Policy-Adaptive Estimator Selection for Off-Policy Evaluation

Takuma Udagawa[1], **Haruka Kiyohara**[2*], **Yusuke Narita**[3], **Yuta Saito**[4], **Kei Tateno**[1]
[1]Sony Group Corporation
[2]Tokyo Institute of Technology
[3]Yale University
[4]Cornell University
Takuma.Udagawa@sony.com, kiyohara.h.aa@m.titech.ac.jp, yusuke.narita@yale.edu,
ys552@cornell.edu, Kei.Tateno@sony.com

Abstract

Off-policy evaluation (OPE) aims to accurately evaluate the performance of counterfactual policies using only offline logged data. Although many estimators have been developed, there is no single estimator that dominates the others, because the estimators' accuracy can vary greatly depending on a given OPE task such as the evaluation policy, number of actions, and noise level. Thus, the data-driven *estimator selection* problem is becoming increasingly important and can have a significant impact on the accuracy of OPE. However, identifying the most accurate estimator using only the logged data is quite challenging because the ground-truth estimation accuracy of estimators is generally unavailable. This paper thus studies this challenging problem of ***estimator selection for OPE*** for the first time. In particular, we enable an estimator selection that is *adaptive* to a given OPE task, by appropriately subsampling available logged data and constructing *pseudo policies* useful for the underlying estimator selection task. Comprehensive experiments on both synthetic and real-world company data demonstrate that the proposed procedure substantially improves the estimator selection compared to a non-adaptive heuristic. Note that complete version with technical appendix is available on arXiv: **http://arxiv.org/abs/2211.13904**.

1 Introduction

Off-Policy Evaluation (OPE) has widely been acknowledged as a crucial technique in search and recommender systems (Gilotte et al. 2018). This is because OPE accurately evaluates the performance of counterfactual policies without performing costly A/B tests (Saito and Joachims 2021). This is made possible by leveraging the logged data naturally collected by some *logging* or *behavior* policies. For example, a music recommender system usually records which songs it presented and how the users responded as feedback valuable for estimating the performance of counterfactual policies (Gruson et al. 2019; Kiyohara et al. 2022). Exploiting logged data is, however, often challenging, as the reward is only observed for the chosen action, but not for all the other actions that the system could have taken (Swaminathan and Joachims 2015a). Moreover, the logged data is biased due to the distribution shift between the behavior and evaluation policies (Levine et al. 2020).

To deal with the difficult statistical estimation involving *counterfactuals* and *distributional shift*, there has been a range of estimators with good theoretical properties – some estimators ensure unbiasedness under the identification assumptions (Strehl et al. 2010; Precup, Sutton, and Singh 2000; Dudík et al. 2014; Jiang and Li 2016; Thomas and Brunskill 2016), some reduce the variance while being consistent (Swaminathan and Joachims 2015b; Kallus and Uehara 2019), some minimize an upper bound of the mean-squared-error (MSE) (Wang, Agarwal, and Dudik 2017; Su et al. 2020; Metelli, Russo, and Restelli 2021). Intuitively, having more estimators makes it easier to achieve an accurate OPE. However, this also implies that practitioners now have to carefully solve the *estimator selection* problem to pick the most accurate estimator for their particular task. If we fail to identify an appropriate estimator, OPE may favor a poor-performing policy that should not be deployed in the field. Indeed, empirical studies have shown that the estimators' MSE and the most accurate estimator can change greatly depending on task-specific configurations such as the evaluation policy (Voloshin et al. 2019), the size of logged data (Saito et al. 2021b), and the number of actions (Saito and Joachims 2022). Moreover, Saito et al. (2021b) indicate that advanced estimators such as DRos (Su et al. 2020) may still produce an inaccurate OPE compared to the typical estimators such as IPS in certain scenarios. These empirical observations suggest the need of an accurate *estimator selection for OPE*. However, this estimator selection problem has remained completely unaddressed in the existing literature despite its practical relevance.

This paper explores the crucial problem of ***Estimator Selection for OPE*** for the first time. Specifically, our goal is ***to select the most accurate estimator among several candidates adaptive to a given OPE task***. One possible approach to conduct an estimator selection is to first estimate the estimators' MSE using only the logged data and then choose the most accurate. However, estimating the MSE is quite challenging, because it depends on the ground-truth performance of the evaluation policy, which is unavailable to us. To overcome this issue, we propose a novel estimator selection procedure called ***P**olicy-**A**daptive Estimator **S**election via **I**mportance **F**itting* (PAS-IF). A key trick lies in PAS-IF

*This work was done at Hanjuku-Kaso Co., Ltd.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

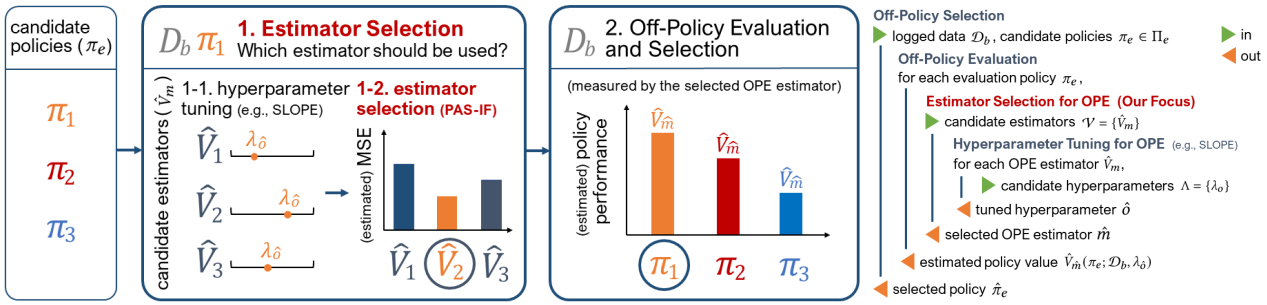

Figure 1: Workflow of Off-Policy Evaluation (OPE) and Selection (OPS) with an Estimator Selection Procedure

is to subsample and divide the logged data into psuedo subpopulations so that one of the subpopulations is deemed to be generated from the evaluation policy. Once we succeed in creating such subpopulations, we can estimate the candidate estimators' MSE based on the empirical average of the reward in that subpopulation. PAS-IF synthesizes appropriate subpopulations by minimizing the squared distance between the importance ratio induced by the true evaluation policy and that induced by the pseudo evaluation policy, which we call the *importance fitting* step. A fascinating feature of our PAS-IF is that it can optimize the subsampling rule differently for different evaluation policies. In this way, PAS-IF is able to find an accurate estimator *adaptive* to a given OPE task. This feature is particularly beneficial for the *Off-Policy Policy Selection* (OPS) task where we aim to identify the best policy among several candidates. Typically, OPS has been solved by applying a single OPE estimator to all candidate policies and picking the best-performing policy based on the OPE results (Doroudi, Thomas, and Brunskill 2017; Kuzborskij et al. 2021). However, our PAS-IF enables us to use the most accurate estimator for each candidate policy, thereby contributing to a much more accurate OPS.

In addition to developing PAS-IF, we empirically compare it to a *non-adaptive* heuristic, which estimates the MSE by naively regarding one of the behavior policies as a pseudo evaluation policy (Saito et al. 2021a,b). In particular, we demonstrate that our PAS-IF substantially improves the estimator selection accuracy by picking different estimators depending on a given evaluation policy. In contrast, the *non-adaptive* heuristic often fails to identify an accurate estimator when applied to a range of evaluation policies. We also demonstrate that PAS-IF enables a much more accurate OPS compared to the non-adaptive heuristic in both synthetic and real-world experiments.

2 Related Work

Here, we summarize some notable existing works.

Off-Policy Evaluation OPE has extensively been studied aiming at evaluating counterfactual policies without requiring risky and costly online interactions (Gilotte et al. 2018; Levine et al. 2020; Saito and Joachims 2021; Kiyohara, Kawakami, and Saito 2021). Direct Method (DM) (Beygelzimer and Langford 2009), Inverse Propensity Scoring (IPS) (Strehl et al. 2010; Precup, Sutton, and Singh 2000), and Doubly Robust (DR) (Dudík et al. 2014; Jiang and Li 2016; Thomas and Brunskill 2016) are the common baselines for OPE studies. DM uses some machine learning algorithms to regress the reward to estimate the policy performance. DM performs reasonably well when the reward estimation is accurate, however, it is vulnerable to the bias due to model mis-specification (Voloshin et al. 2019). In contrast, IPS enables an unbiased estimation by applying the importance sampling technique. However, IPS can suffer from a high variance particularly when the behavior and evaluation policies deviate greatly (Dudík et al. 2014). DR is a hybrid of DM and IPS, and often reduces the variance of IPS while remaining unbiased. However, DR can still suffer from a high variance especially when the action space is large (Saito and Joachims 2022). With the goal of achieving a better bias-variance trade-off, researchers have produced a number of estimators (Wang, Agarwal, and Dudik 2017; Kallus and Uehara 2019; Su et al. 2020). For example, Su et al. (2019) combine DM and IPS via adaptive weighting while Metelli, Russo, and Restelli (2021) modify the importance ratio to improve the typical exponential concentration rate of the error bound to a subgaussian rate. We build on the vast literature on OPE, but rather focus on the relatively under-explored problem of selecting the most accurate estimator among many options in a data-driven way.

Hyperparameter Tuning for OPE Hyperparameter tuning (which is fundamentally different from our estimator selection task) often plays a crucial role in OPE. The bias-variance trade-off of many OPE estimators such as Switch (Wang, Agarwal, and Dudik 2017) and DRos (Su et al. 2020) depends heavily on their built-in hyperparameters (Saito et al. 2021a,b). A naive way to tune such hyperparameters is to estimate the bias and variance of a given estimator and construct some MSE surrogates (Thomas and Brunskill 2016; Wang, Agarwal, and Dudik 2017; Su et al. 2020). However, estimating the bias is equally difficult as OPE itself because the bias depends on the ground-truth performance of the evaluation policy (Su, Srinath, and Krishnamurthy 2020). To tackle this issue, Su, Srinath, and Krishnamurthy (2020) propose a procedure called SLOPE based on the Lepski's principle, which was originally proposed for non-parametric bandwidth selection (Lepski and Spokoiny 1997). Since SLOPE avoids estimating the bias, it theoretically and empirically works better than the naive tuning pro-

cedure relying on some MSE surrogates (Su, Srinath, and Krishnamurthy 2020). However, SLOPE is only applicable to the hyperparameter tuning of a particular estimator class due to its *monotonicity* assumption. In contrast, our estimator selection problem aims at comparing different estimator classes and select the most accurate one, possibly after applying SLOPE within each estimator class.[1]

Estimator Selection for OPE Given several candidate OPE estimators, the estimator selection problem aims to identify the most accurate one. Although this problem has remained unexplored in the existing literature, many empirical studies have shown that the quality of estimator selection can have a significant impact on the accuracy of the downstream OPE. For instance, Voloshin et al. (2019) thoroughly investigate the estimation accuracy of OPE estimators in a range of simulated environments. Their empirical results demonstrate that the estimators' MSE and the most accurate estimator can change greatly depending on the data size, divergence between the behavior and evaluation policies, and many other environmental configurations. Saito et al. (2021b) also demonstrate that recent estimators with better theoretical guarantees can sometimes produce an inaccurate OPE compared to the typical estimators, suggesting the necessity of identifying and using appropriate estimators adaptive to a given OPE task. We are the first to formally formulate the important problem of *estimator selection for OPE*. We also propose a method to enable an *adaptive* estimator selection and empirically demonstrate that the accuracy of estimator selection makes a non negligible difference in the accuracy of the downstream OPE and OPS.

Off-Policy Policy Selection OPS aims to select the best-performing *policy* among several candidate policies using only the logged bandit data (Paine et al. 2020; Tang and Wiens 2021; Zhang and Jiang 2021). When applied to OPS, OPE estimators should work reasonably well among a range of candidate evaluation policies. However, this desideratum is often hard to achieve, because an estimator's accuracy usually vary substantially when applied to different evaluation policies (Voloshin et al. 2019). Existing works on OPS deal with this instability of OPE by leveraging some high probability bounds on policy value estimates (Thomas, Theocharous, and Ghavamzadeh 2015a,b; Kuzborskij et al. 2021; Hao et al. 2021; Yang et al. 2020). In particular, Doroudi, Thomas, and Brunskill (2017) validate whether a fair policy comparison is possible based on a concentration inequality. Yang et al. (2021) estimate a pessimistic policy performance to alleviate an overestimation, which could lead to an inaccurate OPS. However, there is no existing work attempting to *switch* estimators adaptive to each evaluation policy among the set of candidates. As a potential application of our proposed *estimator selection* procedure, we show, in our experiments, that adaptive estimator selection can also significantly improve the accuracy of the *policy selection* task.

[1]For example, SLOPE can be used to tune the hyperparameters of Switch and DRos, however, it does not tell us which estimator is better. Our interest is thus to select the better estimator class.

3 Problem Formulation

This section first formulates OPE of contextual bandits and then the corresponding estimator selection problem, which is our primary interest.

3.1 Off-Policy Evaluation (OPE)

Let $x \in \mathcal{X}$ be a context vector (e.g., user demographics) that the decision maker observes when choosing an action. Let $r \in [0, r_{max}]$ be a reward (e.g., whether a coupon assignment results in an increase in revenue). Context and reward are sampled from some unknown distributions $p(x)$ and $p(r|x,a)$, where $a \in \mathcal{A}$ is a discrete action (i.e., a coupon). We call a function $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ a policy, where $\pi(a|x)$ is the probability of taking action a given context x.

In OPE, we are interested in accurately estimating the following *(policy) value*:

$$V(\pi) := \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}[r].$$

The most reliable way to estimate the policy value of *evaluation policy* π_e is to actually deploy π_e in an online environment (a.k.a. A/B tests). However, such an *on-policy* evaluation is often limited in practice, as it incurs large implementation costs (Matsushima et al. 2021) and there is the risk of deploying poor policies (Gilotte et al. 2018). Therefore, it is often desirable to evaluate the policy value via OPE at first and then pick only a small number of promising policies to be evaluated via A/B tests (Irpan et al. 2019).

To estimate the policy value, OPE leverages the logged bandit data $\mathcal{D}_b := \{(x_i, a_i, r_i)\}_{i=1}^n$ collected by a behavior policy as follows.

$$\{(x_i, a_i, r_i)\}_{i=1}^n \sim \prod_{i=1}^n p(x_i) \underbrace{p(j_i)\pi_{j_i}(a_i|x_i)}_{\pi_b(a_i|x_i)} p(r_i|x_i, a_i),$$

where π_b is the behavior policy, which may consist of $l\,(\geq 1)$ different data collection policies $\pi_1, \cdots, \pi_l$.[2] Here, $\mathcal{D}_b = \bigcup_{j=1}^l \mathcal{D}_j$ can be seen as an aggregate of several logged datasets, each of which contains n_j observations collected by the j-th data collecting policy π_j.

The goal of OPE is then to estimate the aforementioned policy value of evaluation policy using only the logged data: $V(\pi_e) \approx \hat{V}(\pi_e; \mathcal{D}_b)$. The accuracy of an estimator $\hat{V}$ is typically quantified by the *mean-squared-error* (MSE):

$$\begin{aligned} &\mathrm{MSE}(\hat{V}; \pi_e, \pi_b, n) \qquad (1)\\ &:= \mathbb{E}_{\mathcal{D}_b}\left[(\hat{V}(\pi_e; \mathcal{D}_b) - V(\pi_e))^2\right] \\ &= (\mathrm{Bias}(\hat{V}; \pi_e, \pi_b, n))^2 + \mathbb{V}_{\mathcal{D}_b}(\hat{V}; \pi_e, \pi_b, n), \qquad (2)\end{aligned}$$

As suggested in Eq. (2), achieving a reasonable bias-variance tradeoff is critical in enabling an accurate OPE. This motivates many estimators to be developed, including

[2]This is a general formulation, including the standard setting with a single data collection policy ($l = 1$) as a special case. Moreover, our setting is fundamentally different from the multiple logger setting of Agarwal et al. (2017); Kallus, Saito, and Uehara (2021), which assume the deterministic behavior policy assignment.

DM (Beygelzimer and Langford 2009), IPS (Precup, Sutton, and Singh 2000), DR (Dudík et al. 2014), Switch (Wang, Agarwal, and Dudik 2017), DRos (Su et al. 2020), and DR-λ (Metelli, Russo, and Restelli 2021).[3] However, as we have already argued, many empirical studies imply that there is no estimator that is universally the best (Voloshin et al. 2019; Saito et al. 2021a,b). This empirical evidence leads us to study the data-driven *estimator selection* problem for OPE, which we describe in detail below.

3.2 Estimator Selection for OPE

The goal of estimator selection is to select the most accurate estimator (which may change depending on a given OPE task) from a candidate pool of estimator classes $\mathcal{V} := \{\hat{V}_m\}_{m=1}^M$. An ideal strategy for this estimator selection task would be to pick the estimator achieving the lowest MSE:

$$m^* := \underset{m\in\{1,\dots,M\}}{\arg\min}\ \mathrm{MSE}(\hat{V}_m; \pi_e, \pi_b, n). \tag{3}$$

Unfortunately, however, Eq. (3) is infeasible because the MSE depends on the policy value of the evaluation policy, which is arguably unknown. Therefore, we instead consider performing an *offline* estimator selection based on an estimated MSE.

$$\hat{m} := \underset{m\in\{1,\dots,M\}}{\arg\min}\ \widehat{\mathrm{MSE}}(\hat{V}_m; \mathcal{D}_b).$$

Although there is no existing literature that formally discusses this estimator selection problem, the following describes a (non-adaptive) heuristic as a reasonable baseline.

Non-Adaptive Heuristic One possible approach to estimate the MSE is to naively regard one of the data collection policies π_j as a *pseudo* evaluation policy. Then, the *non-adaptive* heuristic estimates the MSE of a given estimator $\hat{V}$ as follows.[4]

$$\widehat{\mathrm{MSE}}(\hat{V}; \mathcal{D}_b) := \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\left(\hat{V}(\pi_j^{(s)}; \mathcal{D}_{b\setminus j}^{*(s)}) - V_{\mathrm{on}}(\pi_j^{(s)}; \mathcal{D}_j^{(s)})\right)^2,$$

where π_j is a pseudo evaluation policy. $V_{\mathrm{on}}(\pi_j; \mathcal{D}_j) := \sum_{i=1}^{n_j} r_i/n_j$ is its on-policy policy value estimate. $\mathcal{D}_{b\setminus j} := \mathcal{D}_b \setminus \mathcal{D}_j$ is the logged data collected by the corresponding (pseudo) behavior policy. $\mathcal{D}_b^*$ indicates bootstrapped samples of $\mathcal{D}_b$, and $\mathcal{S}$ is a set of random seeds for bootstrap. π_j is either randomly picked among available data collection policies (Saito et al. 2021b) or is fixed (Saito et al. 2021a; Saito, Udagawa, and Tateno 2021) for every random seed.

A critical pitfall of this heuristic is that it cannot accurately estimate the MSE when the data collection policies (one of which is used as the pseudo evaluation policy) are totally different from π_e. In fact, as we will show in the synthetic experiment, the non-adaptive heuristic often fails to choose an appropriate OPE estimator when there is a large divergence between the true evaluation policy and the pseudo evaluation policy. Unfortunately, such an undesirable situation is often the case in practice, because we usually want to evaluate counterfactual policies that have never been deployed. This motivates us to develop a novel *adaptive* estimator selection procedure that can estimate the estimators' MSE by taking the task-specific configurations (such as the evaluation policy) into account.

[3] We provide the definition and important statistical properties of these estimators in Appendix D.

[4] This heuristic has been used in some empirical studies of OPE (Saito et al. 2021a; Saito, Udagawa, and Tateno 2021; Saito et al. 2021b), however, the estimator selection problem itself has not yet been formally formulated in the OPE literature.

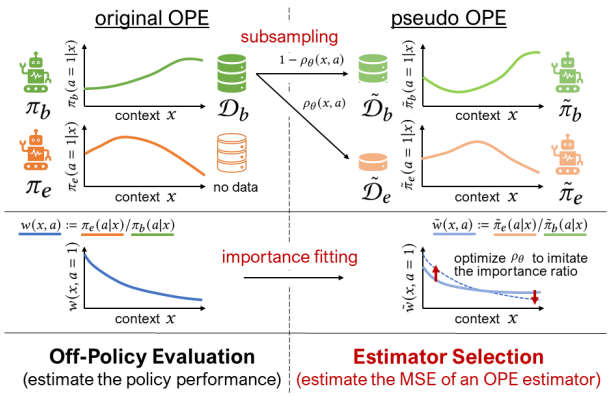

Figure 2: High level overview of Policy-Adaptive Estimator Selection via Importance Fitting (PAS-IF)

4 Our Adaptive Approach

This section proposes a new estimator selection procedure called ***P**olicy-**A**daptive Estimator **S**election via **I**mportance **F**itting* (PAS-IF).

Our key idea is to subsample the logged data and generate subpopulations to accurately estimate the MSE for a range of evaluation policies. For this, we introduce a subsampling rule $\rho_\theta : \mathcal{X}\times\mathcal{A}\to(0,1)$ parameterized by θ. ρ_θ allocates each observation in $\mathcal{D}_b$ to pseudo evaluation dataset $\tilde{\mathcal{D}}_e$ or pseudo behavior dataset $\tilde{\mathcal{D}}_b$ where $\rho_\theta(x,a)\in(0,1)$ is the probability of the data $(x,a,\cdot)$ being allocated to $\tilde{\mathcal{D}}_e$. Under this formulation, the pseudo datasets can be seen as generated by the following pseudo policies:

$$\tilde{\pi}_e(a|x) := \pi_b(a|x)\frac{\rho_\theta(x,a)}{\mathbb{E}_{\pi_b(a|x)}[\rho_\theta(x,a)]},$$
$$\tilde{\pi}_b(a|x) := \pi_b(a|x)\frac{1-\rho_\theta(x,a)}{1-\mathbb{E}_{\pi_b(a|x)}[\rho_\theta(x,a)]},$$

This formulation allows us to control the data generation process of $\tilde{\mathcal{D}}_e$ and $\tilde{\mathcal{D}}_b$ *adaptive* to a given OPE task by appropriately optimizing the subsampling rule ρ_θ. In particular, when optimizing ρ_θ, we try to imitate the true importance ratio $w(x,a) := \pi_e(a|x)/\pi_b(a|x)$, as it plays a significant role in determining the bias-variance tradeoff of OPE (Voloshin et al. 2019). Specifically, we minimize the following squared distance between the importance ratio induced by the true policies and that induced by the pseudo policies:

$$D(\pi,\tilde{\pi}) := \mathbb{E}_{p(x)\pi_b(a|x)}[(w(x,a)-\tilde{w}(x,a))^2],$$

Algorithm 1: Policy-Adaptive Estimator Selection via Importance Fitting (PAS-IF)

Input: a candidate set of OPE estimators $\mathcal{V}$, logged bandit dataset $\mathcal{D}_b$, evaluation policy π_e, target partition rate k, learning rate η, regularization coefficient λ, maximum steps T, a set of random seeds $\mathcal{S}$

Output: selected OPE estimator $\hat{V}_{\hat{m}}$

for $m = 1, \ldots, M$ **do**
  for $s \in \mathcal{S}$ **do**
    $\mathcal{D}_b^{*(s)} \leftarrow \mathrm{Bootstrap}(\mathcal{D}_b; s)$
    Initialize subsampling parameters θ
    for $t = 1, \ldots, T$ **do**
      $\theta \leftarrow \theta - \eta \left(\frac{\partial D}{\partial \theta} + \lambda \frac{\partial R}{\partial \theta} \right)$ ▷ importance fitting
    end for
    $\tilde{\mathcal{D}}_e^{*(s)}, \tilde{\mathcal{D}}_b^{*(s)} \leftarrow \mathrm{Subsample}(\mathcal{D}_b^{*(s)}; \rho_\theta)$
    $z_s \leftarrow (\hat{V}_m(\tilde{\pi}_e; \tilde{\mathcal{D}}_b^{*(s)}) - V_{\mathrm{on}}(\tilde{\pi}_e; \tilde{\mathcal{D}}_e^{*(s)}))^2$
  end for
  $\widehat{\mathrm{MSE}}(\hat{V}_m; \mathcal{D}_b) \leftarrow \sum_{s \in \mathcal{S}} z_s / |\mathcal{S}|$
end for
$\hat{m} \leftarrow \arg\min_{m \in \{1, \ldots, M\}} \widehat{\mathrm{MSE}}(\hat{V}_m; \mathcal{D}_b)$

where we define $\tilde{w}(x,a) := \tilde{\pi}_e(a|x)/\tilde{\pi}_b(a|x)$. We minimize this *importance fitting* objective by gradient decent where the gradient of $d(x,a) := (w(x,a) - \tilde{w}(x,a))^2$ is given as:[5]

$$\frac{\partial d(x,a)}{\partial \rho_\theta} \propto \frac{(\tilde{w}(x,a) - w(x,a))}{(1-\rho_\theta(x,a))^2} \left(\frac{1}{\mathbb{E}_{\pi_b}[\rho_\theta(x,a)]} - 1 \right) \cdot \left(1 - \frac{\pi_b(x,a)\rho_\theta(x,a)(1-\rho_\theta(x,a))}{(\mathbb{E}_{\pi_b}[\rho_\theta(x,a)])(1-\mathbb{E}_{\pi_b}[\rho_\theta(x,a)])} \right),$$

This importance fitting step enables PAS-IF to estimate the MSE adaptive to a range of evaluation policies while preserving a bias-variance tradeoff of a given OPE task, thereby improving the quality of estimator selection compared to the non-adaptive heuristic. A potential concern is that the size of the pseudo behavior dataset $\tilde{\mathcal{D}}_b$ may deviate greatly from that of the original logged data $\mathcal{D}_b$ depending on ρ_θ. If the size of $\tilde{\mathcal{D}}_b$ is much smaller than that of $\mathcal{D}_b$, PAS-IF may prioritize the variance more than necessary when performing estimator selection. To alleviate this potential failure, we set a target partition rate k and apply the following regularization when optimizing ρ_θ.

$$R(\tilde{\pi}, k) := \mathbb{E}_{p(x)} \left[\left(\mathbb{E}_{\pi_b(a|x)} [\rho_\theta(x,a)] - k \right)^2 \right],$$

where we impose regularization on every context to preserve $p(x)$ between $\tilde{\mathcal{D}}_e$ and $\tilde{\mathcal{D}}_b$.

Combining the importance fitting objective $D(\cdot)$ and regularization $R(\cdot)$, our PAS-IF optimizes the subsampling rule ρ_θ by iterating the following gradient update.

$$\theta \leftarrow \theta - \eta \left(\frac{\partial D}{\partial \theta} + \lambda \frac{\partial R}{\partial \theta} \right),$$

where λ is a regularization coefficient and η is a learning rate. Note that, in our experiments, we pick the regularization coefficient λ so that $\mathbb{E}_{\pi_b(a|x)} [\rho_\theta(x,a)] \in [0.18, 0.22]$. By doing this, we ensure that $|\tilde{\mathcal{D}}_b|$ becomes sufficiently close to $|\mathcal{D}_b|$, while making $V_{\mathrm{on}}(\pi_e; \tilde{\mathcal{D}}_e) \approx V(\pi_e)$ reasonably accurate. We later demonstrate that this heuristic tuning procedure works reasonably well in a range of settings.

Once we optimize ρ_θ for a given OPE task, we subsample the logged data to create pseudo datasets and estimate the MSE of an estimator $\hat{V}$ as follows.

$$\widehat{\mathrm{MSE}}(\hat{V}; \mathcal{D}_b) := \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\hat{V}(\tilde{\pi}_e^{(s)}; \tilde{\mathcal{D}}_b^{*(s)}) - V_{\mathrm{on}}(\tilde{\pi}_e^{(s)}; \tilde{\mathcal{D}}_e^{*(s)}) \right)^2.$$

where $\mathcal{S}$ is a set of random seeds for bootstrap sampling, and $\mathcal{D}_b^{*(s)}$ is the s-th bootstrapped logged dataset sampled from $\mathcal{D}_b$. $\tilde{\mathcal{D}}_e^{*(s)}$ and $\tilde{\mathcal{D}}_b^{*(s)}$ are subsampled from $\mathcal{D}_b^{*(s)}$ based on $\tilde{\pi}_e^{(s)}$ and $\tilde{\pi}_b^{(s)}$, respectively. This bootstrapping procedure aims to stabilize PAS-IF. Algorithm 1 summarizes the whole estimator selection procedure based on PAS-IF.[6]

5 Synthetic Experiments

This section compares our PAS-IF with the non-adaptive heuristic in terms of estimator selection and OPS. Note that our synthetic experiment is implemented on top of *OpenBanditPipeline* (Saito et al. 2021a).[7] Our experiment code is available at **https://github.com/sony/ds-research-code/tree/master/aaai23-pasif** and Appendix E describes some additional experiment details.

5.1 Setup

Basic setting. To generate synthetic data, we first randomly sample 10-dimensional context x, independently and normally distributed with zero mean. We also set $|\mathcal{A}| = 10$. The binary rewards r are sampled from the Bernoulli distribution as $r \sim Bern(q(x,a))$ where $q(x,a)$ is `obp.dataset.logistic_reward_function`.

Data Collection We define our behavior policy π_b based on the two different data collection policies π_1 and π_2:

$$\pi_j(a|x) := \frac{\exp(\beta_j \cdot q(x,a))}{\sum_{a' \in \mathcal{A}} \exp(\beta_j \cdot q(x,a'))}, \tag{4}$$

where $j \in \{1,2\}$ and $q(x,a) := \mathbb{E}[r|x,a]$ is the expected reward. β_j is an inverse temperature parameter of the softmax function. A positive value of β leads to a near-optimal policy, while a negative value leads to a bad policy. When $\beta = 0$, π_j is identical to uniform random. The logged dataset contains $n = 2,000$ observations with $p(j=1) = p(j=2) = 1/2$, and we try two different sets of data collection policies: (i) $(\beta_1, \beta_2) = (-2, 2)$ and (ii) $(\beta_1, \beta_2) = (3, 7)$.

[5] Appendix A provides how we derive these gradients.

[6] In addition to the primary benefit (i.e., *adaptive* estimator selection), PAS-IF is also able to relax some assumptions about the data collection policies compared to the non-adaptive heuristic. Appendix B discusses this additional benefit of PAS-IF in detail.

[7] **https://github.com/st-tech/zr-obp**

Evaluation Policies We also follow Eq. (4) to define evaluation policies and vary $\beta_e \in \{-10, -9, \cdots, 10\}$ in the estimator selection task to evaluate the estimator selection accuracy for various evaluation policies. For the policy selection task, we prepare 20 candidate policies learned by different policy learning methods to make OPS reasonably hard. We describe how to define the candidate policies for the policy selection experiment in detail in Appendix E.

Candidate OPE Estimators We construct the candidate set of estimator classes ($\mathcal{V}$) by DM, IPSps, DRps, SNIPS, Switch, DRos, IPS-λ, and DR-λ.[8] We also use three different models to constract $\hat{q}$ for model-based estimators such as DM and DR, which results in a total of $M = 21$ candidate OPE estimators. Note that we perform SLOPE (Su, Srinath, and Krishnamurthy 2020) to tune each estimator's built-in hyperparameter before performing estimator selection to simulate a realistic situation where we combine SLOPE and PAS-IF to improve the end-to-end OPE pipeline.

Compared Methods We compare PAS-IF with the non-adaptive heuristic. For PAS-IF, we set $\mathcal{S} = \{0, 1, \ldots, 9\}$, $k = 0.2$, $\eta = 0.001$, and $T = 5,000$, and select the regularization coefficient λ from $\{10^{-1}, 10^0, 10^1, 10^2, 10^3\}$ by a procedure described in Section 4.

Evaluation Metrics We quantify the estimator selection accuracy of PAS-IF and the non-adaptive heuristic by the following metrics.

- **Relative Regret (e)** (lower is better): This evaluates the accuracy of the estimator selected by each method compared to the most accurate estimator.

$$\mathrm{rRegret}^{(e)} := \frac{\mathrm{MSE}(\hat{V}_{\hat{m}}; \pi_e, \pi_b, n)) - \mathrm{MSE}(\hat{V}_{m^*}; \pi_e, \pi_b, n)}{\mathrm{MSE}(\hat{V}_{m^*}; \pi_e, \pi_b, n)}.$$

- **Rank Correlation (e)** (higher is better): This is the Spearman's rank correlation between the true and estimated MSE. This metric evaluates how well each method preserves the ranking of the candidate estimators in terms of their MSE.

In addition to the above metrics regarding estimator selection, we also evaluate PAS-IF and the non-adaptive heuristic in terms of the quality of the resulting OPS. Note that OPS is an important application of estimator selection whose aim is to select the best-performing policy based on OPE as:

$$\hat{\pi}_e := \underset{\pi_e \in \Pi_e}{\arg\max}\, \hat{V}_{\hat{m}}(\pi_e; \mathcal{D}_b),$$

where $\hat{V}_{\hat{m}}$ is the OPE estimator selected either by PAS-IF or the non-adaptive heuristic. We can evaluate the estimator selection methods with respect to their resulting OPS quality using the following metrics (Paine et al. 2020).

- **Relative Regret (p)** (lower is better): This metric measures the performance of the *policy* $\hat{\pi}_e$ selected based on PAS-IF or the non-adaptive heuristic compared to the best policy π_e^* among the candidates in Π_e.

$$\mathrm{rRegret}^{(p)} := \frac{V(\pi_e^*) - V(\hat{\pi}_e)}{V(\pi_e^*)}.$$

- **Rank Correlation (p)** (higher is better): This is the Spearman's rank correlation between the ground-truth performance of the candidate policies ($\{V(\pi_e)\}_{\pi_e \in \Pi_e}$) and those estimated by the selected estimator ($\{\hat{V}_{\hat{m}}(\pi_e)\}_{\pi_e \in \Pi_e}$).

[8] Appendix D defines and describes these OPE estimators in detail.

5.2 Result

Figure 3 shows Relative Regret (e) in estimator selection where the results are averaged over 100 simulations performed with different seeds. The result clearly demonstrates that our PAS-IF substantially improves relative regret compared to the non-adaptive heuristic for a range of evaluation policies (β_e). In particular, we observe that PAS-IF is able to identify an accurate OPE estimator (that has a low regret) even in cases where the behavior and evaluation policies deviate greatly (i.e., β_e is different from β_1 and β_2), while the non-adaptive heuristic fails dramatically. We also observe the similar trends in terms of the rank correlation metric, which we report in Appendix E. These results demonstrate that being adaptive to a given OPE task such as evaluation policy is crucial for effective estimator selection.

Table 2 compares PAS-IF and the non-adaptive heuristic in terms of their OPS quality when $(\beta_1, \beta_2) = (-2.0, 2.0)$.[9] The results indicate that PAS-IF is better in terms of both relative regret and rank correlation. Moreover, PAS-IF has a smaller standard deviation for both metrics, suggesting its stability. These observations indicate that an adaptive estimator selection also has a substantial positive benefit on the downstream OPS task.

6 Real-World Experiment

In this section, we apply PAS-IF to a real e-commerce application and demonstrate its practical benefits.

Setup For this experiment, we conducted a data collection A/B test in April 2021 on an e-commerce platform whose aim is to optimize a coupon assignment policy that facilitates user consumption. Thus, a is a coupon assignment variable where there are six different types of coupons ($|\mathcal{A}| = 6$). r is the revenue within the 7-day period after the coupon assignment. During data collection, we deployed three different coupon assignment policies (π_1, π_2, and π_3) and assign them (almost) equally to the users, resulting in $n = 40,985$.[10]

We compare PAS-IF with the non-adaptive heuristic for three different settings ($j' = 1, 2, 3$). Each setting regards $\pi_{j'}$ as the evaluation policy and $\mathcal{D}_b = \mathcal{D} \setminus \mathcal{D}_{j'}$ as the logged data. For example, when we consider π_1 as an evaluation policy, $\mathcal{D}_b$ is $\mathcal{D}_2 \cup \mathcal{D}_3$. Then, we use $V_{\mathrm{on}}(\pi_{j'}; \mathcal{D}_{j'})$ (which

[9] We observe the similar results when $(\beta_1, \beta_2) = (3.0, 7.0)$ in Appendix E.

[10] $p(j = 1) \approx p(j = 2) \approx p(j = 3) \approx 1/3$.

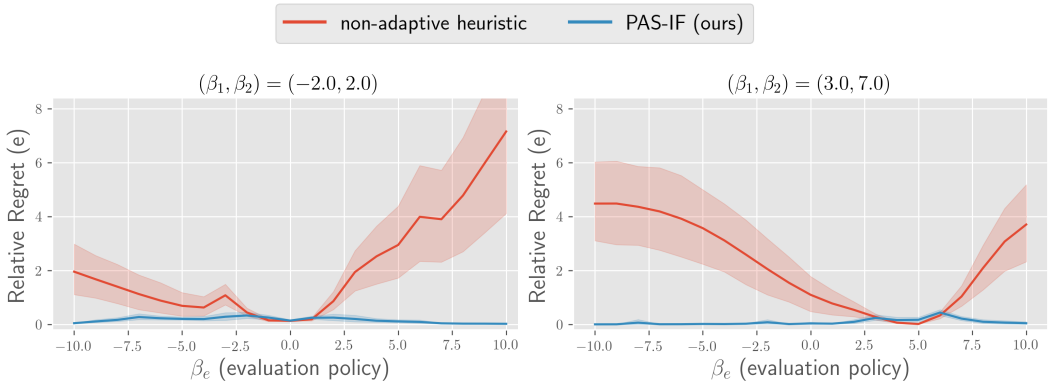

Figure 3: Relative Regret (e) with varying evaluation policies (β_e) in the synthetic experiment.

policy	$j = 1$		$j = 2$		$j = 3$	
metric	**rRegret (e)**	**Correlation (e)**	**rRegret (e)**	**Correlation (e)**	**rRegret (e)**	**Correlation (e)**
Heuristic	1.755 ($\pm$ 1.03)	-0.034 ($\pm$ 0.50)	1.827 ($\pm$ 3.04)	0.192 ($\pm$ 0.37)	3.580 ($\pm$ 1.67)	0.015 ($\pm$ 0.23)
PAS-IF	**1.609** ($\pm$ 0.66)	**0.002** ($\pm$ 0.28)	**0.691** ($\pm$ 0.70)	**0.342** ($\pm$ 0.54)	**1.889** ($\pm$ 1.32)	**0.124** ($\pm$ 0.31)

Table 1: Relative Regret (e) and Rank Correlation (e) in estimator selection in the real-world experiment. A lower value is better for Relative Regret (e), while a higher value is better for Rank Correlation (e). The values in parentheses indicate the standard deviation of the metrics estimated over 10 experiment runs.

	rRegret (p)	**Correlation (p)**
Heuristic	0.0341 ($\pm$ 0.062)	0.7452 ($\pm$ 0.621)
PAS-IF	**0.0077** ($\pm$ 0.014)	**0.9619** ($\pm$ 0.021)

Table 2: Relative Regret (p) and Rank Correlation (p) in OPS in the synthetic experiment: $(\beta_1, \beta_2) = (-2.0, 2.0)$. A lower value is better for Relative Regret (p), while a higher value is better for Rank Correlation (p). The values in parentheses indicate the standard deviation of the metrics.

is not available to PAS-IF and the non-adaptive heuristic) to estimate the true MSE of an estimator $\hat{V}$ as follows.

$$\mathrm{MSE}_{\mathrm{on}}(\hat{V}; \mathcal{D}) := \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\hat{V}(\pi_{j'}; \mathcal{D}_b^{*(s)}) - V_{\mathrm{on}}(\pi_{j'}; \mathcal{D}_{j'}) \right)^2,$$

where $\mathcal{S} := \{0, \ldots, 9\}$ is a set of random seeds for bootstrap sampling. We evaluate the estimator selection performance of PAS-IF and the non-adaptive heuristic by calculating the metrics regarding estimator selection using $\mathrm{MSE}_{\mathrm{on}}(\cdot)$. The other experimental setups are the same as in Section 5.1.

Result Table 1 summarizes the mean and standard deviation of **Relative Regret (e)** and **Rank Correlation (e)** for the three evaluation policies ($j' = 1, 2, 3$). Similarly to the synthetic experiment, the results demonstrate that PAS-IF improves both relative regret and rank correlation in estimator selection for every evaluation policy. Moreover, the standard deviation of PAS-IF is smaller than the non-adaptive heuristic in almost all cases, indicating the stability of the proposed method. These results provide promising empirical evidence that our PAS-IF also works fairly well in practical situations by being adaptive to an OPE problem instance.

7 Conclusion

We explored the problem of *estimator selection for OPE*, which aims to identify the most accurate estimator among many candidate OPE estimators using only the logged data. Our motivation comes from the fact that the non-adaptive heuristic becomes virulent when applied to a range of evaluation policies, which is especially problematic in OPS. With the goal of enabling a more accurate estimator selection, we proposed PAS-IF, which subsamples the logged data and imitates the importance ratio induced by the true evaluation policy, resulting in an *adaptive* estimator selection. Comprehensive synthetic experiments demonstrate that PAS-IF significantly improves the accuracy of OPE and OPS compared to the non-adaptive heuristic, particularly when the evaluation policies are substantially different from the data collection policies. The real-world experiment provides additional evidence that PAS-IF enables a reliable OPE in a real bandit application. We hope that this work would serve as a building block for future studies of estimator selection for OPE.

References

Agarwal, A.; Basu, S.; Schnabel, T.; and Joachims, T. 2017. Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers. *KDD*, 687–696.

Beygelzimer, A.; and Langford, J. 2009. The Offset Tree for Learning with Partial Labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 129–138.

Doroudi, S.; Thomas, P. S.; and Brunskill, E. 2017. Importance Sampling for Fair Policy Selection. *Grantee Submission*.

Dudík, M.; Erhan, D.; Langford, J.; and Li, L. 2014. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29(4): 485–511.

Gilotte, A.; Calauzènes, C.; Nedelec, T.; Abraham, A.; and Dollé, S. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 198–206.

Gruson, A.; Chandar, P.; Charbuillet, C.; McInerney, J.; Hansen, S.; Tardieu, D.; and Carterette, B. 2019. Offline Evaluation to Make Decisions About Playlist Recommendation Algorithms. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, 420–428.

Hao, B.; Ji, X.; Duan, Y.; Lu, H.; Szepesvari, C.; and Wang, M. 2021. Bootstrapping Fitted Q-Evaluation for Off-Policy Inference. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 4074–4084. PMLR.

Irpan, A.; Rao, K.; Bousmalis, K.; Harris, C.; Ibarz, J.; and Levine, S. 2019. Off-Policy Evaluation via Off-Policy Classification. In *Advances in Neural Information Processing Systems*.

Jiang, N.; and Li, L. 2016. Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 652–661. PMLR.

Kallus, N.; Saito, Y.; and Uehara, M. 2021. Optimal Off-Policy Evaluation from Multiple Logging Policies. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 5247–5256. PMLR.

Kallus, N.; and Uehara, M. 2019. Intrinsically Efficient, Stable, and Bounded Off-Policy Evaluation for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 32.

Kiyohara, H.; Kawakami, K.; and Saito, Y. 2021. Accelerating Offline Reinforcement Learning Application in Real-Time Bidding and Recommendation: Potential Use of Simulation. *arXiv preprint arXiv:2109.08331*.

Kiyohara, H.; Saito, Y.; Matsuhiro, T.; Narita, Y.; Shimizu, N.; and Yamamoto, Y. 2022. Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, 487–497.

Kuzborskij, I.; Vernade, C.; Gyorgy, A.; and Szepesvári, C. 2021. Confident Off-Policy Evaluation and Selection through Self-Normalized Importance Weighting. In *International Conference on Artificial Intelligence and Statistics*, 640–648. PMLR.

Lepski, O. V.; and Spokoiny, V. G. 1997. Optimal Pointwise Adaptive Methods in Nonparametric Estimation. *The Annals of Statistics*, 25(6): 2512–2546.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643*.

Matsushima, T.; Furuta, H.; Matsuo, Y.; Nachum, O.; and Gu, S. 2021. Deployment-Efficient Reinforcement Learning via Model-Based Offline Optimization. In *International Conference on Learning Representations*.

Metelli, A. M.; Russo, A.; and Restelli, M. 2021. Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning. In *Advances in Neural Information Processing Systems*, volume 34.

Paine, T. L.; Paduraru, C.; Michi, A.; Gulcehre, C.; Zolna, K.; Novikov, A.; Wang, Z.; and de Freitas, N. 2020. Hyperparameter Selection for Offline Reinforcement Learning. *arXiv preprint arXiv:2007.09055*.

Precup, D.; Sutton, R. S.; and Singh, S. P. 2000. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, 759–766.

Saito, Y.; Aihara, S.; Matsutani, M.; and Narita, Y. 2021a. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Saito, Y.; and Joachims, T. 2021. Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 828–830.

Saito, Y.; and Joachims, T. 2022. Off-Policy Evaluation for Large Action Spaces via Embeddings. In *International Conference on Machine Learning*, 19089–19122. PMLR.

Saito, Y.; Udagawa, T.; Kiyohara, H.; Mogi, K.; Narita, Y.; and Tateno, K. 2021b. Evaluating the Robustness of Off-Policy Evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 114–123.

Saito, Y.; Udagawa, T.; and Tateno, K. 2021. Data-Driven Off-Policy Estimator Selection: An Application in User Marketing on An Online Content Delivery Service. *arXiv preprint arXiv:2109.08621*.

Strehl, A.; Langford, J.; Li, L.; and Kakade, S. M. 2010. Learning from Logged Implicit Exploration Data. In *Advances in Neural Information Processing Systems*, volume 23, 2217–2225.

Su, Y.; Dimakopoulou, M.; Krishnamurthy, A.; and Dudík, M. 2020. Doubly Robust Off-Policy Evaluation with Shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 9167–9176. PMLR.

Su, Y.; Srinath, P.; and Krishnamurthy, A. 2020. Adaptive Estimator Selection for Off-Policy Evaluation. *arXiv preprint arXiv:2002.07729*.

Su, Y.; Wang, L.; Santacatterina, M.; and Joachims, T. 2019. CAB: Continuous Adaptive Blending for Policy Evaluation

and Learning. In *International Conference on Machine Learning*, volume 84, 6005–6014.

Swaminathan, A.; and Joachims, T. 2015a. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research*, 16: 1731–1755.

Swaminathan, A.; and Joachims, T. 2015b. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, volume 28, 3231–3239.

Tang, S.; and Wiens, J. 2021. Model Selection for Offline Reinforcement Learning: Practical Considerations for Healthcare Settings. In *Machine Learning for Healthcare Conference*, 2–35. PMLR.

Thomas, P.; and Brunskill, E. 2016. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 2139–2148. PMLR.

Thomas, P.; Theocharous, G.; and Ghavamzadeh, M. 2015a. High Confidence Policy Improvement. In *Proceedings of the 32th International Conference on Machine Learning*, 2380–2388.

Thomas, P. S.; Theocharous, G.; and Ghavamzadeh, M. 2015b. High-Confidence Off-Policy Evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 3000–3006.

Voloshin, C.; Le, H. M.; Jiang, N.; and Yue, Y. 2019. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning. *arXiv preprint arXiv:1911.06854*.

Wang, Y.-X.; Agarwal, A.; and Dudik, M. 2017. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 3589–3597.

Yang, C.-H. H.; Qi, Z.; Cui, Y.; and Chen, P.-Y. 2021. Pessimistic Model Selection for Offline Deep Reinforcement Learning. *arXiv preprint arXiv:2111.14346*.

Yang, M.; Dai, B.; Nachum, O.; Tucker, G.; and Schuurmans, D. 2020. Offline Policy Selection under Uncertainty. *arXiv preprint arXiv:2012.06919*.

Zhang, S.; and Jiang, N. 2021. Towards Hyperparameter-free Policy Selection for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34.