Self-Normalized Off-Policy Estimators for Ranking

BEN LONDON, ALEXANDER BUCHHOLZ, GIUSEPPE DI BENEDETTO, JAN MALTE LICHTENBERG, YANNIK STEIN, and THORSTEN JOACHIMS, Amazon Music

We propose two new estimators for off-policy evaluation of ranking policies, based on the idea of self-normalization. Importantly, these estimators are parameter-free and asymptotically unbiased. Experiments with synthetic data demonstrate that our estimators can be more accurate than other importance weighting estimators, owing to their ability to control variance, while adding minimal bias. From this, we conclude that self-normalization offers an optimal balance of accuracy and practicality for off-policy ranker evaluation.

ACM Reference Format:
Ben London, Alexander Buchholz, Giuseppe Di Benedetto, Jan Malte Lichtenberg, Yannik Stein, and Thorsten Joachims. 2023. Self-Normalized Off-Policy Estimators for Ranking. In . ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Off-policy evaluation—using the data collected by an existing policy to evaluate the performance of a new policy—is a cornerstone of today's search, recommendation and advertising systems. In these applications, a policy typically ranks (and truncates) a set of available items. To date, off-policy evaluation of ranking policies usually involves structural assumptions about how users engage with rankings—so-called *click models* [4] of user behavior. From a statistical perspective, click models, and their associated estimators, can be analyzed in terms of their *bias* and *variance*. In this paper, we focus on reducing variance, since this is often easier than reducing bias.

There are several popular methods to reduce variance, and all of them make some trade-off between bias, variance and practicality. Importance weight *clipping* [3, 9] and *doubly-robust* [6, 8, 11, 17, 19] estimation are both effective at reducing variance, but require the practitioner to either tune a hyper-parameter or estimate a reward model. A third method of variance reduction, commonly referred to as *self-normalization*, uses a multiplicative *control variate* to rescale the standard importance-weighting estimator [1, 12, 13, 15, 21]. This method has the advantage of being parameter-free, requiring no additional tuning or estimation. Moreover, it is asymptotically unbiased [1, 12, 21]. For these reasons, we (as well as others [20]) argue that self-normalization offers an optimal balance of bias, variance and practicality.

While self-normalization is well known, and has been applied in the "standard" contextual bandit setting, it has yet to be applied to ranking problems. We therefore propose two methods of self-normalization applicable to a variety of off-policy estimators for ranking. The first method is based on a *local* control variate for each position in the ranking; the second is based on a *global* control variate, averaged across all positions. Focusing on the *item-position* click model [14], we analyze the bias of each self-normalization method and prove upper bounds on their finite-sample behavior, showing in both cases that the bias decays at a rate of $\mathrm{O}(k/n)$, where k is the number of positions and n the number of records. We empirically demonstrate their effectiveness on synthetic ranking data, which illustrates how self-normalization can indeed reduce variance, while maintaining an acceptable level of bias, thereby improving estimation accuracy. These results, when considered with the practical benefits, suggest that self-normalization should be a go-to method for off-policy evaluation of rankers.

Manuscript submitted to ACM

2 PRELIMINARIES

A contextual bandit problem consists of interactions between a *policy*, π, and an *environment*. In each interaction, the environment generates a *context*, $x \in \mathcal{X}$, which quantifies its current conditions and defines which *actions*, $\mathcal{A}$, the policy can take. The environment also generates a stochastic *reward* function, which quantifies the contextual utility of each action. In a ranking scenario, the policy returns a sorted list of actions (alternatively, *items*), $A \triangleq (a_1, \dots, a_k)$, containing some subset of $\mathcal{A}$.[1] We consider a *semi-bandit* setting in which reward can be observed for each action (item) in the ranking, and use $r(x, A; j)$ to denote the reward for item a at position j in context x. A policy can be stochastic, so we denote its probability of selecting a ranking, given context x, by $\pi(A \mid x)$; and the *marginal* probability of ranking item a at position j in context x, by $\pi(a \mid x, j)$. The quantity that we are interested in is a policy's expected reward over draws of contexts, reward functions and rankings: $R(\pi) \triangleq \mathbb{E}_{x,r}\, \mathbb{E}_{A \sim \pi(\cdot \mid x)} \left[\sum_{j=1}^{k} r(x, A; j) \right]$.

2.1 Off-Policy Evaluation

Assume that we have collected a dataset of contextual bandit interactions using an existing policy, which we call the *logging policy*, π_0. Let $S \triangleq (x_i, (a_{ij}, r_{ij})_{j=1}^{k})_{i=1}^{n}$ denote the logged contexts, actions and rewards ($r_{ij} \triangleq r(x_i, A_i; j)$). We may also log the *propensities*, $\pi_0(a_{ij} \mid x_i, j)$, or the logging policy itself. Using this data to estimate the expected reward of a new policy—typically referred to as the *target policy*—we face the fundamental challenge in off-policy evaluation: the distribution of rankings produced by π may not be the same as π_0. This discrepancy creates a counterfactual conundrum; how can we reason about what would have happened had we deployed π instead of π_0?

Broadly speaking, there are two main approaches to this problem. The first approach, called the *direct method*, uses a reward function estimate to predict the target policy's expected reward in each logged context. This estimator has low variance (assuming bounded rewards), but can be significantly biased if the reward predictions are inaccurate. The alternative approach, known as *importance weighting* (a.k.a. *importance sampling*), re-weights the logged rewards using ratios of probabilities of observing the logged actions. Under certain conditions, importance weighting estimators are unbiased; however, they can have high variance. This tension between bias and variance is the central trade-off in designing off-policy estimators. An estimator's bias is, ultimately, a function of the environment's unknown dynamics, but its variance is largely determined by properties of the estimator and the number of samples. For this reason, we focus on controlling the variance of importance weighting.

The standard importance weighting estimator can have extremely high variance with ranking policies, due to the combinatorially large space of rankings. Thus, for ranking, one typically makes structural assumptions on user behavior—so-called *click models* [4]—which leads to importance weights with a tighter range; and hence, lower variance. We consider two click models and their associated estimators. One simple and general click model is the *item-position model* (IPM) [14], which assumes that clicks at a given position depend only on the context and the item displayed at said position. This assumption leads to the estimator in Eq. 8 (see Appendix B.3), which is unbiased—provided $\pi_0(a \mid x, j)) > 0$ whenever $\pi(a \mid x, j)) > 0$. Since the IPM importance weights involve marginal item-position propensities, the estimator has lower variance than standard importance weighting, but can nonetheless have high variance. Another popular click model is the *position-based model* (PBM) [5], which further assumes that click probabilities factorize as a product of relevance and *position bias*, ρ_j, which is the probability that a user examines position j, independent of the item displayed there. The corresponding estimator (Eq. 9 in Appendix B.3) typically has lower variance than IPM, but potentially much higher bias, and requires an additional step of estimating the position biases [2, 7, 10, 18, 22].

[1] For simplicity, we will assume that the length of a ranking (e.g., number of display positions) is always k, and that $\mathcal{A}$ always contains at least k items.

Beyond click models, there are other ways to tune the bias-variance trade-off in importance weighting estimators. The most straightforward method is to *clip* (alternatively, *cap* or *truncate*) the importance weights so that their magnitudes never exceed a certain value [3, 9]. Though this effectively controls the variance, it can cause a nontrivial increase in bias, and it requires the practitioner to select a clipping threshold. Another option is *doubly-robust* (DR) estimation [6], which uses importance weighting to de-bias the direct method, while simultaneously reducing the variance of importance weighting. To date, DR has only recently been applied to ranking applications [8, 11, 17, 19]. Despite its attractive theoretical properties, DR can be somewhat burdensome to apply in practice as it requires estimating the reward predictor, as well as several other parameters.

A third variance reduction technique is *self-normalization* (a.k.a. *weighted importance sampling*) [1, 12, 13, 15, 21]. Like DR, self-normalization uses control variates; but in this case, the control variate is the average of importance weights that *divides* the importance weighted average. This introduces bias, but the bias vanishes asymptotically as $n \to \infty$. The primary advantage of this approach, over clipping and DR, is that it is parameter-free; there is nothing to tune or estimate. This quality makes it particularly attractive to practitioners, who may not wish (or be able) to tune their estimator's bias-variance trade-off. In the next section, we introduce two self-normalized estimators for ranking.

3 SELF-NORMALIZATION FOR THE IPM

Our first self-normalized estimator is based on the observation that the IPM estimator can be written as a sum of estimators in the standard contextual bandit setting, with one estimator for each position in the ranking. Accordingly, we can apply self-normalization to each position. Let $w_{ij} \triangleq \frac{\pi(a_{ij} \,|\, x_i, j)}{\pi_0(a_{ij} \,|\, x_i, j)}$ denote the importance weight for the item at the j^{th} position of the i^{th} logged ranking. Let $\Phi_j \triangleq \frac{1}{n}\sum_{i=1}^{n} w_{ij}$ denote a *local* control variate for the j^{th} position. It is easily verified that the expected value of each control variate is exactly one (see Lemma 1)—a useful property of control variates that we exploit in our bias analysis. With these definitions, we introduce our first self-normalized IPM estimator, *SNIPM*, in Eq. 1.

While SNIPM uses a separate control variate to normalize each positional reward estimate, one can also normalize the entire IPM estimator with a single, *global* control variate. Let $\bar{\Phi} \triangleq \frac{1}{k}\sum_{j=1}^{k} \Phi_j$ denote the average of positional control variates. Since $\mathbb{E}[\Phi_j] = 1$ for all j, we have that $\mathbb{E}[\bar{\Phi}] = 1$; that is, $\bar{\Phi}$ is also a control variate. Accordingly, we define another self-normalized estimator, *SNIPM-G* (Eq. 2)—which is simply the IPM estimator divided by $\bar{\Phi}$.

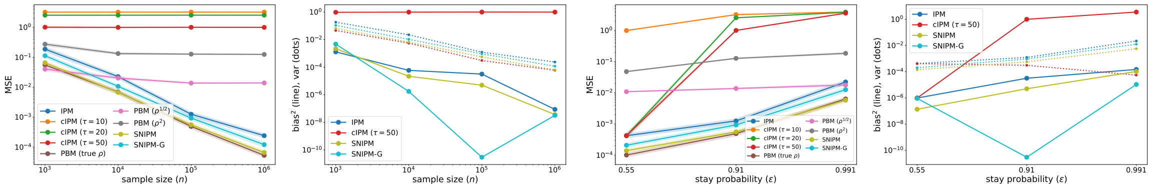
$$\hat{R}_{\text{SNIPM}}(\pi, S) \triangleq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{w_{ij} r_{ij}}{\Phi_j}, \qquad (1) \qquad \hat{R}_{\text{SNIPM-G}}(\pi, S) \triangleq \frac{1}{n\bar{\Phi}}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} r_{ij} = \frac{1}{\bar{\Phi}}\hat{R}_{\text{IPM}}(\pi, S) \qquad (2)$$

3.1 Bias Analysis

Self-normalization adds bias to the IPM estimator, but we will show that the bias decreases as a function of *n*; meaning, both SNIPM and SNIPM-G are *asymptotically unbiased.* All proofs for this section are deferred to Appendix A.

Proposition 1. *Assume the following: (common support) for every $x \in \mathcal{X}$, $a \in \mathcal{A}$ and $j \in \{1, \ldots, k\}$, if $\pi(a \,|\, x, j) > 0$ then $\pi_0(a \,|\, x, j) > 0$; (bounded rewards) $\sup r(\cdot, \cdot) - \inf r(\cdot, \cdot) \leq M < \infty$; (importance weights have finite variance) for each $j \in \{1, \ldots, k\}$, $\sigma^2_{w_j} < \infty$. Then, with $W \triangleq \frac{1}{k}\sum_{j=1}^{k} w_j$,*

$$\mathbb{E}[\hat{R}_{\mathit{SNIPM}}(\pi, S)] - R(\pi) \leq \sum_{j=1}^{k} \frac{M(6\sigma^2_{w_j} + 2\sigma_{w_j})}{n} \quad (3) \quad \textit{and} \quad \mathbb{E}[\hat{R}_{\mathit{SNIPM\text{-}G}}(\pi, S)] - R(\pi) \leq \frac{kM(6\sigma^2_W + 2\sigma_W)}{n}. \quad (4)$$



(a) Sample Size vs. MSE (b) Sample Size vs. Bias/Var (c) Stochasticity vs. MSE (d) Stochasticity vs. Bias/Var

Fig. 1. Results of the synthetic data experiments. Figs. 1a and 1b fix $\epsilon = 0.91$ and vary n, while Figs. 1c and 1d fix $n = 10^5$ and vary ϵ.

For both estimators, the bias is of order $\mathrm{O}(k/n)$. Since we often assume that k is a fixed constant, $\mathrm{O}(k/n)$ vanishes as $n \to \infty$. Note that Eqs. 3 and 4 are zero whenever the logging and target policies are the same, since in that case the importance weights will have zero variance (they will all equal one). This supports the intuition that the estimator should be unbiased when used for *on-policy* evaluation. Comparing the two bounds, we note that the bias bound of SNIPM-G is less than that of SNIPM, due to the fact that $\sigma_W^2 \leq \max_j \sigma_{w_j}^2$.

4 EXPERIMENTS

To validate our proposed estimators empirically, we conduct experiments using a synthetic data generator. This allows us to compare off-policy reward estimates to “ground truth" reward for the target policy. The details of our data generator and experimental methodology are given in Appendix B. We compare our proposed estimators, SNIPM and SNIPM-G, to several natural baselines: IPM; cIPM (clipped IPM, with clipping threshold τ)[2]; and PBM with either the true position bias curve, ρ, or an exponentiated copy of it, $\rho^{1/2}$ or ρ^2, to simulate an incorrectly estimated curve. Note that PBM with the true curve is unattainable in practice; we provide it only for reference.

Fig. 1a plots the *mean squared error* (MSE), with 95% confidence interval, as a function of the size of the dataset. Both cIPM and the incorrectly specified PBM estimators exhibit error that does not decrease with more data, owing to their respective biases (see Fig. 2a in Appendix B.4). In contrast, the error of IPM and SNIPM(-G) decreases, since this error is mainly dominated by variance (see Fig. 1b), which vanishes as $n \to \infty$. SNIPM and SNIPM-G consistently show lower variance than IPM, with SNIPM providing a bit more variance reduction, as expected. Note also that the bias of SNIPM(-G) decreases with more data, which concurs with Proposition 1.[3] Ignoring the unrealistic PBM baseline (with true ρ), SNIPM achieves the lowest error for $n \geq 10^4$, followed by SNIPM-G and then IPM.

We also plot MSE as a function of the stochasticity of the logging policy, which is controlled by a parameter called the *stay probability*, ϵ (see Appendix B.1). Decreasing the stochasticity (higher ϵ) increases the variance of the importance weights—which, in theory, should increase the variance and MSE of the estimators. We find that this is indeed the case, but (ignoring the unrealistic PBM) SNIPM and SNIPM-G still have the lowest error and reduce the variance of IPM.

5 CONCLUSIONS AND FUTURE WORK

We have presented theoretical and empirical evidence that self-normalization is an easy, effective tool for reducing variance in off-policy ranker evaluation, while incurring only a small, asymptotically decreasing bias. We conclude by noting that self-normalization may be compatible with estimators other than just IPM; and it is further composable with other variance reduction techniques, such as DR. We plan to investigate these possibilities in future work.

[2] While clipping works with any importance weighting estimator, we only apply it to the IPM estimator, since PBM already reduces variance significantly.
[3] The fact that IPM appears to have higher bias than SNIPM(-G) is likely due to error (caused by variance) in estimating the bias from a fixed sample.

REFERENCES

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431, 2017.

[2] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. Estimating position bias without intrusive interventions. In *Web Search and Data Mining*, 2019.

[3] L. Bottou, J. Peters, J. Qui nonero Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[4] Aleksandr Chuklin, Ilya Markov, and M. de Rijke. Click models for web search. In *Click Models for Web Search*, 2015.

[5] Nick Craswell, Onno Zoeter, Michael J. Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Web Search and Data Mining*, 2008.

[6] M. Dudik, J. Langford, and L. Lihong. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

[7] Zhichong Fang, Aman Agarwal, and Thorsten Joachims. Intervention harvesting for context-dependent examination-bias estimation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

[8] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. Enhanced doubly robust learning for debiasing post-click conversion rate estimation. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[9] E. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[10] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Web Search and Data Mining*, 2017.

[11] Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, N. Shimizu, and Yasuo Yamamoto. Doubly robust off-policy evaluation for ranking policies under the cascade behavior model. *Web Search and Data Mining*, 2022.

[12] A. Kong. A note on importance sampling using standardized weights. Technical Report 348, University of Chicago, Dept. of Statistics, 1992.

[13] Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *Artificial Intelligence and Statistics*, 2021.

[14] Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, Shan Muthukrishnan, Vishwa Vinay, and Zheng Wen. Offline evaluation of ranking policies with click models. In *Knowledge Discovery and Data Mining*, 2018.

[15] Ashique Rupam Mahmood, H. V. Hasselt, and Richard S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Neural Information Processing Systems*, 2014.

[16] H. Oosterhuis and M. de Rijke. Policy-aware unbiased learning to rank for top-k rankings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[17] Harrie Oosterhuis. Doubly robust estimation for correcting position bias in click feedback for unbiased learning to rank. *ACM Transactions on Information Systems*, 41(3), 2023.

[18] Matteo Ruffini, Vito Bellini, Alexander Buchholz, Giuseppe Di Benedetto, and Yannik Stein. Modeling position bias ranking for streaming media services. In *The Web Conference*, 2022.

[19] Yuta Saito. Doubly robust estimator for ranking metrics with post-click conversions. In *Conference on Recommender Systems*, 2020.

[20] Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. Evaluating the robustness of off-policy evaluation. *Conference on Recommender Systems*, 2021.

[21] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems*, 2015.

[22] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Web Search and Data Mining*, 2018.

A DEFERRED PROOFS

A.1 Proof of Eq. 3

Before proceeding to the proof, we note two things. First, observe that the expected reward, $R(\pi)$, decomposes as a sum of positional expected rewards:

$$R(\pi) = \sum_{j=1}^{k} \mathop{\mathbb{E}}_{x,r} \mathop{\mathbb{E}}_{a \sim \pi(\cdot \,|\, x,j)} [r(x, A; j)] \triangleq \sum_{j=1}^{k} R(\pi; j).$$

Second, recall that SNIPM can be written as a sum of self-normalized estimators for the standard contextual bandit setting, with one estimator per position:

$$\hat{R}_{\text{SNIPM}}(\pi, S) = \sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{w_{ij} r_{ij}}{\Phi_j} \right) \triangleq \sum_{j=1}^{k} \hat{R}_{\text{SNIPS}}(\pi, S; j).$$

This particular estimator is commonly referred to as *self-normalized inverse propensity scoring*, which we refer to as SNIPS, with an 'S', to distinguish it from SNIPM. All of the properties of SNIPS should transfer over to SNIPM, albeit with a correction for the number of display positions. Thus, to upper-bound the bias of SNIPM, we will first upper-bound the bias of SNIPS, which immediately yields a bound for SNIPM.

Prior work by Agapiou et al. [1] has established that the bias of SNIPS decreases at a rate of $\mathrm{O}(n^{-1})$, which vanishes as the dataset grows. Their result is stated for more general conditions than those we consider herein, and is thus not as optimized as it could be. We therefore include our own proof, which is based on theirs, but with some minor corrections and improvements.

In the following, we omit j from our notation, since the results hold for any position.

Lemma 1. *If, for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $\pi(a \,|\, x) > 0$ implies $\pi_0(a \,|\, x) > 0$, then $\mathbb{E}[\Phi] = 1$.*

Proof. Expanding the definition of Φ and applying via linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}[\Phi] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i \,|\, x_i)}{\pi_0(a_i \,|\, x_i)} \right] \\
&= \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{a \sim \pi_0(\cdot \,|\, x)} \left[\frac{\pi(a \,|\, x)}{\pi_0(a \,|\, x)} \right] \\
&= \mathop{\mathbb{E}}_{x} \sum_{a \in \mathcal{A}} \pi_0(a \,|\, x) \frac{\pi(a \,|\, x)}{\pi_0(a \,|\, x)} \\
&= \mathop{\mathbb{E}}_{x} \sum_{a \in \mathcal{A}} \pi(a \,|\, x) \\
&= 1.
\end{aligned}$$

The common support assumption ensures that the importance weights are finite. □

Proposition 2. *Assume: (1) the range of the reward function is bounded by some finite constant, $\sup r(\cdot, \cdot) - \inf r(\cdot, \cdot) \leq M < \infty$; (2) the variance of the importance weights, σ_w^2, is finite. Then, for any constant $t \in (0, 1)$, the SNIPS estimator has bias*

$$\mathbb{E}[\hat{R}_{SNIPS}(\pi, S)] - R(\pi) \leq \frac{M\sigma_w^2}{n(t-1)^2} + \frac{M(\sigma_w^2 + \sigma_w)}{t\,n}. \tag{5}$$

In particular, for $t = 1/2$,

$$\mathbb{E}[\hat{R}_{SNIPS}(\pi, S)] - R(\pi) \leq \frac{6M\sigma_w^2 + 2M\sigma_w}{n}. \tag{6}$$

Proof. We first define a function,

$$\beta(\pi, S) \triangleq \left(\frac{1}{\Phi} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n} w_i\big(r_i - R(\pi)\big)\right),$$

and note that its expected value,

$$\begin{aligned}
\mathbb{E}[\beta(\pi, S)] &= \mathbb{E}\left[\left(\frac{1}{\Phi} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n} w_i\big(r_i - R(\pi)\big)\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \frac{w_i}{\Phi}\big(r_i - R(\pi)\big)\right)\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} w_i\big(r_i - R(\pi)\big)\right] \\
&= \mathbb{E}\left[\hat{R}_{\text{SNIPS}}(\pi, S)\right] - R(\pi) - \mathbb{E}\left[\hat{R}_{\text{IPS}}(\pi, S)\right] - R(\pi) \\
&= \mathbb{E}\left[\hat{R}_{\text{SNIPS}}(\pi, S)\right] - R(\pi) - 0,
\end{aligned}$$

is the bias of the SNIPS estimator. We will use a proof technique in which we decompose the bias into two parts, using the event $\Phi \leq t$, for any constant $t \in (0, 1)$. Via linearity of expectation,

$$\mathbb{E}[\beta(\pi, S)] = \mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi \leq t\}] + \mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi > t\}]. \tag{7}$$

We can now bound each of the righthand terms separately.

For $\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi \leq t\}]$, we will upper-bound $\Pr\{\Phi \leq t\}$ and use the fact that the rewards are range-bounded to upper-bound the expression inside the expectation. Note that $\mathbb{1}\{\Phi \leq t\}(\Phi^{-1} - 1)$ is nonzero only when $\Phi \leq t$, which is when $\Phi^{-1} - 1 \geq t^{-1} - 1 > 0$. This means that

$$0 \leq \left|\mathbb{1}\{\Phi \leq t\}(\Phi^{-1} - 1)\right| = \mathbb{1}\{\Phi \leq t\}(\Phi^{-1} - 1) \leq \mathbb{1}\{\Phi \leq t\}\Phi^{-1} \leq \Phi^{-1}.$$

The last inequality holds because $\Phi \geq 0$. Thus, with $\bar{r}_i \triangleq r_i - R(\pi)$, we have that

$$\begin{aligned}
\beta(\pi, S)\mathbb{1}\{\Phi \leq t\} &= \mathbb{1}\{\Phi \leq t\}\left(\frac{1}{\Phi} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n} w_i\bar{r}_i\right) \\
&\leq \left|\mathbb{1}\{\Phi \leq t\}\left(\frac{1}{\Phi} - 1\right)\right|\left|\frac{1}{n}\sum_{i=1}^{n} w_i\bar{r}_i\right| \\
&\leq \frac{1}{\Phi}\left|\frac{1}{n}\sum_{i=1}^{n} w_i\bar{r}_i\right| = \left|\frac{1}{n}\sum_{i=1}^{n} \frac{w_i}{\Phi}\bar{r}_i\right|.
\end{aligned}$$

The expression inside the absolute value is the SNIPS estimator with rewards $\bar{r}_i$. Since $\sup r(\cdot,\cdot) - \inf r(\cdot,\cdot) \leq M$, we have that $|\bar{r}_i| \leq M$, and

$$\beta(\pi, S)\mathbb{1}\{\Phi \leq t\} \leq \left|\frac{1}{n}\sum_{i=1}^{n} \frac{w_i}{\Phi}\bar{r}_i\right| \leq M.$$

Then, via Chebyshev's inequality, for $t < 1$, we have that

$$\begin{aligned}\Pr\{\Phi \leq t\} &= \Pr\{\Phi - 1 \leq t - 1\} \\ &= \Pr\{\Phi - \mathbb{E}[\Phi] \leq -(1-t)\} \\ &\leq \Pr\{|\Phi - \mathbb{E}[\Phi]| \geq (1-t)\} \\ &\leq \frac{\mathbb{E}[(\Phi - \mathbb{E}[\Phi])^2]}{(t-1)^2} \\ &= \frac{\sigma_w^2}{n(t-1)^2}.\end{aligned}$$

In the second equality, we used the fact that the mean of the control variate is 1 (Lemma 1); and in the last equality, we used the fact that the variance of the control variate is σ_w^2/n. Thus,

$$\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi \leq t\}] \leq M \Pr\{\Phi \leq t\} \leq \frac{M\sigma_w^2}{n(t-1)^2}.$$

Now, turning to the other side, $\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi > t\}]$, we have that

$$\begin{aligned}\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi > t\}] &= \mathbb{E}\left[\mathbb{1}\{\Phi > t\}\left(\frac{1}{\Phi} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right)\right] \\ &= \mathbb{E}\left[\frac{\mathbb{1}\{\Phi > t\}}{\Phi}(1-\Phi)\left(\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right)\right] \\ &\leq \mathbb{E}\left[\frac{\mathbb{1}\{\Phi > t\}}{\Phi}|1-\Phi|\left|\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right|\right] \\ &\leq \mathbb{E}\left[\frac{\mathbb{1}\{\Phi > t\}}{t}|1-\Phi|\left|\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right|\right] \\ &\leq \frac{1}{t}\mathbb{E}\left[|1-\Phi|\left|\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right|\right] \\ &\leq \frac{1}{t}\sqrt{\mathbb{E}[(1-\Phi)^2]}\sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right)^2\right]}.\end{aligned}$$

The first inequality uses Jensen's inequality and the fact that $\mathbb{1}\{\Phi > t\}\Phi^{-1} \geq 0$; the second inequality follows from $\Phi > t$, which implies $\Phi^{-1} \leq t^{-1}$; the third inequality removes $\mathbb{1}\{\Phi > t\}$ from the expectation because the remaining terms are all nonnegative; and the final inequality follows from Cauchy-Schwarz. We are left with a product of standard deviations. First, recall that

$$\mathbb{E}[(1-\Phi)^2] = \mathbb{E}[(\Phi - \mathbb{E}[\Phi])^2] = \frac{\sigma_w^2}{n}.$$

Further,

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right)^2\right] = \frac{\mathbb{E}\left[(w\bar{r})^2\right]}{n} \leq \frac{M^2\,\mathbb{E}[w^2]}{n} = \frac{M^2(\sigma_w^2 + 1)}{n}.$$

Therefore,

$$\begin{aligned}
\frac{1}{t}\sqrt{\mathbb{E}[(1-\Phi)^2]}\sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} w_i \bar{r}_i\right)^2\right]} &\le \frac{1}{t}\sqrt{\frac{\sigma_w^2}{n}}\sqrt{\frac{M^2(\sigma_w^2+1)}{n}} \\
&= \frac{M}{t\,n}\sqrt{\sigma_w^4+\sigma_w^2} \\
&\le \frac{M}{t\,n}(\sigma_w^2+\sigma_w).
\end{aligned}$$

Putting it all together, we have

$$\mathbb{E}[\beta(\pi,S)] \le \frac{M\sigma_w^2}{n(t-1)^2} + \frac{M(\sigma_w^2+\sigma_w)}{t\,n},$$

which completes the proof of Eq. 5. Eq. 6 follows from plugging in $t = 1/2$ and simplifying. □

Remark 1. One could optimize t in Eq. 5, but this would require accounting for the constraint that $t \in (0,1)$, which can be accomplished via Lagrange multipliers. It is unclear whether optimizing t would result in a significantly tighter or interpretable bound. △

Since

$$\mathbb{E}[\hat{R}_{\text{SNIPM}}(\pi,S)] - R(\pi) = \sum_{j=1}^{k} \mathbb{E}[\hat{R}_{\text{SNIPS}}(\pi,S;j)] - R(\pi;j),$$

the proof of Eq. 3 follows directly from Proposition 2 (Eq. 6).

A.2 Proof of Eq. 4

Before proceeding to the proof, we first note that for any constant, $C \in \mathbb{R}$,

$$\frac{1}{n\bar{\Phi}}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}C = \frac{C\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}}{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n} w_{ij}} = kC.$$

We will use this fact in the proof.

Since the proof follows that of Proposition 2, we will overload some of the previous notation. Let

$$\beta(\pi,S) \triangleq \left(\frac{1}{\bar{\Phi}} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\big(r_{ij} - R(\pi)\big)\right),$$

and note that its expected value,

$$\begin{aligned}
\mathbb{E}[\beta(\pi,S)] &= \mathbb{E}\left[\left(\frac{1}{\bar{\Phi}} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\big(r_{ij} - R(\pi)\big)\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n\bar{\Phi}}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\big(r_{ij} - R(\pi)\big)\right)\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\big(r_{ij} - R(\pi)\big)\right] \\
&= \mathbb{E}\left[\hat{R}_{\text{SNIPM-G}}(\pi,S)\right] - kR(\pi) - \mathbb{E}\left[\hat{R}_{\text{IPM}}(\pi,S)\right] - kR(\pi) \\
&= \mathbb{E}\left[\hat{R}_{\text{SNIPM-G}}(\pi,S)\right] - R(\pi) - 0,
\end{aligned}$$

is the bias of the SNIPM-G estimator. Using the decomposition from Eq. 7, we will bound $\mathbb{E}[\beta(\pi,S)\mathbb{1}\{\bar{\Phi} \le t\}]$ and $\mathbb{E}[\beta(\pi,S)\mathbb{1}\{\bar{\Phi} > t\}]$ separately, for a value of $t \in (0,1)$ to be specified later.

For $\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\bar{\Phi} \leq t\}]$, we first upper-bound the magnitude of $\beta(\pi, S)\mathbb{1}\{\bar{\Phi} \leq t\}$ using the fact that the rewards are range-bounded. Using the same logic as in Proposition 2, with $\bar{r}_{ij} \triangleq r_{ij} - R(\pi)$, we have that

$$\begin{aligned}
\beta(\pi, S)\mathbb{1}\{\bar{\Phi} \leq t\} &= \mathbb{1}\{\bar{\Phi} \leq t\}\left(\frac{1}{\bar{\Phi}} - 1\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\bar{r}_{ij}\right) \\
&\leq \left|\mathbb{1}\{\bar{\Phi} \leq t\}\left(\frac{1}{\bar{\Phi}} - 1\right)\right|\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\bar{r}_{ij}\right| \\
&\leq \left|\frac{1}{\bar{\Phi}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\bar{r}_{ij}\right| \\
&\leq \frac{1}{n\bar{\Phi}}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}M = kM.
\end{aligned}$$

Further, using Chebyshev's inequality and $\mathbb{E}[\bar{\Phi}] = 1$, we have for $t < 1$ that

$$\begin{aligned}
\Pr\{\bar{\Phi} \leq t\} &= \Pr\{\bar{\Phi} - \mathbb{E}[\bar{\Phi}] \leq -(1 - t)\} \\
&\leq \frac{\mathbb{E}[(\bar{\Phi} - \mathbb{E}[\bar{\Phi}])^2]}{(t - 1)^2} \\
&= \frac{\sigma_W^2}{n(t - 1)^2}.
\end{aligned}$$

In the last line, we used the fact that $\bar{\Phi}$ is an average of n i.i.d. instantiations of the random variable W; thus, its variance is σ_W^2/n. Combining the above inequalities, we have

$$\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\bar{\Phi} \leq t\}] \leq kM \Pr\{\bar{\Phi} \leq t\} \leq \frac{kM\sigma_W^2}{n(t - 1)^2}.$$

Moving on to $\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\Phi > t\}]$, we have that

$$\mathbb{E}[\beta(\pi, S)\mathbb{1}\{\bar{\Phi} > t\}] \leq \frac{1}{t}\sqrt{\mathbb{E}[(1 - \bar{\Phi})^2]}\sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\bar{r}_{ij}\right)^2\right]},$$

using the same reasoning as in Proposition 2. Then,

$$\mathbb{E}[(1 - \bar{\Phi})^2] = \mathbb{E}[(\bar{\Phi} - \mathbb{E}[\bar{\Phi}])^2] = \frac{\sigma_W^2}{n};$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\bar{r}_{ij}\right)^2\right] &= \frac{\mathbb{E}\left[\left(\sum_{j=1}^{k} w_j \bar{r}_j\right)^2\right]}{n} \\ &\leq \frac{M^2\,\mathbb{E}\left[\left(\sum_{j=1}^{k} w_j\right)^2\right]}{n} \\ &= \frac{k^2 M^2\,\mathbb{E}[W^2]}{n} \\ &= \frac{k^2M^2(\sigma_W^2+1)}{n},\end{aligned}$$

where the last line uses the fact that $\sigma_W^2 = \mathbb{E}[W^2] - \mathbb{E}[W]^2 = \mathbb{E}[W^2] - 1$. Thus,

$$\begin{aligned}\frac{1}{t}\sqrt{\mathbb{E}[(1-\bar{\Phi})^2]}\sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\bar{r}_{ij}\right)^2\right]} &\leq \frac{1}{t}\sqrt{\frac{\sigma_W^2}{n}}\sqrt{\frac{k^2M^2(\sigma_W^2+1)}{n}} \\ &= \frac{kM}{t\,n}\sqrt{\sigma_W^4+\sigma_W^2} \\ &\leq \frac{kM}{t\,n}(\sigma_W^2+\sigma_W).\end{aligned}$$

Putting it all together, we have

$$\begin{aligned}\mathbb{E}[\beta(\pi,S)] &= \mathbb{E}[\beta(\pi,S)\mathbb{1}\{\bar{\Phi}\leq t\}] + \mathbb{E}[\beta(\pi,S)\mathbb{1}\{\bar{\Phi} > t\}] \\ &\leq \frac{kM\sigma_W^2}{n(t-1)^2} + \frac{kM(\sigma_W^2+\sigma_W)}{t\,n},\end{aligned}$$

Finally, setting $t = 1/2$ and reducing completes the proof.

B EXPERIMENT DETAILS

This appendix provides details of our experiments that were omitted from the main paper, due to space limitations.

B.1 Synthetic Data Generator

To simulate a ranking scenario, we generate data from a simple synthetic environment, with tunable parameters. In this environment, there are $|\mathcal{A}| = 10$ available actions (items), of which $k = 5$ can be displayed. Actions are represented by contextual feature vectors, which are drawn from a normal distribution with mean equal to the 1-hot encoding of the action, and standard deviation $\sigma = 0.1$. (Thus, context is defined implicitly via the action features.) To test situations in which the PBM is (in)correctly specified, we define the reward function as satisfying the PBM assumptions, with a possibly unknown position bias curve, $\rho_j = 1/j$ for $j = 1, \ldots, k$. Relevance is defined as a linear threshold function, $\mathrm{rel}(a\,|\,x, j) \triangleq \mathbb{1}\{\mathbf{a}\cdot\boldsymbol{\theta} \geq 0\}$, where $\mathbf{a}$ is the contextual feature vector for action a, and $\boldsymbol{\theta}$ is a weight vector. As such, a click is computed for action a at position j as $e_j \cdot \mathrm{rel}(a\,|\,x, j)$, where $e_j \in \{0, 1\}$ is a Bernoulli random variable with mean ρ_j. Note that this calculation is independent of other actions and positions, thereby satisfying the IPM assumption.

Since we control the relevance function's parameters, we define them in a way that favors certain actions. Without loss of generality, we set $\boldsymbol{\theta} = [-1, 1, 1, -1, 1, -1, -1, 1, -1, -1]$, such that actions $(2, 3, 5, 8)$ are usually relevant, and the remaining actions are usually irrelevant. (We say "usually" because there is randomness in the contextual features that could cause relevance to change.)

We use ranking policies that score actions according to a linear model, $\boldsymbol{\vartheta} \cdot \mathbf{a}$, then sort by score in descending order (without truncating the list). In the case of the logging policy, we further randomize this rankings using a *generalized Fisher-Yates* algorithm, which applies random swaps to the ranking. For a parameter $\epsilon \in (0, 1)$, which we refer to as the *stay probability*, the probability that an item ranked at position j is kept at that position is equal to ϵ; and with probability $1 - \epsilon$, it is swapped with one of the other items, uniformly at random. Thus, the probability that the item at j is equal to any of the items not originally ranked there is equal to $\frac{1-\epsilon}{|\mathcal{A}|-1}$. Finally, the (randomized) ranking is truncated to the top k items.

Without loss of generality, we define the logging and target policies' respective model weights as:

$$\boldsymbol{\vartheta}_{\text{LOG}} = [3, 1, -1, 2, -2, 0, 0, 4, 0, 0] \quad \text{and} \quad \boldsymbol{\vartheta}_{\text{TGT}} = [-1, 2, 3, -2, 4, 0, 0, 1, 0, 0].$$

As such, the logging policy tends to place items $(8, 1, 4, 2)$ at the top of the ranking and $(3, 5)$ at the bottom, prior to swapping. (Recall that there is randomness in the contextual features, which creates randomness in the scores.) Similarly, the target policy tends to rank $(5, 3, 2, 8)$ at the top and $(1, 4)$ at the bottom. By design, there is some overlap between the top-ranked items, but the policies are not identical.

B.2 Methodology

For each experiment, we perform 100 trials, wherein each trial we: generate a new dataset of contexts, actions and rewards; execute the logging policy on said data to generate a simulated log dataset; execute the target policy on the fully-observed data and average its earned rewards to generate "ground truth"; and finally, execute each estimator on the log data and compare its estimated reward with the target policy's ground truth.

When generating mean squared error, we average over the errors of the experiment trials. The corresponding confidence intervals are computed using the normal approximation.

B.3 Baselines

We compare our estimators to several baselines. The first baseline is the IPM estimator [14],

$$\hat{R}_{\text{IPM}}(\pi, S) \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \frac{\pi(a_{ij} \,|\, x_i, j)}{\pi_0(a_{ij} \,|\, x_i, j)}. \tag{8}$$

The clipped version of this estimator is given by

$$\hat{R}_{\text{cIPM}}(\pi, S) \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \min \left\{ \frac{\pi(a_{ij} \,|\, x_i, j)}{\pi_0(a_{ij} \,|\, x_i, j)}, \tau \right\},$$

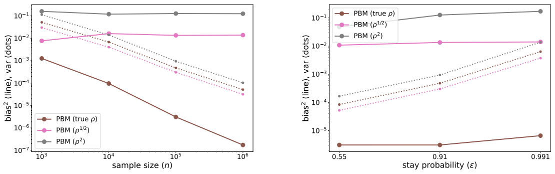
where $\tau \geq 1$ is the clipping threshold (which, in practice, must be tuned). We also compare to the *policy-aware* PBM estimator [16],

$$\hat{R}_{\text{PBM}}(\pi, S) \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \frac{\sum_{\ell=1}^{k} \pi(a_{ij} \,|\, x_i, \ell) \rho_\ell}{\sum_{\ell=1}^{k} \pi_0(a_{ij} \,|\, x_i, \ell) \rho_\ell} \tag{9}$$

where ρ_j is the bias at position j (which, in practice, must be estimated). In our experiments, we use either the true position bias from the data generator (Appendix B.1), or an exponentiated copy of it, $\rho^{1/2}$ or ρ^2, to simulate an incorrectly estimated curve.

B.4 Additional Plots

This section contains plots deferred from the main paper. Fig. 2 plots bias-variance decompositions for the PBM estimators, both as a function of n (fixing $\epsilon = 0.91$) and ϵ (fixing $n = 10^5$).



(a) Sample Size vs. Bias/Var

(b) Stochasticity vs. Bias/Var

Fig. 2. Bias-variance decompositions for the PBM estimators.