Off-Policy Evaluation with Deficient Support Using Side Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Off-Policy Evaluation (OPE) problem consists in evaluating the performance of new policies from the data collected by another one. OPE is crucial when evaluating a new policy online is too expensive or risky. Many of the state-of-the-art OPE estimators are based on the Inverse Propensity Scoring (IPS) technique, which provides an unbiased estimator when the full support assumption holds, i.e., when the logging policy assigns a non-zero probability to each action. However, there are several scenarios where this assumption does not hold in practice, i.e., there is deficient support, and the IPS estimator is biased in the general case. In this paper, we consider two alternative estimators for the deficient support OPE problem. We first show how to adapt an estimator that was originally proposed for a different domain to the deficient support setting. Then, we propose another estimator, which is a novel contribution of this paper. These estimators exploit additional information about the actions, which we call side information, in order to make reliable estimates on the unsupported actions. Under alternative assumptions that do not require full support, we show that the considered estimators are unbiased. We also provide a theoretical analysis of the concentration when relaxing all the assumptions. Finally, we provide an experimental evaluation showing how the considered estimators are better suited for the deficient support setting than the IPS baseline.

1 Introduction

Many real-world decision-making problems can be viewed through the lens of the contextual bandit framework. Some prominent examples are medical treatments [15], recommendation systems [26, 14, 43], search engines [27], ad-placement systems [8], and many others. In any of these problems, we have a decision-maker who repeatedly observes a context (e.g., a profile of a patient), samples an action according to its policy (e.g., provides a medical treatment), and collects a reward (e.g., +1 if the patient survives). Also, in many of those applications, we have access to additional information about the actions, which we call *side information*. For example, in movie recommendation, we may know the director and the actors of each movie; in a clinical trial, we may have access to a set of characteristics of each drug.

While deploying and evaluating a new policy on a real system may be prohibitively expensive (if not unfeasible), logged data of contextual bandit problems is cheap to obtain. Hence, it is desirable to use them to evaluate new policies, without the need to collect new data. This problem is called *Off-Policy Evaluation* (OPE). In this case, we have data collected by a given logging policy, and we want to evaluate a different policy, called evaluation policy. One way to think about this problem is that we are

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

trying to answer the following counterfactual question: "*What would have happened if the evaluation policy was deployed instead of the logging one?*". One of the most widely employed estimators for OPE is the *Inverse Propensity Score* (IPS), along with its variants [38, 51, 56, 43]. The IPS technique tries to de-bias the collected rewards by accounting for the propensity of the logging policy. One of the reasons for the widespread adoption of the IPS estimator is that it is unbiased under the *full support* assumption. The full support assumption states that no action has zero probability under the logging policy. Unfortunately, we argue that this assumption is unrealistic in many real-world systems. A practical example is a recommendation system that adopts a user-based pre-processing of the items to recommend, discarding some items for a given user to improve scalability. In this way, some actions (i.e., items) will never be proposed to a given context (i.e., a user), violating the full support assumption. Another example is the *new action* problem, namely, when the action space is expanded after the logging phase. This problem is typical of many real-world use cases, such as recommendation systems [46, 24, 47, 4], drug interaction evaluation [28], clinical trials [35], etc. In general, whenever the full support assumption is not valid, we say that we have an OPE problem with *deficient support*. If this is the case, the IPS estimator can be highly biased [42]. A possible mitigation for this issue is using a model-based approach, which means training a regression model that aims to approximate the reward function and extrapolate the rewards for the unsupported actions [5, 10, 6, 42, 30]. The drawback of this approach is that model misspecification can lead to a high bias [10, 30, 45].

In this paper, we focus on estimators without regression model for deficient support that exploit side information about the actions. After introducing the necessary background (Section 2), we consider two alternative estimators for Off-Policy Evaluation with deficient support in Section 3. While relaxing the full support assumption, we offer alternative assumptions that hold even with deficient support. In Section 3.1, we present the *PseudoInverse* estimator. This estimator was introduced by Swaminathan et al. [58] in the context of slate recommendation. We show how this estimator can be adapted to the presence of side information, and we show that it is unbiased under two alternative assumptions: *full support on side information* and *linearity*. Full support on side information is a milder assumption than full support: it requires the logging policy to be able to select each feature. Linearity, instead, requires that the rewards are linear combinations of the action features. This assumption is already used in other contexts, such as recommendation systems [22] and online linear bandits [2, 9]. In Section 3.2, we propose a novel estimator, which we call *Similarity* estimator. It is based on the assumption that expected rewards of similar actions are similar. Under this assumption, we prove its unbiasedness, even in the presence of deficient support. In Section 3.3, we relax all the assumptions and derive finite-sample concentration inequalities of the considered estimators for deficient support and the traditional IPS to better understand the theoretical guarantees of each of them. In Section 4, we provide an experimental evaluation, showing that the considered estimators consistently outperform the IPS baseline on a real-world dataset with deficient support from the recommendation systems domain. Finally, in Section 5 we review related work, and in Section 6 we draw conclusions.

Societal Impacts Improving OPE may be beneficial to society as it allows decision-makers to assess potentially dangerous policies without testing them on a real system. This is even more relevant in the presence of deficient support, when there are context-action pairs that are never observed in the data.

2 Background and Preliminaries

First of all, we introduce the *Off-Policy Evaluation* (OPE) problem in the contextual bandit setting with side information on the actions, and then, we describe the deficient support problem.

Off-Policy Evaluation Let x be a context sampled from a prior (unknown) context distribution p over the context space X, i.e., $x \sim p(\cdot)$. An action a is sampled from a probability π (called *policy*) over the action space A, conditioned on the observed context x, i.e., $a \sim \pi(\cdot|x)$. After this choice, a

reward r is observed from the reward distribution conditioned on the observed context and chosen action, i.e., $r \sim p(\cdot|x, a)$. Without loss of generality, we consider as reward space $[0, 1]$. To evaluate a policy, we use the *policy value* $R(\pi) := \mathbb{E}_{x\sim p(\cdot)} \mathbb{E}_{a\sim\pi(\cdot|x)} \mathbb{E}_{r\sim p(\cdot|x,a)}[r]$. Sometimes, it is useful to define the expected reward given a context and an action: $\delta(x, a) := \mathbb{E}_{r\sim p(\cdot|x,a)}[r]$. With this function, we can simplify the policy value formulation: $R(\pi) = \mathbb{E}_{x\sim p(\cdot)} \mathbb{E}_{a\sim\pi(\cdot|x)}[\delta(x, a)]$. In this paper, we focus on the contextual bandit problem with side information about the actions. Therefore, we have access to action features, which are represented as vectors: $\mathrm{f}(a)$ for any $a \in A$. In particular, we will focus on binary features: $\mathrm{f}(a) \in \{0, 1\}^F$. We focus on the *Off-Policy Evaluation* (OPE) problem. We call the logging policy π_0. Therefore, the collected dataset will be in the following form: $\mathcal{D} := \{x_i, a_i, r_i, \pi_0(a_i|x_i)\}_{i=1}^n$. Given this dataset, we would like to evaluate another policy by estimating its value function. The new policy is called the *evaluation policy* π_e. Thus, we want to find a plausible estimator $\hat{R}(\pi_e)$ using the given dataset ($\hat{R}(\pi_e) = \hat{R}(\pi_e; \mathcal{D})$) such that $\hat{R}(\pi_e) \approx R(\pi_e)$. One of the most used estimators for OPE is the *Inverse Propensity Score* (IPS), along with its variants. The IPS estimator is defined as $\hat{R}_{\text{IPS}}(\pi_e) = \frac{1}{n} \sum_{i=1}^n \frac{\pi_e(a_i|x_i)}{\pi_0(a_i|x_i)} r_i$. A fundamental property of such estimator is that it is an *unbiased* estimator of the evaluation policy value if the *full support assumption* holds [26]. Let us define $supp(f) := \{z \mid f(z) > 0\}$, $A_0(x) := supp(\pi_0(\cdot|x))$, and $A_e(x) := supp(\pi_e(\cdot|x))$, for any x. The full support assumption is stated in the following.

Assumption 1 (Full Support)**.** *The off-policy evaluation problem satisfies the full support assumption if $A_e(x) \subseteq A_0(x)$ with probability one for $x \sim p(\cdot)$.*

Deficient Support Whenever the full support assumption is not valid, we say that we have an OPE problem with *deficient support* [42]. If this is the case, the IPS estimator is no longer unbiased. Let us define the set of unsupported actions for context x under π_0 as $\mathrm{Un}(x, \pi_0) := A_e(x) \setminus A_0(x)$. The bias of the IPS depends on this set, as illustrated by the following proposition.

Proposition 1 ([42], Proposition 1)**.** *In an off-policy evaluation problem, the bias of $\hat{R}_{IPS}(\pi_e)$ is equal to the negative expected reward on the unsupported action set:*

$$bias(\hat{R}_{IPS}(\pi_e)) = \mathop{\mathbb{E}}_{x\sim p(\cdot)} \left[- \sum_{a\in Un(x,\pi_0)} \pi_e(a|x)\delta(x, a) \right]$$

This finding is intuitive: if the unsupported set is empty (i.e., the full support assumption is valid), the IPS estimator has no bias. Otherwise, if there are unsupported actions, the IPS estimator has a bias caused by the blind areas in the dataset, where there are actions for which we never observe the reward.

3 Estimators for Off-Policy Evaluation with Side Information

In this section, we first introduce two estimators for a better off-policy evaluation when we have deficient support and side information about the actions, and then we provide finite-sample error bounds. The proofs of the mathematical statements can be found in Appendix A.

3.1 PseudoInverse Estimator

The *PseudoInverse* (PI) estimator was initially proposed by Swaminathan et al. [58] for the problem of off-policy evaluation for *slate* recommendation [19]. In the following, we will show that we can use an analogous estimator for the OPE problem with side information.

We begin by describing two assumptions, which are relatively mild conditions and can also hold in the presence of deficient support.

Assumption 2 (Full Support on Side Information)**.** *The off-policy evaluation problem satisfies the full support on side information if, whenever $\pi_e(a|x) > 0$, this implies that, for every feature of a (i.e., $\forall j \in \{j|\mathrm{f}(a)_j = 1\}$) there exists an action $a' \in A_0(x)$ that has the same feature (i.e., $\mathrm{f}(a')_j = 1$) with probability one over $x \sim p(\cdot)$.*

We notice that this assumption can also be satisfied in the presence of deficient support, as long as the logging policy can propose each feature. Now we make an assumption on the reward structure.

Assumption 3 (Linearity)**.** *For each context $x \in X$ there exists an (unknown) intrinsic reward vector $\phi_x \in \mathbb{R}^F$ such that $\delta(x, a) = \mathrm{f}(a)^T \phi_x$.*

Intuitively, this assumption says that each action feature contributes linearly to the final reward. This assumption found applications in different fields: in recommendation systems, it constitutes the basis of the Matrix Factorization algorithm [22], where we have fixed the latent item factors to be the feature vectors; in bandits literature, it corresponds to the linear bandit formulation [2, 9].

Hence, we can view this problem as a regression problem, where we interpret the ϕ_x vector as the weight vector w to learn and we want to minimize the following error (for each x): $\mathbb{E}_{a\sim\pi_0(\cdot|x)} \mathbb{E}_{r\sim p(\cdot|x,a)}[(\mathrm{f}(a)^T w - r)^2]$. The PseudoInverse estimator derives from the minimization of this error, and it is unbiased under the two previously described assumptions.

Theorem 1. *Consider the off-policy evaluation problem where Assumptions 2 and 3 hold. Then, we can define an unbiased estimator $\hat{R}_{PI}(\pi_e)$ of the expected reward of the evaluation policy as:*

$$\hat{R}_{PI}(\pi_e) := \frac{1}{n} \sum_{i=1}^{n} r_i \cdot \mathbf{q}_{\pi_e,x_i}^T \mathbf{\Gamma}_{\pi_0,x_i}^{\dagger} \mathrm{f}(a_i),$$

where $\mathbf{\Gamma}_{\pi_0,x} := \mathbb{E}_{a\sim\pi_0(\cdot|x)}[\mathrm{f}(a)\mathrm{f}(a)^T] \in \mathbb{R}^{F\times F}$, $\mathbf{q}_{\pi_e,x} := \mathbb{E}_{a\sim\pi_e(\cdot|x)}[\mathrm{f}(a)] \in \mathbb{R}^F$, and $M^\dagger$ is the Moore-Penrose pseudoinverse of a generic matrix M.

For the details of the derivation, we refer to [58]. Inspired by [56], we can also use a *Self-Normalized* version of the estimator, which is a biased but consistent estimator with a reduced variance:

$$\hat{R}_{SN-PI}(\pi_e) := \frac{1}{\sum_{i=1}^{n} \mathbf{q}_{\pi_e,x_i}^T \mathbf{\Gamma}_{\pi_0,x_i}^{\dagger} \mathrm{f}(a_i)} \sum_{i=1}^{n} r_i \cdot \mathbf{q}_{\pi_e,x_i}^T \mathbf{\Gamma}_{\pi_0,x_i}^{\dagger} \mathrm{f}(a_i).$$

3.2 Similarity Estimator

In some cases, the calculation of the pseudoinverse of a matrix can be computationally expensive. Thus, in the following, we propose an estimator based on computationally cheaper operations, which stems from a different assumption. Such assumption comes from the recommendation systems research literature [41]. Two of the most common approaches in recommendation systems are *collaborative filtering* [40] and *content-based filtering* [31] methods. Both of these methods are based on the assumption that the reward of an item (i.e., an action in the contextual bandit setting) given a user (i.e., a context) can be approximated by a weighted sum of rewards observed from different items given the same user. In our setting, we can define this assumption as follows:

Assumption 4. *The off-policy evaluation problem satisfies the similarity assumption if, given any context $x \in X$, the following condition holds:*

$$\exists w_x : A_e(x) \times A_0(x) \to \mathbb{R} \ \textit{ such that } \ \forall a \in A_e(x),\ \delta(x, a) = \sum_{a' \in A_0(x)} w_x(a, a')\delta(x, a').$$

This indicates that we can approximate the reward for any action with a weighted sum of rewards observed from the actions supported by the logging policy. The weighting function has no constraint, but intuitively it should indicate how much two actions are perceived as similar for a given context.

The aforementioned assumption is also strictly tied with the *Lipschitz assumption*, which is widely exploited in the research field of online bandits with side information [36, 20, 50, 49, 21]. The connection between the two assumptions is exhibited by the Proposition 3, in Appendix A.

By exploiting the similarity assumption, we can get an unbiased estimator of the expected reward of the evaluation policy even with deficient support. We explain the procedure in the following.

Theorem 2. *Consider the off-policy evaluation problem where the similarity assumption holds. Then, we can define an unbiased estimator $\hat{R}_S(\pi_e)$ of the expected reward of the evaluation policy as:*

$$\hat{R}_S(\pi_e) := \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{\pi}(a_i|x_i)}{\pi_0(a_i|x_i)} r_i,$$

where $\bar{\pi}(a_i|x_i) := \mathbb{E}_{a\sim\pi_e(\cdot|x_i)}\left[w_{x_i}(a, a_i)\right]$.

We notice that the meaning of the estimator is somehow intuitive. Each observed reward r_i is divided by the propensity of the logging policy to eliminate the bias introduced by the data collection procedure (as in IPS). At the same time, it is amplified by how much, on average, the evaluation policy proposes an action similar to a_i for the context x_i.

3.3 Concentration Analysis

Up until now, we have shown that the previous estimators are unbiased under some assumptions. In the following, we will analyze the concentration behavior of the PseudoInverse, Similarity, and traditional IPS estimators. We will also relax every assumption and investigate the bias-variance trade-off that arises. To do so, we provide finite-sample error bounds. In order to lighten the notation, we will drop the distributions of the expected value whenever the expectation is for $(x, a, r) \sim p(\cdot)\pi_0(\cdot|x)p(\cdot|x, a)$. We will refer to the absolute value of the bias of an estimator as $\epsilon := |\,\mathbb{E}[\hat{R}(\pi_e)] - R(\pi_e)|$. We introduce two different functions to measure the difference between the conditional distributions induced by two policies. The first one is a premetric called *Support Divergence*, introduced by Sachdeva et al. [42]:

$$d^{supp}(\pi_e||\pi_0) := \mathop{\mathbb{E}}_{x\sim p(\cdot)}\left[\sum_{a\in \mathrm{Un}(x,\pi_0)} \pi_e(a|x)\right].$$

The second one is the *Exponentiated Rényi Divergence* [39, 32]:

$$d_2(\pi_e||\pi_0) := \mathop{\mathbb{E}}_{\substack{x\sim p(\cdot)\\ a\sim\pi_0(\cdot|x)}}\left[\left(\frac{\pi_e(a|x)}{\pi_0(a|x)}\right)^2\right].$$

Now, we are ready to introduce finite-sample concentration inequalities.

Proposition 2. *Let us consider a generic off-policy evaluation problem with side information. Let $\hat{R}_{IPS}$ be the IPS estimator, $\hat{R}_S$ the similarity estimator with weighting function w_x, $\hat{R}_{PI}$ the pseudo-inverse estimator. Then, for any $\gamma \in (0, 1)$, the following inequalities hold with probability at least $1 - \gamma$:*

$$\left|\hat{R}_{IPS}(\pi_e) - R(\pi_e)\right| \leq d^{supp}(\pi_e||\pi_0) + \sqrt{\frac{d_2(\pi_e||\pi_0)}{n\gamma}},$$

$$\left|\hat{R}_S(\pi_e) - R(\pi_e)\right| \leq \epsilon^S + \sqrt{\frac{d_2(\bar{\pi}||\pi_0)}{n\gamma}},$$

$$\left|\hat{R}_{PI}(\pi_e) - R(\pi_e)\right| \leq \epsilon^{PI} + \sqrt{\frac{2\sigma_{PI}^2 \ln(2/\gamma)}{n}} + \frac{2(\rho_{PI}+1)\ln(2/\gamma)}{3n},$$

where $\bar{\pi}(a|x) := \mathbb{E}_{a'\sim\pi_e(\cdot|x)}[w_x(a', a)]$, $\sigma_{PI}^2 := \mathbb{E}_{x\sim p(\cdot)}[\mathbf{q}_{\pi_e,x}^T \mathbf{\Gamma}_{\pi_0,x}^{\dagger} \mathbf{q}_{\pi_e,x}]$, $\rho_{PI} := \sup_x \sup_a \left|\mathbf{q}_{\pi_e,x}^T \mathbf{\Gamma}_{\pi_0,x}^{\dagger} \mathrm{f}(a)\right|$, *and* $\mathbf{q}_{\pi_e,x}, \mathbf{\Gamma}_{\pi_0,x}$ *are defined as in Theorem 1.*

The error bounds for both the IPS and the Similarity estimators are derived following the main idea by Metelli et al. [32]. In particular, they provide a bound for the standard IPS estimator based on a Chebyshev-like inequality using the Rényi divergence [39]. The bound for the PI estimator derives from Theorem 1 by Swaminathan et al. [58] after having taken the bias into account.

We can notice how all the bounds exhibit a similar pattern. Each inequality has a component that depends on the approximation error of the estimator ($d^{supp}(\pi_e||\pi_0), \epsilon^S, \epsilon^{PI}$, respectively). In particular cases, the approximation error of such estimators can be zero ($d^{supp}(\pi_e||\pi_0) = 0$ when there is full support; $\epsilon^S = 0$ when the similarity assumption holds; $\epsilon^{PI} = 0$ when there is full support on side information and the linearity assumption holds). Still, in general, we have to consider the error resulting from the possible violation of the assumptions. Furthermore, we can notice that this type of error does not decrease with the sample size n. Therefore, if present, it is an irreducible error.

On the other hand, all the bounds have a remaining component that decreases with the sample size and depends on the difference between the evaluation and logging policies. In particular, the first two inequalities display a *polynomial* concentration (the dependence on n and γ is $\mathcal{O}(\sqrt{1/n\gamma})$). This bound has been proven to be tight in [34] for the IPS estimator. For the third inequality, we have a better behavior due to the *exponential* concentration (when the variance σ^2_{PI} is small the dependence on n and γ is $\mathcal{O}(\sqrt{\ln(1/\gamma)/n})$).

Finally, we notice that the first two inequalities are very similar since the IPS estimator and the Similarity one have a comparable structure, but the former depends on $d_2(\pi_e||\pi_0)$, while the latter on $d_2(\bar{\pi}||\pi_0)$. This indicates how the Similarity estimator has an additional degree of freedom for regulating the bias-variance trade-off. Indeed, by inducing a complex but accurate $\bar{\pi}$, it may decrease the bias factor ϵ^S, while increasing $d_2(\bar{\pi}||\pi_0)$. On the other hand, with a simple model of $\bar{\pi}$, it may reduce the $d_2(\bar{\pi}||\pi_0)$ factor, but it may incur in a high bias.

4 Experiments

In this section, we empirically evaluate the considered estimators in an OPE problem with real logged bandit feedback. In particular, we will show how the estimators that exploit side information are competitive in practice with traditional IPS techniques, especially in the presence of deficient support. The code used for the experiments can be found in the Supplemental Material.

4.1 Setup

We mainly make use of two Python packages: *Open Bandit Pipeline* [44] and *PyIEOE* [45]. These two packages are open source and freely available[1,2]. Open Bandit Pipeline is a library with many state-of-the-art Off-Policy estimators already implemented, while PyIEOE is a collection of scripts for evaluating and comparing OPE estimators.

The dataset that we use is the *Open Bandit Dataset* (OBD), released with Open Bandit Pipeline. OBD contains logged bandit feedback from a real-world application (a large-scale fashion e-commerce platform). There are three campaigns available, namely "ALL", "Men", and "Women". We select the "ALL" campaign. The dimension of the context vector x is 20, and the number of actions $|A| = 80$. Importantly, this dataset contains side information regarding the actions (inside the dataset, called *action context*). The action context has 4 features for each action, among which 3 categorical and one real. We discard the last one and apply one-hot encoding on the first 3. In the end, we get a binary feature vector $\mathrm{f}(a)$ for each a with dimension 40.

4.2 Deficient Support Evaluation

OBD consists of two separate datasets collected with two different policies: one with a uniform random policy $\mathcal{D}_r = \{x_i, a_i, r_i, \pi_r(a_i|x_i) = \frac{1}{|A|}\}_{i=1}^{n_r}$, and the other with a Bernoulli Thompson Sampling (BTS) policy $\mathcal{D}_{\text{bts}} = \{x_i, a_i, r_i, \pi_{\text{bts}}(a_i|x_i)\}_{i=1}^{n_{\text{bts}}}$. Since both policies satisfy the full support assumption, we have to pre-process the data in order to simulate a deficient support scenario.

Specifically, we select the random policy as the logging one ($\pi_0 = \pi_r$), and the BTS as the evaluation policy ($\pi_e = \pi_{\text{bts}}$). Since we have a dataset collected with the BTS, we can easily compute its policy

[1]Open Bandit Pipeline available with Apache License 2.0: `https://github.com/st-tech/zr-obp`
[2]PyIEOE available with MIT License: `https://github.com/sony/pyIEOE`

value by on-policy estimation, which we call $R_{\text{on}}(\pi_e) := \frac{1}{n_{\text{bts}}} \sum_{r \in \mathcal{D}_{\text{bts}}} r$. We select a set of random seeds, $\mathcal{S}$. We select a deficient support rate p. This rate represents the percentage of the actions that should have deficient support for each context under the logging policy.

For each random seed $s \in \mathcal{S}$, our pre-processing algorithm proceeds as follows: (i) Select a random sub-sample of $n^* = 100,000$ rows from the logging dataset $\mathcal{D}_r$; (ii) Select p actions randomly for each context, obtaining $\text{Un}(x, \pi_0)$; (iii) For each context x, discard the data points $(x, a, r, \pi_0(a|r))$ where $a \in \text{Un}(x, \pi_0)$; (iv) Modify the logged propensity for each data point: since we know that the logging policy was a uniform random, we need to set $\pi_0(a|x) = 1/|A|(100\% - p)$; (v) Now, we can use the resulting dataset to get the estimate $\hat{R}(\pi_e)$ and get the squared error w.r.t. $R_{\text{on}}(\pi_e)$. This pre-processing protocol is summarized in Algorithm 1, presented in Appendix B.

4.3 Hyperparameter Selection

The proposed Similarity estimator has a weighting function to be set. In the following experimental evaluation, we present two possible variants. The crucial aspect is that they both exploit side information of the actions so that we have a weighting for unsupported actions. Other weighting functions can be proposed, but we leave this analysis as future work.

The first proposal is to have a weighting function proportional to the *cosine* similarity among side information: $w_x(a, a') \propto s_{\cos}(a, a') := f(a)^T f(a') / \|f(a)\|_2 \|f(a')\|_2$. The cosine similarity is typical in fields like recommendation systems [12] and information retrieval [48]. We scale the weighting function such that the resulting $\bar{\pi}$ satisfies $\sum_{a \in A_0(x)} \bar{\pi}(a|x) = 1$, so that the Self-Normalized version of the estimator is consistent under the same assumptions of Theorem 2 [56].

The second proposal is to create a clustering of the actions, which induces a partition of A. This is done by applying *K-Means Clustering* [13, 29] (we set $k = 30$) applied on the normalized action feature vectors $f(a)/\|f(a)\|_2$. Let us define the function $c(a)$ that returns, for each action, its corresponding cluster. Then, we can set the weighting function $w_x(a, a') \propto \mathbf{1}(a \in c(a'))$, which is 1 if both a and a' are in the same cluster, and 0 otherwise. Again, we scale the weighting function such that $\sum_{a \in A_0(x)} \bar{\pi}(a|x) = 1$. This type of estimator can also be derived starting from an alternative assumption, as we show in Appendix C.

4.4 Results

We repeat our bootstrap evaluation for various deficient support rates: $\{10\%, 20\%, \ldots, 90\%\}$, with random seeds $\mathcal{S} = \{1, 2, \ldots, 200\}$, for a total of 1800 different combinations. Further details on the infrastructure used and computation time can be found in Appendix B. We compare three estimators (PseudoInverse, Similarity with cosine, Similarity with clustering) with the IPS baseline. We use Self-Normalized variants for each estimator because they outperformed non-normalized ones. We compute the Mean Squared Error (MSE) for each deficient support rate by averaging the squared errors obtained with different seeds. Figure 1 compares the MSE of the evaluated estimators, varying the deficient support rate. From Figure 1, we see how the Similarity estimators dominate the IPS baseline for each deficient support rate. For a low rate (10%), the estimator with cosine similarity has more or less the same error as the IPS. Nonetheless, increasing the deficient support, the proposed Similarity estimators are clear winners. The clustering estimator has a lower MSE with respect to the cosine one for low deficient support rates. For high deficient support rates (>50%), it has a higher error than the cosine one but is still lower with respect to IPS. This shows how the Similarity estimator with clustering displays a hybrid behavior between the vanilla IPS and the cosine Similarity estimator. This suggests that whenever we have deficient support with low rates, the clustering estimator is preferable; conversely, the one with cosine similarity performs better with high deficient support rates. Regarding the PseudoInverse, it has the best MSE up until 50% of deficient support rate. Instead, for high deficient support (higher than or equal to 70%), it displays an unstable behavior with a high variance. This effect may be caused by the violation of the full support on side information assumption, which is shown in Table 2, Appendix B.

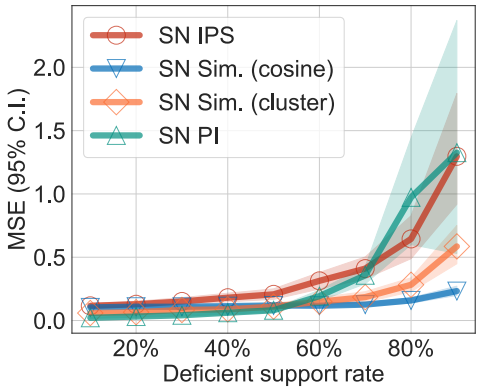

Figure 1: MSE ($\times 10^5$) of the estimators for each deficient support rate.

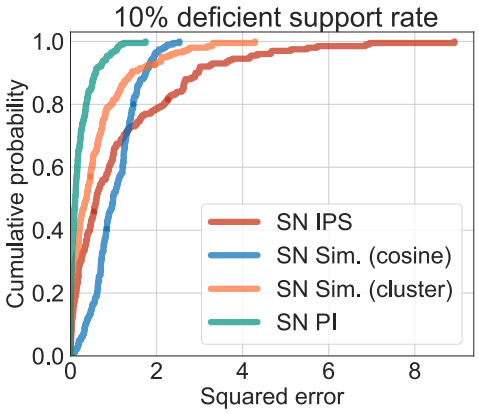

Figure 2: Cumulative Distribution Functions of the squared error ($\times 10^6$) for 10% deficient support rate.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
SN IPS	5.22	3.83	3.50	2.91	2.78	3.43	4.11	5.17	6.72
SN Sim. (cosine)	2.96	2.16	1.72	1.28	1.13	**1.00**	**1.00**	**1.00**	**1.00**
SN Sim. (cluster)	2.59	1.95	1.93	1.39	1.47	1.68	1.88	2.24	2.90
SN PI	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	2.04	3.39	8.30	7.29

Table 1: $\text{CVaR}_{0.7}$ with various deficient support rates. Values are normalized for each deficient support rate, the best result is highlighted in bold.

For a more interpretable evaluation, we compute the Cumulative Distribution Function (CDF) of the squared errors of the estimators for each deficient support rate, as done by, e.g., [56, 63, 43]. The plot of the computed CDF[3] for a deficient support rate of 10% is in Figure 2. This curve sheds light on a noteworthy aspect of the evaluated estimators, as we explain in the following. Looking solely at Figure 1, we see how the MSE of the Similarity estimator with cosine is almost equal to the value of the standard IPS for a deficient support rate of 10%. The MSE, though, is not a desirable metric when we are in a risk-sensitive scenario [7], i.e., whenever we want to minimize the error in the worst case. From Figure 2 we notice that, while having a similar MSE, our Similarity estimator with cosine is highly preferable to IPS in the worst case (IPS has a much longer tail).

For a better evaluation of the worst case, we show the *Conditional Value-at-Risk*$_\alpha$ (CVaR_α) metric [1] ($\alpha \in [0, 1)$). This metric computes the average error in the worst $(1-\alpha) \cdot 100\%$ case. It has been widely used, e.g., in financial portfolio optimization [23, 64]. We set $\alpha = 0.7$ (following [45]), meaning that we evaluate the worst 30% outcomes of the squared error[4]. The results for the various deficient support rates are shown in Table 1. We can see how both Similarity variants outperform the IPS baseline, particularly when we have a strong presence of deficient actions. This indicates that the Similarity estimators are suitable also in risk-sensitive applications with deficient support. The PseudoInverse estimator has an even lower $\text{CVaR}_{0.7}$ than Similarity estimators until 50% of deficient support. For higher deficient support rates, it suffers a strong performance degradation.

5 Related Work

We focus on the mathematical framework of *contextual bandits* [62, 25], which is a specific instance of the more general *Reinforcement Learning* framework [54]. We address the *Off-Policy Evaluation*

[3] The CDF plots for all the deficient support rates are in the Appendix B.

[4] Alternative choices of α can be found in Appendix B.

(OPE) problem, which consists in evaluating a new policy from data logged by a different policy. This problem traces its roots in causal inference [37, 18], and it is related to the estimation of the average treatment effect [15, 16]. Several Off-Policy estimators have been proposed recently for contextual bandit logged data, as this is a very active research area [53, 57, 52, 60, 10, 11, 63, 45]. One of the most used approaches is the Inverse Propensity Score (IPS) estimator [38, 59, 55], which constitutes the heart of many other proposed state-of-the-art estimators [51, 10, 11, 56, 63, 52, 61]. These estimators all rely on the full support assumption, which we argued is challenging to obtain in a real-world application. Thus, our paper focuses on OPE with deficient support. Despite its relevance in real-world applications, this setting is not very explored in the research literature.

London and Joachims [30] suggested an approach for improved OPE estimators whenever there is a *"new action"* problem, i.e., when the action set is expanded after the logging phase. This problem is a specialization of the deficient support problem because it means that there are actions that the logging policy could not take for any context. Instead, in the deficient support scenario, the action set with no support may be different for different contexts. Therefore, our approach is more general. Furthermore, they focus only on improving regression-based estimators using theoretical tools from domain adaptation [17, 3], while we concentrate on estimators with no regression model.

Sachdeva et al. [42] are the first to address the problem of deficient support with contextual bandit feedback data. In particular, they propose three techniques for learning in this scenario. Their most successful technique is a learning method based on the IPS estimator. This means that they do not rely on regression models, as done in this paper. However, they focus only on Off-Policy Learning. Hence, their method can not be used for evaluating other policies but only for learning one. Our work, instead, proposes solutions for Off-Policy Evaluation. OPE can be the first step in a learning procedure that maximizes the given estimator.

In the estimators proposed in this paper, we use side information about the actions. Within the scope of bandits, the term *"side information"* is often used with different meanings. In principle, the contextual bandit problem was called *"bandit with side information"* [62, 25] to distinguish it from the standard multi-armed bandit problem. In our paper, we differentiate contexts and side information about the actions: a context consists of information given at each round before selecting an action, while the side information about the actions is known to the decision-maker and fixed for each action. This kind of side information is actively exploited in some *online* bandit algorithms, which make the assumption that there is a metric space on the actions with a Lipschitz distance function of the rewards [36, 20, 50, 49, 21]. Our approach, instead, focuses on the *offline* scenario.

For the concentration analysis, we took inspiration from Metelli et al. [32] and their follow-up papers [33, 34], where they analyzed the theoretical properties of the IPS-like estimators using theoretical tools from information theory, such as the Rényi divergence [39].

6 Conclusions

In this paper, we focused on Off-Policy Evaluation, paying special attention to the scenario where the full support assumption is violated. We presented two estimators for a better evaluation by exploiting side information about the actions. We evaluated their theoretical guarantees under alternative assumptions without relying on full support. We also provided finite-sample concentration inequalities, relaxing all the assumptions. The experimental evaluation with real-world data showed that the estimators using side information outperform (both in the average-case and in the worst-case scenario) the standard IPS approach in the presence of deficient support. For future work, a natural direction to follow is to extend this work to the Off-Policy Learning setting by using the proposed estimators. Furthermore, we presented a novel estimator with hyperparameters to be selected. So, it would be valuable to have a data-driven procedure for the hyperparameter tuning.

References

[1] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2002. URL http://jmlr.org/papers/v3/auer02a.html.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79 (1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

[4] Cesare Bernardis and Paolo Cremonesi. Nfc: a deep and hybrid item-based model for item cold-start recommendation. *User Modeling and User-Adapted Interaction*, pages 1–34, 2021. ISSN 1573-1391. doi: 10.1007/s11257-021-09303-w. URL https://doi.org/10.1007/s11257-021-09303-w.

[5] Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki, editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 129–138. ACM, 2009. doi: 10.1145/1557019.1557040. URL https://doi.org/10.1145/1557019.1557040.

[6] Aurélien Bibaut, Ivana Malenica, Nikos Vlassis, and Mark J. van der Laan. More efficient off-policy evaluation through regularized targeted learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 654–663. PMLR, 2019. URL http://proceedings.mlr.press/v97/bibaut19a.html.

[7] Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. Risk-averse trust region optimization for reward-volatility reduction. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4583–4589. ijcai.org, 2020. doi: 10.24963/ijcai.2020/632. URL https://doi.org/10.24963/ijcai.2020/632.

[8] Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, 14(1): 3207–3260, 2013. URL http://dl.acm.org/citation.cfm?id=2567766.

[9] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 355–366. Omnipress, 2008. URL http://colt2008.cs.helsinki.fi/papers/80-Dani.pdf.

[10] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1097–1104. Omnipress, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.

[11] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

[12] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Trans. Inf. Syst.*, 39(2), January 2021. ISSN 1046-8188. doi: 10.1145/3434185. URL https://doi.org/10.1145/3434185.

[13] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[14] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline A/B testing for recommender systems. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, editors, *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 198–206. ACM, 2018. doi: 10.1145/3159652.3159687. URL https://doi.org/10.1145/3159652.3159687.

[15] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[16] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[17] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.*, 26:101–126, 2006. doi: 10.1613/jair.1872. URL https://doi.org/10.1613/jair.1872.

[18] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[19] Ray Jiang, Sven Gowal, Yuqiu Qian, Timothy A. Mann, and Danilo J. Rezende. Beyond greedy ranking: Slate optimization via list-cvae. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1xX42R5Fm.

[20] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In Cynthia Dwork, editor, *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 681–690. ACM, 2008. doi: 10.1145/1374376.1374475. URL https://doi.org/10.1145/1374376.1374475.

[21] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *J. ACM*, 66(4):30:1–30:77, 2019. doi: 10.1145/3299873. URL https://doi.org/10.1145/3299873.

[22] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263. URL https://doi.org/10.1109/MC.2009.263.

[23] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[24] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In Won Kim and Hyung-Jin Choi, editors, *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, ICUIMC 2008, Suwon, Korea, January 31 - February 01, 2008*, pages 208–211. ACM, 2008. doi: 10.1145/1352793.1352837. URL https://doi.org/10.1145/1352793.1352837.

[25] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 817–824. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/4b04a686b0ad13dce35fa99fa4161c65-Abstract.html.

[26] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In Irwin King, Wolfgang Nejdl, and Hang Li, editors, *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 297–306. ACM, 2011. doi: 10.1145/1935826.1935878. URL https://doi.org/10.1145/1935826.1935878.

[27] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 929–934. ACM, 2015. doi: 10.1145/2740908.2742562. URL https://doi.org/10.1145/2740908.2742562.

[28] Zun Liu, Xing-Nan Wang, Hui Yu, Jian-Yu Shi, and Wen-Min Dong. Predict multi-type drug–drug interactions in cold start scenario. *BMC bioinformatics*, 23(1):1–13, 2022.

[29] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

[30] Ben London and Thorsten Joachims. Offline policy evaluation with new arms. In *Offline Reinforcement Learning Workshop at Neural Information Processing Systems*, 2020.

[31] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer, 2011. doi: 10.1007/978-0-387-85820-3_3. URL https://doi.org/10.1007/978-0-387-85820-3_3.

[32] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5447–5459, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/6aed000af86a084f9cb0264161e29dd3-Abstract.html.

[33] Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141:1–141:75, 2020. URL http://jmlr.org/papers/v21/20-124.html.

[34] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8119–8132, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4476b929e30dd0c4e8bdbcc82c6ba23a-Abstract.html.

[35] Huseyin Naci and Alec B O'Connor. Assessing comparative effectiveness of new drugs before approval using prospective network meta-analyses. *Journal of clinical epidemiology*, 66(8): 812–816, 2013.

[36] Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for taxonomies: A model-based approach. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA*, pages 216–227. SIAM, 2007. doi: 10.1137/1.9781611972771.20. URL https://doi.org/10.1137/1.9781611972771.20.

[37] Judea Pearl. *Causality*. Cambridge university press, 2009.

[38] Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 759–766. Morgan Kaufmann, 2000.

[39] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

[40] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In John B. Smith, F. Donelson Smith, and Thomas W. Malone, editors, *CSCW '94, Proceedings of the Conference on Computer Supported Cooperative Work, Chapel Hill, NC, USA, October 22-26, 1994*, pages 175–186. ACM, 1994. doi: 10.1145/192844.192905. URL https://doi.org/10.1145/192844.192905.

[41] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer, 2015. ISBN 978-1-4899-7636-9. doi: 10.1007/978-1-4899-7637-6. URL https://doi.org/10.1007/978-1-4899-7637-6.

[42] Noveen Sachdeva, Yi Su, and Thorsten Joachims. Off-policy bandits with deficient support. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 965–975. ACM, 2020. doi: 10.1145/3394486.3403139. URL https://doi.org/10.1145/3394486.3403139.

[43] Yuta Saito and Thorsten Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge, editors, *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, pages 828–830. ACM, 2021. doi: 10.1145/3460231.3473320. URL https://doi.org/10.1145/3460231.3473320.

[44] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/33e75ff09dd601bbe69f351039152189-Abstract-round2.html.

[45] Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. Evaluating the robustness of off-policy evaluation. In Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge, editors, *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, pages 114–123. ACM, 2021. doi: 10.1145/3460231.3474245. URL https://doi.org/10.1145/3460231.3474245.

[46] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In Kalervo Järvelin, Micheline Beaulieu, Ricardo A. Baeza-Yates, and Sung-Hyon Myaeng, editors, *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 253–260. ACM, 2002. doi: 10.1145/564376.564421. URL https://doi.org/10.1145/564376.564421.

[47] Sulthana Shams, Daron Anderson, and Douglas J. Leith. Cluster-based bandits: Fast cold-start for recommender system new users. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1613–1616. ACM, 2021. doi: 10.1145/3404835.3463033. URL https://doi.org/10.1145/3404835.3463033.

[48] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24 (4):35–43, 2001.

[49] Aleksandrs Slivkins. Contextual bandits with similarity information. *J. Mach. Learn. Res.*, 15 (1):2533–2568, 2014. URL http://dl.acm.org/citation.cfm?id=2670330.

[50] Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. Learning optimally diverse rankings over large document collections. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 983–990. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/446.pdf.

[51] Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 2217–2225. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/c0f168ce8900fa56e57789e2a2f2c9d0-Abstract.html.

[52] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. CAB: continuous adaptive blending for policy evaluation and learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6005–6014. PMLR, 2019. URL http://proceedings.mlr.press/v97/su19a.html.

[53] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9167–9176. PMLR, 2020. URL http://proceedings.mlr.press/v119/su20a.html.

[54] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[55] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 814–823. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/swaminathan15.html.

[56] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3231–3239, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/39027dfad5138c9ca0c474d71db915c3-Abstract.html.

[57] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *J. Mach. Learn. Res.*, 16:1731–1755, 2015. URL http://dl.acm.org/citation.cfm?id=2886805.

[58] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3632–3642, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/5352696a9ca3397beb79f116f3a33991-Abstract.html.

[59] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3000–3006. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10042.

[60] Aaron David Tucker and Thorsten Joachims. Variance-optimal augmentation logging for counterfactual evaluation in contextual bandits. *CoRR*, abs/2202.01721, 2022. URL https://arxiv.org/abs/2202.01721.

[61] Nikos Vlassis, Aurélien Bibaut, Maria Dimakopoulou, and Tony Jebara. On the design of estimators for bandit off-policy evaluation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6468–6476. PMLR, 2019. URL http://proceedings.mlr.press/v97/vlassis19a.html.

[62] Chih-Chun Wang, Sanjeev R Kulkarni, and H Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

[63] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597. PMLR, 2017. URL http://proceedings.mlr.press/v70/wang17a.html.

[64] Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research*, 57(5):1155–1168, 2009.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the final part of Section 1
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Supplemental Material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4 and Appendix B
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3
   (b) Did you mention the license of the assets? [Yes] See Section 3
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Supplemental Material
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix B
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix B
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]