PERSONALIZED ESTIMATOR AGGREGATION FOR OFF-POLICY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Off-policy evaluation (OPE) enables the assessment of counterfactual policies using only logged interaction data. Although numerous OPE estimators have been developed, their estimation accuracy depends strongly on factors such as the logging policy, the evaluation policy, and the available sample size, making it difficult for any single estimator to perform reliably across diverse practical settings. Existing approaches address this challenge through data-driven estimator selection or aggregation, but typically remain context-agnostic. In many applications, however, the most effective estimator can vary substantially across contexts or user segments. We therefore propose *Personalized Estimator Aggregation for Off-Policy Evaluation* (PerOPE), a framework that adaptively aggregates multiple base estimators according to contextual information. By accounting for contextual variation, PerOPE captures heterogeneity in estimator accuracy across the context space and learns personalized aggregation rules based on user features. We conduct extensive experiments on synthetic data and a real-world production dataset, demonstrating that PerOPE consistently outperforms existing estimator selection and aggregation methods across a broad range of experimental conditions.

1 INTRODUCTION

Off-policy evaluation (OPE) has become an essential tool in many practical domains, including personalized recommendation, online advertising, and search systems (Gilotte et al., 2018). It enables practitioners to estimate the performance of a new policy using only historical data collected under a different policy. This offline approach provides a safe and cost-effective alternative to online A/B testing (Saito & Joachims, 2021), which can be expensive, risky, and time-consuming. By enabling reliable policy assessment before deployment, OPE supports the development of improved decision-making systems while reducing the cost and risk of online experimentation.

Despite its broad utility, reliable OPE requires practitioners to make the difficult practical choice of selecting an appropriate estimator (Udagawa et al., 2023). A wide range of estimators with desirable theoretical properties have been proposed, from classical approaches (Precup et al., 2000; Beygelzimer & Langford, 2009; Thomas & Brunskill, 2016; Jiang & Li, 2016) to more sophisticated methods with additional tuning parameters (Wang et al., 2017; Su et al., 2020a; Sakhi et al., 2024). However, no single estimator performs best across all settings (Saito et al., 2021a). Estimation accuracy can depend strongly on properties of the data-generating process, including the evaluation policy (Voloshin et al., 2021), the sample size (Saito et al., 2021a), and the number of available actions (Saito & Joachims, 2022; Saito et al., 2023; Sachdeva et al., 2024). For example, the direct method can be relatively more reliable when the sample size is small, whereas doubly robust or inverse propensity scoring estimators often become advantageous as the sample size grows (Saito & Joachims, 2021). Selecting an appropriate estimator is therefore a crucial step toward reliable OPE.

To address the estimator-selection problem, several data-driven methods have been developed. Early work proposed heuristic strategies (Saito et al., 2021b) and methods that identify robust estimators based on their stability under data perturbations (Saito et al., 2021a). A substantial step forward was PAS-IF (Udagawa et al., 2023), which formulated estimator selection as a learning problem and enabled a more principled selection procedure. More recent methods further improve how estimators are selected or combined. OCV (Cief et al., 2025), for example, applies ideas from cross-validation to select a single estimator for the entire dataset by optimizing an estimate of OPE

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

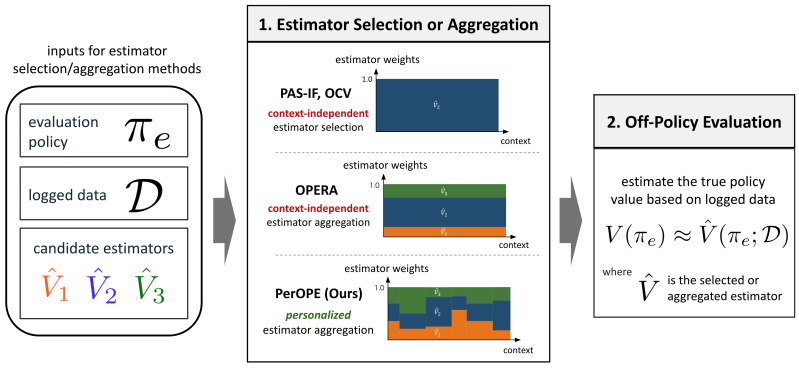


Figure 1: Overview of estimator selection and aggregation for OPE. Given an evaluation policy π_e, logged data $\mathcal{D}$, and a set of candidate estimators, conventional methods first determine a single estimator or a global aggregation of estimators. PAS-IF and OCV select a single estimator, whereas OPERA assigns context-*independent* weights to multiple estimators. In contrast, PerOPE (Ours) assigns estimator weights adaptively across the context space, enabling personalized estimator aggregation. The resulting estimator is then used to perform off-policy evaluation.

accuracy. In contrast, OPERA (Nie et al., 2024) constructs an estimator through a weighted aggregation of multiple base estimators rather than selecting only one. Nevertheless, its aggregation remains *context-independent*; the same aggregation weights are applied to all data points. This overlooks an important source of heterogeneity, i.e., the most suitable estimator may vary across users or contexts. Different user segments can exhibit substantially different characteristics, such as the divergence between the logging and evaluation policies or the level of noise in observed rewards. Ignoring such contextual variation can therefore leave substantial room for improving OPE accuracy.

We challenge this context-agnostic assumption by introducing a personalized perspective on estimator aggregation. Focusing on user features as contextual information, we propose *Personalized Estimator Aggregation for Off-Policy Evaluation* (PerOPE), which adaptively aggregates multiple base OPE estimators conditioned on the user context. PerOPE learns a decision tree whose splitting rules are chosen to reduce estimated mean squared error (Udagawa et al., 2023). Each leaf represents a data-driven segment of the context space and is associated with an aggregation of base estimators optimized for that segment. As illustrated in Figure 1, this formulation generalizes both global estimator selection and context-independent aggregation by allowing estimator weights to vary across contexts, thereby enabling more accurate and reliable OPE under heterogeneous data.

We conduct extensive experiments on synthetic data and an A/B test dataset collected from a real-world production system. Across a wide range of experimental settings, PerOPE consistently and substantially improves upon existing estimator selection and aggregation methods, demonstrating the practical value of context-aware personalization of OPE estimators.

Our main contributions are threefold. **(1)** We introduce personalized estimator aggregation for OPE, allowing estimator choice and aggregation to vary across the context space rather than being fixed globally. **(2)** We develop PerOPE, a practical algorithm that realizes this idea through a decision tree optimized using estimated OPE accuracy. **(3)** We provide extensive empirical evidence that PerOPE improves upon existing selection and aggregation methods on both synthetic and production data.

2 Related Work

This section reviews prior work and clarifies how our approach differs from these lines of research.

Off-Policy Evaluation (OPE). OPE aims to estimate the expected performance of an evaluation policy π_e using logged data collected under a different logging policy π_0. Foundational estimators include Inverse Propensity Scoring (IPS) (Precup et al., 2000; Strehl et al., 2010), the Direct Method (DM) (Beygelzimer & Langford, 2009), and Doubly Robust (DR) estimators (Thomas & Brunskill, 2016; Jiang & Li, 2016; Dudík et al., 2014). IPS is unbiased when the propensity scores, i.e., the action probabilities under π_0, are correctly specified, but can suffer from high variance. DM relies

on a learned reward model and can achieve low variance, but may incur substantial bias when the model is misspecified. DR combines IPS and DM and remains unbiased when either the propensity model or the reward model is correctly specified.

The accuracy of an OPE estimator is fundamentally governed by its bias–variance trade-off. A substantial body of work has therefore developed more sophisticated estimators that improve this trade-off through clipping, shrinkage, and other regularization techniques (Wang et al., 2017; Su et al., 2020a; Metelli et al., 2021). While designing improved estimators is an important research direction, our work addresses a complementary practical problem: given a collection of candidate estimators, how should they be selected or aggregated to achieve accurate OPE for a particular task?

Estimator Selection/Aggregation for OPE. Estimator selection is motivated by the fact that no single estimator performs best across all OPE settings. The appropriate choice depends on properties of the logging and evaluation policies, the sample size, noise level, and the underlying data distribution, making data-driven selection important for reliable OPE. Early work proposed heuristic strategies motivated by practical applications such as marketing (Saito et al., 2021b), as well as approaches that identify robust estimators based on their stability under data perturbations (Saito et al., 2021a). However, these methods have limited ability to adapt to the specific structure of the OPE problem and the observed logged data.

PAS-IF (Udagawa et al., 2023) provided a more principled formulation of data-driven estimator selection by explicitly learning which estimator is most appropriate for a given problem configuration. More recent approaches include OCV (Cief et al., 2025) and OPERA (Nie et al., 2024). OCV applies ideas from cross-validation to estimate the mean squared error of each candidate estimator and selects the estimator with the smallest estimated error. OPERA instead uses a bootstrap-based procedure to construct a weighted aggregation of multiple estimators, allowing it to improve upon methods restricted to selecting a single estimator.

Existing approaches can therefore be broadly divided into methods that select a single estimator and methods that learn a single global aggregation of estimators. Although both can adapt to properties of the overall OPE problem, they remain *context-agnostic*; the selected estimator or aggregation weights are shared across all users and contexts. This implicitly assumes that the same estimator is appropriate throughout the context space, despite the fact that estimator accuracy may vary substantially across user segments or contexts. Our work addresses this limitation through *personalized aggregation*, allowing estimator weights to vary across context-dependent segments.

Hyperparameter Tuning for OPE. Estimator selection is also closely related to hyperparameter tuning for OPE. Many modern estimators, including DR with optimistic shrinkage (DRos) (Su et al., 2020a), Switch-DR (Wang et al., 2017), DR with clipping or pessimistic shrinkage (DRps) (Su et al., 2020a), and DR with subgaussian importance sampling (DR-λ) (Metelli et al., 2021), introduce hyperparameters that control the bias–variance trade-off and consequently affect estimation accuracy. Data-driven estimator selection methods can naturally be extended to this setting by treating each hyperparameter configuration as a distinct candidate estimator.

SLOPE (Su et al., 2020b) is specifically designed for hyperparameter tuning in OPE and uses Lepski's principle to select a hyperparameter without explicitly estimating mean squared error. However, its monotonicity assumption restricts it to tuning a single hyperparameter and prevents direct application to estimators with multiple hyperparameters. In contrast, our framework can treat arbitrary hyperparameter configurations as candidate estimators, enabling personalized selection or aggregation even when multiple hyperparameters are tuned jointly.

3 Problem Formulation

This section formalizes the conventional OPE problem and then introduces the estimator selection and aggregation problem underlying existing approaches.

3.1 Off-Policy Evaluation (OPE)

We consider a standard contextual bandit setting. Let $\mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^n$ denote a logged dataset, where $x_i \in \mathcal{X}$ is a context vector, such as user demographics, $a_i \in \mathcal{A}$ is a discrete action, such

as a promotional email or music recommendation, and $r_i \in \{0, r_{\max}\}$ is an observed reward, such as a click or conversion. The logged data are generated as follows. A context x is drawn from an unknown distribution $p(x)$, an action a is selected according to a logging policy $\pi_0(a \mid x)$, and a reward r is sampled from a conditional distribution $p(r \mid x, a)$.

The goal of OPE is to estimate the value of a new evaluation policy π_e, defined as

$$V(\pi_e) := \mathbb{E}_{x\sim p(x),\, a\sim\pi_e(\cdot|x),\, r\sim p(\cdot|x,a)}[r] = \mathbb{E}_{x\sim p(x),\, a\sim\pi_e(\cdot|x)}[q(x,a)],$$

where $q(x,a) := \mathbb{E}_{r\sim p(\cdot|x,a)}[r \mid x, a]$ denotes the expected reward for a context–action pair. An OPE estimator $\hat{V}$ takes the logged dataset $\mathcal{D}$ collected under π_0 as input and returns an estimate $\hat{V}(\pi_e; \mathcal{D})$ of $V(\pi_e)$. The accuracy of an estimator $\hat{V}$ is measured by its mean squared error (MSE),

$$\mathrm{MSE}(\hat{V}) := \mathbb{E}_{\mathcal{D}}\left[\left(\hat{V}(\pi_e; \mathcal{D}) - V(\pi_e)\right)^2\right] = \mathrm{Bias}(\hat{V})^2 + \mathbb{V}_{\mathcal{D}}(\hat{V}).$$

The objective of OPE is therefore to obtain an estimate $\hat{V}(\pi_e; \mathcal{D})$ that is as close as possible to the true policy value $V(\pi_e)$, or equivalently, to minimize its MSE. However, prior work has shown that no single estimator is universally optimal (Voloshin et al., 2021; Saito et al., 2021a; Saito & Joachims, 2022). This motivates data-driven approaches to estimator selection and aggregation.

3.2 Estimator Selection/Aggregation for OPE

Suppose we are given a set of candidate estimators $\mathcal{V} = \{\hat{V}_1, \hat{V}_2, \ldots\}$. Depending on the selection or aggregation method, this set may consist of individual estimators such as IPS, DM, and DR, as in OCV (Cief et al., 2025), or aggregations of multiple estimators, such as linear combinations of IPS, DM, and DR, as in OPERA (Nie et al., 2024). The goal is to identify an estimator from $\mathcal{V}$ that performs well for the OPE task at hand. Ideally, one would select the oracle estimator

$$\hat{V}^{\mathrm{best}^*} = \arg\min_{\hat{V}\in\mathcal{V}} \mathrm{MSE}(\hat{V}).$$

This oracle choice is infeasible because evaluating the true MSE requires knowledge of $V(\pi_e)$, which is precisely the quantity OPE seeks to estimate. The same difficulty arises in OPE hyperparameter tuning, where $\mathcal{V}$ may instead represent multiple configurations of a single estimator.

Practical methods therefore approximate the oracle by minimizing an estimated MSE computed from the logged data as

$$\hat{V}^{\widehat{\mathrm{best}}} = \arg\min_{\hat{V}\in\mathcal{V}} \widehat{\mathrm{MSE}}(\hat{V}; \mathcal{D}).$$

The central challenge is thus to accurately estimate the MSE of each candidate estimator from $\mathcal{D}$.

Off-Policy Cross-Validation (OCV). OCV (Cief et al., 2025) selects a single estimator from the candidate set using cross-validation. The logged dataset $\mathcal{D}$ is repeatedly divided into training and validation subsets, denoted by $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{val}}$, respectively. For each candidate $\hat{V} \in \mathcal{V}$, OCV first computes $\hat{V}(\pi_e; \mathcal{D}_{\mathrm{train}})$. It then constructs a proxy for the true policy value, V_{proxy}, using an unbiased estimator such as IPS or DR. The squared difference $(\hat{V}(\pi_e; \mathcal{D}_{\mathrm{train}}) - V_{\mathrm{proxy}})^2$ serves as an estimate of the squared error, and averaging this quantity across splits yields an estimated MSE. To reduce the risk of selecting an estimator that appears favorable only because of sampling variability, OCV additionally applies the one-standard-error rule.

Automatic Offline Policy Evaluation with Re-weighted Aggregates of Multiple Estimators (OPERA). OPERA (Nie et al., 2024) considers aggregated estimators rather than restricting the candidate set to individual estimators. Specifically, its search space is

$$\mathcal{V} = \left\{ \sum_{m=1}^{M} \alpha_m \hat{V}_m \;\middle|\; \sum_{m=1}^{M} \alpha_m = 1 \right\},$$

where $\hat{V}_m$ denotes the m-th base estimator and α_m is its aggregation weight. OPERA uses bootstrap resampling of $\mathcal{D}$ to approximate the MSE of candidate aggregations, using the variance of the resulting estimates as a proxy for estimation error. It then optimizes the aggregation weights $\{\alpha_m\}_{m=1}^{M}$ to minimize the estimated MSE. By combining estimators with complementary error characteristics, such as estimators that tend to overestimate and underestimate the policy value, OPERA can achieve more accurate estimates than methods restricted to selecting a single estimator.

4 OUR PERSONALIZED OPE APPROACH

As discussed above, existing estimator selection and aggregation methods are context-agnostic; they use the same estimator or aggregation weights across the entire dataset. This design overlooks the fact that estimator accuracy can vary substantially across contexts. We address this limitation by introducing personalized estimator aggregation, which allows the aggregation rule to adapt to individual users or regions of the context space.

The key idea is that the most suitable OPE estimator can itself be context-dependent. By assigning different estimators or aggregation weights to different regions of the context space, we can exploit heterogeneity in estimator accuracy to improve the resulting policy-value estimate. To instantiate this idea, we propose *Personalized Estimator Aggregation for Off-Policy Evaluation* (PerOPE), a framework that formulates personalized estimator aggregation as a nonparametric learning problem.

We first express an OPE estimator in terms of its per-sample contributions, i.e.,

$$\hat{V}(\pi_e) := \frac{1}{n}\sum_{i=1}^{n} \hat{v}(\pi_e, x_i),$$

where $\hat{v}(\pi_e, x_i)$ denotes the contribution of context x_i to the resulting policy-value estimate. This representation covers a broad class of commonly used OPE estimators and makes explicit how their behavior can vary across contexts. For example, for IPS, $\hat{v}_{\mathrm{IPS}}(\pi_e, x_i) := \frac{\pi_e(a_i|x_i)}{\pi_0(a_i|x_i)} r_i$, where a_i is the action selected by the logging policy π_0 for context x_i. This per-sample representation allows PerOPE to adapt the estimator aggregation rule across different regions of the context space.

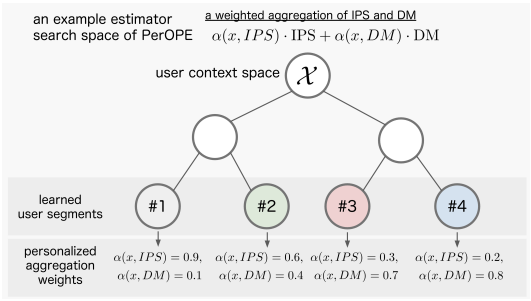


Figure 2: **Illustration of PerOPE.** PerOPE partitions the context space $\mathcal{X}$ into user segments using a learned decision tree and assigns a context-dependent set of aggregation weights to each segment. The illustration shows an example combining IPS and DM as $\alpha(x, \mathrm{IPS}) \cdot \mathrm{IPS} + \alpha(x, \mathrm{DM}) \cdot \mathrm{DM}$. By learning different aggregation rules across segments, PerOPE adapts OPE to heterogeneous user characteristics.

Given M base estimators with per-sample contributions $\{\hat{v}_m\}_{m=1}^{M}$, PerOPE considers the following personalized estimator class.

$$\mathcal{V}_{\mathrm{PerOPE}} = \left\{ \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} \alpha(x_i, m)\hat{v}_m(\pi_e, x_i) \,\middle|\, \forall x_i,\ \alpha(x_i, m) \geq 0,\ \sum_{m=1}^{M} \alpha(x_i, m) = 1 \right\}.$$

Here, $\alpha(x_i, m)$ denotes the weight assigned to the m-th base estimator for context x_i. The constraints ensure that the resulting estimator is a context-dependent convex combination of the candidate estimators. Unlike conventional aggregation methods that use a single global weight vector, PerOPE allows these weights to vary across the context space.

As illustrated in Figure 2, PerOPE parameterizes the context-dependent weights $\alpha(x_i, m)$ using a decision tree that recursively partitions the context space $\mathcal{X}$ into disjoint segments. Each leaf corresponds to a data-driven context segment and is associated with a single set of aggregation weights learned by minimizing an estimated MSE. At each node, PerOPE searches for a split that yields the largest reduction in estimated MSE, and recursive partitioning continues until the maximum tree depth K is reached or no candidate split provides further improvement. For a new context x_i, the aggregation weights are determined by the leaf to which x_i is assigned.

This tree-based formulation provides both personalization and interpretability while retaining flexibility in how estimation error is assessed. In particular, PerOPE is modular with respect to the MSE estimator; any suitable MSE estimation procedure can be used as a plug-in component, including those underlying OCV or OPERA. The resulting personalized estimator, denoted by $\hat{V}^{\widehat{\mathrm{best}}}_{\mathrm{PerOPE}}$, adapts estimator aggregation to the observed context. The complete procedure of PerOPE is provided in pseudocode in Appendix C.

Theoretical Analysis. We next establish two basic properties that clarify when personalization preserves desirable guarantees and why it can improve upon context-independent aggregation.

Theorem 1 (Unbiasedness) *Suppose that each base contribution is conditionally unbiased, i.e.,*

$$\mathbb{E}[\hat{v}_m(\pi_e, X) \mid X = x] = \sum_{a \in \mathcal{A}} \pi_e(a \mid x) q(x, a)$$

for every m and x. If the aggregation rule $\alpha(x, m)$ is fixed independently of the evaluation sample conditional on x, satisfies $\alpha(x, m) \geq 0$, and obeys $\sum_{m=1}^{M} \alpha(x, m) = 1$, then the resulting personalized estimator is unbiased for $V(\pi_e)$.

Theorem 1 shows that context-dependent aggregation does not inherently introduce bias when combining conditionally unbiased base estimators. In particular, the condition is satisfied when the aggregation rule is learned on data independent of the sample used to construct the final policy-value estimate, as can be implemented through sample splitting or cross-fitting.

Theorem 2 (Non-negative Oracle Improvement) *Let $\hat{V}^{\text{best}^*}_{\text{PerOPE}}$ denote the oracle estimator minimizing MSE over $\mathcal{V}_{\text{PerOPE}}$, and let $\hat{V}^{\text{best}^*}_{\text{no-per}}$ denote the oracle estimator minimizing MSE over the corresponding class of context-independent aggregations $\mathcal{V}_{\text{no-per}}$. Then,*

$$\text{MSE}\Big(\hat{V}^{\text{best}^*}_{\text{PerOPE}}(\pi_e)\Big) \leq \text{MSE}\Big(\hat{V}^{\text{best}^*}_{\text{no-per}}(\pi_e)\Big).$$

The result follows from $\mathcal{V}_{\text{no-per}} \subseteq \mathcal{V}_{\text{PerOPE}}$, since any context-independent aggregation can be recovered by setting $\alpha(x, m)$ to the same value for every x. Thus, at the oracle level, personalization cannot increase the minimum achievable MSE and can strictly improve it whenever different regions of the context space favor different aggregation rules. This provides a theoretical motivation for the empirical improvements observed in our experiments.

5 Synthetic Experiment

This section evaluates PerOPE on synthetic data under controlled settings designed to vary the degree of context dependence in estimator performance.

5.1 Setup

Data Generation Process. We generate a logged dataset $\mathcal{D}$ with sample size $n = 1000$. Each context is 5-dimensional, with features drawn independently from a standard normal distribution, and the action space contains $|\mathcal{A}| = 10$ actions. The reward function $q(x, a)$ consists of context-only, action-only, and context–action interaction terms; full details are provided in Appendix D. The binary reward is sampled from a Bernoulli distribution with mean $q(x, a)$. Logged data are collected under a softmax logging policy π_0 with inverse temperature β_0 defined as

$$\pi_0(a \mid x) = \frac{\exp(\beta_0 q(x, a))}{\sum_{a' \in \mathcal{A}} \exp(\beta_0 q(x, a'))},$$

where $q(x, a)$ is the true expected reward. We set $\beta_0 = 2$ throughout the experiments.

Evaluation Policies. To study how context-dependent heterogeneity affects estimator selection and aggregation, we construct evaluation policies whose stochasticity varies across contexts. Specifically, we define a softmax evaluation policy

$$\pi_e(a \mid x) = \frac{\exp(\beta(x) q(x, a))}{\sum_{a' \in \mathcal{A}} \exp(\beta(x) q(x, a'))},$$

where $\beta(x) = \beta_{\text{cons}} + \beta_{\text{coef}}(b^\top x)$. Here, β_{cons} is a constant, b is a coefficient vector whose entries are sampled from $\mathcal{N}(0, 1)$, and β_{coef} controls the strength of context dependence. We consider $\beta_{\text{cons}} \in \{2, 5\}$ and $\beta_{\text{coef}} \in \{0, 2.5, 5, 10, 20\}$. When $\beta_{\text{coef}} = 0$, policy stochasticity is constant across contexts. As β_{coef} increases, the degree of determinism varies more strongly across the context space, creating settings in which the most accurate estimator or estimator aggregation is more likely to depend on the context.

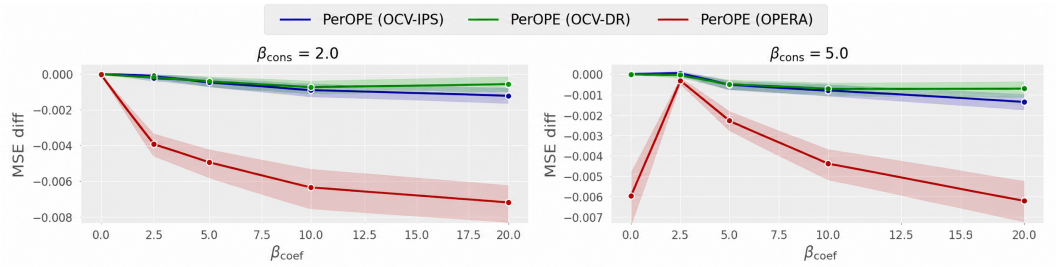


Figure 3: Difference in MSE between each PerOPE variant and its corresponding context-agnostic baseline on the synthetic dataset, computed as $\mathrm{MSE}(\mathrm{PerOPE}) - \mathrm{MSE}(\mathrm{Baseline})$. The x-axis denotes the context-dependence parameter β_{coef}; negative values indicate that PerOPE achieves lower MSE. The performance gap generally increases with β_{coef}, indicating that personalization becomes increasingly beneficial as estimator performance varies more strongly across contexts. Error bars denote 95% confidence intervals.

Candidate OPE Estimators. We use five standard hyperparameter-free estimators as candidates: $\{\mathrm{IPS}, \mathrm{DM}, \mathrm{DR}, \mathrm{SNIPS}, \mathrm{SNDR}\}$. For estimators requiring a reward model (DM, DR, and SNDR), we use logistic regression to estimate $\hat{q}(x, a)$. By restricting the candidate set to estimators without tuning parameters, we isolate the estimator selection and aggregation problem from the separate issue of OPE hyperparameter tuning.

Compared Methods. We compare PerOPE against two representative estimator selection and aggregation baselines, OCV and OPERA. For OCV, we consider two variants, OCV-IPS and OCV-DR, which use IPS and DR, respectively, as the internal estimator for MSE estimation. Hyperparameters for OCV and OPERA are provided in Appendix D. To ensure a controlled comparison, each PerOPE variant uses the same underlying MSE estimation procedure as its corresponding baseline. We therefore evaluate PerOPE (OCV-IPS), PerOPE (OCV-DR), and PerOPE (OPERA), where the term in parentheses denotes the underlying MSE estimation method. Unless otherwise specified, we set the maximum tree depth to $K = 4$ and use Optuna (Akiba et al., 2019) with 50 trials to efficiently search for the best split at each node.

5.2 Results

Figure 3 summarizes the synthetic results by reporting the difference in MSE between each PerOPE variant and its corresponding baseline as a function of β_{coef}, which controls the degree of context dependence in the evaluation policy.

Overall Estimator Selection/Aggregation Performance. As shown in Figure 3, PerOPE consistently improves upon its corresponding baseline across all values of β_{coef}. Moreover, the performance gap generally widens as β_{coef} increases. When β_{coef} is small or zero, $\beta(x)$ varies little across contexts, and the characteristics of the evaluation policy are therefore relatively homogeneous throughout the context space. In this regime, similar estimators tend to perform well across different contexts, leaving limited room for personalization and causing PerOPE to perform comparably to its context-agnostic counterparts. In contrast, larger values of β_{coef} induce stronger context-dependent variation in the evaluation policy, making the most suitable estimator increasingly likely to differ across regions of the context space. This is precisely the regime targeted by our PerOPE, and the observed gains become correspondingly larger. **These results demonstrate that adapting estimator aggregation to context-specific heterogeneity can provide substantial improvements in OPE accuracy over applying a single estimator or aggregation rule globally.**

It should be noted that OPERA exhibits substantially higher MSE than the OCV-based methods in this synthetic setting, as shown in Appendix D, which also explains the comparatively large gains of PerOPE (OPERA) in Figure 3. We attribute this behavior to the underlying MSE estimation procedure. OPERA relies on bootstrap resampling and primarily uses variation across bootstrap estimates as a proxy for estimation error. Our candidate set, however, contains biased estimators such as DM, SNIPS, and SNDR, for which the bias component can be non-negligible. Consequently, the result-

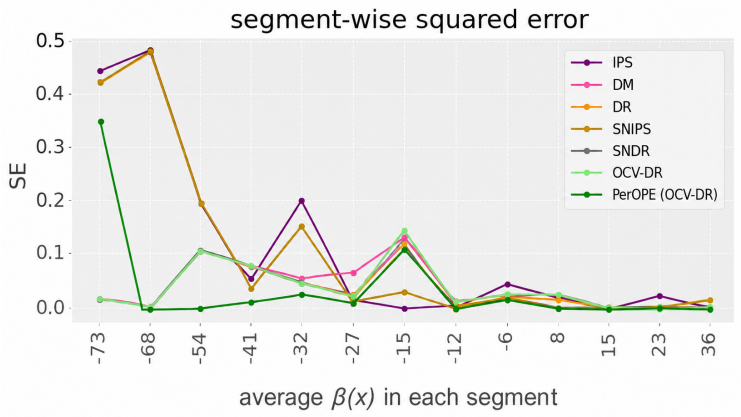


Figure 4: Case study of a learned PerOPE tree and segment-wise performance. The figure reports the squared error (SE) of different estimators within each leaf of a PerOPE (OCV-DR) tree trained with $\beta_{\text{cons}} = 5$ and $\beta_{\text{coef}} = 10$. Each group on the x-axis represents a distinct context segment, and the y-axis reports SE computed using samples belonging to that segment. The average value of $\beta(x)$ within each leaf is shown for reference.

ing MSE estimates can be less accurate in this setting. Nevertheless, replacing the global OPERA aggregation with context-dependent aggregation through PerOPE (OPERA) still yields substantial reductions in the resulting MSE. **This suggests that the benefits of personalization by our proposed method are not specific to a particular MSE estimation procedure.**

Inspecting a Learned PerOPE Tree. To better understand how PerOPE exploits contextual heterogeneity, we examine a tree learned by PerOPE (OCV-DR) with $\beta_{\text{cons}} = 5$ and $\beta_{\text{coef}} = 10$. Figure 4 reports segment-wise estimation performance across the learned leaves. PerOPE achieves the lowest squared error (SE) in most segments, indicating that the learned tree successfully adapts estimator aggregation across the context space rather than relying on a single global estimator.

The learned segments can also be interpreted in terms of the evaluation policy. In segments where the average $\beta(x)$ is positive, the evaluation policy tends to be more similar to the logging policy (recall that $\beta_0 = 2$), making the OPE problem relatively easier. Accordingly, most estimators achieve low SE in these regions. By contrast, when the average $\beta(x)$ is negative, the evaluation policy can differ substantially from the logging policy, making OPE more challenging. In these segments, the performance of individual base estimators varies considerably, while PerOPE maintains relatively low SE by adapting its estimator aggregation. This case study illustrates that PerOPE successfully identifies context segments with distinct estimation characteristics and learns aggregation rules that exploit these differences automatically in a data-driven manner.

Ablation Study. We additionally conduct several ablation studies, with full results provided in Appendix E. These experiments demonstrate that PerOPE extends naturally to OPE hyperparameter tuning, remains effective across different choices of the maximum tree depth K, and continues to improve upon its corresponding baselines across different logged-data sample sizes. Together, these results demonstrate that the benefits of personalized estimator aggregation are robust across a range of experimental configurations.

6 Real-World Experiment

This section evaluates PerOPE using logged data collected from a real-world production system.

Setup. We use logged data from a real A/B test conducted in April 2024 as part of a personalized email marketing system operated by a machine-learning policy at a private company. In the experiment, users received a targeted marketing email, and the action corresponds to one of three possible delivery timings. Each user is represented by a 30-dimensional feature vector, and the reward is a binary indicator of whether the user clicked the promotional email. The dataset $\mathcal{D}$ consists of logs

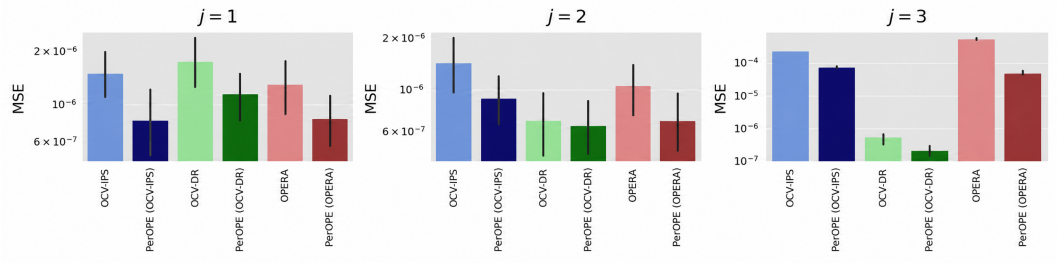


Figure 5: Comparison of the resulting MSE after estimator aggregation on the real-world dataset. Error bars denote 95% confidence intervals. Each baseline and its corresponding PerOPE variant share the same underlying MSE estimation procedure and are shown using the same color.

collected under three distinct policies, π_1, π_2, and π_3, which were deployed simultaneously during the A/B test. The total sample size is 204,191, partitioned into $\mathcal{D}_1$ ($n_1 = 87{,}878$), $\mathcal{D}_2$ ($n_2 = 89{,}193$), and $\mathcal{D}_3$ ($n_3 = 27{,}120$). Thus, $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$, allowing us to treat data collected under two policies as hypothetical offline logs and evaluate the remaining policy via OPE.

We use the same candidate OPE estimators $\mathcal{V}$ and the same baseline methods (OCV, OPERA, and PerOPE) as in the synthetic experiments. To capture richer relationships among context, action, and reward, we use a Random Forest classifier, rather than logistic regression, to estimate the reward model $\hat{q}(x, a)$ for DM, DR, and SNDR. The same Random Forest model is also used in the MSE estimation procedures for OCV-DR and PerOPE (OCV-DR).

To evaluate performance across different logging and evaluation policies, we consider three scenarios indexed by $j \in \{1, 2, 3\}$. In each scenario, π_j is treated as the evaluation policy, while $\mathcal{D}' = \mathcal{D} \setminus \mathcal{D}_j$ is used as the offline logged dataset for OPE. Because the true policy value $V(\pi_j)$ is unavailable in the production environment, we use the corresponding on-policy estimate from the A/B test as a proxy for the ground-truth policy value. We then evaluate each estimator-selection or aggregation method using the resulting estimated MSE and compute 95% confidence intervals.

Results. Figure 5 reports the estimated MSE of each method across the three real-world evaluation scenarios (three different evaluation policies). PerOPE consistently achieves lower MSE than its corresponding context-agnostic baseline. In particular, PerOPE (OCV-IPS) outperforms OCV-IPS in every evaluation setting. **These consistent improvements across different combinations of logging and evaluation policies show that the benefits of personalized estimator aggregation extend beyond the controlled synthetic setting to real-world production data.**

The favorable results obtained by our proposed method also provide empirical evidence that estimator accuracy can vary meaningfully across contexts in practice. Context-agnostic methods use a single estimator or aggregation rule across the entire context space and therefore cannot exploit this heterogeneity. In contrast, PerOPE learns context-dependent aggregation rules and leverages such variation to produce more accurate off-policy value estimates.

7 Conclusion

We introduced PerOPE, a personalized estimator aggregation framework for off-policy evaluation. Existing selection and aggregation methods typically use a single estimator or a fixed set of aggregation weights across the entire dataset, despite the fact that estimator accuracy can vary across users and contexts. PerOPE addresses this limitation by partitioning the context space into data-driven segments and learning an appropriate estimator aggregation for each segment.

Experiments on synthetic data and a real-world A/B test dataset demonstrate that PerOPE consistently improves upon conventional estimator selection and aggregation methods across a broad range of settings. The gains become particularly substantial when estimator performance exhibits stronger context dependence, while the real-world results demonstrate that the benefits of personalization also arise in practical production environments. These findings highlight context-dependent estimator aggregation as a promising direction for improving the reliability of off-policy evaluation.

AI USE STATEMENT

Generative AI tools were used for language editing, wording refinement, and title brainstorming. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pp. 129–138, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557040. URL https://doi.org/10.1145/1557019.1557040.

Matej Cief, Branislav Kveton, and Michal Kompan. Cross-validated off-policy evaluation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i15.33765. URL https://doi.org/10.1609/aaai.v39i15.33765.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29(4):485 – 511, 2014. doi: 10.1214/14-STS500. URL https://doi.org/10.1214/14-STS500.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 198–206, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159687. URL https://doi.org/10.1145/3159652.3159687.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 652–661. JMLR.org, 2016.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Allen Nie, Yash Chandak, Christina J. Yuan, Anirudhan Badrinath, Yannis Flet-Berliac, and Emma Brunskill. Opera: automatic offline policy evaluation with re-weighted aggregates of multiple estimators. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Noveen Sachdeva, Lequn Wang, Dawen Liang, Nathan Kallus, and Julian McAuley. Off-policy evaluation for large action spaces via policy convolution. In *Proceedings of the ACM Web Conference 2024*, pp. 3576–3585, 2024.

Yuta Saito and Thorsten Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pp. 828–830, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3473320. URL https://doi.org/10.1145/3460231.3473320.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19089–19122. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saito22a.html.

Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. Evaluating the robustness of off-policy evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pp. 114–123, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474245. URL https://doi.org/10.1145/3460231.3474245.

Yuta Saito, Takuma Udagawa, and Kei Tateno. Data-driven off-policy estimator selection: An application in user marketing on an online content delivery service, 2021b. URL https://arxiv.org/abs/2109.08621.

Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. In *International Conference on Machine learning*, pp. 29734–29759. PMLR, 2023.

Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. *Advances in Neural Information Processing Systems*, 37:80706–80755, 2024.

Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pp. 2217–2225, Red Hook, NY, USA, 2010. Curran Associates Inc.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9167–9176. PMLR, 2020a.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9196–9205. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/su20d.html.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 3231–3239, Cambridge, MA, USA, 2015. MIT Press.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2139–2148. JMLR.org, 2016.

Takuma Udagawa, Haruka Kiyohara, Yusuke Narita, Yuta Saito, and Kei Tateno. Policy-adaptive estimator selection for off-policy evaluation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26195. URL https://doi.org/10.1609/aaai.v37i8.26195.

Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3589–3597. JMLR.org, 2017.

A Off-Policy Estimators

Here, we summarize the definition and the important statistical properties of OPE estimators used in both synthetic and real-world data experiments.

A.1 Direct Method (DM)

DM (Beygelzimer & Langford, 2009) first trains a supervised machine learning model $\hat{q}(x,a)$ to estimate the mean reward function $q(x,a)$. DM estimates the policy value as follows.

$$\hat{V}_{\mathrm{DM}} := \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\pi_e(a|x_i)\hat{q}(x_i,a),$$

where $\hat{q}(x_i,a)$ is the predicted reward. DM tends to have low variance. It performs well when $\hat{q}$ is accurate. However, it suffers from high bias when $\hat{q}$ is inaccurate.

A.2 Inverse Propensity Scoring (IPS)

IPS (Precup et al., 2000) (Strehl et al., 2010) uses importance sampling to address the distribution shift between the logging and evaluation policies as follows:

$$\hat{V}_{\mathrm{IPS}} := \frac{1}{n}\sum_{i=1}^{n} w(x_i,a_i)r_i,$$

where $w(x_i,a_i) := \pi_e(a_i|x_i)/\pi_0(a_i|x_i)$ is the importance weight. IPS is unbiased under some identification assumptions such as full support ($\pi_e(a|x) > 0 \rightarrow \pi_0(a|x) > 0, \forall(x,a)$). However, when the logging and evaluation policies are dissimilar or the action space is large, IPS often suffers from high variance (Saito & Joachims, 2022).

A.3 Doubly Robust (DR)

DR (Thomas & Brunskill, 2016) (Jiang & Li, 2016)(Dudík et al., 2014) is a combination of IPS and DM to reduce the variance of IPS while retaining unbiasedness,

$$\hat{V}_{\mathrm{DR}} := \\ \frac{1}{n}\sum_{i=1}^{n}\left(w(x_i,a_i)(r_i-\hat{q}(x_i,a_i)) + \sum_{a\in\mathcal{A}}\pi_e(a|x_i)\hat{q}(x_i,a)\right).$$

When $\hat{q}$ is accurate, DR reduces the variance of IPS, remaining unbiased. However, DR can suffer from high variance when the action space is large (Saito & Joachims, 2022).

A.4 Clipping estimators (IPSps, DRps).

One possible approach to deal with the high variance of DR and IPS is to clip the large importance weight as follows.

$$w_{\mathrm{ps}}(x_i,a_i) := \min\{w(x_i,a_i),\lambda\},$$

where $\lambda \geq 0$ is the clipping threshold. When using the clipped importance weight, IPS and DR are no longer unbiased, but we can reduce their variance by capping large importance weights. We call these estimators pessimistic shrinkage estimators (IPSps and DRps) following (Su et al., 2020a).

A.5 Self-Normalized estimators (SNIPS, SNDR)

The self-normalized estimators (Swaminathan & Joachims, 2015) normalize the importance weight to gain some stability as follows.

$$w_{\mathrm{norm}}(x_i,a_i) := \frac{w(x_i,a_i)}{\sum_{j=1}^{n} w(x_j,a_j)/n}.$$

SN estimators are biased, but are still consistent.

A.6 SWITCH

Switch (Wang et al., 2017) aims to reduce the variance of DR by interpolating between DM and DR. Specifically, Switch uses DR only when the importance weight is not large. Otherwise, it uses DM, which has a low variance. Specifically, this estimator alternatively defines the importance weight as follows.

$$w_{\mathrm{switch}}(x_i, a_i) := w(x_i, a_i)\mathbb{I}\{w(x_i, a_i) \leq \lambda\},$$

where $\lambda \geq 0$ is the threshold hyperparameter. When $\lambda = 0$, Switch is identical to DM, while $\lambda \to \infty$ yields DR.

A.7 DR WITH OPTIMISTIC SHRINKAGE (DROS)

DRos (Su et al., 2020a) is motivated by the goal of reducing the variance of DR, which is similar to that of Switch, but does so by directly minimizing the sharp bound of MSE by applying the following shrunk version of the importance weight.

$$w_{\mathrm{os}}(x_i, a_i) := \frac{\lambda}{w(x_i, a_i)^2 + \lambda} w(x_i, a_i),$$

where $\lambda \geq 0$ is a parameter that controls the bias and variance. When $\lambda = 0$, $w_{\mathrm{os}}(x_i, a_i) = 0$ leading to DM. In contrast, when $\lambda \to \infty$, $w_{\mathrm{os}}(x_i, a_i) = w(x_i, a_i)$ leading to DR.

A.8 SUBGAUSSIAN IMPORTANCE SAMPLING ESTIMATORS (IPS-λ, DR-λ).

To improve the error concentration rate from the exponential rate to subgaussian rate, λ-estimators (Metelli et al., 2021) apply smooth shrinkage to the importance weight as follows.

$$w_\lambda(x_i, a_i) := ((1 - \lambda)w(x_i, a_i)^s + \lambda)^{\frac{1}{s}}$$

where $\lambda \in [0, 1]$ and $s \in (-\infty, \infty)$ are the hyperparameters.

B PROOF OF THEOREM 1

Here, we give the proof of theorem 1.

We assume that every estimator in $\{\hat{v}_m\}_{m=1}^M$ is unbiased. Then, the expected value of the PerOPE estimator is

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\hat{V}_{\mathrm{PerOPE}}(\pi_e; \mathcal{D})] \\
&= \mathbb{E}_{\mathcal{D}}[\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\alpha(x_i, m)\hat{v}_m(\pi_e, x_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}[\sum_{m=1}^{M}\alpha(x, m)\hat{v}_m(\pi_e, x)] \\
&= \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}[\sum_{m=1}^{M}\alpha(x, m)\hat{v}_m(\pi_e, x)] \\
&= \mathbb{E}_{p(x)}[\sum_{m=1}^{M}\alpha(x, m)\mathbb{E}_{\pi_0(a|x)p(r|x,a)}[\hat{v}_m(\pi_e, x)]].
\end{aligned}$$

Since every estimator in $\{\hat{v}_m\}_{m=1}^M$ is unbiased, $\hat{v}_m$ satisfies the following equation.

$$\mathbb{E}_{p(x)}[\mathbb{E}_{\pi_0(a|x)p(r|x,a)}[\hat{v}_m(\pi_e, x)]] = \mathbb{E}_{p(x)}[\sum_{a\in\mathcal{A}}\pi_e(a|x)q(x, a)].$$

By definition of unbiasedness, this equation is valid for $\forall p(x)$. This means that

$$\mathbb{E}_{\pi_0(a|x)p(r|x,a)}[\hat{v}_m(\pi_e, x)] = \sum_{a\in\mathcal{A}}\pi_e(a|x)q(x, a).$$

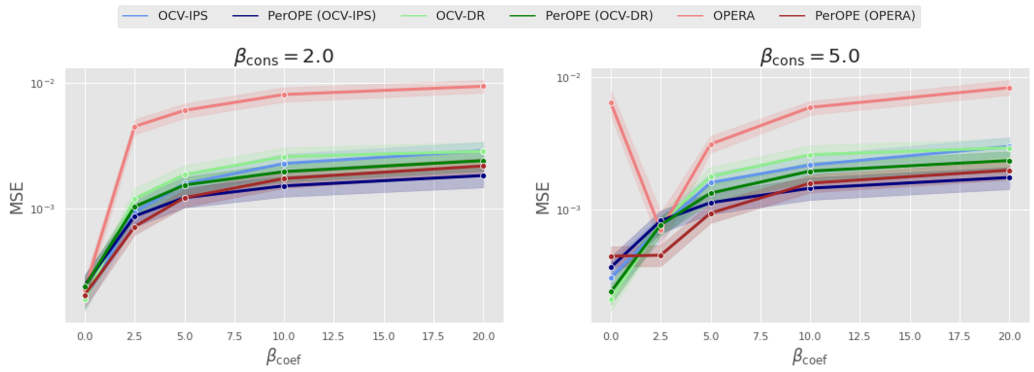


Figure 6: Comparison of MSE on the synthetic data. Error bars denote 95% confidence intervals over 200 trials.

Therefore, we can derive

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\hat{V}_{\mathrm{PerOPE}}(\pi_e;\mathcal{D})] \\
&= \mathbb{E}_{p(x)}[\sum_{m=1}^{M} \alpha(x,m)\mathbb{E}_{\pi_0(a|x)p(r|x,a)}[\hat{v}_m(\pi_e,x)]] \\
&= \mathbb{E}_{p(x)}[\sum_{a\in\mathcal{A}} \pi_e(a|x)q(x,a) \underbrace{\sum_{m=1}^{M} \alpha(x,m)}_{=1}] \\
&= \mathbb{E}_{p(x)}[\sum_{a\in\mathcal{A}} \pi_e(a|x)q(x,a)] \\
&= V(\pi_e).
\end{aligned}
$$

C Algorithmic Details

This section provides the algorithmic details of PerOPE. Algorithm 1 describes the full procedure of our method for personalized estimator aggregation. The algorithm takes as input a set of candidate base estimators $\mathcal{V}$, a logged dataset $\mathcal{D}$, an evaluation policy π_e, and a maximum tree depth K. It additionally relies on two modular components: an MSE estimation module $\mathrm{algo}_{\mathrm{MSE}}$, which estimates the mean squared error of candidate estimators, and a split search module $\mathrm{algo}_{\mathrm{search}}$, which proposes candidate splits of the context space together with the corresponding estimator weights.

The procedure starts by initializing a single segment that covers the entire context space, with all weight mass assigned to a single estimator. At each depth level, the algorithm iterates over the current set of segments and applies the split search module to each segment. Given the context information of a segment, the split search module evaluates possible splits using the provided MSE estimation module and returns two child segments along with estimator weights optimized to minimize the estimated MSE. The set of segments is updated recursively, and this process continues until the maximum depth K is reached.

After the recursive partitioning is completed, the algorithm assigns personalized estimator weights to each data point based on the context segment it belongs to. Finally, PerOPE outputs a personalized OPE estimator as a weighted combination of the per-sample estimator contributions, where the weights vary across contexts according to the learned segmentation of the context space.

Algorithm 1 PerOPE

Input: candidate estimators $\{\hat{v}_m\}_{m=1}^M$, logged dataset $\mathcal{D}$, evaluation policy π_e, maximum depth K, MSE estimation module $\mathrm{algo}_{\mathrm{MSE}}$, split search module $\mathrm{algo}_{\mathrm{search}}$ which receives the context information of the target segment and outputs the context information and estimator weights for split segments

Output: personalized estimator $\hat{V}_{\mathrm{PerOPE}}^{\widehat{\mathrm{best}}}$

Initialize segment dictionary $S := \{[\mathrm{None}] : \{\hat{V}_1 : 1, \hat{V}_2 : 0, \ldots, \hat{V}_M : 0\}\}$ where the key is context information defining the segment and the value is the weight of each estimator in that segment.
Initialize empty dictionary $S^{\mathrm{new}} := \{\}$
for $k = 1, \ldots, K$ **do**
  for $(S_{\mathrm{context}}, S_{\mathrm{weight}})$ in $S.\mathrm{items}()$ **do**
    $S_{\mathrm{context1}}, S_{\mathrm{weight1}}, S_{\mathrm{context2}}, S_{\mathrm{weight2}} \leftarrow \mathrm{algo}_{\mathrm{search}}(S_{\mathrm{context}}, \mathrm{algo}_{\mathrm{MSE}}, \{\hat{v}_m\}_{m=1}^M, \mathcal{D}, \pi_e)$
    $S^{\mathrm{new}}[S_{\mathrm{context1}}] \leftarrow S_{\mathrm{weight1}}$
    $S^{\mathrm{new}}[S_{\mathrm{context2}}] \leftarrow S_{\mathrm{weight2}}$
  end for
  $S \leftarrow S^{\mathrm{new}}, \quad S^{\mathrm{new}} \leftarrow \{\}$
end for
Initialize personalized weight $\alpha(x_i, m) := 0$ for $\forall i, \forall m$.
for $i = 1, \ldots, n$ **do**
  for $m = 1, \ldots, M$ **do**
    $\alpha(x_i, m) \leftarrow S[S_{\mathrm{context}}][\hat{V}_m]$ such that $x_i \in S_{\mathrm{context}}$
  end for
end for
return $\hat{V}_{\mathrm{PerOPE}}^{\widehat{\mathrm{best}}} = \frac{1}{n}\sum_{i=1}^n \sum_{m=1}^M \alpha(x_i, m)\, \hat{v}_m(\pi_e, x_i)$

D Details of Synthetic Experiments

D.1 Reward function

The reward function consists of three components on context only, action only, and their interaction. We use three coefficients; coefficient vector c_x of the same dimension as x, coefficient vector c_a of the $|\mathcal{A}|$ dimension, and coefficient matrix c_{both} of the dimension (dim-x, $|\mathcal{A}|$). All elements of these coefficients are sampled from a standard normal distribution individually. The binary rewards r are sampled from the Bernoulli distribution as $r \sim \mathrm{Bern}(q(x, a))$. The expected reward function q used in the synthetic experiments is defined as

$$q(x, a) = \mathrm{sigmoid}(c_x^T x + c_a^T a_{\mathrm{onehot}} + x^T c_{\mathrm{both}} a_{\mathrm{onehot}}),$$

where a_{onehot} is a one-hot vector of $a \in \mathcal{A}$.

D.2 Hyperparameters in MSE estimation

OCV estimates the MSE of candidate estimators using a cross-validation procedure. Therefore, it has the number of folds as a hyperparameter. In our experiments, we used 10-fold cross-validation for OCV-IPS and OCV-DR, as well as for the MSE estimation modules in PerOPE (OCV-IPS) and PerOPE (OCV-DR). In OCV, the training/validation split ratio is determined in a data-driven way (Cief et al., 2025) for each estimator. On the other hand, in PerOPE, we use the average of all estimators' training/validation split ratio.

Since OPERA relies on bootstrap sampling to estimate the MSE, the number of bootstrap samples is a hyperparameter. We set it to 10. This setting was used for both OPERA and the MSE estimation module within PerOPE (OPERA).

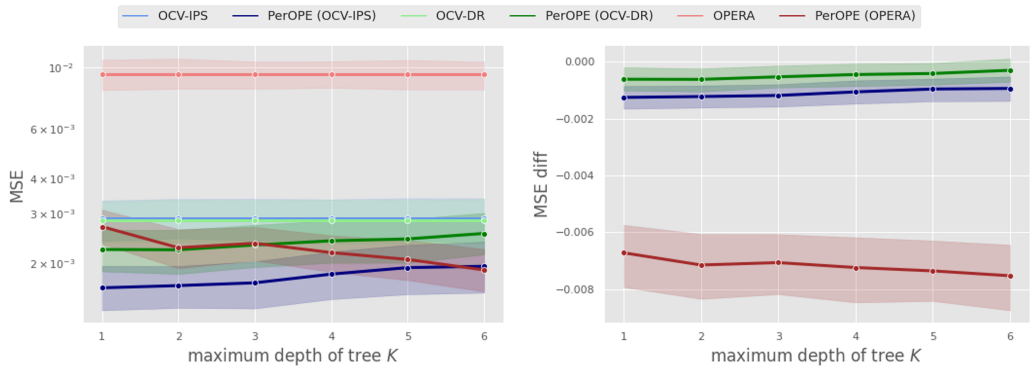


Figure 7: Comparison of MSE on the synthetic data with varying the maximum depth of tree K. Error bars denote 95% confidence intervals over 200 trials.

D.3 Ground-truth policy value

The evaluation metric MSE requires the ground-truth policy value of the evaluation policy, $V(\pi_e)$. Since it is impossible to calculate the true expected value precisely, we calculate it from a sufficiently large sample of 1,000,000 observations that are independent of $\mathcal{D}$.

D.4 MSE of each method

MSE of each estimator selection method is shown in Figure 6. We can see the consistent improvement by utilizing the idea of personalization.

E Ablation Study with Synthetic Data

In this section, we conduct an ablation study to investigate the performance of PerOPE from several technical viewpoints.

E.1 Hyperparameter in PerOPE

PerOPE has a hyperparameter, the maximum depth of the tree K . The experiment with changing this hyperparameter helps us understand the sensitivity of our method to K and demonstrates its robustness.

We vary K from 1 to 6. The experiment is conducted under a setting with high context-dependency ($\beta_0 = 2, \beta_{\text{cons}} = 2, \beta_{\text{coef}} = 20$) to clearly illustrate the effect of personalization. All other settings are identical to those in the synthetic experiments mentioned in the main text.

The results are presented in Figure 7. The key takeaway is that PerOPE consistently outperforms the baseline methods across all tree depths, demonstrating the benefit of our personalized approach and the robustness to the selection of K.

For PerOPE (OCV), we observe a significant performance improvement even with a very shallow tree of depth $K = 1$. The performance remains strong and relatively stable for $K = 1$ and $K = 2$. This indicates that even a single split of the context space allows PerOPE to find a better, personalized combination of estimators compared to the single global choice made by OCV. The fact that substantial gains are achieved with a shallow tree is a significant practical advantage, as it suggests that PerOPE can be highly effective without the need to learn a complex, deep tree. While performance slightly degrades for deeper trees in this specific setting, this suggests that a value of $K = 1$ or 2 is optimal only in this specific setting. One possible reason is that partitions may become too granular in deeper trees, leading to less reliable MSE estimation.

The results for PerOPE (OPERA) provide even clearer evidence for the value of personalization. Since OPERA is an ensemble method, the performance gap between OPERA and PerOPE (OPERA)

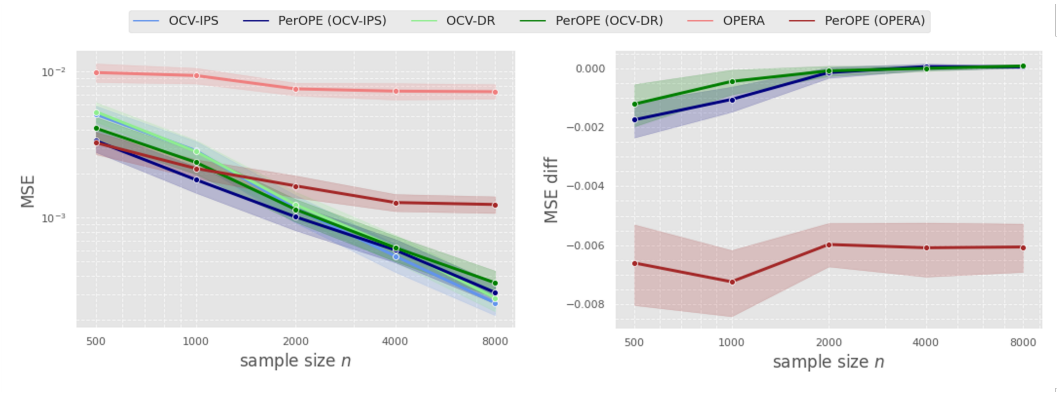


Figure 8: Comparison of MSE on the synthetic data with varying the sample size n. Error bars denote 95% confidence intervals over 200 trials.

highlights the pure benefit derived from context-aware personalization. As shown in the figure, the MSE consistently decreases as the tree depth K increases, with the best performance achieved at the deepest tested level ($K = 6$). This suggests that $K = 6$ is optimal in this setting, leading to a superior overall evaluation.

In summary, this study demonstrates that our method is robust to the choice of K; even a shallow tree with $K = 1$ delivers significant improvements, making PerOPE a practical and powerful tool for achieving more accurate off-policy evaluation.

E.2 Sample size of log data

To further understand the characteristics of PerOPE, we investigate its performance by varying the sample size of logged data $\mathcal{D}$. This study examines the robustness of our method and its advantages over baselines across different sample sizes.

We conduct experiments with the sample size n varying across the set $\{500, 1000, 2000, 4000, 8000\}$. The experiment is run in a setting with high context-dependency ($\beta_0 = 2, \beta_{\mathrm{cons}} = 2, \beta_{\mathrm{coef}} = 20$) to ensure that the benefits of personalization are observable. All other settings remain identical to those in the main synthetic experiments.

The results, shown in Figure 8, illustrate how the MSE of each method changes with the sample size. The performance of all methods generally improves (i.e., MSE decreases) as the sample size increases, since more data makes the OPE task less challenging. However, we observe distinct trends between the OCV-based and OPERA-based methods.

For the OCV and PerOPE (OCV), the performance gap narrows as the sample size grows. This is an intuitive result. A larger sample size improves the accuracy of each base estimator. Consequently, the choice of estimator becomes less critical, and the relative advantage gained from context-aware personalization diminishes. Nevertheless, PerOPE consistently maintains a performance edge, especially at smaller sample sizes where estimator selection is more critical.

In contrast, for OPERA and PerOPE (OPERA), the performance gap remains substantial and consistent across all tested sample sizes. This is because the MSE improvement of both methods appears to saturate at larger sample sizes. This phenomenon can likely be attributed to the limitations of OPERA's bootstrap-based MSE estimation technique in this setting. As mentioned in the main text, this MSE estimation method primarily captures variance and does not fully account for the bias. Since bias does not necessarily decrease with sample size, the MSE estimation for OPERA may remain inaccurate, limiting its performance. Even under these challenging conditions, PerOPE consistently improves upon the baseline by finding better, context-specific ensembles, demonstrating the robustness and consistent value of our personalization approach.

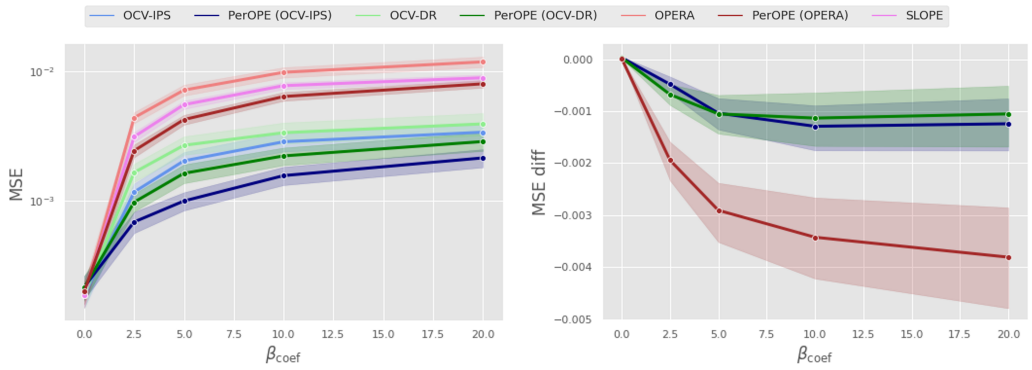


Figure 9: Comparison of MSE on the synthetic data for hyperparameter tuning of DRos. Error bars denote 95% confidence intervals over 200 trials.

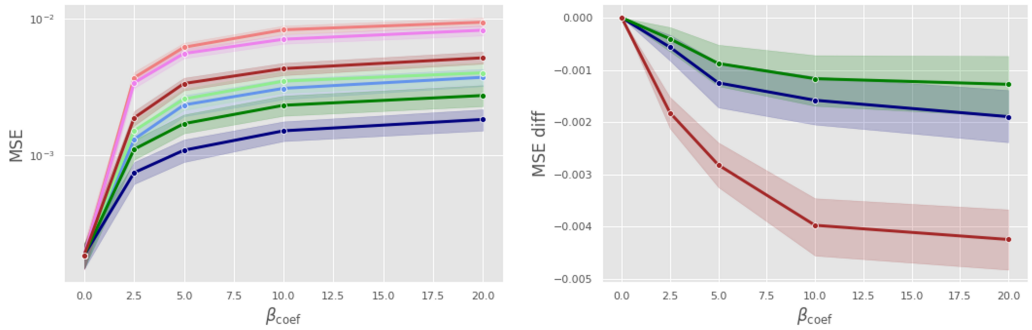


Figure 10: Comparison of MSE on the synthetic data for hyperparameter tuning of SwitchDR. Error bars denote 95% confidence intervals over 200 trials.

Overall, this study confirms that PerOPE is effective across a wide range of data regimes. It provides significant benefits, particularly when data is limited or when the underlying MSE estimation of the baseline method has inherent limitations.

E.3 Hyperparameter tuning of OPE estimator

While the main text focuses on estimator selection, our PerOPE framework is general enough to be applied to the closely related problem of hyperparameter tuning for OPE estimators. This can be

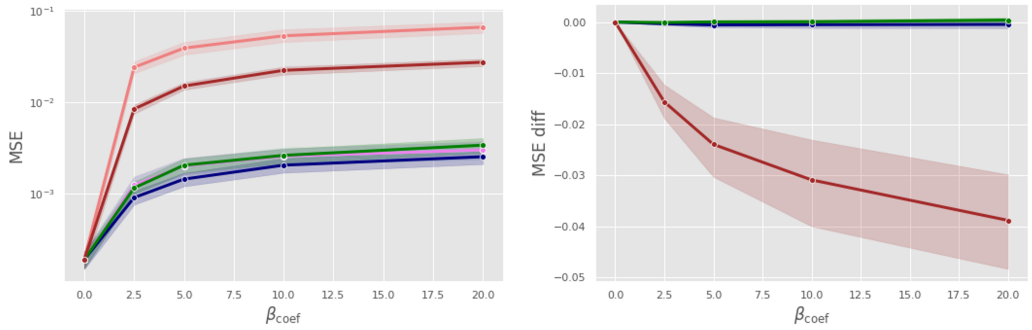


Figure 11: Comparison of MSE on the synthetic data for hyperparameter tuning of DRps. Error bars denote 95% confidence intervals over 200 trials.

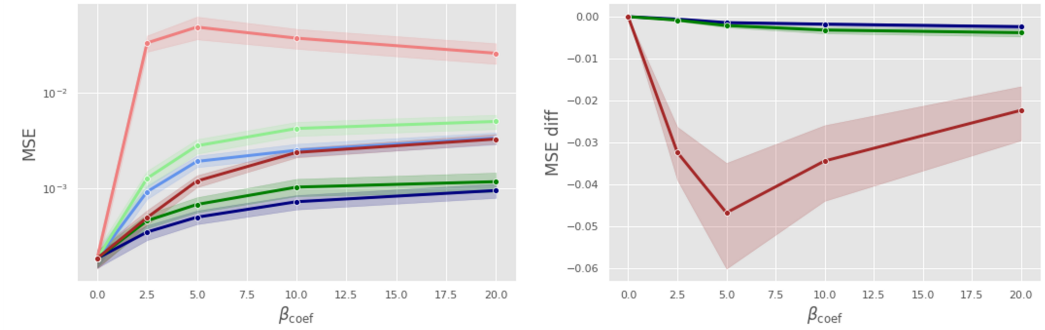


Figure 12: Comparison of MSE on the synthetic data for 2 tunable hyperparameter tuning of DR-λ . Error bars denote 95% confidence intervals over 200 trials.

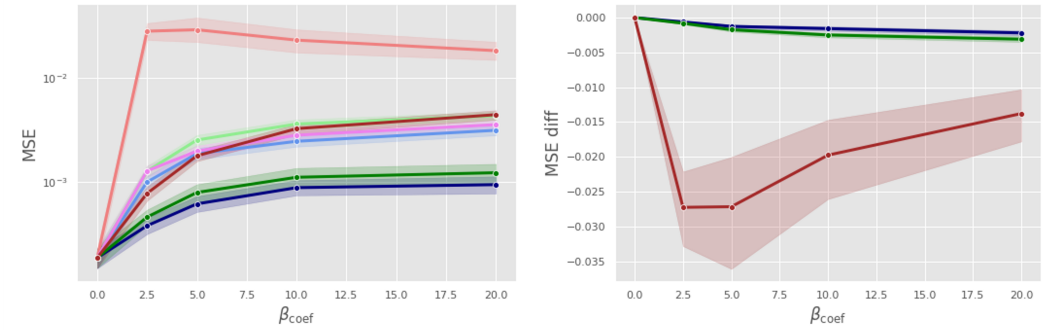


Figure 13: Comparison of MSE on the synthetic data for 1 tunable hyperparameter tuning of DR-λ . Error bars denote 95% confidence intervals over 200 trials.

achieved by treating each unique combination of hyperparameters as a distinct candidate estimator. In this section, we empirically validate the effectiveness of PerOPE in this setting.

The experimental setup is largely identical to that of the main synthetic experiments, with $\beta_0 = 2, \beta_{\text{cons}} = 2, \beta_{\text{coef}} = 20$. We compare PerOPE with OCV and OPERA, and additionally include SLOPE (Su et al., 2020b), a method specifically designed for hyperparameter tuning in OPE, as a baseline.

We conduct experiments on several estimators with hyperparameters:

1. DRos with shrinkage parameter $\lambda \in \{10^0, 10^2, 10^4, 10^6\}$.
2. Switch-DR with threshold parameter $\lambda \in \{10^0, 10^2, 10^4, 10^6\}$.
3. DRps with clipping parameter $\lambda \in \{10^0, 10^2, 10^4, 10^6\}$.
4. DR-λ with two hyperparameters: $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $s \in \{-1, -0.5\}$, resulting in 10 candidates.
5. DR-λ with a single hyperparameter: $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and a fixed $s = -1$.

We note that SLOPE is designed for tuning a single hyperparameter and thus is excluded from the comparison in the two-hyperparameter case (scenario 4).

The results are presented in Figure 9, 10, 11, 12, and 13. The findings clearly demonstrate the effectiveness of PerOPE for hyperparameter tuning.

First, consistent with our estimator selection experiments, PerOPE significantly improves upon its corresponding baselines (OCV and OPERA) across all scenarios. This indicates that the benefit of personalization is not limited to choosing between different types of estimators but is equally potent for selecting the optimal hyperparameter value in a context-dependent manner.

Second, when compared to SLOPE, PerOPE demonstrates highly competitive or superior performance. In particular, PerOPE (OCV-IPS) and PerOPE (OCV-DR) consistently achieve a lower MSE than SLOPE across almost all applicable scenarios. This is a remarkable result, as it shows that our general-purpose framework can outperform a method specifically tailored for hyperparameter tuning.

Furthermore, the case of tuning two hyperparameters for DR-λ highlights a key advantage of PerOPE's flexibility. While SLOPE cannot be applied, PerOPE handles this multi-hyperparameter scenario seamlessly, personalizing the selection from the 10 candidates.

In conclusion, this study provides strong evidence that PerOPE is a powerful tool, not only for estimator selection but also for hyperparameter tuning. Its ability to personalize the choice of hyperparameters allows it to achieve a new level of accuracy, even surpassing SLOPE as a specialized state-of-the-art method.

F Broader Impact

Off-policy evaluation is widely used to assess new decision-making policies before online deployment, particularly in applications such as recommendation, advertising, and personalized services. By improving the accuracy of OPE through context-dependent estimator aggregation, PerOPE may help practitioners make more reliable deployment decisions while reducing the need for costly or risky online experimentation. In particular, the ability to adapt estimator aggregation across heterogeneous user contexts may be beneficial in production environments where different user segments exhibit substantially different logging-policy overlap, reward noise, or data availability.

At the same time, personalization introduces potential risks that should be considered carefully. PerOPE relies on contextual features to partition the data and determine estimator aggregation rules. If these features contain sensitive or protected attributes, the learned partitions may reflect or reinforce undesirable distinctions between user groups. Moreover, improved estimation accuracy does not by itself guarantee that the policy being evaluated is fair, safe, or socially beneficial. PerOPE should therefore be viewed as a tool for improving statistical evaluation rather than as a mechanism for ensuring fairness or responsible policy design.

There are also privacy considerations when applying personalized OPE to real-world user data. Practitioners should minimize the use of unnecessary personal information and follow appropriate data-governance, access-control, and privacy-preserving practices. In our real-world experiment, the method is evaluated on logged interaction data from an existing A/B test, and the analysis focuses on aggregate estimator performance rather than identifying or targeting individual users.

Finally, the reliability of PerOPE depends on the quality of the logged data, candidate estimators, and the MSE estimation procedure used within the framework. Poor overlap between logging and evaluation policies, misspecified reward models, or unreliable MSE estimates may still lead to inaccurate conclusions. We therefore recommend that PerOPE be used together with standard OPE diagnostics, robustness checks, and domain-specific safety considerations when informing real-world deployment decisions.