A Simulation Framework for Off-Policy Evaluation of Ranking Models

Dan Qiao[1], Mohsen Malmir[2], Houssam Nassif[2], Murat Bayir[2]
[1]University of California, Santa Barbara [2]Meta

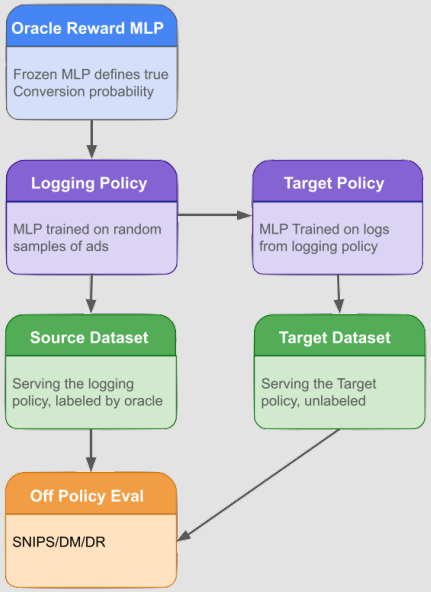
Introduction

- **Ranking models** drive search and recommendation systems, but testing new policies online is costly and risky.
- **Off-policy evaluation** (OPE) enables offline assessment of new ranking policies using logs collected under production models.
- Common estimators include Direct Method (DM), Inverse Propensity Score (IPS), Self-Normalized IPS (SNIPS), and Doubly Robust (DR) methods.
- **Our prior work** introduced a domain-adapted reward model to reduce bias when learning from source-domain logs.
- **This study** builds on that foundation using a controlled simulation environment to systematically compare OPE estimators and analyze how domain adaptation affects policy evaluation.

Simulation Environment

- **Goal**: Build a controlled, reproducible testbed that mimics a production ads-ranking system while retaining ground-truth access.
- **Oracle**: A fixed MLP maps each ad's feature vector to its true conversion probability $r^*(a) \in [0,1]$; this defines the environment and enables exact policy value computation.
- **Logging policy**: Point-wise MLP trained on synthetic interactions $r \sim Bernoulli(r^*(a))$
- **Domain (target) policy**: Another MLP trained on logs from the logging policy, simulating models retrained from production data.
- May differ in architecture or feature subsets, inducing covariate and policy shift.

Simulation Overview

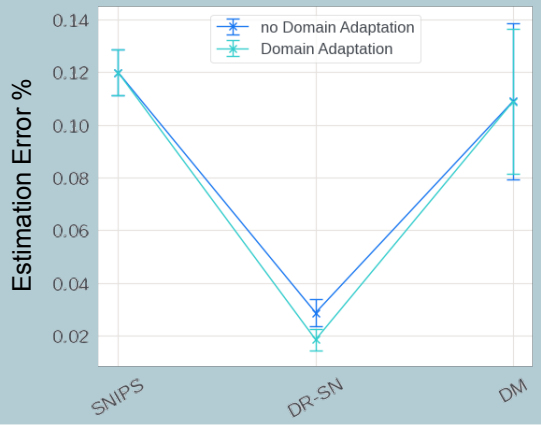


Evaluation Methods

- **Ground Truth** (Oracle Benchmark)
- Uses the true conversion probabilities r*(a) from the oracle to compute the exact policy value.
- **Direct Method (DM)**: Trains a reward model $\hat{r}(a)$ on source-domain data: $\hat{V}_{DM}(\pi_T) = \sum_a P_T(a)\hat{r}(a)$
- **Self-Normalized Inverse Propensity Scoring** (SNIPS): Reuses logged data from the source policy and reweights samples by target-to-source propensity ratio: $w_i = P_T(a_i)/P_S(a_i)$
- **Doubly Robust with Self-Normalization** (DR-SNIPS), Combines DM's reward model (low variance) and SNIPS's correction term (unbiasedness).

Results

- **Doubly Robust (DR-SNIPS)** with domain adaptation achieves the lowest estimation error, confirming strong robustness under distribution shift.



Conclusions

- We developed a **controlled simulation framework** that enables ground-truth benchmarking of off-policy estimators for ranking models.
- **Domain-adaptive reward models** effectively reduce estimation bias and improve robustness to distribution shift.
- The **doubly robust estimator with domain adaptation** achieves the most accurate and stable offline evaluation across all tested settings.

Contacts

[1]danqiao@ucsb.edu
[2]{mohsenm, houssamn, mbayir}@meta.com