Unifying Online and Counterfactual Learning to Rank

A Novel Counterfactual Estimator that Effectively Utilizes Online Interventions

Harrie Oosterhuis[1], Maarten de Rijke[2,3]

March 11, 2021

Radboud University[1], University of Amsterdam[2], Ahold Delhaize[3]
harrie.oosterhuis@ru.nl, derijke@uva.nl
https://twitter.com/HarrieOos

Introduction

Unbiased Learning to Rank:

- **Learning from clicks** while correcting for interaction biases.

Online Learning to Rank:

- Correct for bias by **randomizing** results through **online interventions**.

Counterfactual Learning to Rank:

- Infer a **model of bias**, use it to correct when learning from **historical click data**.

Position Bias, Item-Selection Bias and Trust Bias

Position Bias:

- Users are more likely to **examine higher ranked** results (Craswell et al., 2008).
- **Solution:** **Inverse Propensity Scoring** (Joachims et al., 2017).

Item-Selection Bias:

- Users **cannot** examine items that are **not displayed** (Ovaisi et al., 2020).
- **Solution:** Policy-Aware Propensities (Oosterhuis and de Rijke, 2020).

Trust Bias:

- Users are more likely to **incorrectly presume relevance** at higher ranked results (Agarwal et al., 2019).
- **Solution:** Apply **inverse transformation** (Vardasbi et al., 2020).

Intervention-Oblivious Estimator

Starting assumption: clicks follow an **affine model**, for item d displayed at rank k:

$$P(C = 1 \mid d, k) = \alpha_k P(R = 1 \mid d) + \beta_k.$$

Starting assumption: clicks follow an **affine model**, for item d displayed at rank k:

$$P(C = 1 \mid d, k) = \alpha_k P(R = 1 \mid d) + \beta_k.$$

We **condition** the click probability on the **logging policy** π:

$$\begin{aligned} P(C = 1 \mid d, \pi) &= \sum_{k=1}^{K} \pi(k \mid d)(\alpha_k P(R = 1 \mid d) + \beta_k) \\ &= \mathbb{E}_k[\alpha_k \mid d, \pi] P(R = 1 \mid d) + \mathbb{E}_k[\beta_k \mid d, \pi]. \end{aligned}$$

Intervention-Oblivious Estimator

Starting assumption: clicks follow an **affine model**, for item d displayed at rank k:

$$P(C = 1 \mid d, k) = \alpha_k P(R = 1 \mid d) + \beta_k.$$

We **condition** the click probability on the **logging policy** π:

$$\begin{aligned} P(C = 1 \mid d, \pi) &= \sum_{k=1}^{K} \pi(k \mid d)(\alpha_k P(R = 1 \mid d) + \beta_k) \\ &= \mathbb{E}_k[\alpha_k \mid d, \pi] P(R = 1 \mid d) + \mathbb{E}_k[\beta_k \mid d, \pi]. \end{aligned}$$

The **intervention-oblivious estimator** is based on the **inverse** of this transformation:

$$P(R = 1 \mid d) = \frac{P(C = 1 \mid d, \pi) - \mathbb{E}_k[\beta_k \mid d, \pi]}{\mathbb{E}_k[\alpha_k \mid d, \pi]}.$$

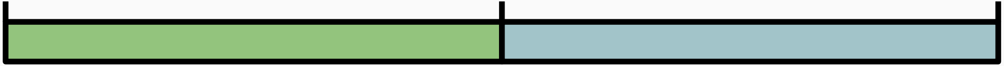
Intervention-Oblivious Estimator: Visualization

deployed: π_1

$$\mathbb{E}_k[\alpha_k \mid d, \pi_1] = 0.1$$
$$\mathbb{E}_k[\alpha_k \mid d', \pi_1] = 0.5$$
$$\mathbb{E}_k[\beta_k \mid d, \pi_1] = 0$$
$$\mathbb{E}_k[\beta_k \mid d', \pi_1] = 0$$

deployed: π_2

$$\mathbb{E}_k[\alpha_k \mid d, \pi_2] = 0.5$$
$$\mathbb{E}_k[\alpha_k \mid d', \pi_2] = 0.5$$
$$\mathbb{E}_k[\beta_k \mid d, \pi_2] = 0$$
$$\mathbb{E}_k[\beta_k \mid d', \pi_2] = 0$$



$$\frac{1}{\mathbb{E}_k[\alpha_k | d, \pi_1]} = 10$$
$$\frac{1}{\mathbb{E}_k[\alpha_k | d', \pi_1]} = 2$$

$$\frac{1}{\mathbb{E}_k[\alpha_k | d, \pi_2]} = 2$$
$$\frac{1}{\mathbb{E}_k[\alpha_k | d', \pi_2]} = 2$$

Intervention-Aware Estimator

Intervention-Aware Estimator

Due to **interventions** the logging policy is **updated** during data-gathering.

Let Π contain **all logging policies** for each timestep t:

$$\Pi = \{\pi_1, \pi_2, \ldots\}.$$

Due to **interventions** the logging policy is **updated** during data-gathering.

Let Π contain **all logging policies** for each timestep t:

$$\Pi = \{\pi_1, \pi_2, \ldots\}.$$

We can **condition** the click probability on the **set** Π:

$$\begin{aligned} P(C = 1 \mid d, \Pi) &= \frac{1}{|\Pi|} \sum_{\pi_t \in \Pi} \sum_{k=1}^{K} \pi_t(k \mid d)(\alpha_k P(R = 1 \mid d) + \beta_k) \\ &= \mathbb{E}_k[\alpha_k \mid d, \Pi] P(R = 1 \mid d) + \mathbb{E}_k[\beta_k \mid d, \Pi]. \end{aligned}$$

Due to **interventions** the logging policy is **updated** during data-gathering.

Let Π contain **all logging policies** for each timestep t:

$$\Pi = \{\pi_1, \pi_2, \ldots\}.$$

We can **condition** the click probability on the **set** Π:

$$\begin{aligned} P(C = 1 \mid d, \Pi) &= \frac{1}{|\Pi|} \sum_{\pi_t \in \Pi} \sum_{k=1}^{K} \pi_t(k \mid d)(\alpha_k P(R = 1 \mid d) + \beta_k) \\ &= \mathbb{E}_k[\alpha_k \mid d, \Pi] P(R = 1 \mid d) + \mathbb{E}_k[\beta_k \mid d, \Pi]. \end{aligned}$$

The **intervention-aware estimator** is based on the **inverse**:

$$P(R = 1 \mid d) = \frac{P(C = 1 \mid d, \Pi) - \mathbb{E}_k[\beta_k \mid d, \Pi]}{\mathbb{E}_k[\alpha_k \mid d, \Pi]}.$$

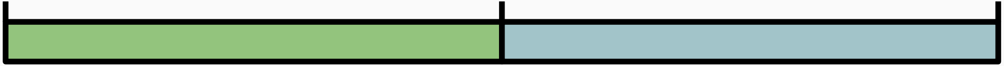
Intervention-Aware Estimator: Visualization

deployed: π_1

$$\mathbb{E}_k[\alpha_k \mid d, \pi_1] = 0.1$$
$$\mathbb{E}_k[\alpha_k \mid d', \pi_1] = 0.5$$
$$\mathbb{E}_k[\beta_k \mid d, \pi_1] = 0$$
$$\mathbb{E}_k[\beta_k \mid d', \pi_1] = 0$$

deployed: π_2

$$\mathbb{E}_k[\alpha_k \mid d, \pi_2] = 0.5$$
$$\mathbb{E}_k[\alpha_k \mid d', \pi_2] = 0.5$$
$$\mathbb{E}_k[\beta_k \mid d, \pi_2] = 0$$
$$\mathbb{E}_k[\beta_k \mid d', \pi_2] = 0$$



$$\frac{1}{\mathbb{E}_k[\alpha_k|d,\Pi]} = \frac{2}{\mathbb{E}_k[\alpha_k|d,\pi_1]+\mathbb{E}_k[\alpha_k|d,\pi_2]} = 3 + \frac{1}{3}$$
$$\frac{1}{\mathbb{E}_k[\alpha_k|d',\Pi]} = \frac{2}{\mathbb{E}_k[\alpha_k|d',\pi_1]+\mathbb{E}_k[\alpha_k|d',\pi_2]} = 2$$

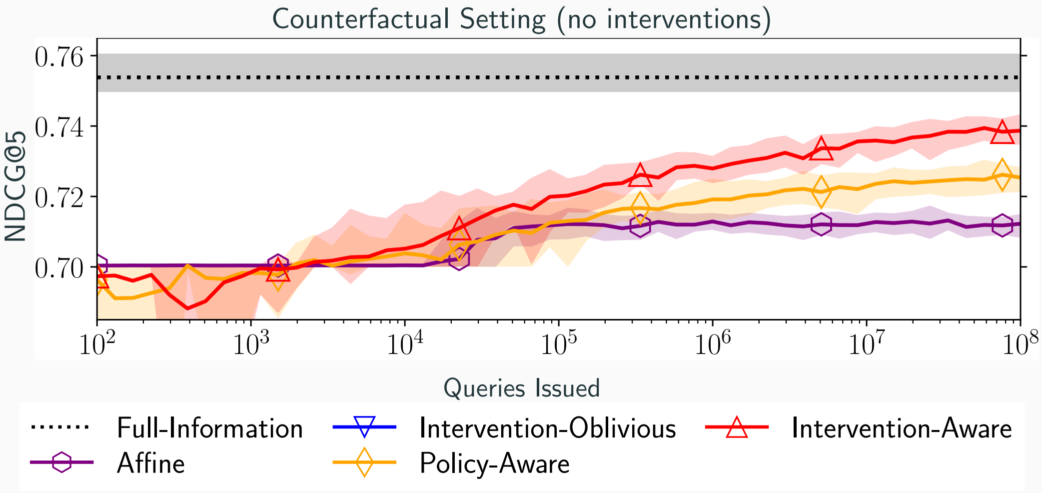
Experiments and Results

Experimental Setup

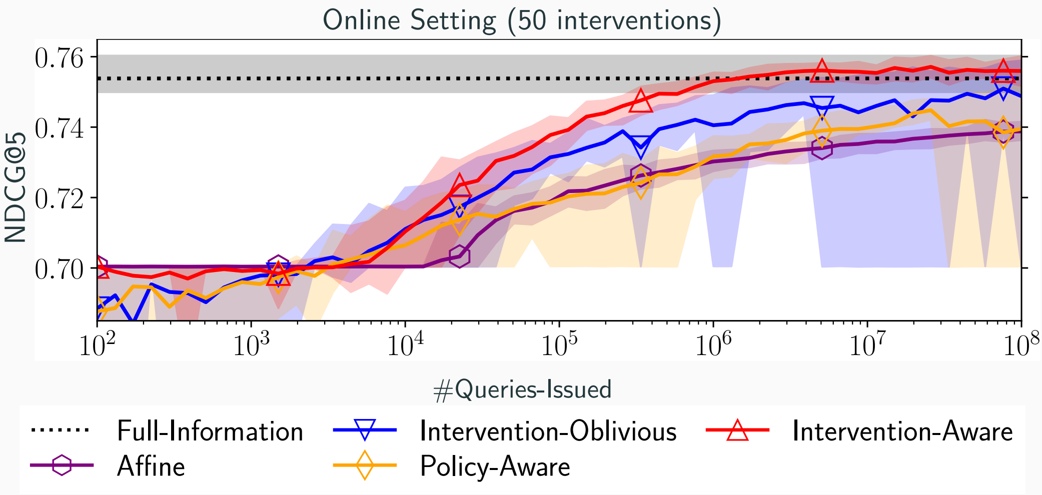
Semi-synthetic experiments on the Yahoo! Webscope dataset (Chapelle and Chang, 2011).

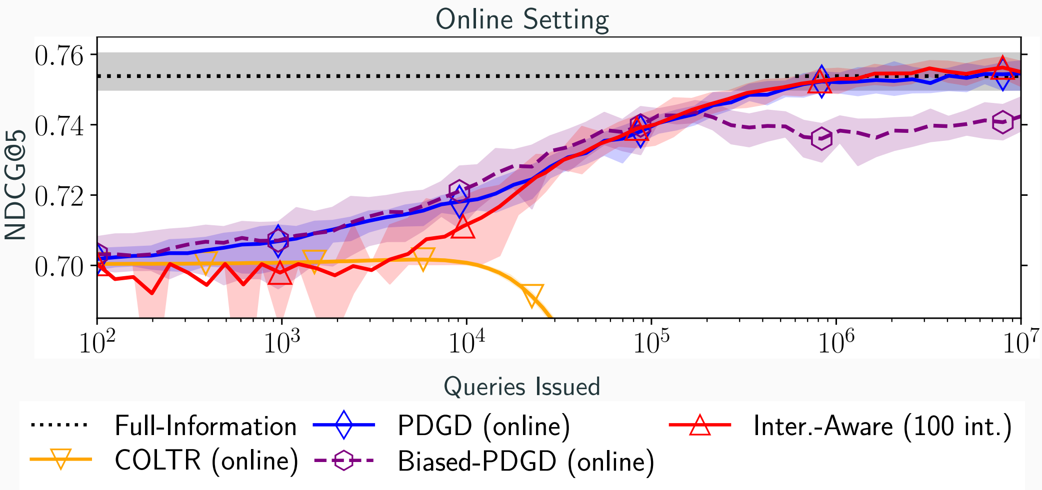
Affine top-5 click model based parameters inferred by Agarwal et al. (2019).

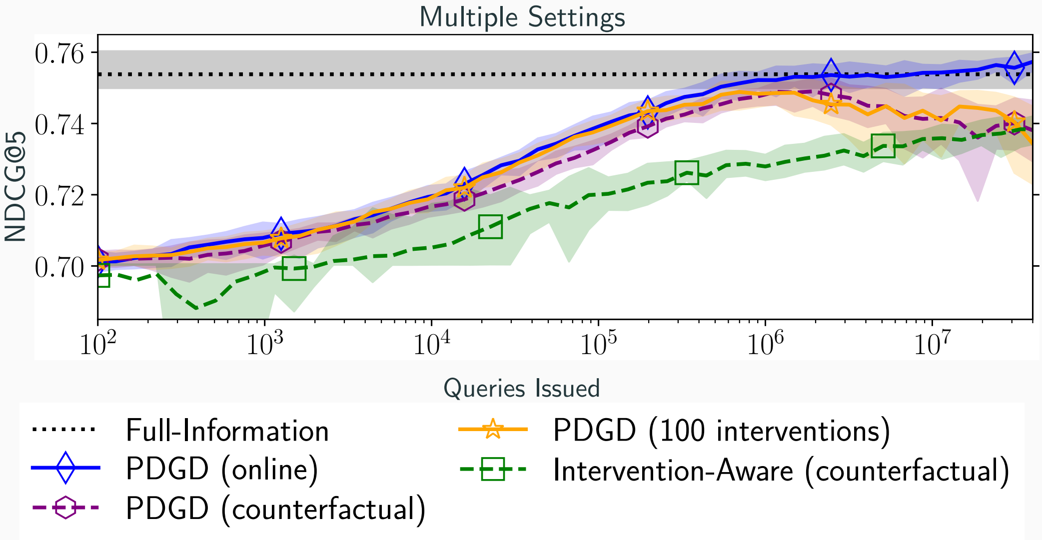
Both counterfactual and online experiments,
online interventions are spread evenly on a logarithmic scale.


Counterfactual Setting (no interventions)
NDCG@5
0.76
0.74
0.72
0.70
10^2
10^3
10^4
10^5
10^6
10^7
10^8
Queries Issued
Full-Information
Intervention-Oblivious
Intervention-Aware
Affine
Policy-Aware

Counterfactual Methods: Online Comparison




Online Setting
NDCG@5
0.76
0.74
0.72
0.70
10^2
10^3
10^4
10^5
10^6
10^7
Queries Issued
Full-Information
COLTR (online)
PDGD (online)
Biased-PDGD (online)
Inter.-Aware (100 int.)


Multiple Settings
0.76
0.74
0.72
0.70
NDCG@5
10^2
10^3
10^4
10^5
10^6
10^7
Queries Issued
Full-Information
PDGD (online)
PDGD (counterfactual)
PDGD (100 interventions)
Intervention-Aware (counterfactual)

Conclusion

Conclusion

- **Intervention-Aware Estimator:**
  - Novel **counterfactual/online** estimator.
  - **Most reliable** choice for counterfactual learning.
  - Online performance **comparable to state-of-the-art**.

- **Intervention-Aware Estimator:**
  - Novel **counterfactual/online** estimator.
  - **Most reliable** choice for counterfactual learning.
  - Online performance **comparable to state-of-the-art**.

- **PDGD** is **not reliable** when **not** applied **fully online**.

Conclusion

- **Intervention-Aware Estimator:**
  - Novel **counterfactual/online** estimator.
  - **Most reliable** choice for counterfactual learning.
  - Online performance **comparable to state-of-the-art**.
- **PDGD** is **not reliable** when **not** applied **fully online**.
- A **single method** that is the **best choice** for **both online and counterfactual** learning to rank.
- Continue our work: `https://github.com/HarrieO/2021wsdm-unifying-LTR`

A. Agarwal, X. Wang, C. Li, M. Bendersky, and M. Najork. Addressing trust bias for unbiased learning-to-rank. In *The World Wide Web Conference*, pages 4–14. ACM, 2019.

O. Chapelle and Y. Chang. Yahoo! Learning to Rank Challenge Overview. *Journal of Machine Learning Research*, 14:1–24, 2011.

N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789. ACM, 2017.

H. Oosterhuis and M. de Rijke. Policy-aware unbiased learning to rank for top-k rankings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 489–498. ACM, 2020.

Z. Ovaisi, R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva. Correcting for selection bias in learning-to-rank systems. In *Proceedings of The Web Conference 2020*, pages 1863–1873, 2020.

A. Vardasbi, H. Oosterhuis, and M. de Rijke. When inverse propensity scoring does not work: Affine corrections for unbiased learning to rank. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2020.

Acknowledgments

All content represents the opinion of the author(s), which is not necessarily shared or endorsed by their employers and/or sponsors.

Extra Results: Effect of Online Interventions

Radboud University