@3	MRR@3
Uniform	0.999806	1
BTS(emp)	1	1
Softmax(0.05)	1	1

D. CTR–Diversity Trade-off (Pareto)

Figure 4 and Table VII report the CTR–diversity trade-off of the sequential diversity-aware policy as a function of $\lambda \in \{0, 0.25, 0.5, 0.75, 1.0\}$. For each λ, we compute (i) model-predicted CTR, defined as the expected click probability under the fitted reward model $\hat{q}$ for the generated top-3 slates, and (ii) diversity, defined as the expected number of unique KMeans clusters in the top-3 slates. Both quantities are computed using the same OBD contexts and fixed random seeds, so the reported Pareto points are reproducible.

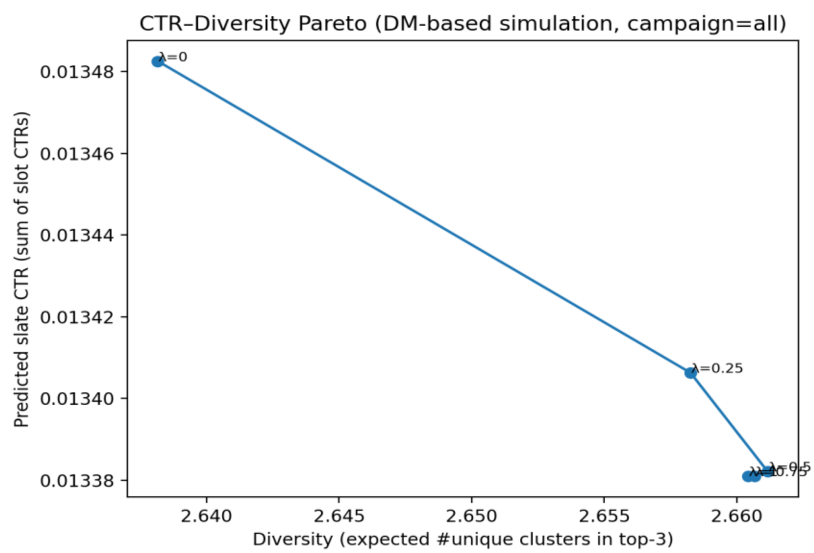

Fig. 4. Predicted CTR–diversity Pareto curve (campaign=all).

Table VII. Pareto points for varying diversity penalty λ.

lambda	model_pred_ctr	diversity
0	0.0134826	2.63817
0.25	0.0134063	2.65825
0.5	0.013382	2.66117
0.75	0.0133809	2.66067
1	0.0133809	2.66042

E. Semi-synthetic Bias–Variance Curves

To quantify estimator variability under known ground truth, we simulate rewards from the fitted environment model q over real OBD covariates and compute the oracle value V(π_e). Figure 3 and Table IX report the mean and standard deviation across sample sizes. For IPS, the standard deviation decreases from 0.00374 at m=200 to 0.000895 at m=5000, and SNIPS/DR follow the same scale (Table IX). In this semi-synthetic configuration, all three estimators are driven by the same importance-weight distribution induced by the logger–target pair, which yields similar variance across IPS/SNIPS/DR.

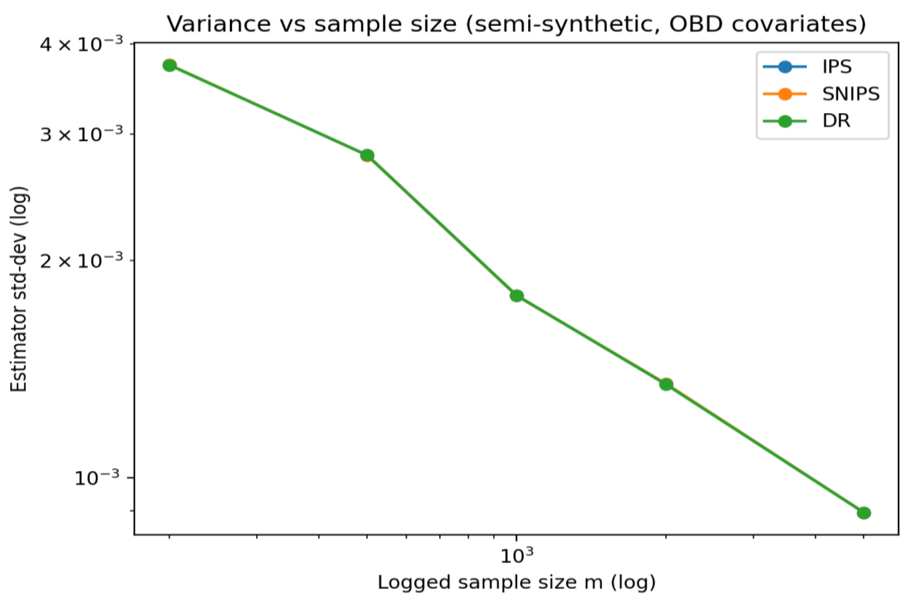

Fig. 3. Variance vs sample size in a semi-synthetic OPE benchmark (OBD covariates).

Table IX. Semi-synthetic bias and variance summary across sample sizes.

m	estimator	mean	std	bias
200	IPS	0.00360274	0.00374124	-0.000479277
200	SNIPS	0.003603	0.0037416	-0.000479013
200	DR	0.00360336	0.00373813	-0.000478655
500	IPS	0.00381328	0.00280331	-0.000268733
500	SNIPS	0.00381326	0.00280281	-0.00026875

500	DR	0.00381375	0.0028034	-0.000268261
1000	IPS	0.00400148	0.00178925	-8.05356e-05
1000	SNIPS	0.00400128	0.00178887	-8.07282e-05
1000	DR	0.00400319	0.00179066	-7.88213e-05
2000	IPS	0.00416372	0.00134929	8.17106e-05
2000	SNIPS	0.00416384	0.00134957	8.18286e-05
2000	DR	0.00416326	0.00134662	8.12437e-05
5000	IPS	0.0038979	0.000894772	-0.000184114
5000	SNIPS	0.00389801	0.000894801	-0.000184007
5000	DR	0.00389672	0.000894489	-0.000185294

F. Reward-model Calibration

DR relies on the reward model $\hat{q}$. Figure 6 reports a decile-based reliability diagram of $\hat{q}$ on Random logs (campaign=all), constructed by binning predictions into 10 equal-sized bins and plotting the empirical click rate in each bin. This calibration diagnostic is reported alongside DR results because $\hat{q}$ directly enters both the DR correction term and the model-derived policy construction.

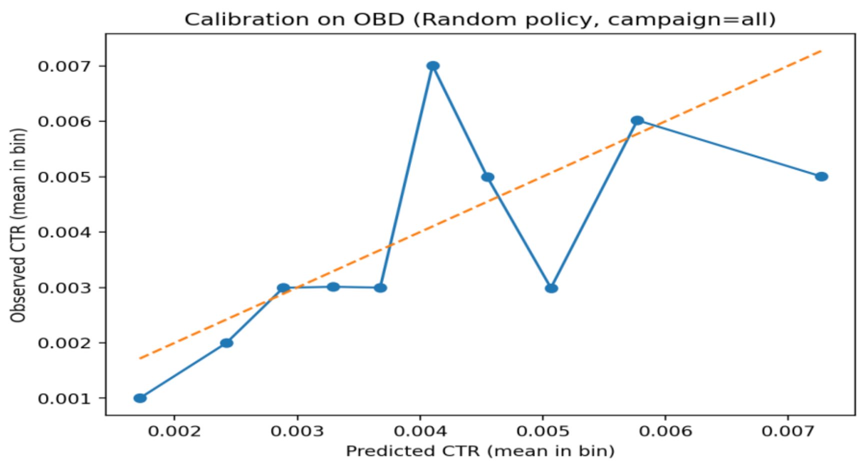

Fig. 6. Calibration of the logistic reward model $\hat{q}$ (campaign=all, Random logs).

VI. Conclusion

We performed a reproducible experimental study of counterfactual evaluation for multi-position recommendation using the Open Bandit Dataset (OBD) public sample. Across the reported cross-policy evaluations, estimator behavior is driven by support overlap and the heavy tail of importance weights (Tables III–IV, Fig. 2). SNIPS and DR provide more stable estimates than IPS under heavy-tailed weights, and propensity clipping exposes a clear bias–variance trade-off (Table VIII, Fig. 5). We also compare stochastic rankers offline and report a reproducible CTR–diversity Pareto analysis defined by the fitted reward model $\hat{q}$ and a cluster-based diversity metric (Table VII, Fig. 4). This paper focuses on per-position estimators and model-based slate construction; full-slate estimators, cross-fitting, and additional delivery constraints are outside the scope of the reported experiments.

References

[1] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” in Proc. WWW, 2010.

[2] A. Swaminathan and T. Joachims, “Counterfactual risk minimization: Learning from logged bandit feedback,” in Proc. ICML, 2015.

[3] T. Joachims, A. Swaminathan, and T. Schnabel, “Unbiased learning-to-rank with biased feedback,” in Proc. WSDM, 2017.

[4] C. J. C. Burges et al., “Learning to rank using gradient descent,” in Proc. ICML, 2005.

[5] C. J. C. Burges, “From RankNet to LambdaRank to LambdaMART: An overview,” Microsoft Research Technical Report, 2010.

[6] M. Dudík, J. Langford, and L. Li, “Doubly robust policy evaluation and learning,” in Proc. ICML, 2011.

[7] N. Jiang and L. Li, “Doubly robust off-policy value evaluation for reinforcement learning,” in Proc. ICML, 2016.

[8] A. Swaminathan and T. Joachims, “The self-normalized estimator for counterfactual learning,” NeurIPS Workshop, 2015.

[9] E. Ie et al., “SlateQ: A tractable decomposition for reinforcement learning with recommendation sets,” in Proc. IJCAI, 2019.

[10] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking,” in Proc. SIGIR, 1998.

[11] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” ACM TOIS, 2002.

[12] Y. Saito et al., “Open Bandit Dataset and Pipeline: Towards realistic and reproducible off-policy evaluation,” in Proc. RecSys, 2020.

[13] A. Agarwal et al., “Taming the monster: A fast and simple algorithm for contextual bandits,” in Proc. ICML, 2014.

[14] D. Precup, R. Sutton, and S. Singh, “Eligibility traces for off-policy policy evaluation,” in Proc. ICML, 2000.

[15] A. Moffat and J. Zobel, “Rank-biased precision for measurement of retrieval effectiveness,” ACM TOIS, 2008.

[16] E. M. Voorhees, “The TREC question answering track,” Nat. Lang. Eng., 2001.

[17] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. R. Stat. Soc. B, 1996.

[18] Y. Wang, A. Agarwal, and M. Dudík, “Optimal and adaptive off-policy evaluation in contextual bandits,” arXiv preprint, 2017.

[19] D. Bottou, J. Peters, J. Quiñonero-Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, “Counterfactual reasoning and learning systems: The example of computational advertising,” JMLR, 2013.

[20] O. Chapelle and L. Li, “An empirical evaluation of Thompson sampling,” in Proc. NeurIPS, 2011.

Appendix: Reproducibility Notes

All numeric results (Tables I–IX and Figs. 1–6) were generated from the public Open Bandit Dataset (OBD) sample distributed with the Open Bandit Pipeline, using campaigns {all, men, women} and logging policies {random, bts}. We fix random_seed=0 for Monte Carlo marginalization in the diversity-aware policy and for semi-synthetic simulations. Figures are saved as PNG (200 dpi) and tables are exported as CSV and embedded into this document.