Handling Confounding for Realistic Off-Policy Evaluation

Saurabh Sohoney
Amazon
sohoneys@amazon.com

Nikita Prabhu
Amazon
niprabhu@amazon.com

Vineet Chaoji
Amazon
vchaoji@amazon.com

ABSTRACT

Inverse Propensity Score estimator (IPS) is a basic, unbiased, off-policy evaluation technique to measure the impact of a user-interactive system without serving live traffic. We present our work on applying IPS to real-world settings by addressing some practical challenges, thereby enabling successful policy evaluation. In particular, we show that off-policy evaluation can be impossible in the absence of a complete context and we describe a systematic way of defining the context.

1 INTRODUCTION

A/B test is a standard technique to evaluate user-interactive systems on live traffic. However, an A/B test has a high turn-around time and is expensive to conduct. This becomes a bottle-neck for a fast-moving company like Amazon and necessitates off-policy evaluation.

Off-policy evaluation seeks to exploit the logs generated from past user interactions to empirically estimate the value of a new policy. For example, the CTR of a new ad-ranker could be estimated using past clicks, without having the ranker actually serve ads to the users. A popular and fundamental technique for off-policy evaluation is Inverse Propensity Score (IPS) estimator [2, 3]. In a contextual bandit framework, where the domain of data is defined by distribution $(x, \vec{r}) \in D$ (x: context, $\vec{r}$: reward vector over actions), IPS defines the estimate of the value a policy, $\pi : x \rightarrow a$, as

$$\hat{\mathbb{V}}_H(\pi) = \frac{1}{|H|} \sum_{(x_t, a_t, r_t, p_t) \in H} \frac{r_t * \mathbb{I}(a_t = \pi(x_t))}{p_t} \quad (1)$$

Here, H represents historical logs of quadruples (context, action, reward, probability), indexed by t. Reward r_t is a measure of user-satisfaction, e.g., click, like or purchase. Probability $p_t = p(a_t|x_t)$ represents the propensity with which the action a_t is selected for the context x_t in H. Division by action propensities neutralizes the sampling bias in H and makes IPS an unbiased estimator [1, 4]. Note that unbiasedness is crucial for simulating an A/B test. An important assumption for IPS to work is that all the actions should be explored sufficiently often [1]. Next, we discuss three practical challenges faced while applying IPS to real-world datasets.

(1) **High variance:** IPS is sensitive to the probability values in the denominator of Eq. 1 and often has large estimator variance. This problem has been addressed in literature and a standard solution is to trade some bias in order to limit the variance. [4] achieves this by bounding the denominator with a minimum value, while [5] suggests forms of additive and multiplicative biases to accomplish the same.
(2) **Handling too many actions**: When the pool of actions is large or dynamic, IPS performs poorly due to under-exploration. This problem has been previously addressed by limiting the effective number of distinct actions. [3] & [1] recommend fuzzy matching of actions and [3] also suggests partitioning of the action set through clustering.
(3) **Computing propensities**: When action probabilities are not recorded in H, they need to be estimated. In [3] & [1], authors suggest estimating them empirically as count ratios, $p(a_t|x_t) = \#(a_t, x_t)/\#(x_t)$ from H. Note that this is only possible when a clear definition of context (x_t) is available.

Our contribution in this work is two-fold - 1) We present results for successful policy evaluation on real-world datasets, 2) We prove that IPS estimates can be invalid when the context is incomplete and we recommend a systematic way to define the context. As IPS forms the building block for many state-of-the-art off-policy evaluation techniques [4, 5], our solution automatically becomes relevant to all of them. To the best of our knowledge the problem of defining context, although crucial, has not been addressed previously for IPS, as authors have always assumed that either the propensities are recorded in H [2] or a clear definition of context is known [1, 3].

2 THE EFFECT OF CONFOUNDING

As mentioned previously, count ratio estimates of propensities are dependent on the definition of the context. Propensity values can become incorrect when the context is incomplete. This lack of information is referred to as confounding in the data. Proposition 1 states the ill-effect of confounding on IPS estimates.

Proposition 1: Policy evaluation using IPS can be impossible if the propensities are computed using incomplete context.

Proof sketch: We prove this through a constructive example. Suppose, the exploration data H, consists of six tuples as shown in Table 1. y, z represent two context features and the action propensities are not recorded. We assume, without loss of generality, that these six tuples represent the entire domain. Let P be a policy, defined as $\{y_1 z_1 \rightarrow a_1; y_1 z_2 \rightarrow a_2\}$. The average reward, if P was run for generating H would be 6/6. Referring to the estimated action propensities (column 2 and 3, in Table 1), the IPS estimates of the value of P as per Eq. 1, becomes $\frac{1}{6} * (\frac{1}{2/3} + 0 + 0 + \frac{1}{2/3} + \frac{1}{2/3} + \frac{1}{2/3})$=6/6, when y & z together form the context and 8/6, when only y forms the context. Clearly, the estimate is incorrect when z is ignored.

Note that, if the last two tuples in H are eliminated, the propensities become independent of z (all equal to 1/2). As a result, the estimate without z included in the context is also correct, i.e., 4/4.

Having understood the importance of an accurate context, we now discuss a systematic way to achieve it.

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '18 Companion, April 23–27, 2018, Lyon, France
© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.
ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186915

Table 1: Constructive example to demonstrate confounding

H: $<x_t, a_t, r_t>$	$\hat{p}(a_t \mid x_t = (y,z))$	$\hat{p}(a_t \mid x_t = y)$
$<y_1z_1, a_1, 1>$	2/3	1/2
$<y_1z_1, a_2, 0>$	1/3	1/2
$<y_1z_2, a_1, 0>$	1/3	1/2
$<y_1z_2, a_2, 1>$	2/3	1/2
$<y_1z_1, a_1, 1>$	2/3	1/2
$<y_1z_2, a_2, 1>$	2/3	1/2

3 HANDLING CONFOUNDING

In most real-world situations, the propensities are not recorded during logging and a valid context definition is needed to compute the propensities, i.e., to model $p(a|x)$. Defining context can be challenging due to lack of visibility into the preceding systems or due to multiple systems simultaneously logging user interactions[1]. Moreover, choosing the full feature vector as context may not be feasible as it can result in under-exploration. We recommend a systematic approach, similar to wrapper-based feature-selection techniques [6], to choose the right context for IPS.

(1) Start with a maximal set of features that look relevant to the logic of action selection in exploration data.
(2) Keep eliminating features until negative log-likelihood on a hold-out set starts increasing.

Increase in negative log-likelihood implies higher confounding. We suggest two guidelines in addition to the mentioned steps - 1) Numeric features should be binned in order to arrive at a discrete context, 2) For datasets with limited exploration logs, early stopping can be employed to trade confounding for better exploration. The described procedure is greedy and can also be followed in the reverse order by adding features. We note that more sophisticated alternatives can be employed for improving the performance further.

4 EXPERIMENTATION

Considering space limitations, we restrict ourselves to discussing experiments on only two datasets obtained from Amazon.

4.1 Dataset Description

4.1.1 Payments. This dataset comes from Amazon's payment gateway page, where a user selects a payment method (action) to check-out the cart. A successful transaction results in a non-zero (=1) reward. Along with the action and reward, the system also logs customer, device, location and order related features. This data is collected when no recommendation policy is in place. Therefore, the logic for action selection is completely driven by users' choice and we do not have sufficient information to model $p(a|x)$.

4.1.2 Search. This dataset comes from the Product Search page of Amazon where a search ranker maps a user query to a ranked list of products within a specified category. Unlike Payments, here the context is well-defined, i.e., a combination of query and category. Reward is 1 if the product is purchased within the session.

[1]Collecting logs from multiple policies is desired for IPS as it results in better randomization [1, 4]

Table 2: IPS based evaluation for Payments

Context	Neg LL	% Deviation
Feats123	**1.301**	**-0.77**
Feats12	1.307	+31.47
Feats1	1.312	+104.94

Table 3: IPS estimates for Search policies.

Context	P1	P2	P3	P4
Query	23.37	23.27	23.80	23.65
Query+Category	**20.07**	**20.21**	**20.67**	**20.39**

4.2 Results

We aim to highlight two aspects of our results - 1) Analysis of the ill-effects of confounding on off-policy evaluation (Payments) and off-policy comparison (Search), 2) Successful policy evaluation and A/B test simulation by minimizing the effect of confounding.

4.2.1 Payments. We define context by following the steps discussed in Section 3 and evaluate a payment recommender policy by estimating the success rate using IPS (Feats123 in Table 2). We further delete two sets of features, one at a time, from the defined context, to introduce confounding (Feats12 and Feats1). It is clear that the estimate is closest to the ground truth (measured using % deviation) when the confounding is minimum, i.e., with $Feats123$. On the other hand, estimates using incomplete contexts are sub-optimal. This proves the merit of the suggested approach.

4.2.2 Search. We introduce confounding in Search data by eliminating Category field from the context. Table 3 compares four competing policies based on the IPS estimates of the purchase rate (numbers are scaled to maintain confidentiality). The true pair-wise ordering as per three independent A/B tests is P1<P2, P2<P3 and P3>P4. IPS using the complete context is able to predict the outcome of all the three A/B tests correctly, while the order is reversed for the pair P1-P2, when context is incomplete. We also compare IPS estimate of one of the policies with the ground truth and observe that the estimate with the complete context is 20% closer to the target than that with the incomplete context.

5 CONCLUSIONS

We demonstrated that defining context is a crucial step for applying IPS in a real-world setting. By providing a systematic way of choosing the context and by handling other practical challenges associated with IPS, we successfully evaluated one policy (in Payments) and simulated three A/B tests (in Search) offline.

REFERENCES

[1] J. Langford, A. Strehl, and J. Wortman. 2008. Exploration Scavenging. In *ICML*.
[2] L. Li, S. Chen, J. Kleban, and A. Gupta. 2015. Counterfactual Estimation and Optimization of Click Metrics in Search Engines: A Case Study. In *WWW '15 Companion*.
[3] L. Li, J. Young Kim, and I. Zitouni. 2015. Toward Predicting the Outcome of an A/B Experiment for Search Relevance. In *WSDM*.
[4] A. Strehl, J. Langford, L. Li, and S. Kakade. 2010. Learning from Logged Implicit Exploration Data.. In *NIPS*.
[5] A. Swaminathan and T. Joachims. 2015. The self-normalized estimator for counterfactual learning. In *NIPS*.
[6] J. Tang, A. Salem, and L. Huan. 2014. Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*.